Appendix and Supplement Material

We provide the technical proofs, experiment details as well as the relegated discussions mentioned in the paper. The appendix for Section 4, 5, 6 are provided in A.1, A.2, A.3, respectively. The auxiliary lemmas in our proofs are summarized in A.4. The additional experiment details are provided in A.5.

A.1 Material for Section 4

When optimized by the gradient descent: $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} - \eta \nabla \mathcal{L}(\boldsymbol{\theta}^{(t-1)})$ using an infinitesimal learning rate, the updates in the parameter space can be equivalently described by the *gradient flow*:

$$\frac{d\boldsymbol{\theta}^{(t)}}{dt} = -\nabla_{\boldsymbol{\theta}} \mathcal{L}\big(\boldsymbol{\theta}^{(t)}\big).$$

A nice property of gradient flow is that if $\mathcal{L}$ is smooth, then the objective function is non-increasing during the updates since:

$$\frac{d\mathcal{L}\big(\boldsymbol{\theta}^{(t)}\big)}{dt} = -\Big\langle \nabla \mathcal{L}\big(\boldsymbol{\theta}^{(t)}\big), \frac{d\boldsymbol{\theta}^{(t)}}{dt} \Big\rangle = -\Big\| \frac{d\boldsymbol{\theta}^{(t)}}{dt} \Big\|_2^2, \tag{A.1}$$

which is non-negative. Therefore, it saves the discussion of choosing the proper learning rate to ensure the same property in gradient descent.

Another preparation work is to reformulated NCF, especially the input (which are essentially the users and items embeddings), into a standard form of FFN: $\mathbf{W}_1 \sigma\big(\mathbf{W}_2 \mathbf{x}_{u,i}\big)$ where $\mathbf{x}_{u,i}$, are fixed and do not depend on the unknown embeddings.

We use $\mathbf{e}_u \in \mathbb{R}^{|\mathcal{U}|}$ and $\mathbf{e}_i \in \mathbb{R}^{|\mathcal{I}|}$ to denote the one-hot encoding of the user and item id. Also, we use $\mathbf{e}_{u,i} \in \mathbb{R}^{|\mathcal{U}||\mathcal{I}|}$ to denote the one-hot encoding of user+item id combined. Therefore, NCF with addition can be efficiently represented as:

$$f^{\text{NCF-a}}(u,i) = \mathbf{W}_1 \sigma\big(\mathbf{W}_2 \mathbf{x}_{u,i}\big), \text{ with } \mathbf{W}_2 = \big[\mathbf{Z}_U^\intercal, \mathbf{Z}_I^\intercal\big]^\intercal \text{ and } \mathbf{x}_{u,i} = [\mathbf{e}_u, \mathbf{e}_i]^\intercal; \tag{A.2}$$

and NCF with concatenation is:

$$f^{\text{NCF-c}}(u,i) = \mathbf{W}_1 \sigma\big(\mathbf{W}_2 \mathbf{x}_{u,i}\big), \text{ with } \mathbf{W}_2 = \begin{bmatrix} \mathbf{z}_{i_1} & \mathbf{z}_{u_1} \\ \mathbf{z}_{i_1} & \mathbf{z}_{u_2} \\ \dots & \cdots \end{bmatrix}^\intercal \text{ and } \mathbf{x}_{u,i} = \mathbf{e}_{u,i}^\intercal. \tag{A.3}$$

Recall the linearization from Section 4, where we denote the first-order Taylor approximation of $f(\boldsymbol{\theta}; \cdot)$ by $\tilde{f}(\boldsymbol{\theta}; \cdot)$ such that:

$$\tilde{f}\big(\boldsymbol{\theta}; (u,i)\big) := f\big(\boldsymbol{\theta}^{(0)}; (u,i)\big) + \Big\langle \boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}, \nabla f\big(\boldsymbol{\theta}^{(0)}; (u,i)\big) \Big\rangle.$$

Also, we use d, d_1 to denote the embedding dimension and the dimension of the first hidden layer in the FFN for NCF, and assume $d_1 = d$ w.l.o.g. We still consider the scaled initialization $N(0, \alpha/d)$ where α is a constant. Under the infinite-width limit, we can show that the NTK converges to a fixed kernel at initialization, which we referred to as the collaborative filtering kernel.

Lemma A.1. *For the MCF and NCF we described in Section 3.1, the neural tangent kernel* $K\big((u,i),(u',i')\big) = \big\langle \nabla f\big(\boldsymbol{\theta}; (u,i)\big), \nabla f\big(\boldsymbol{\theta}; (u',i')\big) \big\rangle$ *have the following convergence result:*

$$\lim_{d \to \infty} K\big((u,i),(u',i')\big) = K_{CF}\big((u,i),(u',i')\big) := a + b \cdot \mathbb{1}[i = i'] + c \cdot \mathbb{1}[u = u'],$$

where under the $N(0,1/d)$ initializations, $a=0$, $b=c=1$ for K_{MCF}; $a=1/\pi$, $b=c=\frac{1}{2}-\frac{1}{2\pi}$ for $K_{NCF\text{-}c}$, and $a=1/\pi$, $b=c=\frac{1}{2}-\frac{(2-\sqrt{3})}{2\pi}$ for $K_{NCF\text{-}a}$.

Proof. We first consider NCF. We first reformulate NCF's formulation in (A.2) and (A.3) as:

$$f(\mathbf{x}_{u,i}) = \sqrt{\frac{2}{d}}\mathbf{W}_1\sigma\big(\mathbf{W}_2\mathbf{x}_{u,i}\big), \mathbf{W}_1, \mathbf{W}_2 \sim N(0,1),$$

where we extract the $1/d$ variance to the front, and add the $\sqrt{2}$ factor for convenience.

Notice that: $\frac{\partial f(\mathbf{x}_{u,i})}{\partial \mathbf{W}_{1,j}} = \sqrt{2/d}\big(\mathbf{W}_{2,j}^\intercal \mathbf{x}_{u,i}\big)_+$, and

$$\nabla_{\mathbf{W}_{2,j}} f(\mathbf{x}_{u,i}) = \sqrt{2/d}\mathbf{W}_{1,j}\mathbf{x}_{u,i}\mathbb{1}\big[\mathbf{W}_{2,j}^\intercal \mathbf{x}_{u,i} \geq 0\big],$$

where $\mathbf{W}_{1,j}$ is the j^{th} element of the vector $\mathbf{W}_1$, and $\mathbf{W}_{2,j}$ is the j^{th} column of the matrix $\mathbf{W}_2$. For notation simplicity, we define $v_j = \mathbf{W}_{1,j}$ and $\mathbf{w}_j = \mathbf{W}_{2,j}$.

Therefore, the NTK for NCF is given by:

$$\nabla f(\mathbf{x}_{u,i})^\intercal \nabla f(\mathbf{x}_{u',i'}) = \frac{2}{d}\sum_{j=1}^{d}\big(\mathbf{w}_j^\intercal \mathbf{x}_{u,i}\big)_+\big(\mathbf{w}_j^\intercal \mathbf{x}_{u',i'}\big)_+ + \frac{2}{d}\sum_{j=1}^{d}(v_j\mathbf{x}_{u,i})^\intercal(v_j\mathbf{x}_{u',i'})\mathbb{1}\big[\mathbf{w}_j^\intercal \mathbf{x}_{u,i}\geq 0\big]\mathbb{1}\big[\mathbf{w}_j^\intercal \mathbf{x}_{u',i'}\geq 0\big]. \tag{A.4}$$

Using the mean and variance formula of truncated normal distribution, following the setup in (A.2) and (A.3), for NCF with concatenation we have:

- When $u \neq u'$ and $i \neq i'$, we have:

$$K\big((u,i),(u',i')\big) = \frac{2}{d}\sum_{j=1}^{d}(\mathbf{w}_j)_+(\mathbf{w}_j^*)_+, \quad \mathbf{w}_j^* \text{ is an i.i.d copy of } \mathbf{w}_j$$
$$\overset{d\to\infty}{=} 2\mathbb{E}\big[(\mathbf{w}_j)_+(\mathbf{w}_j^*)_+\big] = \frac{4}{\pi}$$

- When $u = u'$ or $i = i'$, we have:

$$K\big((u,i),(u',i')\big) = \frac{1}{d}\sum_{j=1}^{d}(\mathbf{w}_j)_+^2 + \frac{2}{d}\sum_{j=1}^{d}(\mathbf{w}_j)_+(\mathbf{w}_j^*)_+, \quad \mathbf{w}_j^* \text{ is a copy of } \mathbf{w}_j$$
$$\overset{d\to\infty}{=} var\big((\mathbf{w}_j)_+\big) + 2\mathbb{E}\big[(\mathbf{w}_j)_+(\mathbf{w}_j^*)_+\big] = 2 + \frac{2}{\pi}$$

- When $u = u'$ and $i = i'$, we leverage the integral formulation of arc-cosine kernel K_0 in Lemma A.3 such that:

$$K\big((u,i),(u',i')\big) = K_0(\mathbf{x}_{u,i},\mathbf{x}_{u',i'}) - \frac{2}{d}\sum_{j=1}^{d}(\mathbf{w}_j)_+(\mathbf{w}_j^*)_+ + \frac{2}{d}\sum_{j=1}^{d}(\mathbf{w}_j)_+^2$$
$$\overset{d\to\infty}{=} 8 - \frac{16}{\pi}.$$

The results for NCF under addition is obtained using basically the same computations. For MCF, on the other hand, we reformulate the predictor as: $f(u,i) = \frac{1}{d}\big\langle \mathbf{Z}_U\mathbf{Z}_I, \mathbf{X}_{u,i}\big\rangle$ where $\mathbf{X}_{u,i} = \mathbf{e}_u\mathbf{e}_i^\intercal$, and the embeddings follow $N(0,1)$ initializations. Then it holds that:

$$\nabla f(\mathbf{X}_{u,i})^\intercal \nabla f(\mathbf{X}_{u',i'}) = \frac{1}{d}\Big(\langle \mathbf{z}_u, \mathbf{z}_{u'}\rangle\mathbb{1}[i=i'] + \langle \mathbf{z}_i, \mathbf{z}_{i'}\rangle\mathbb{1}[u=u']\Big), \tag{A.5}$$

which directly leads to the stated results of K_{CF}.

