Supplementary Materials for the Paper: "On Dropout and Nuclear Norm Regularization"

A. Proofs of the Main Results

In this section, we provide the complete proofs of our main results. For notational simplicity, we define

$$\mathrm{W}_{i\to j} := \mathrm{W}_i \mathrm{W}_{i-1} \cdots \mathrm{W}_{j+1} \mathrm{W}_j,$$
$$\bar{\mathrm{W}}_{i\to j} := \frac{1}{\theta^{i-j}} \mathrm{W}_i \operatorname{diag}(\mathrm{b}_{i-1}) \mathrm{W}_{i-1} \cdots \operatorname{diag}(\mathrm{b}_{j+1}) \mathrm{W}_{j+1} \operatorname{diag}(\mathrm{b}_j) \mathrm{W}_j.$$

Since the Bernoulli random vectors are i.i.d., it holds that $\mathbb{E}_{\mathrm{b}_i}[\bar{\mathrm{W}}_{i\to j}] = \mathrm{W}_{i\to j}$. A quantity that shows up when analyzing dropout training under squared error is $\mathbb{E}[\|\mathrm{U} \operatorname{diag}(\mathrm{b}) \mathrm{Vx}\|^2]$, where the expectation is taken with respect to b, which is a Bernoulli random vector with parameter θ. The following lemma gives the closed form of this expectation.

Lemma A.1. *Let $U \in \mathbb{R}^{d_2 \times r}$, $V \in \mathbb{R}^{d_1 \times r}$, and $\mathrm{C} := \mathbb{E}[\mathrm{x}\mathrm{x}^\top]$. It holds that*

$$\mathbb{E}[\|U \operatorname{diag}(b) V\mathrm{x}\|^2] = \theta^2 \mathbb{E}[\|UV\mathrm{x}\|^2] + (\theta - \theta^2) \sum_{j=1}^{r} \|u_{:j}\|^2 \|\mathrm{C}^{\frac{1}{2}} v_{j:}\|^2.$$

The proof can be found in (Cavazza et al., 2018; Mianjy et al., 2018). Nonetheless, we provide a proof here for completeness.

Proof of Lemma A.1.

$$\mathbb{E}[\|\mathrm{U} \operatorname{diag}(\mathrm{b}) \mathrm{Vx}\|^2] = \mathbb{E} \sum_{i=1}^{d_2} \mathbb{E}_{\mathrm{b}} \left(\sum_{j=1}^{r} u_{ij} b_j \mathrm{v}_{j:}^\top \mathrm{x} \right)^2 = \mathbb{E} \sum_{i=1}^{d_2} \mathbb{E}_{\mathrm{b}} [\sum_{j,k=1}^{r} u_{ij} u_{ik} b_j b_k (\mathrm{v}_{j:}^\top \mathrm{x})(\mathrm{v}_{k:}^\top \mathrm{x})]$$
$$= \mathbb{E} \sum_{i=1}^{d_2} \sum_{j,k=1}^{r} u_{ij} u_{ik} (\theta^2 1_{j\neq k} + \theta 1_{j=k}) (\mathrm{v}_{j:}^\top \mathrm{x})(\mathrm{v}_{k:}^\top \mathrm{x}) = \theta^2 \mathbb{E}[\|\mathrm{UVx}\|^2] + (\theta - \theta^2) \mathbb{E} \sum_{i=1}^{d_2} \sum_{j=1}^{r} u_{ij}^2 (\mathrm{v}_{j:}^\top \mathrm{x})^2$$
$$= \theta^2 \mathbb{E}[\|\mathrm{UVx}\|^2] + (\theta - \theta^2) \sum_{j=1}^{r} \mathbb{E}[\mathrm{v}_{j:}^\top \mathrm{x}\mathrm{x}^\top \mathrm{v}_{j:}] \sum_{i=1}^{d_2} u_{ij}^2 = \theta^2 \mathbb{E}[\|\mathrm{UVx}\|^2] + (\theta - \theta^2) \sum_{j=1}^{r} \|\mathrm{C}^{\frac{1}{2}} \mathrm{v}_{j:}\|^2 \|\mathrm{u}_{:j}\|^2.$$