□

Remark 1 (The parameterization of K_{CF} and the intializations of MCF, NCF)**.** *It is evident from the above proof that the relative scale of a, b and c in K_{CF} can depend on the constant term α in the intializations of $N(0, \alpha/d)$. For instance, if the infinite-width MCF initializes the user embeddings $\mathbf{Z}_U$ with $N(0, \alpha_1/d)$ and the item embeddings $\mathbf{Z}_I$ with $N(0, \alpha_2/d)$, then by (A.5) we immediately have $b = \alpha_1$ and $c = \alpha_2$.*

Also, the parameterization of K_{CF} for NCF is also dependent on the initialization. We observe from (A.4) that a would not change as long as the initializations are i.i.d., but b and c again depend on the individual α. The exact derivations for the NTK of FFN is stuided by [2], and in [23] the author provides the NTK formulation for a broad range of neural networks.

Other than the convergence to a fixed kernel at initialization, the infinite-width limit also suggests that the parameters varies little during the gradient flow updates, and the linearization of $\tilde{f}$ has a good approximation:

$$\frac{\big\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(0)}\big\|_2}{\big\|\boldsymbol{\theta}^{(0)}\big\|_2} = \mathcal{O}(\sqrt{1/d}), \text{ and } \tilde{f}\big(\boldsymbol{\theta}^{(t)}; (u, i)\big) = f\big(\boldsymbol{\theta}^{(t)}; (u, i)\big) + \mathcal{O}(\sqrt{1/d}).$$

We formalize the above arguments in the following lemma.

Lemma A.2. *Let the gradient flow updates under $\tilde{f}$ be denoted by $\tilde{\boldsymbol{\theta}}^{(t)}$. Under the exponential loss or log loss, when the predictor $f(\boldsymbol{\theta}; \cdot)$ is local Lipschitz and admits chain rule, the corresponding decision boundaries for the two gradient flow trajectories satisfy the following result for any $T > 0$:*

$$\lim_{d\to\infty} \sup_{t\le T} \big\|F(\boldsymbol{\theta}^{(t)}) - \tilde{F}(\tilde{\boldsymbol{\theta}}^{(t)})\big\|_2 = 0. \tag{A.6}$$

To the best of our knowledge, the similar infinite-width convergence results were studied under the squared loss [2, 14], and we extend them to the classification setting.

Proof. By the chain rule, for any step $T > 0$, we have:

$$\int_0^T \big\|\boldsymbol{\theta}^{(t)}\big\|_2 dt = \int_0^T \big\|\nabla\mathcal{L}(\boldsymbol{\theta}^{(t)})\big\|_2 dt \le \sqrt{T}\Big(\int_0^T \big\|\nabla\mathcal{L}(\boldsymbol{\theta}^{(t)})\big\|_2^2 dt\Big)^{1/2},$$

by the Hölder's inequality. According to (A.1), $d\mathcal{L}\big(\boldsymbol{\theta}^{(t)}\big)/dt = -\Big\|\nabla\mathcal{L}(\boldsymbol{\theta}^{(t)})\Big\|_2^2$, so we have:

$$\sup_{t\le T} \big\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(0)}\big\|_2 \le \sqrt{T\mathcal{L}\big(\boldsymbol{\theta}^{(t)}\big)} \lesssim \sqrt{1/d},$$

due to the scaled initializations. Denote the risk associated with a predictor by $R\big(f(\boldsymbol{\theta})\big)$. It can then be deduced that $\sup_{t\le T} \big\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(0)}\big\|_2 \le C_1$ and $\sup_{t\le T} \big\|\nabla R\big(f(\boldsymbol{\theta}^{(t)})\big)\big\|_2 \le C_2$, for some constants C_1 and C_2.

Define $\delta(t) := \|f(\boldsymbol{\theta}^{(t)}) - \tilde{f}(\boldsymbol{\theta}^{(t)})\|_2$, and the NTK under a specific $\boldsymbol{\theta}$ as $K(\boldsymbol{\theta})$. Since $\|K(\boldsymbol{\theta})^{(t)} - K(\boldsymbol{\theta})^{(0)}\|_2 = \mathcal{O}(\sqrt{1/d})$ according to [14, 2], we have:

$$\begin{aligned}
\frac{d\delta(t)}{dt} &\le \Big\|K(\boldsymbol{\theta}^{(t)})\nabla R\big(f(\boldsymbol{\theta}^{(t)})\big) - K(\boldsymbol{\theta}^{(0)})R\big(\tilde{f}(\boldsymbol{\theta}^{(t)})\big)\Big\|_2 \\
&\le \Big\|\big(K(\boldsymbol{\theta}^{(t)}) - K(\boldsymbol{\theta}^{(0)})\big)\nabla R\big(f(\boldsymbol{\theta}^{(t)})\Big\|_2 + \Big\|K(\boldsymbol{\theta}^{(0)})\big(R\big(f(\boldsymbol{\theta}^{(t)}) - R\big(\tilde{f}(\boldsymbol{\theta}^{(t)})\big)\Big\|_2 \\
&\le C_3/\sqrt{d} + C_4\delta(t),
\end{aligned}$$

for some constants C_3 and C_4. Since $\delta(0) = 0$, then δ_t is a sub-solution to the ordinary differential equation: $\frac{d\delta(t)}{dt} = C_3/\sqrt{d} + C_4\delta(t)$ with $\delta(0) = 0$. It then follows: $\delta(t) \le \frac{C_3(\exp(C_4 t)-1)}{\sqrt{d}C_4}$, so we conclude that: $\lim_{d\to\infty} \sup_{t\le T} \|F(\boldsymbol{\theta}^{(t)}) - \tilde{F}(\tilde{\boldsymbol{\theta}}^{(t)})\|_2 = 0$. □

Proof for Theorem 1

Proof. According to Corollary A.1 from A.2, under the exponential or log loss, the linearization $\tilde{f}(\tilde{\boldsymbol{\theta}};\,\cdot)$ satisfies condition **C1**, **C2** and **C3**, so the gradient flow optimization of $\tilde{\boldsymbol{\theta}}^{(t)}$ converges to the stationary points of:

$$\min\ \|\tilde{\boldsymbol{\theta}}\|_2\ \text{ s.t. } y_{u,i}\tilde{f}(\tilde{\boldsymbol{\theta}};(u,i))\geq 1,\ \forall(u,i)\in\mathcal{D}_{\text{train}}.$$

Combining the results from Lemma A.1 and Lemma A.2, under the gradient flow optimization, the response surface of MCF and NCF converges to the minimum RKHS norm solution:

$$\lim_{t\to\infty}\lim_{d\to\infty}F\Big(\frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2}\Big)\overset{\text{stationary points of}}{\to}\underset{f:(\mathcal{U},\mathcal{I})\to\mathbb{R}}{\arg\min}\big\|f\big\|_{K_{\text{CF}}}\text{ s.t. }y_{u,i}f(u,i)\geq 1,\ \forall(u,i)\in\mathcal{D}_{\text{train}}.$$

□

A.2 Material for Section 5

We first discuss the implications of condition **C1**, **C2** and **C3**. Recall that:

C1. The loss function has the exponential-tail behavior such as the exponential loss and log loss;

C2. Both the MCF and NCF in are L-homogeneous, i.e. $f(\boldsymbol{\theta};\,\cdot)=\|\boldsymbol{\theta}\|_2^L\cdot f\big(\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2;\,\cdot\big)$ for some $L>0$, and have some smoothness property;

C3. The data is separable with respect to the overparameterized MCF and NCF.

The exponential decay on the tail of the loss function is important for the inductive bias of gradient descent as $\ell(u)$ behaves like $\exp(-u)$ when $u\to\infty$. Soudry et al. [20] first propose the notion of *tight exponential tail*, where the negative loss derivative $-\ell'(u)$ behave like:

$$-\ell'(u)\lesssim\big(1+\exp(-c_1u)\big)e^{-u}\text{ and }-\ell'(u)\gtrsim\big(1-\exp(-c_2u)\big)e^{-u},$$

for sufficiently large u, where c_1 and c_2 are positive constants. There is also a smoothness assumption on $\ell(\cdot)$. There is a more general (and perhaps more direct) definition of exponential-tail loss function [18], where $\ell(u)=\exp(-f(u))$, such that:

- f is smooth and $f'(u)\geq 0,\forall u$;
- there exists $c>0$ such that $f'(u)u$ is non-decreasing for $u>c$ and $f'(u)u\to\infty$ as $u\to\infty$.

It is easy to verify that the exponential loss, log loss and cross-entropy loss satisfy both requirements.

The predictor of MCF is obviously homogeneous, but for NCF to be homogeneous, the bias terms cannot be used for the hidden layers in the FFN. The requirement on the activation function is relative mild, since ReLU, LeakyReLU and some other common activation functions all preserve the homogeneity of the predictor.

On the other hand, the smoothness condition, which includes the locally Lipschitz condition and differentiability. Notice that Lipschitz condition is rather mild assumption for neural networks, and several recent paper are dedicated to obtaining the Lipschitz constant of deep learning models using activation such as ReLU [10, 22]. The differentiablity condition is more technical-driven such that we can analyze the gradients. In practice, neural networks with ReLU activation do not satisfy the condition. We point out that there do exist smooth homogeneous activation functions, such as the quadratic activation $\sigma(x)=x^2$. Nevertheless, the ReLU activation admits the chain rule, so the same analysis using gradients can be carried out by the sub-differentials. Therefore, in our experiments, we use ReLU as the activation function for convenience, and assume differentiablity to provide a more straightforward analysis.

Finally, by implying the separability, we also assume that there exists t_0 such that:

$$\forall t>t_0,\quad\frac{1}{|\mathcal{D}_{\text{train}}|}\sum_{(u,i)\in\mathcal{D}}\exp\Big(-y_{u,i}f\big(\boldsymbol{\theta}^{(t)};(u,i)\big)\Big)<1.\tag{A.7}$$

In the following corollary, we prove a general result for gradient flow converging to the max-margin solution under condition.

Corollary A.1. *For gradient flow optimization with the exponential loss, under condition **C1**, **C2** and **C3**, $\lim_{t\to\infty} \boldsymbol{\theta}^{(t)}/\|\boldsymbol{\theta}^{(t)}\|_2$ converges to the KKT points of:*

$$\min \|\boldsymbol{\theta}\|_2 \quad \textit{s.t.} \quad y_{u,i} f\big(\boldsymbol{\theta}; (u,i)\big) \geq 1,\ \forall (u,i) \in \mathcal{D}_{\textit{train}}.$$

Compared with the results in [13], we do not assume the loss function already converged in direction. Although we use the same type of constraint quality idea as in [18], their results focus on the convergence of the normalized margin and our result emphasize the dynamics in the parameter space. The interests in this result is also beyond the content of this paper.

Proof. First notice that the KKT condition for the original problem (where we add a $\frac{1}{2}$ factor for convenience): $\min \frac{1}{2}\|\boldsymbol{\theta}\|_2$ s.t. $y_{u,i} f\big(\boldsymbol{\theta}; (u,i)\big) \geq 1,\ \forall (u,i) \in \mathcal{D}_{\text{train}}$ is given by:

$$\exists \lambda_{u,i} \geq 0, (u,i) \in \mathcal{D}_{\text{train}},\ \text{s.t.} \begin{cases} \boldsymbol{\theta} + \sum_{(u,i)\in\mathcal{D}_{\text{train}}} \lambda_{u,i} y_{u,i} \nabla f\big(\boldsymbol{\theta}; (u,i)\big) = 0 \\ \lambda_{u,i}\Big(y_{u,i} f\big(\boldsymbol{\theta}; (u,i)\big) - 1\Big) = 0,\ \forall (u,i) \in \mathcal{D}_{\text{train}}. \end{cases} \tag{A.8}$$

As we mentioned in Section 5, when the constraints are non-convex, the stationarity of local or global optimum does not equal KKT optimality, but when the Guignard constraint qualification (GCQ) is satisfied [12], those two become exchangeable. GCQ might be the weakest constraint qualification in some sense, but it is very difficult to check in practice.

On the other hand, the Mangasarian-Fromovitz constraint qualification (MFCQ), though stronger than GCQ (and thus imply GCQ), is easier to examine. Specifically, it states for the stationary points that:

$$\exists\, \mathbf{d} \quad \text{s.t.} \quad \Big\langle y_{u,i} \nabla f\big(\boldsymbol{\theta}; (u,i)\big), \mathbf{d} \Big\rangle > 0, \text{ for all } (u,i) \in \mathcal{D}_{\text{train}}.$$

Notice that L-homogeneous functions all satisfy MFCQ, because according to Lemma A.4, for any stationary point $\boldsymbol{\theta}^*$ satisfying $y_{u,i} f\big(\boldsymbol{\theta}^*; (u,i)\big) = 1$, we let $\mathbf{d}^* = \boldsymbol{\theta}^*$ and it holds that:

$$\Big\langle y_{u,i} \nabla f\big(\boldsymbol{\theta}^*; (u,i)\big), \mathbf{d}^* \Big\rangle = L > 0.$$

As a consequence, the stationary points for the L-homogeneous predictors are indeed the KKT points. Then we show the convergence of the gradient flow optimization path to the KKT points. We first define the quantity:

$$\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big) := \frac{-\log \sum_{(u,i)} \exp\big(-y_{u,i} f\big(\boldsymbol{\theta}^{(t)}; (u,i)\big)\big)}{\|\boldsymbol{\theta}^{(t)}\|_2^2} = \|\boldsymbol{\theta}^{(t)}\|_2^2 \cdot \log \frac{1}{\mathcal{L}(\boldsymbol{\theta}^{(t)})},$$

which is smoothed version of the average margin normalized by the $\|\boldsymbol{\theta}^{(t)}\|_2^2$.

We show the convergence by the following three steps.

S1. Under the gradient flow optimization, $\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)$ is non-decreasing for $t \geq t_0$ (A.7), together with $\mathcal{L}(\boldsymbol{\theta}^{(t)}) \to 0$ and $\|\boldsymbol{\theta}^{(t)}\|_2 \to \infty$.

S2. With a scaling factor $\alpha > 0$, it holds that: $\exists \lambda_{u,i}(t) \geq 0, (u,i) \in \mathcal{D}_{\text{train}},$ s.t.

$$\begin{cases} \Big\|\alpha\boldsymbol{\theta}^{(t)} - \sum_{(u,i)\in\mathcal{D}_{\text{train}}} \lambda_{u,i} y_{u,i} \nabla f\big(\alpha\boldsymbol{\theta}^{(t)}; (u,i)\big)\Big\|_2 \lesssim \left(1 - \Big\langle \frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2}, \frac{d\boldsymbol{\theta}^{(t)}/dt}{\|d\boldsymbol{\theta}^{(t)}/dt\|_2} \Big\rangle\right) \frac{1}{\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)} & \textbf{(S2.1)} \\ \lambda_{u,i}(t)\Big(y_{u,i} f\big(\alpha\boldsymbol{\theta}^{(t)}; (u,i)\big) - 1\Big) \lesssim \frac{1}{\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)\|\boldsymbol{\theta}^{(t)}\|_2^2} & \textbf{(S2.2)} \end{cases} \tag{A.9}$$

We mention that **S2.2** $\overset{t\to\infty}{\to}$ 0 will be a consequence of **S1**, and to show **S2.1** $\overset{t\to\infty}{\to}$ 0, we need to prove $\Big\langle \frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2}, \frac{d\boldsymbol{\theta}^{(t)}/dt}{\|d\boldsymbol{\theta}^{(t)}/dt\|_2} \Big\rangle \to 1$. Once we have **S2.1** $\overset{t\to\infty}{\to}$ 0 + **S2.2** $\overset{t\to\infty}{\to}$ 0 + MFCQ, it holds that $\lim_{t\to\infty} \boldsymbol{\theta}^{(t)}/\|\boldsymbol{\theta}^{(t)}\|_2$ are proportional to the KKT points.

S3. Finally, the goal is show that $\Big\langle \frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2}, \frac{d\boldsymbol{\theta}^{(t)}/dt}{\|d\boldsymbol{\theta}^{(t)}/dt\|_2} \Big\rangle \to 1$.

We follow our game plan and first prove the results in **S1**. To show that $\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)$ is a non-decreasing function of t, we derive $\frac{d\log\tilde{\gamma}\left(\boldsymbol{\theta}^{(t)}\right)}{dt}$ and leverage (A.1) to show it's non-negative:

$$\begin{aligned}\frac{d\log\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)}{dt} &= \frac{d\log\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}^{(t)})}}{dt} - L\frac{d\log\|\boldsymbol{\theta}^{(t)}\|_2}{dt} \\ &= \frac{1}{\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}^{(t)})}\mathcal{L}(\boldsymbol{\theta}^{(t)})}\bigg(-\frac{d\mathcal{L}(\boldsymbol{\theta}^{(t)})}{dt}\bigg) - L\cdot\frac{d\log\|\boldsymbol{\theta}^{(t)}\|_2}{dt}.\end{aligned} \tag{A.10}$$

Define $q_{u,i}(t) := y_{u,i}f(\boldsymbol{\theta}^{(t)};(u,i))$ to be the margin of each data point during optimization. Notice that:

- For all $(u,i)\in\mathcal{D}_{\text{train}}$, $\min_{u,i}q_{u,i}(t)\geq\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}^{(t)})}$, therefore:

$$\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}^{(t)})}\mathcal{L}(\boldsymbol{\theta}^{(t)}) \leq \sum_{(u,i)}\exp\big(-q_{u,i}(t)\big)q_{u,i}(t), \tag{A.11}$$

  and we denote the RHS by $Q(t)$. By the separability assumption, we have $\mathcal{L}(\boldsymbol{\theta}^{(t)})<1$ for $t>t_0$, which indicates $Q(t)>0$ for $t>t_0$.

- It holds that:

$$\begin{aligned}\frac{d\|\boldsymbol{\theta}^{(t)}\|_2}{dt} &= 2\Big\langle\boldsymbol{\theta}^{(t)},\frac{d\boldsymbol{\theta}^{(t)}}{dt}\Big\rangle \\ &= 2\Big\langle\boldsymbol{\theta}^{(t)},\nabla\mathcal{L}\big(\boldsymbol{\theta}^{(t)}\big)\Big\rangle \quad \text{by (A.1)} \\ &= 2\Big\langle\boldsymbol{\theta}^{(t)},\sum_{(u,i)}\exp\big(-y_{u,i}f\big(\boldsymbol{\theta}^{(t)};(u,i)\big)\big)\cdot y_{u,i}\nabla f\big(\boldsymbol{\theta}^{(t)};(u,i)\big)\Big\rangle \\ &= 2L\sum_{(u,i)}\exp\big(-q_{u,i}(t)\big)q_{u,i}(t) \quad \text{by Lemma A.4.}\end{aligned}$$

  Hence, we have:

$$\begin{aligned}\frac{d\log\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)}{dt} &= \frac{1}{2\|\boldsymbol{\theta}^{(t)}\|_2^2}\frac{d\|\boldsymbol{\theta}^{(t)}\|_2^2}{dt} = \frac{L}{\|\boldsymbol{\theta}^{(t)}\|_2^2}\exp\big(-q_{u,i}(t)\big)q_{u,i}(t) \\ &= \frac{\big\langle\boldsymbol{\theta}^{(t)},d\boldsymbol{\theta}^{(t)}/dt\big\rangle}{\|\boldsymbol{\theta}^{(t)}\|_2^2}.\end{aligned} \tag{A.12}$$

Combining (A.1), (A.10), (A.11) and (A.12):

$$\begin{aligned}\frac{d\log\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)}{dt} &\geq \frac{1}{Q(t)}\bigg(\Big\|\frac{d\boldsymbol{\theta}^{(t)}}{dt}-\Big\langle\frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2^2},\frac{d\boldsymbol{\theta}^{(t)}}{dt}\Big\rangle\Big\|_2^2\bigg) \\ &= \frac{\|\boldsymbol{\theta}^{(t)}\|_2^2}{Q(t)}\bigg\|\frac{d\boldsymbol{\theta}^{(t)}/\|\boldsymbol{\theta}^{(t)}\|_2^2}{dt}\bigg\|_2^2 = L\bigg(\frac{d\log\|\boldsymbol{\theta}^{(t)}\|_2}{dt}\bigg)^{-1}\bigg\|\frac{d\tilde{\boldsymbol{\theta}^{(t)}}}{dt}\bigg\|_2^2\geq 0\end{aligned}$$

where we defined $\tilde{\boldsymbol{\theta}^{(t)}}=\frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2}$.