□

A.1. Properties of the explicit regularizer

Recall that training a network with dropout aims at minimizing the following *dropout objective*

$$L_\theta(\{\mathrm{W}_i\}_{i=1}^{k+1}) = \mathbb{E}_{\substack{\mathrm{b}_i \sim \mathrm{Bern}(\theta) \\ (\mathrm{x},\mathrm{y}) \sim \mathcal{D}}} [\|\mathrm{y} - \frac{1}{\theta^k} \mathrm{W}_{k+1} \operatorname{diag}(\mathrm{b}_k) \mathrm{W}_k \ldots \operatorname{diag}(\mathrm{b}_1) \mathrm{W}_1 \mathrm{x}\|^2].$$

In Proposition 2.1 we show that this objective can be decomposed into a summation of the population loss plus an *explicit regularizer*, i.e. $L_\theta(\cdot) = L(\cdot) + R(\cdot)$, and give the closed form expression for the explicit regularizer.

Proof of Proposition 2.1. We start by expanding the squared error:

$$\begin{aligned} L_\theta(\{\mathrm{W}_i\}_{i=1}^{k+1}) &= \mathbb{E}_{\substack{\mathrm{b}_i\sim\text{Bern}(\theta)\\(\mathrm{x},\mathrm{y})\sim\mathcal{D}}}[\|\mathrm{y}-\bar{\mathrm{W}}_{k+1\to 1}\mathrm{x}\|^2] \\ &= \mathbb{E}[\|\mathrm{y}\|^2] - 2\mathbb{E}[\langle \bar{\mathrm{W}}_{k+1\to 1}\mathrm{x},\mathrm{y}\rangle] + \mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to 1}\mathrm{x}\|^2] \\ &= \mathbb{E}[\|\mathrm{y}\|^2] - 2\mathbb{E}\langle \mathrm{W}_{k+1\to 1}\mathrm{x},\mathrm{y}\rangle + \frac{1}{\theta^{2k}}\mathbb{E}[\|\mathrm{W}_{k+1}\,\mathrm{diag}\,(\mathrm{b}_k)\,\mathrm{W}_k\ldots\mathrm{diag}\,(\mathrm{b}_1)\,\mathrm{W}_1\mathrm{x}\|^2] \end{aligned} \tag{10}$$

We now focus on the last term in the right hand side of Equation (10).

$$\begin{aligned} \mathbb{E}[\|\mathrm{W}_{k+1}\,\mathrm{diag}\,(\mathrm{b}_k)\,\mathrm{W}_k\ldots\mathrm{diag}\,(\mathrm{b}_1)\,\mathrm{W}_1\mathrm{x}\|^2] &= \mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to 2}\,\mathrm{diag}\,(\mathrm{b}_1)\,\mathrm{W}_1\mathrm{x}\|^2] \\ &= \theta^2\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to 2}\mathrm{W}_1\mathrm{x}\|^2] + (\theta-\theta^2)\sum_{j=1}^{d_1}\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to 2}(:,j)\|^2]\|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_1(j,:)\|^2 \end{aligned} \tag{11}$$

The second equality follows from Lemma A.1. Similarly, the first term on the right hand side of Equation (11) can be expressed as:

$$\begin{aligned} \mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to 2}\mathrm{W}_1\mathrm{x}\|^2] &= \mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to 3}\,\mathrm{diag}\,(\mathrm{b}_2)\,\mathrm{W}_{2\to 1}\mathrm{x}\|^2] \\ &= \theta^2\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to 3}\mathrm{W}_{2\to 1}\mathrm{x}\|^2] + (\theta-\theta^2)\sum_{j=1}^{d_2}\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to 3}(:,j)\|^2]\|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{2\to 1}(j,:)\|^2 \end{aligned}$$

By recursive application of the above identity and plugging the result into Equation (11), we obtain:

$$\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to 1}\mathrm{x}\|^2] = \theta^{2k}\mathbb{E}[\|\mathrm{W}_{k+1\to 1}\mathrm{x}\|^2] + (1-\theta)\sum_{i=1}^{k}\sum_{j=1}^{d_i}\theta^{2i-1}\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to i+1}(:,j)\|^2]\|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{i\to 1}(j,:)\|^2 \tag{12}$$