Then we show $\mathcal{L}(\boldsymbol{\theta}^{(t)})\to 0$ and $\|\boldsymbol{\theta}^{(t)}\|_2\to 0$ using the monotonicity of $\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)$. Note that:

$$\frac{-d\mathcal{L}(\boldsymbol{\theta}^{(t)})}{dt} = \Big\|\frac{d\boldsymbol{\theta}^{(t)}}{dt}\Big\|_2^2 \geq \Big\langle\frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2},\frac{d\boldsymbol{\theta}^{(t)}}{dt}\Big\rangle^2 = L^2\frac{Q(t)^2}{\|\boldsymbol{\theta}^{(t)}\|_2^2} \quad \text{by (A.12),}$$

where the inequality is by applying the Cauchy-Schwartz inequality. By (A.11): $Q(t)\geq\mathcal{L}(\boldsymbol{\theta}^{(t)})\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}^{(t)})}$, and by the definition, we have: $\|\boldsymbol{\theta}^{(t)}\|_2=\Big(\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}^{(t)})}/\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)\Big)^{1/L}$. As a consequence,

$$\begin{aligned}\frac{-d\mathcal{L}(\boldsymbol{\theta}^{(t)})}{dt} &\geq L^2\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)^{2/L}\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}^{(t)})}^{2-2/L}\mathcal{L}(\boldsymbol{\theta}^{(t)}) \\ &\geq L^2\tilde{\gamma}\big(\boldsymbol{\theta}^{(t_0)}\big)^{2/L}\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}^{(t)})}^{2-2/L}\mathcal{L}(\boldsymbol{\theta}^{(t)}),\end{aligned}$$

which indicates that:

$$\frac{d\big(1/\mathcal{L}(\boldsymbol{\theta})\big)}{dt}\frac{1}{\log\frac{1}{\mathcal{L}(\boldsymbol{\theta}^{(t)})}^{2-2/L}} \geq l^2\tilde{\gamma}\big(\boldsymbol{\theta}^{(t_0)}\big)^{2/L}.$$

Taking the intergral on both sides from t_0 to t, we immediate have:

$$\int_{1/\mathcal{L}(\boldsymbol{\theta}^{(t_0)})}^{1/\mathcal{L}(\boldsymbol{\theta}^{(t)})}\big(\log u\big)^{2/L-2}du \geq L^2\tilde{\gamma}\big(\boldsymbol{\theta}^{(t_0)}\big)^{2/L}(t-t_0)\overset{t\to\infty}{\to}\infty.$$

Therefore, $1/\mathcal{L}(\boldsymbol{\theta}^{(t)})$ must diverge when $L \geq 1$, which implies $\mathcal{L}(\boldsymbol{\theta}^{(t)}) \to 0$. Due to the L-homogeneous property of $f(\boldsymbol{\theta};\cdot)$, we must also have $\|\boldsymbol{\theta}^{(t)}\|_2 \to \infty$, which completes **S1**.

Now we show the results for **S2**. We design the scaling factor to be: $\alpha(t) = \min_{u,i} q_{u,i}(\boldsymbol{\theta}^{(t)})^{1/L}$. Consequently, $\nabla_{\boldsymbol{\theta}} q_{u,i}(\boldsymbol{\theta}^{(t)})/\alpha(t)^{L-1} = \nabla_{\boldsymbol{\theta}} q_{u,i}\big(\alpha(t)\boldsymbol{\theta}^{(t)}\big)$.

We then construct the Lagrange multipliers as:

$$\lambda_{u,i}(t) = \alpha(t)^{L-2}\|\boldsymbol{\theta}^{(t)}\|_2\exp\big(-q_{u,i}(\boldsymbol{\theta}^{(t)})\big)/\|\frac{d\boldsymbol{\theta}^{(t)}}{dt}\|_2.$$

Using the results in S1, and by straightforward calculations, we obtain:

$$\Big\|\alpha(t)\boldsymbol{\theta}^{(t)} - \sum_{(u,i)\in\mathcal{D}_{\text{train}}}\lambda_{u,i}y_{u,i}\nabla f\big(\alpha(t)\boldsymbol{\theta}^{(t)};(u,i)\big)\Big\|_2^2 \leq \frac{2}{\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)^{2/L}}\left(1-\Big\langle\frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2},\frac{d\boldsymbol{\theta}^{(t)}/dt}{\|d\boldsymbol{\theta}^{(t)}/dt\|_2}\Big\rangle\right) \tag{A.13}$$

and

$$\lambda_{u,i}(t)\Big(y_{u,i}f\big(\alpha(t)\boldsymbol{\theta}^{(t)};(u,i)\big)-1\Big) \lesssim \frac{2e|\mathcal{D}_{\text{train}}|}{L\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)^{2/L+1}\|\boldsymbol{\theta}^{(t)}\|_2^2}. \tag{A.14}$$

According to the game plan, we then need to show $\Big\langle\frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2},\frac{d\boldsymbol{\theta}^{(t)}/dt}{\|d\boldsymbol{\theta}^{(t)}/dt\|_2}\Big\rangle \to 1$ to show (A.13)$\to 0$, since (A.14)$\to$ 0 is implied by S1. First notice that $\big\langle\frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2},\frac{d\boldsymbol{\theta}^{(t)}/dt}{\|d\boldsymbol{\theta}^{(t)}/dt\|_2}\Big\rangle \leq 1$ by the Cauchy-Schwartz inequality. We then need to show it is ≥ 1 as $t\to\infty$. Using the results in **S1**, we have:

$$\begin{aligned}\frac{d\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)}{dt} &\geq \frac{\|\boldsymbol{\theta}^{(t)}\|_2^2}{Q(t)}\cdot\Big\|\frac{d\tilde{\boldsymbol{\theta}^{(t)}}}{dt}\Big\|_2^2 = L\Big\|\frac{\|\boldsymbol{\theta}^{(t)}\|_2^2}{LQ(t)}\frac{d\tilde{\boldsymbol{\theta}^{(t)}}}{dt}\Big\|_2^2\cdot\frac{LQ(t)}{\|\boldsymbol{\theta}^{(t)}\|_2^2}\\ &= L\Big\|\frac{\|\boldsymbol{\theta}^{(t)}\|_2^2}{LQ(t)}\cdot\frac{d\tilde{\boldsymbol{\theta}^{(t)}}}{dt}\Big\|_2^2\cdot\frac{d\log\|\boldsymbol{\theta}^{(t)}\|_2}{dt}.\end{aligned} \tag{A.15}$$

Since

$$\begin{aligned}\Big\|\frac{\|\boldsymbol{\theta}^{(t)}\|_2^2}{LQ(t)}\frac{d\tilde{\boldsymbol{\theta}^{(t)}}}{dt}\Big\|_2^2 &= \Big\|\frac{\|\boldsymbol{\theta}^{(t)}\|_2^2}{LQ(t)}\cdot\frac{1}{\|\boldsymbol{\theta}^{(t)}\|_2}\Big(\mathbf{I}-\tilde{\boldsymbol{\theta}^{(t)}}\tilde{\boldsymbol{\theta}^{(t)}}^{\intercal}\Big)\frac{d\boldsymbol{\theta}^{(t)}}{dt}\Big\|_2^2\\ &= \frac{\Big\|\frac{d\boldsymbol{\theta}^{(t)}}{dt}\Big\|_2^2-\Big\langle\tilde{\boldsymbol{\theta}^{(t)}},\frac{d\boldsymbol{\theta}^{(t)}}{dt}\Big\rangle^2}{\Big\langle\tilde{\boldsymbol{\theta}^{(t)}},\frac{d\boldsymbol{\theta}^{(t)}}{dt}\Big\rangle^2}\\ &= \Big\langle\frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2},\frac{d\boldsymbol{\theta}^{(t)}/dt}{\|d\boldsymbol{\theta}^{(t)}/dt\|_2}\Big\rangle^{-2}-1,\end{aligned} \tag{A.16}$$

combining (A.15) and (A.16), we have:

$$\Big\langle\frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2},\frac{d\boldsymbol{\theta}^{(t)}/dt}{\|d\boldsymbol{\theta}^{(t)}/dt\|_2}\Big\rangle \geq \sqrt{1+\frac{d\log\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)/dt}{L\cdot d\log\|\boldsymbol{\theta}^{(t)}\|_2/dt}} \geq \sqrt{1+\frac{\epsilon(t)}{L}} \text{ for some } \epsilon(t)\geq 0,$$

because both $d\log\|\boldsymbol{\theta}^{(t)}\|_2/dt \geq 0$ and $d\log\tilde{\gamma}\big(\boldsymbol{\theta}^{(t)}\big)/dt \geq 0$. Hence, by the previous argument, we have $\lim_{t\to\infty}\Big\langle\frac{\boldsymbol{\theta}^{(t)}}{\|\boldsymbol{\theta}^{(t)}\|_2},\frac{d\boldsymbol{\theta}^{(t)}/dt}{\|d\boldsymbol{\theta}^{(t)}/dt\|_2}\Big\rangle = 1$. By showing the resutls in **S1**, **S2** and **S3**, we see that $\alpha\boldsymbol{\theta}^{(t)}$ converges in direction to the KKT points, which completes the proof.

□

Proof for Theorem 2.

Proof. The first part of the results for NCF is a direct consequence of Corollary A.1. To show the second part for MCF, we need to consider the symmetrized setting with:

$$\mathbf{W} := \mathbf{Z}_U \mathbf{Z}_I^\intercal, \quad \tilde{\mathbf{W}} = \begin{bmatrix} \mathbf{M}_1 & \mathbf{W} \\ \mathbf{W} & \mathbf{M}_2 \end{bmatrix}, \quad \tilde{\mathbf{X}}_{u,i} = \mathbf{e}_u \mathbf{e}_i^\intercal + \mathbf{e}_i \mathbf{e}_u^\intercal,$$

where $\mathbf{M}_1$ and $\mathbf{M}_2$ are two p.s.d matrices that are irrelevant for the objective, and the definition of $\mathbf{e}_u, \mathbf{e}_i$ are provided in A.1. Notice that in the main paper, we use $\mathbf{X}$ to denote the predictor which is now given by $\mathbf{W}$. Hence, the MF parameterization can be considered by: $\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\intercal = \tilde{\mathbf{W}}$, and the objective becomes:

$$\min_{\tilde{\mathbf{Z}}} \mathcal{L}(\tilde{\mathbf{Z}}) = \sum_{(u,i)\in\mathcal{D}_{\text{train}}} \ell\Big(- y_{u,i} \Big\langle \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\intercal, \tilde{\mathbf{X}}_{u,i} \Big\rangle \Big).$$

It is easy to verify that the symmetrized MCF corresponds exactly to the original problem instance, and satisfy the conditions in Corollary A.1 under exponential or log loss. Since $\lim_{t\to\infty} \frac{\mathbf{W}^{(t)}}{\left\|\mathbf{W}^{(t)}\right\|_*} = \lim_{t\to\infty} \frac{\tilde{\mathbf{Z}}^{(t)}}{\left\|\tilde{\mathbf{Z}}^{(t)}\right\|_F} \frac{\tilde{\mathbf{Z}}^{(t)}}{\left\|\tilde{\mathbf{Z}}^{(t)}\right\|_F}^\intercal$, and the MCF predictor is convex, we conclude that the predictor of MCF converges in direction to the stationary point of:

$$\min \|\mathbf{W}\|_* \quad s.t. \quad y_{u,i}\mathbf{W}_{u,i} \geq 1, \forall (u,i) \in \mathcal{D}_{\text{train}}. \tag{A.17}$$

□

A.3 Material for Section 6

The major tools we use to show the generalization results are the Rademacher complexities. The procedure of bounding the inductive generalization error via the symmetrization technique and Talagrand's contraction inequalities are more often encountered in the literature [4]. The similar ideas can also be applied to bound the transductive generalization error, but specific modifications are required [9].