Plugging back the above equality into Equation (10), we get

$$\begin{aligned} L_\theta(\{\mathrm{W}_i\}) &= \|\mathrm{y}\|^2 - 2\mathbb{E}\langle W_{k+1\to 1}\mathrm{x},\mathrm{y}\rangle + \mathbb{E}[\|\mathrm{W}_{k+1\to 1}\mathrm{x}\|^2] + \frac{1-\theta}{\theta^{2k}}\sum_{i=1}^{k}\sum_{j=1}^{d_i}\theta^{2i-1}\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to i+1}(:,j)\|^2]\|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{i\to 1}(j,:)\|^2 \\ &= \mathbb{E}_{\mathrm{x}}[\|\mathrm{y}-\mathrm{W}_{k+1\to 1}\mathrm{x}\|^2] + (1-\theta)\sum_{i=1}^{k}\sum_{j=1}^{d_i}\theta^{2(i-k)-1}\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to i+1}(:,j)\|^2]\|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{i\to 1}(j,:)\|^2. \end{aligned} \tag{13}$$

It remains to calculate the terms of the form $\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to i+1}(:,j)\|^2]$ in the right hand side of Equation (13). We introduce the variable $x\sim\mathcal{N}(0,1)$ so that we can use Lemma A.1 again:

$$\begin{aligned} \mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to i+1}(:,j)\|^2] &= \mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to i+2}\,\mathrm{diag}\,(\mathrm{b}_{i+1})\,\mathrm{W}_{i+1}(:,j)x\|^2] \\ &= \theta^2\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to i+2}\mathrm{W}_{i+1}(:,j)\|^2] + (\theta-\theta^2)\mathbb{E}\sum_{l=1}^{d_{i+1}}\|\bar{\mathrm{W}}_{k+1\to i+2}(:,l)\|^2\mathrm{W}_{i+1}(l,j)^2. \end{aligned} \tag{14}$$

The first term on the right hand side of Equation (14) can be expanded as:

$$\begin{aligned} \mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to i+2}\mathrm{W}_{i+1}(:,j)\|^2] &= \mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to i+3}\,\mathrm{diag}\,(\mathrm{b}_{i+2})\,\mathrm{W}_{i+2\to i+1}(:,j)x\|^2] \\ &= \theta^2\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to i+3}\mathrm{W}_{i+2\to i+1}(:,j)\|^2] + (\theta-\theta^2)\mathbb{E}\sum_{l=1}^{d_{i+2}}\|\bar{\mathrm{W}}_{k+1\to i+3}(:,l)\|^2\mathrm{W}_{i+2\to i+1}(l,j)^2 \end{aligned}$$

By recursive application of the above equality and plugging the results into Equation (14), we get

$$\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to i+1}(:,j)\|^2] = \theta^{2(k-i)}\|\mathrm{W}_{k+1\to i+1}(:,j)\|^2 + (1-\theta)\sum_{m=1}^{k-i}\theta^{2m-1}\mathbb{E}\sum_{l=1}^{d_{i+m}}\|\bar{\mathrm{W}}_{k+1\to i+1+m}(:,l)\|^2\mathrm{W}_{i+m\to i+1}(l,j)^2$$

Plugging back the above identity into Equation (13) we get

$$
\begin{aligned}
R(\{\mathrm{W}_i\}) &= (1-\theta)\sum_{i=1}^{k}\sum_{j=1}^{d_i}\theta^{2(i-k)-1}\|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{i\to 1}(j,:)\|^2\mathbb{E}[\|\bar{\mathrm{W}}_{k+1\to i+1}(:,j)\|^2] \\
&= \frac{1-\theta}{\theta}\sum_{i=1}^{k}\sum_{j=1}^{d_i}\|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{i\to 1}(j,:)\|^2\|\mathrm{W}_{k+1\to i+1}(:,j)\|^2 \\
&+ (\frac{1-\theta}{\theta})^2\sum_{i=1}^{k}\sum_{j=1}^{d_i}\|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{i\to 1}(j,:)\|^2\sum_{m=1}^{k-i}\theta^{2(i+m-k)}\mathbb{E}\sum_{l=1}^{d_{i+m}}\mathrm{W}_{i+m\to i+1}(l,j)^2\|\bar{\mathrm{W}}_{k+1\to i+m+1}(:,l)\|^2 \\
&= \frac{1-\theta}{\theta}\sum_{i=1}^{k}\sum_{j=1}^{d_i}\|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{i\to 1}(j,:)\|^2\|\mathrm{W}_{k+1\to i+1}(:,j)\|^2 \\
&+ (\frac{1-\theta}{\theta})^2\sum_{i=1}^{k}\sum_{j=1}^{d_i}\|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{i\to 1}(j,:)\|^2\sum_{m=1}^{k-i}\sum_{l=1}^{d_{i+m}}\mathrm{W}_{i+m\to i+1}(l,j)^2\|\mathrm{W}_{k+1\to i+m+1}(:,l)\|^2 \\
&+ (\frac{1-\theta}{\theta})^3\sum_{i=1}^{k}\sum_{j=1}^{d_i}\|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{i\to 1}(j,:)\|^2\sum_{m=1}^{k-i}\sum_{l=1}^{d_{i+m}}\mathrm{W}_{i+m\to i+1}(l,j)^2\left(\sum_{mm=1}^{k-i-m}\theta^{2(i+m+mm)}\right. \\
&\left.\sum_{ll=1}^{d_{i+m+mm}}\mathrm{W}_{i+m+mm\to i+m+1}(ll,l)^2\mathbb{E}\|\bar{\mathrm{W}}_{k+1\to i+1+m+mm}(:,ll)\|^2\right) \\
&= \cdots \\
&= \sum_{l=1}^{k}\lambda^l\sum_{\substack{(j_l,\ldots,j_1)\\ \in\binom{[k]}{l}}}\sum_{\substack{(i_l,\ldots,i_1)\\ \in[d_{j_l}]\times\cdots\times[d_{j_1}]}}\|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{j_1\to 1}(i_1,:)\|^2\prod_{p=1\cdots l-1}\mathrm{W}_{j_{p+1}\to j_p+1}(i_{p+1},i_p)^2\|\mathrm{W}_{k+1\to j_l+1}(:,i_l)\|^2,
\end{aligned}
$$

which completes the proof. □

Lemma A.2. *[Properties of R and Θ] The following statements hold true:*

1. *All sub-regularizers, and hence the explicit regularizer, are re-scaling invariant.*
2. *The infimum in Equation (2) is always attained.*
3. *If $\mathrm{C}=I$, then $\Theta(\mathrm{M})$ is a* spectral function*, i.e. if M and M' have the same singular values, then $\Theta(\mathrm{M})=\Theta(\mathrm{M}')$.*

Proof of Lemma A.2. First, it is easy to see that the explicit regularizer and the sub-regularizers are all *rescaling invariant*. For any sequence of scalars $\{\alpha_i\}$ such that such that $\prod_{i=1}^{k+1}\alpha_i=1$, let $\bar{\mathrm{W}}_i:=\alpha_i\mathrm{W}_i$. Then it holds that:

$$
\begin{aligned}
R_l(\{\bar{\mathrm{W}}_i\}) &= \sum_{\substack{(j_l,\ldots,j_1)\\ \in\binom{[k]}{l}}}\sum_{\substack{(i_l,\ldots,i_1)\\ \in[d_{j_l}]\times\cdots\times[d_{j_1}]}}\|\prod_{q=1}^{j_1}\alpha_q\mathrm{W}_{j_1\to 1}(i_1,:)\|^2\prod_{p\in[l-1]}\prod_{q=j_p+1}^{j_{p+1}}\alpha_q^2\mathrm{W}_{j_{p+1}\to j_p+1}(i_{p+1},i_p)^2\|\prod_{q=j_l+1}^{k+1}\alpha_q\mathrm{W}_{k+1\to j_l+1}(:,i_l)\|^2 \\
&= \prod_{q=1}^{k+1}\alpha_q^2\sum_{\substack{(j_l,\ldots,j_1)\\ \in\binom{[k]}{l}}}\sum_{\substack{(i_l,\ldots,i_1)\\ \in[d_{j_l}]\times\cdots\times[d_{j_1}]}}\|\mathrm{W}_{j_1\to 1}(i_1,:)\|^2\prod_{p\in[l-1]}\mathrm{W}_{j_{p+1}\to j_p+1}(i_{p+1},i_p)^2\|\mathrm{W}_{k+1\to j_l+1}(:,i_l)\|^2 \\
&= R_l(\{\mathrm{W}_i\})
\end{aligned}
$$