The different meaning of generalization decides the distinctive definitions of Rademacher complexities. We use $\mathcal{X}$ to denote the domain (of user and items) for the CF predictors, n_1 to denote $|\mathcal{D}_{\text{train}}|$ and n_2 to denote $|\mathcal{D}_{\text{test}}|$. We first provide the definitions of Rademacher complexities, and briefly discuss their different implications for the transductive and inductive learning.

Definition 1. *Recall that $n_1 = |\mathcal{D}_{train}|$ and $n_2 = |\mathcal{D}_{test}|$.*

- ***Transductive Rademacher complexity**. Let $\mathcal{V} \in \mathbb{R}^{n_1+n_2}$ and $p \in [0, 1/2]$, and $\epsilon_i(p)$ be i.i.d random variables such that:*

$$\epsilon_i(p) = \begin{cases} 1 & \textit{with probability} \quad p \\ -1 & \textit{with probability} \quad p \\ 0 & \textit{with probability} \quad 1-2p, \end{cases}$$

  then the trasductive Rademacher complexity of $\mathcal{V}$ is:

$$R_{n_1+n_2}(\mathcal{V}, p) = \Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)\mathbb{E}\Big\{ \sup_{\mathbf{v}\in\mathcal{V}} \boldsymbol{\epsilon}(p)^\intercal \mathbf{v} \Big\}, \tag{A.18}$$

  where $\boldsymbol{\epsilon}(p) = \big[\epsilon_1(p), \ldots, \epsilon_{n_1+n_2}(p)\big]^\intercal$.

- ***Inductive Rademacher complexity**. Let $\mathcal{F}$ be a function class with domain $\mathcal{X}$, and $\{X_i\}$ be a set of samples generated by a distribution $P_{\mathcal{X}}$ on $\mathcal{X}$. Let ϵ_i be shorthand of the same i.i.d random variables as above, with $p = 1/2$. Then the empirical Rademacher complexity of $\mathcal{F}$ is:*

$$\hat{R}_n(\mathcal{F}) = \mathbb{E}\Big\{ \sup_{f\in\mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \Big| X_1, \ldots, X_n \Big\},$$

  and the Rademacher complexity is given by: $R_n(\mathcal{F}) = \mathbb{E}_{P_{\mathcal{X}}} \hat{R}_n(\mathcal{F})$.

A important difference between the two settings is that the transductive complexity is an empirical quantity that does not depend on any underlying distributions, and it depends on both the training and testing data. The other difference is reflected in the specific formulations are:

1. transductive Rademacher complexity depends on both n_1 and n_2 because of the need to bound the test error: $\mathcal{D}_{\text{test}}$, i.e. $\text{Err}_{\mathcal{D}_{\text{test}}}(f) := \sum_{(u,i)\in\mathcal{D}_{\text{test}}} \mathbb{1}\big[y_{u,i}f(u,i) \leq 0\big]$;

2. it depends only on the outcomes' vector space rather than the underlying function space that produces the outcomes.

The different definitions of Rademacher complexity induces the two versions of contraction inequalities, which we provide in Lemma A.5 and Lemma A.6. We first prove the results in Theorem 3 for the transductive setting. Using the idea of symmetrization, the bound on the testing error for the transductive learning can be stated as in the following Corollary.

Corollary A.2 (Adapted from El-Yaniv and Pechyony [9])**.** *Let $\mathcal{V}$ be a set of real-valued vectors in $[-B, B]^{n_1+n_2}$, where $n_1 > n_2$ by our assumption. Define $Q = (1/n_1 + 1/n_2)$, $S = \frac{n_1+n_2}{(n_1+n_2-1/2)(1-n_1/2)}$. Then for all $\mathbf{v} \in \mathcal{V}$, with probability of at least $1-\delta$ over the random permutation of $\mathbf{v}$, which we deonote by $\tilde{\mathbf{v}}$, we have:*

$$\sum_{j=n_1+1}^{n_1+n_2} \tilde{\mathbf{v}}_j \leq \sum_{j=1}^{n_1} \tilde{\mathbf{v}}_j + R_{n_1+n_2}(\mathcal{V}, p_0) + Bc_0Q\sqrt{n_2} + B\sqrt{\frac{S}{2}Q\log\frac{1}{\delta}},$$

where $c_0 = \sqrt{32\log(4e)/3}$ and $p_0 = n_1n_2/(n_1+n_2)^2$.

By defining the $\mathcal{V}$ in the above corollary by the scores of the predictor, and using Lemma A.6 for contraction, we are able to show the results in Theorem 3.

Proof for Theorem 3.

Proof. Define $\mathbf{h} \in \mathcal{H}_{out} \in \mathbb{R}^{n_1+n_2}$ as the output scores of the predictor, and consider $\mathbf{v}$ in Corollary A.2 as $\ell\big(y_{u,i}f(\boldsymbol{\theta};(u,i))\big)$ where $\ell(u) = \mathbb{1}[u<0]$. Define $\ell_\gamma(y_{u,i}f(\boldsymbol{\theta},\mathbf{x}_{u,i}))$ to be the margin loss: $\ell_\gamma(u) = \min\{1, 1-u/\gamma\}$. Note that the margin loss is an upper bound on the classification error.

Therefore, using the results in Corollary A.2 and Lemma A.6, for any fixed $\gamma > 0$ and $\mathbf{h} \in \mathcal{H}_{out}$, with probability of at least $1-\delta$ over the random splits of $\mathcal{D}$:

$$\frac{1}{n_2}\sum_{(u,i)\in\mathcal{D}_{\text{test}}} \mathbb{1}\big[y_{u,i}f\big(\boldsymbol{\theta},(u,i)\big) < 0\big] \leq \frac{1}{n_1}\sum_{(u,i)\in\mathcal{D}_{\text{train}}} \mathbb{1}\big[y_{u,i}f\big(\boldsymbol{\theta},(u,i)\big) < \gamma\big] + \frac{R_{n_1+n_2}(\mathcal{H}_{out},p_0)}{\gamma} + c_0Q\sqrt{n_2} + \sqrt{\frac{S}{2}Q\log\frac{1}{\delta}}.$$

We first show the bound for the transductive Rademacher complexity for NCF. Recall that $\mathcal{H}_{out}$ for NCF is given by the form of: $\mathbf{W}_1\sigma(\mathbf{W}_2\sigma(\ldots\sigma(\mathbf{W}_q\sigma(\mathbf{W}_{q+1}\mathbf{X}_{u,i}))))$, where $\mathbf{W}_{q+1}$ is given by (A.2) or (A.3), with $\max_{(u,i)\in\mathcal{D}}\|\mathbf{z}_u + \mathbf{z}_i\|_2 \leq B_{\text{NCF}}$ for NCF with addition, and $\|\mathbf{W}_i\|_F \leq \lambda_i$ for $i = 1,\ldots,q$. We denote the output of the k^{th} layer by $\mathbf{H}^k_{out} \in \mathcal{H}^k_{out}$. It holds that:

$$R_{n_1+n_2}(\mathcal{H}_{out},p_0) = \big(\frac{1}{n_1}+\frac{1}{n_2}\big)\mathbb{E}\Big\{\sup_{\|\mathbf{W}_1\|_F\leq\lambda_1}\sum_{(u,i)\in\mathcal{D}_{\text{train}}}\epsilon_{u,i}\big[\mathbf{W}_1\mathbf{H}^{(q-1)}_{out}\big]_{u,i}\Big\}$$

$$\leq \lambda_1\big(\frac{1}{n_1}+\frac{1}{n_2}\big)\mathbb{E}\Big\{\sup_{\|\mathbf{W}_2\|_F\leq\lambda_2}\sum_{(u,i)\in\mathcal{D}_{\text{train}}}\epsilon_{u,i}\big[\mathbf{W}_2\mathbf{H}^{(q-2)}_{out}\big]_{u,i}\Big\} \text{ (applying Lemma A.6 on ReLU)}$$

recursively apply the peeling argument

$$\leq \prod_{i=1}^{q}\lambda_i\big(\frac{1}{n_1}+\frac{1}{n_2}\big)\mathbb{E}\Big\{\sup_{\max\|\mathbf{z}_u+\mathbf{z}_i\|_2\leq B_{\text{NCF}}}\big\langle\mathbf{z}_u+\mathbf{z}_i,\sum_{(u,i)\in\mathcal{D}_{\text{train}}}\epsilon_{u,i}\mathbf{X}_{u,i}\big\rangle\Big\}\text{(under addition, for example)}$$

$$\leq B_{\text{NCF}}\prod_{i=1}^{q}\lambda_i\big(\frac{1}{n_1}+\frac{1}{n_2}\big)\mathbb{E}\Big\|\sum_{(u,i)\in\mathcal{D}_{\text{train}}}\epsilon_{u,i}\mathbf{X}_{u,i}\Big\|_2, \tag{A.19}$$

where we use $\epsilon_{u,i}$ as a shorthand for $\epsilon_{u,i}(p_0)$. By Jensen's inequality, the last line is upper-bounded by:

$$B_{\text{NCF}} \prod_{i=1}^{q} \lambda_i \big(\frac{1}{n_1} + \frac{1}{n_2}\big) \sqrt{\sum_{(u,i)\in\mathcal{D}_{\text{train}}} \mathbb{E}\big[\epsilon_{u,i}(p_0)^2\big] \|\mathbf{X}_{u,i}\|_2} \leq B_{\text{NCF}} \prod_{i=1}^{q} \lambda_i \frac{n_1+n_2}{n_1 n_2},$$

where it is easy to compute that: $\mathbb{E}\big[\epsilon_{u,i}(p_0)^2\big] = \frac{2n_1n_2}{(n_1+n_2)^2}$. By plugging in the relation between n_1 and n_2, we obtain the result stated in Theorem 3 for NCF.