Therefore, without loss of generality, we can express the induced regularizer as follows:

$$
\Theta(\mathrm{M}) := \inf_{\substack{\mathrm{W}_{k+1}\cdots\mathrm{W}_1=\mathrm{M}\\ \|\mathrm{W}_i\|_F\le\|\mathrm{M}\|_F}} R(\{\mathrm{W}_i\}) \tag{15}
$$

Note that $R(\{\mathrm{W}_i\})$ is a continuous function and the feasible set $\mathcal{F} := \{(\mathrm{W}_i)_{i=1}^{k+1} : \mathrm{W}_{k+1}\cdots\mathrm{W}_1 = \mathrm{M},\ \|\mathrm{W}_i\|_F \le \|\mathrm{M}\|_F\}$ is compact. Hence, by Weierstrass extreme value theorem, the infimum is attained.

Now let $\mathrm{U} \in \mathbb{R}^{d_{k+1}\times d_{k+1}}$ and $\mathrm{V} \in \mathbb{R}^{d_0\times d_0}$ be a pair of rotation matrices, i.e. $\mathrm{U}^\top\mathrm{U} = \mathrm{U}\mathrm{U}^\top = \mathrm{I}$ and $\mathrm{V}^\top\mathrm{V} = \mathrm{V}\mathrm{V}^\top = \mathrm{I}$. When the data is isotropic, i.e. $\mathrm{C} = \mathrm{I}$, the following equalities hold

$$\begin{aligned}
R(\{\mathrm{W}_i\}) &= \sum_{l=1}^{k}\lambda^l \sum_{\substack{(j_l,\dots,j_1)\\ \in\binom{[k]}{l}}}\ \sum_{\substack{(i_l,\dots,i_1)\\ \in[d_{j_l}]\times\cdots\times[d_{j_1}]}} \|\mathrm{W}_{j_1\to 1}(i_1,:)\|^2 \prod_{p=1\cdots l-1}\mathrm{W}_{j_{p+1}\to j_p+1}(i_{p+1},i_p)^2\|\mathrm{W}_{k+1\to j_l+1}(:,i_l)\|^2\\
&= \sum_{l=1}^{k}\lambda^l \sum_{\substack{(j_l,\dots,j_1)\\ \in\binom{[k]}{l}}}\ \sum_{\substack{(i_l,\dots,i_1)\\ \in[d_{j_l}]\times\cdots\times[d_{j_1}]}} \|\mathrm{W}_{j_1\to 1}(i_1,:)^\top\mathrm{V}\|^2 \prod_{p=1\cdots l-1}\mathrm{W}_{j_{p+1}\to j_p+1}(i_{p+1},i_p)^2\|\mathrm{U}^\top\mathrm{W}_{k+1\to j_l+1}(:,i_l)\|^2\\
&= R(\mathrm{U}^\top\mathrm{W}_{k+1},\mathrm{W}_k,\cdots,\mathrm{W}_2,\mathrm{W}_1\mathrm{V})
\end{aligned}$$