We then show the results for MCF. We use $\mathbf{\Sigma}(p)$ to denote the matrix of transductive Rademacher random variables such that $\mathbf{\Sigma}_{u,i}(p) = \epsilon_{u,i}(p)$. For $\mathbf{H} := \mathbf{Z}_U \mathbf{Z}_I^\intercal \in \mathcal{H}_{out}$ under $\big\|\mathbf{Z}_U \mathbf{Z}_I^\intercal\big\|_* \leq \lambda_{\text{nuc}}$, we have:

$$\begin{aligned}
R_{n_1+n_2}(\mathcal{H}_{out}, p_0) &= \big(\frac{1}{n_1} + \frac{1}{n_2}\big)\mathbb{E}\Big\{ \sup_{\mathbf{H}:\|\mathbf{H}\|_*\leq\lambda_{\text{nuc}}} \sum_{(u,i)\in\mathcal{D}_{\text{train}}} \mathbf{\Sigma}_{u,i}\mathbf{H}_{u,i} \Big\} \\
&\leq \lambda_{\text{nuc}} \big(\frac{1}{n_1} + \frac{1}{n_2}\big)\mathbb{E}\big\|\mathbf{\Sigma}\big\|_{sp} \text{ (by Hölder inequality, where } \|\cdot\|_{sp} \text{ is the spectral norm)} \\
&\lesssim \lambda_{\text{nuc}} \frac{(n_1+n_2)\sqrt{|\mathcal{I}|}\sqrt[4]{\log|\mathcal{U}|}}{n_1 n_2} \text{ (by Lemma A.7).}
\end{aligned} \tag{A.20}$$

The second line holds because nuclear norm is the dual of the spectral norm. Again we plug in the relation between n_1 and n_2 and obtain the stated result for MCF. □

We move on to proving the generalization results for the inductive CF. We first state a useful corollary for inductive learning, when the training and testing distribution are different.

Corollary A.3. *Consider an arbitrary function class $\mathcal{F}$ such that $\forall f \in \mathcal{F}$ we have $\sum_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})| \leq C$. Then, with probability at least $1-\delta$ over the sample, for all margins $\gamma > 0$ and all $f \in \mathcal{F}$ we have,*

$$\begin{aligned}
&P_{\textit{test}}\Big(yf(\mathbf{x}) \leq 0\Big) \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \eta(\mathbf{x}_{u,i}) \mathbb{1}\big(y_i f(\mathbf{x}_{u,i}) < \gamma\big) + 4\frac{R_{n,\boldsymbol{\eta}}(\mathcal{F})}{\gamma} + \sqrt{\frac{\log(\log_2 \frac{4C}{\gamma})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},
\end{aligned} \tag{A.21}$$

where $\eta(\mathbf{x}_{u,i}) = P_{\textit{test}}(\mathbf{x}_{u,i})/P_{\textit{train}}(\mathbf{x}_{u,i})$ gives the importance weighting, and $R_{n,\boldsymbol{\eta}}(\mathcal{F}) = \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \eta(\mathbf{x}_{u,i}) f(\mathbf{x}_{u,i})\epsilon_i\Big]$ is the weighted Rademacher complexity.

Proof. This corollary is adapted from the more general Theorem 1 of [16] by considering the deviation of the testing distribution from the training distribution. The stated result is then obtained following the Theorem 5 of [15]. □

Therefore, the key step for proving the results in Theorem 4 is to bound the weighted Rademacher complexity for NCF and MCF.

Proof for Theorem 4.

Proof. We first show the results for NCF, where we denote the predictor family by $\mathcal{F}_{\text{NCF}}$. Here, using the similar setup from Theorem 1 of [11], and combining the same arguments from the proof of Theorem 3, we arrive at

$$n_1 R_{n_1,\boldsymbol{\eta}}(\mathcal{F}_{\text{NCF}}) \leq \frac{1}{\lambda}\log\Big(2^q \cdot \mathbb{E}_{\boldsymbol{\epsilon}}\Big(M\lambda \|\sum_{i=1}^{n_1} \epsilon_i \eta(\mathbf{x}_{u,i})\mathbf{x}_{u,i}\|\Big)\Big),$$

where $M = B_{\text{NCF}} \prod_{h=1}^{q} \lambda_i$. Consider $Z := M \cdot \|\sum_{i=1}^{n_1} \epsilon_i \eta(\mathbf{x}_{u,i})\mathbf{x}_{u,i}\|$ that is a random function of the n_1 Rademacher variables. Then

$$\frac{1}{\lambda}\log\Big\{2^q \mathbb{E}\exp(\lambda Z)\Big\} = \frac{q\log(2)}{\lambda} + \frac{1}{\lambda}\log\{\mathbb{E}\exp\lambda(Z - \mathbb{E}Z)\} + \mathbb{E}Z.$$

By Jensen's inequality, we have

$$\mathbb{E}[Z] \leq M\sqrt{\mathbb{E}_{\boldsymbol{\epsilon}}\|\sum_{i=1}^{n_1}\epsilon_i\eta(\mathbf{x}_{u,i})\mathbf{x}_{u,i}\|^2} = M\sqrt{\sum_{i=1}^{n_1}\eta(\mathbf{x}_{u,i})^2\|\mathbf{x}_{u,i}\|^2}.$$

In addition, we note that

$$Z(\epsilon_1,\ldots,\epsilon_i,\ldots,\epsilon_{n_1}) - Z(\epsilon_1,\ldots,-\epsilon_i,\ldots,\epsilon_{n_1}) \leq 2M\eta(\mathbf{x}_{u,i})\|\mathbf{x}_{u,i}\|.$$

By the bounded-difference condition [5], Z is a sub-Gaussian with variance factor $v = \frac{1}{4}\sum_{i=1}^{n_1}(2M\eta(\mathbf{x}_{u,i})\|\mathbf{x}_{u,i}\|)^2 = M^2\sum_{i=1}^{n_1}\eta(\mathbf{x}_{u,i})^2\|\mathbf{x}_{u,i}\|^2$. So

$$\frac{1}{\lambda}\{\mathbb{E}\exp\lambda(Z-\mathbb{E}Z)\} \leq \frac{\lambda M^2\sum_{i=1}^{n_1}\eta(\mathbf{x}_{u,i})^2\|\mathbf{x}_{u,i}\|^2}{2}.$$

Taking $\lambda = \frac{\sqrt{2\log(2)q}}{M\sqrt{\sum_{i=1}^{n_1}\eta(\mathbf{x}_{u,i})^2\|\mathbf{x}_{u,i}\|^2}}$, it follows that

$$\begin{aligned}&\frac{1}{\lambda}\{2^q\mathbb{E}\exp\lambda Z\}\\ &\leq M(\sqrt{2\log(2)q}+1)\sqrt{\sum_{i=1}^{n_1}\eta(\mathbf{x}_{u,i})^2\|\mathbf{x}_{u,i}\|^2} \leq \sqrt{n_1}CM(\sqrt{2\log(2)q}+1)\sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}\eta(\mathbf{x}_{u,i})^2},\end{aligned} \tag{A.22}$$

where $C = 1$ for NCF-c and $C = \sqrt{2}$ for NCF-a. By law of large number, $\frac{1}{n_1}\sum_{i=1}^{n_1}\eta(\mathbf{x}_{u,i})^2 = D(P_{\text{test}}||P_{\text{train}}) + 1 + o(\frac{1}{\sqrt{n_1}})$. The desired result for NCF follows.

Then we show the result for MCF. Here, we provide a general result for generalization of using importance weighting under distribution shift. We assume the training distribution is P, testing distribution is Q, and the weight for any (u,i) instance is therefore given by: $w_i = Q(i)/P(i)$. We define $\mathcal{N}(\frac{1}{n},\mathcal{F},\ell_2^n)$ as the $\frac{1}{n}$-covering number for $\mathcal{F}$ in $\|\cdot\|_2$ based on n i.i.d samples from P, and $d(P\|Q) = \int_{\mathcal{S}_Q}(dP/dQ)dP$ is a divergence measure, where $\mathcal{S}$ is used to denote the support of a distribution. We use $\mathbb{E}_Q R(f)$ to denote the testing risk, and use $\mathbb{E}_{P_{n,w}}R(f)$ to denote the weighted empirical training risk.

Our proof leverages the classical "double sampling" technique from Anthony and Bartlett [1]. We use $\vec{\mathbf{z}} = [\mathbf{z}_1,\ldots,\mathbf{z}_n]$ to denote the observed samples, and $\vec{\mathbf{z}}' = [\mathbf{z}'_1,\ldots,\mathbf{z}'_n]$ to denote an i.i.d copy of $\vec{\mathbf{z}}$. We first define by:

$$UB_1(f,\vec{\mathbf{z}},t) = \frac{1}{n}\sum_{i=1}^{n}w_i\ell_f(\mathbf{z}_i) + \frac{3Mt}{n} + \sqrt{\frac{2d(P\|Q)t}{n}},$$

and

$$UB_2(f,\vec{\mathbf{z}},t) = \frac{1}{n}\sum_{i=1}^{n}w_i\ell_f(\mathbf{z}_i) + \frac{9Mt}{n} + \sqrt{\frac{18d(P\|Q)t}{n}}.$$