That is, $R(\mathrm{U}^\top\mathrm{W}_{k+1},\mathrm{W}_k,\dots,\mathrm{W}_2,\mathrm{W}_1\mathrm{V}) = R(\mathrm{W}_{k+1},\mathrm{W}_k,\dots,\mathrm{W}_2,\mathrm{W}_1)$ for all rotation matrices U and V. In particular, let U, V be the left and right singular vectors of M, i.e. $\mathrm{M} = \mathrm{U}\Sigma\mathrm{V}^\top$. To prove that Θ is a spectral function, we need to show that $\Theta(\mathrm{M}) = \Theta(\Sigma)$. Let $\{\bar{\mathrm{W}}_i\},\{\tilde{\mathrm{W}}_i\}$ be such that $\Theta(\mathrm{M}) = R(\{\bar{\mathrm{W}}_i\}), \Theta(\Sigma) = R(\{\tilde{\mathrm{W}}_i\})$. Note that such weight matrices always exist since the infimum is always attained. Then

$$\Theta(\Sigma) = \Theta(\mathrm{U}^\top\mathrm{M}\mathrm{V}) \le R(\mathrm{U}^\top\bar{\mathrm{W}}_{k+1},\bar{\mathrm{W}}_k,\dots,\bar{\mathrm{W}}_2,\bar{\mathrm{W}}_1\mathrm{V}) = R(\bar{\mathrm{W}}_{k+1},\bar{\mathrm{W}}_k,\dots,\bar{\mathrm{W}}_2,\bar{\mathrm{W}}_1) = \Theta(\mathrm{M}).$$

Similarly, we have that $\Theta(\mathrm{M}) \le R(\mathrm{U}^\top\tilde{\mathrm{W}}_{k+1},\tilde{\mathrm{W}}_k,\dots,\tilde{\mathrm{W}}_2,\bar{\mathrm{W}}_1\mathrm{V}) = R(\tilde{\mathrm{W}}_{k+1},\bar{\mathrm{W}}_k,\dots,\tilde{\mathrm{W}}_2,\tilde{\mathrm{W}}_1) = \Theta(\Sigma)$, which completes the proof. □

A.2. The induced regularizer and its convex envelope

Proof of Theorem 2.6. By Lemma 3.1, for any architecture, any dropout rate, and any set of weights $\{\mathrm{W}_i\}$ that implements a network map $\mathrm{W}_{k+1\to 1}$, the explicit regularizer is lower bounded by the effective regularization parameter times the product of the squared nuclear norm of the network map and the principal squared root of the second moment of x, i.e. $R(\{\mathrm{W}_i\}) \ge \nu_{\{d_i\}}\|\mathrm{W}_{k+1\to 1}\mathrm{C}^{\frac{1}{2}}\|_*^2$. Consequently, the induced regularizer can also be lowerbounded as $\Theta(\mathrm{M}) \ge \nu_{\{d_i\}}\|\mathrm{M}\mathrm{C}^{\frac{1}{2}}\|_*^2$. On the other hand, Lemma 3.2 establishes that $\Theta^{**}(\mathrm{M}) \le \nu_{\{d_i\}}\|\mathrm{M}\mathrm{C}^{\frac{1}{2}}\|_*^2$ holds for any network map M. Putting these two inequalities together, we arrive at

$$\Theta^{**}(\mathrm{M}) \le \nu_{\{d_i\}}\|\mathrm{M}\mathrm{C}^{\frac{1}{2}}\|_*^2 \le \Theta(\mathrm{M}).$$

Since $\Theta^{**}(\mathrm{M})$ is the largest convex underestimator of $\Theta(\mathrm{M})$, and the squared nuclear norm is a convex function, we conclude that $\Theta^{**}(\mathrm{M}) = \nu_{\{d_i\}}\|\mathrm{M}\|_*^2$. □

Despite the complex form of the explicit regularizer given in Proposition 2.1, we can show that it is always lower bounded by *effective regularization parameter* times $\|\mathrm{M}\mathrm{C}^{\frac{1}{2}}\|_*^2$. This result is given by Lemma 3.1.