Given $f\in\mathcal{F}$, let $A := \mathbb{E}_Q R(f) + \frac{6Mt}{n} + \sqrt{\frac{8d(P\|Q)t}{n}}$it holds:

$$\begin{aligned}\mathbb{P}\big(UB_2(f,\vec{\mathbf{z}}',t) \leq UB_1(f,\vec{\mathbf{z}},t)\big) &\leq \mathbb{P}\big(UB_2(f,\vec{\mathbf{z}}',t)\leq A\big) + \mathbb{P}\big(UB_1(f,\vec{\mathbf{z}},t)\geq A\big)\\ &\leq 2\mathbb{P}\Big(\big|\mathbb{E}_Q R(f) - \frac{1}{n}\sum w_i\ell_f(\mathbf{z}_i)\big| \geq \frac{3Mt}{n} + \sqrt{\frac{2d(P\|Q)t}{n}}\Big)\\ &\leq 4e^{-t},\end{aligned}$$

where the last line follows from Lemma A.8. Next, we define $\mathcal{C}(\epsilon,\ell\circ\mathcal{F},\ell_1(P_{n,w}))$ be the ϵ-cover of $\ell\circ\mathcal{F}$ with the empirical ℓ_1 norm under $P_{n,w}$ such that for any $f\in\ell\circ\mathcal{F}$, there exists $\tilde{f}$ in $\mathcal{C}(\epsilon,\ell\circ\mathcal{F},\ell_1^n)$: $\big|\frac{1}{n}\sum w_i f(\mathbf{z}_i) - \frac{1}{n}\sum w_i\tilde{f}(\mathbf{z}_i)\big| \leq \epsilon$, for $(\mathbf{z}_1,\ldots,\mathbf{z}_n)$ sampled i.i.d from P. It then

holds:

$$
\begin{aligned}
&\mathbb{P}\big(\exists f \in \mathcal{F} : \mathbb{E}_Q R(f) \geq UB_2(f, \vec{\mathbf{z}}, t) + \epsilon\big) \\
&= \mathbb{E}_{\vec{\mathbf{z}}} \sup_{f \in \mathcal{F}} I[\mathbb{E}_Q R(f) \geq UB_2(f, \vec{\mathbf{z}}, t) + \epsilon] \\
&\overset{(a)}{\leq} \mathbb{E}_{\vec{\mathbf{z}}} \sup_{f \in \mathcal{F}} I[\mathbb{E}_Q R(f) \geq UB_2(f, \vec{\mathbf{z}}, t) + \epsilon] \cdot 2\mathbb{E}_{\vec{\mathbf{z}}'} I[UB_1(f, \vec{\mathbf{z}}', t) \geq \mathbb{E}_Q R(f)] \\
&\leq 2\mathbb{E}_{\vec{\mathbf{z}}, \vec{\mathbf{z}}'} \sup_{f \in \mathcal{F}} I[UB_1(f, \vec{\mathbf{z}}', t) \geq UB_2(f, \vec{\mathbf{z}}, t) + \epsilon] \\
&\overset{(b)}{\leq} 2\mathbb{P}_{\sigma(\vec{\mathbf{z}}, \vec{\mathbf{z}}')}\big(\exists \tilde{f} \in \mathcal{C}(\epsilon, \ell \circ \mathcal{F}, \ell_1(P_{n,w})) : UB_1(f, \sigma(\vec{\mathbf{z}}, \vec{\mathbf{z}}'), t) \geq UB_2(f, \sigma(\vec{\mathbf{z}}, \vec{\mathbf{z}}'), t)\big) \\
&\leq 8\mathcal{N}(\epsilon, \ell \circ \mathcal{F}, \ell_1(P_{n,w})) \cdot e^{-t},
\end{aligned}
$$

where (a) follows from the fact that $\mathbb{E}_{\vec{\mathbf{z}}'} I[UB_1(f, \vec{\mathbf{z}}', t) \geq \mathbb{E}_Q R(f)] \geq \frac{1}{2}$ as suggested by Lemma A.8, and in step (b) we let $\sigma(\vec{\mathbf{z}}, \vec{\mathbf{z}}')_i$ takes the value of $\mathbf{z}_i, \mathbf{z}_i'$ with equal probability, and the inequality follows from the definition of the ϵ cover. Notice that $\mathcal{N}(\epsilon, \ell \circ \mathcal{F}, \ell_1(P_{n,w})) \leq \mathcal{N}(\epsilon/M, \ell \circ \mathcal{F}, \ell_2^n)$. We take $\epsilon = \frac{1}{n}$, which solves for $t = c \log \frac{1}{\delta} + \log \mathcal{N}(\epsilon/M, \ell \circ \mathcal{F}, \ell_2^n)$ for some constant c. We use $\mathcal{N}_2(\epsilon, \mathcal{F})$ as a shorthand to denote the covering number under the empirical ℓ_2 norm.

By rearranging terms, we have that for any $\delta > 0$, with probability at least $1 - \delta$, it holds:

$$
\mathbb{E}_Q R(f) \lesssim \mathbb{E}_{P_{n,w}} R(f) + \frac{M\big(\log \frac{1}{\delta} + \log \mathcal{N}_2(\frac{1}{n}, \mathcal{F})\big)}{n} + \sqrt{\frac{Md(P\|Q)\big(\log \frac{1}{\delta} + \log \mathcal{N}_2(\frac{1}{n}, \mathcal{F})\big)}{n}}, \tag{A.23}
$$

when the loss function ℓ is Lipschitz and we ignore the constants. Hence, the remaining task is to bound the covering number of the matrix factorization class $\mathcal{F}_{\text{MCF}}$ with a bounded nuclear norm. When $\|\mathbf{X}\|_F = 1$, the nunclear norm is strictly a lower bound of the matrix rank. Therefore, we use the covering number of low-rank matrix as a upper bound, which according to Lemma 3.1 of , if $\text{rank}(\mathbf{X}) \leq \lambda_{\text{nuc}}$, then the covering number for $\mathcal{F}_{\text{MCF}}$ under the matrix Frobenius norm obeys:

$$
\mathcal{N}(\epsilon, \mathcal{F}_{\text{MCF}}, \|\cdot\|_F) \leq (9/\epsilon)^{(|\mathcal{U}|+|\mathcal{I}|+1)\lambda_{\text{nuc}}},
$$

which we plug back to (A.23) and obtain the desired result. □

Discussion: the tightness of the generalization bounds.

When proving the bounds for both the transductive and inductive CF, we use the standard generalization results based on Rademacher complexity, according to Bartlett and Mendelson [4] and El-Yaniv and Pechyony [9]. Their results rely on the following components:

- a symmetrization argument to bound the testing error;
- the Mcdiarmid's inequality for bounded difference;
- the Rademacher contraction inequalities (Lemma A.5 and Lemma A.6).

All these results are known to be tight, so the question narrows down to the tightness of our bounds on the Rademacher complexities. To see that the provided result for NCF are tight up to a constant factor of $\sqrt{q}$, we simply consider the following construction: $\mathbf{x}_{u,i} \mapsto \lambda_1 \cdot \lambda_2 \cdots \lambda_q \cdot \sigma(\mathbf{W}_{q+1}\mathbf{x}_{u,i})$, which belongs to the general NCF family, and the worst-case scenario for computing Rademacher complexity is obvious given by:

$$
\lambda_1 \cdot \lambda_2 \cdots \lambda_q \cdot \sigma\Big(\max_{u,i:\|\mathbf{z}_u + \mathbf{z}_i\|_2} \mathbf{z}_u + \mathbf{z}_i\Big),
$$

where we use NCF-a for example. Here, all the training samples are $(u, i) = \arg\max_{u,i:\|\mathbf{z}_u + \mathbf{z}_i\|_2}$. Consequently, the Rademacher complexity is at least $B_{\text{NCF}} \prod_{i=1}^q \lambda_i$.

On the MCF side, it is pointed out by [3] that for the spectral norm of Rademacher matrix, the dependency on $\sqrt{|\mathcal{I}|}$ and $\sqrt[4]{\log |\mathcal{U}|}$ are inevitable, and therefore our result for transdutive MCF is also tight up to constants. As for the inductive setting, we refer to the results in Candès and Recht [6]

that the bound with $\sqrt{(\sqrt{|\mathcal{D}|}+\sqrt{|\mathcal{U}|})}/\sqrt{n_1}$ is not improvable. Notice that they assume a uniform distribution over the matrix indices, where our result is distribution-free. However, despite several minor discrepancies, their setting can be recognized as a special case of our problem, and thus we conjecture that our results for MCF can be further tightened to get rid of the $\log n$ dependency, e.g. by deriving the covering number for nuclear-norm-constraint matrices instead of using the existing result for low-rank matrices.

A.4 Auxiliary Lemmas

Lemma A.3 (Adapted from Cho and Saul [7]). *Define the shorthand $\varsigma(u) := \frac{1}{2}(1+sign(u))$. For $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, the n^{th} order arc-cosine kernel is defined as:*

$$K_n(\mathbf{x},\mathbf{y}) = \frac{1}{\pi}\|\mathbf{x}\|_2^n\|\mathbf{y}\|_2^n J_n(\theta),$$

where $J_n(\theta) = (-1)^n(\sin\theta)^{2n+1}\big(\frac{1}{\sin\theta}\frac{d}{d\theta}\big)^n\big(\frac{\pi-\theta}{\sin\theta}\big)$. Then the arc-cosine kernel has an equivalent integral representation:

$$K_n(\mathbf{x},\mathbf{y}) = 2\int d\mathbf{w}\frac{\exp(-\frac{\|\mathbf{w}\|_2^2}{2})}{(2\pi)^{d/2}}\varsigma(\mathbf{w}^\intercal\mathbf{x})\varsigma(\mathbf{w}^\intercal\mathbf{y})(\mathbf{w}^\intercal\mathbf{x})^n(\mathbf{w}^\intercal\mathbf{y})^n.$$

For instance, when $n=0$, $K_0(\mathbf{x},\mathbf{y}) = 1-\frac{1}{\pi}\cos^{-1}\frac{\mathbf{x}^\intercal\mathbf{y}}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2}$.

Lemma A.4 (Euler's Theorem for homogeneous functions). *If $f(\boldsymbol{\theta},\cdot)$ is L-homogeneous, then:*

- $\nabla f(\alpha\boldsymbol{\theta},\cdot) = \alpha^{L-1}\nabla f(\boldsymbol{\theta},\cdot)$,
- $\big\langle\boldsymbol{\theta},\nabla f(\boldsymbol{\theta},\cdot)\big\rangle = L\cdot f(\boldsymbol{\theta},\cdot)$,

if $f(\boldsymbol{\theta},\cdot)$ is differentiable.

The proof for the Lemma is relatively standard, so we do not repeat it here.

Lemma A.5 (Ledoux and Talagrand [17]). *Let $f:\mathbb{R}_+\to:\mathbb{R}_+$ be convex and increasing. Let $\phi_i:\mathbb{R}\to\mathbb{R}$ satisfy $\phi_i(0)=0$ and is Lipschitz with constant L. Then for any $\mathcal{V}\in\mathbb{R}^n$:*

$$\mathbb{E}f\Big(\frac{1}{2}\sup_{\mathbf{v}\in\mathcal{V}}\Big|\sum_{i=1}^n\epsilon_i\phi_i(\mathbf{v}_i)\Big|\Big)\le\mathbb{E}f\Big(L\cdot\frac{1}{2}\sup_{\mathbf{v}\in\mathcal{V}}\Big|\sum_{i=1}^n\epsilon_i\mathbf{v}_i\Big|\Big),$$

where ϵ_i are the standard Rademacher random variables.

The similar contraction result in the transductive setting is given as below.

Lemma A.6 (Lemma 5 of El-Yaniv and Pechyony [9]). *Consider $\mathcal{V}\in\mathbb{R}^{n_1+n_2}$. Let $f,g:\mathbb{R}\to\mathbb{R}$ be such that for all $1\le i\le n_1+n_2$ and $\mathbf{v},\mathbf{v}'\in\mathcal{V}$, $\big|f(\mathbf{v}_1)-f(\mathbf{v}_1')\big|\le L\big|g(\mathbf{v}_1)-g(\mathbf{v}_1')\big|$, then:*

$$\mathbb{E}\Big\{\sup_{\mathbf{v}\in\mathcal{V}}\sum_{i=1}^{n_1+n_2}\epsilon_i(p)f(\mathbf{v}_i)\Big\}\le\mathbb{E}\Big\{L\cdot\sup_{\mathbf{v}\in\mathcal{V}}\sum_{i=1}^{n_1+n_2}\epsilon_i(p)g(\mathbf{v}_i)\Big\},$$

for any $p\in[0,1/2]$.

Lemma A.7 (Concentration of random matrices.). *Let $\mathbf{X}$ be a $m\times n$ matrix with $m>n$.*

- *By Bandeira et al. [3], if $\mathbf{X}$ is composed of independent Rademacher random variables, then:*
$$\mathbb{E}\|\mathbf{X}\|_{sp}\lesssim\sqrt[4]{\log n}\sqrt{m}.$$
- *By Tropp [21], if $\mathbf{X}$ is composed of independnet zero-mean random variables, then:*
$$\mathbb{E}\|\mathbf{X}\|_{sp}\lesssim\max_i\sqrt{\sum_j\mathbb{E}\mathbf{X}_{i,j}^2}+\max_j\sqrt{\sum_i\mathbb{E}\mathbf{X}_{i,j}^2}+\sqrt[4]{\sum_{i,j}\mathbb{E}\mathbf{X}_{i,j}^4}$$

Lemma A.8. *Let P and Q be the training and target distribution supported on $\mathcal{S}_P, \mathcal{S}_Q \subseteq \mathcal{D}$, and $w(u,i) = \frac{Q(u,i)}{P(u,i)}$, for $(u,i) \in \mathcal{S}_P \cap \mathcal{S}_Q$. $\mathcal{D}_n$ consists of training instances sampled i.i.d from P. Given a single hypothesis $f \in \mathcal{F}$, suppose $w_{ui} \in (0,1)$, for any $\delta > 0$, it holds with probability at least $1-\delta$ that:*

$$\mathbb{E}_Q R(f) \leq \mathbb{E}_{P_{n,w}} R(f) + \frac{2\log\frac{1}{\delta}}{3n} + \sqrt{\frac{2d(P\|Q)\log\frac{1}{\delta}}{n}},$$

where $d(P\|Q) = \int_{\mathcal{S}_Q}(dP/dQ)dP$.

The above result for importance weighting of a single hypothesis is stated in the Theorem 1 of [8].

A.5 Experiment details

Both MCF and NCF are implemented in Tensorflow 2.3, and the computation infrastructure involves a Nvidia Tesla V100 GPU with 32 Gb memory. We provide a kernel SVM implementation using the python Scikit-learn package, and a CVX implementation using the Python CVXOPT API. The code is also provided as a part of the supplementary material.

As we mentioned in the main paper, we use the log loss $\ell(u) = \log(1+\exp(-u))$ for all our experiments. The metrics we consider, i.e. the ranking AUC, top-k hitting rate and NDCG are computed from a scan over all the possible candidates. Since there is only one relevant item (the last interacted item) for each user, then according to Rendle [19], the metric computations are simplified to:

- Suppose the ranking of the relevant item, among the whole set of candidate items: $\tilde{\mathcal{I}}(u) = \mathcal{I} - \{i \,|\, (u,i) \in \mathcal{D}_{\text{train}}\}$, is given by r for user u. Then the ranking AUC for user u is given by: AUC(u)= $\big(|\tilde{\mathcal{I}}(u)| - r\big)/\big(|\tilde{\mathcal{I}}(u)| - 1\big)$;
- The top-k hitting rate for user u is: HR@k(u)= $\mathbb{1}[r \leq k]$;
- the top-k NDCG for user u is: NDCG@k(u)= $\mathbb{1}[r \leq k]\frac{1}{\log_2(r+1)}$.

Then the overall metric is computed by taking the population average.

Since a large proportion of our discussion surrounds the gradient descent, we use the SGD optimizer unless otherwise specified.

Experiment for Figure 1.

We point out that the data is relatively small after the subsampling so that the performance can vary significantly across different sampled datasets. Therefore, we do not repeat the experiments on different sampled datasets, but on the different splits (of generating negative samples). We point out that sampling the movies by popularity is necessary, because otherwise, the obtained data will be very sparse and thus not representative of the original dataset. After sampling 200 movies and 200 users, we end up with approximately 30,000 records for training when setting the number of negative samples to 4. Notice that this already requires a 30,000×30,000 matrix for the kernel method.

The experiment setting follows that of the transductive CF, where the negative samples are constructed via sampling without replacement, and the random splits are conducted before training. As for the positive label, we adopt the standard setting where the last user-item interaction is used for testing, and the rest are used for training. Notice that we do not need validation data in this case, because there are no tuning parameters as we fixed the dimensions and learning rate, and do not use regularizations of any kind.

Experiment for Figure 2.

We first use CVXOPT to obtain the exact solution of the convex nuclear-norm max-margin problem in Theorem 2. The optimality is reached with the duality gap $\leq 1e^{-19}$. We use the same dataset generated for the experiments in Figure 1. As we stated before, we consider the unscaled $N(0, 0.1)$ initialization, set the moderate width of $d = 32$ and the learning rate of 0.1. Again, the repetitions are over the random splits of the negative samples. The normalized margin for the nuclear-norm max-margin problem is obtained via: $\gamma_{u,i}^{\text{SVM}} = y_{u,i}\mathbf{X}_{u,i}/\|\mathbf{X}\|_*$, and the normalized margin for MCF is obtained via: $\gamma_{u,i}^{\text{MCF}} = y_{u,i}\langle \mathbf{z}_u, \mathbf{z}_i\rangle \big/ \|\mathbf{Z}_U\mathbf{Z}_I^\intercal\|_*$.

Experiment for Figure 3.

We use all the data for the inductive CF task, where the last user-item interaction is used for testing, the second-to-last is used for validation, and the rest are used for training. The negative samples are also obtained via sampling without replacement, where we fix the number of negative samples to 4 for each positive interaction. Due to the sampling without replacement, setting the number of negative samples per positive to a high value may not increase the total number of negative samples proportionally (e.g. a user may have watched 300 out of the 1,000 movies). Therefore, we do not tune the number of negative samples per positive.

We select d from $\{16, 32, 48\}$ for MCF, and $d \in \{16, 32, 48\}, d1 \in \{8, 16, 24\}$ for NCF (since we study the two-layer setting). We experiment with a learning rate of $\{0.01, 0.05, 0.1, 0.2\}$, and do not find a significant difference since we study the converged behavior after several thousand epochs. For illustration purpose, we use 0.1 as the learning rate. We make the hyper-parameter selection over one run and fix it during the rest repetitions. We find $d = 32$ and $d = 32, d_1 = 16$ gives the best performance for MCF and NCF, as we reported in Figure 3. The results reported in Figure 1 are the average over 10 random splits of the negative samples (and random initializations).

Experiment for Figure 4 and 5.

The learning of the relevance mechanism, exposure mechanism and the final data generating mechanism are stated in Section 6. When learning the relevance and exposure mechanism, we do not conduct the train/test split, since this step aims to construct the mechanisms according to the data, rather than examining how the models fit the data. When the g_{rel} and g_{expo} are given by the MCF, we use $d = 32$; and when they are given by the NCF, we use $d = 32, d_1 = 8$. We do not tune these hyperparameters due to the same reason stated above. We use the mean squared-root error (MSE) and the binary cross-entropy loss when training the relevance and exposure models. Unlike training for the CF tasks, we use the Adam optimizer with a learning rate of 0.001, which we find to work well with the MSE.

After we settle down with the learnt relevance and exposure mechanism, we generate the observed data according to the click model. Before that, we tune the μ and ρ in the relevance model to ensure the generated data has about the same sparsity as the original data. Since neither MCF nor NCF leverage the sequential information, the order by which we generate the interacted items for a specific user is not important. After we generate the click data for all the user-item pairs, we sample from the positive and negative parts with replacement to construct the training, validation and testing data, according to the empirical data distribution (which is a uniform distribution over the indices).

All the results reported in Figure 4 and 5 about NCF are from the concatenation. We observe that NCF with addition has very similar patterns in the inductive CF experiments, so we do not report its results to avoid repetition. For MCF, we select $d \in \{16, 32, 48\}$, and for NCF we select $d \in \{16, 32, 48\}, d_1 \in \{8, 16, 24\}$. We also do not experiment on using regularizations here. We repeat the generation, training, evaluation process for ten times. Each time, we tune the hyperparameters according to the validation performance. The evaluation metric we report is the biased and unbiased ranking AUC, where the biased AUC is computed in the regular fashion, and the unbiased AUC is computed via: unbiased-AUC(u,i)$= \big(|\tilde{\mathcal{I}}(u)| - r(i)\big)/\big(|\tilde{\mathcal{I}}(u)| - 1\big) \cdot \frac{1}{p(O_{u,i}=1)}$, where $r(i)$ is the ranking of item i in $\tilde{\mathcal{I}}(u)$.

References

[1] M. Anthony and P. L. Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[2] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019.

[3] A. S. Bandeira, R. Van Handel, et al. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Annals of Probability*, 44(4):2479–2506, 2016.

[4] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[5] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[6] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[7] Y. Cho and L. K. Saul. Kernel methods for deep learning. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pages 342–350, 2009.

[8] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Nips*, volume 10, pages 442–450. Citeseer, 2010.

[9] R. El-Yaniv and D. Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.

[10] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 11427–11438, 2019.

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

[12] F. Gould and J. W. Tolle. A necessary and sufficient qualification for constrained optimization. *SIAM Journal on Applied Mathematics*, 20(2):164–172, 1971.

[13] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[14] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[15] S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: risk bounds, margin bounds, and regularization. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pages 793–800, 2008.

[16] V. Koltchinskii, D. Panchenko, et al. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of statistics*, 30(1):1–50, 2002.

[17] M. Ledoux and M. Talagrand. Probability in banach spaces: isoperimetry and processes. 1991.

[18] K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[19] S. Rendle. Evaluation metrics for item recommendation under sampling. *arXiv preprint arXiv:1912.02263*, 2019.

[20] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[21] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[22] A. Virmaux and K. Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, pages 3835–3844, 2018.

[23] G. Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.