Proof of Lemma 3.1. Recall that the explicit regularizer $R(\{\mathrm{W}_i\})$ is composed of k *sub-regularizers*

$$R(\{\mathrm{W}_i\}) = R_1(\{\mathrm{W}_i\}) + R_2(\{\mathrm{W}_i\}) + \cdots + R_k(\{\mathrm{W}_i\}).$$

The l-th sub-regularizer $R_l(\{\mathrm{W}_i\})$ can be written in the form of:

$$R_l(\{\mathrm{W}_i\}) = \lambda^l \sum_{(j_l,\dots,j_1)\in\binom{[k]}{l}} R_{\{j_l,\dots,j_1\}}(\{\mathrm{W}_i\})$$

where

$$R_{\{j_l,\dots,j_1\}}(\{\mathrm{W}_i\}) := \|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{j_1\to 1}(i_1,:)\|^2 \prod_{p=1\cdots l-1}\mathrm{W}_{j_{p+1}\to j_p+1}(i_{p+1},i_p)^2\|\mathrm{W}_{k+1\to j_l+1}(:,i_l)\|^2.$$

The following set of equalities hold true:

$$
\begin{aligned}
R_{\{j_l,\dots,j_1\}}(\{\mathrm{W}_i\}) &= \sum_{i_l,\cdots,i_1} \|\mathrm{W}_{k+1\to j_l+1}(:,i_l)\|^2 \mathrm{W}_{j_l\to j_{l-1}+1}(i_l,i_{l-1})^2 \cdots \mathrm{W}_{j_2\to j_1+1}(i_2,i_1)^2 \|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{j_1\to 1}(i_1,:)\|^2 \\
&\geq \frac{1}{\prod_{i\in[l]} d_{j_i}} \left(\sum_{i_l,\dots,i_1} \|\mathrm{W}_{k+1\to j_l+1}(:,i_l)\| |\mathrm{W}_{j_l\to j_{l-1}+1}(i_l,i_{l-1})| \cdots |\mathrm{W}_{j_2\to j_1+1}(i_2,i_1)| \|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{j_1\to 1}(i_1,:)\| \right)^2 \\
&= \frac{1}{\prod_{i\in[l]} d_{j_i}} \left(\sum_{i_l,\dots,i_1} \|\mathrm{W}_{k+1\to j_l+1}(:,i_l)\mathrm{W}_{j_l\to j_{l-1}+1}(i_l,i_{l-1}) \cdots \mathrm{W}_{j_2\to j_1+1}(i_2,i_1)\mathrm{W}_{j_1\to 1}(i_1,:)^\top \mathrm{C}^{\frac{1}{2}}\|_* \right)^2 \\
&\geq \frac{1}{\prod_{i\in[l]} d_{j_i}} \| \sum_{i_l,\dots,i_1} \mathrm{W}_{k+1\to j_l+1}(:,i_l)\mathrm{W}_{j_l\to j_{l-1}+1}(i_l,i_{l-1}) \cdots \mathrm{W}_{j_2\to j_1+1}(i_2,i_1)\mathrm{W}_{j_1\to 1}(i_1,:)^\top \mathrm{C}^{\frac{1}{2}}\|_*^2 \\
&= \frac{1}{\prod_{i\in[l]} d_{j_i}} \| \sum_{i_l,i_1} \mathrm{W}_{k+1\to j_l+1}(:,i_l) \left(\sum_{i_{l-1},\dots,i_2} \mathrm{W}_{j_l\to j_{l-1}+1}(i_l,i_{l-1}) \cdots \mathrm{W}_{j_2\to j_1+1}(i_2,i_1) \right) \mathrm{W}_{j_1\to 1}(i_1,:)^\top \mathrm{C}^{\frac{1}{2}}\|_*^2 \\
&= \frac{1}{\prod_{i\in[l]} d_{j_i}} \| \sum_{i_l,i_1} \mathrm{W}_{k+1\to j_l}(:,i_l)\mathrm{W}_{j_l-1\to j_1}(i_l,i_1)\mathrm{W}_{j_1\to 1}(i_1,:)^\top \mathrm{C}^{\frac{1}{2}}\|_*^2 \\
&= \frac{1}{\prod_{i\in[l]} d_{j_i}} \|\mathrm{W}_{k+1}\cdots\mathrm{W}_1\mathrm{C}^{\frac{1}{2}}\|_*^2
\end{aligned}
$$

where the first inequality follows due to the Cauchy-Schwartz inequality, and the second inequality follows from the triangle inequality for the matrix norms. The inequality holds with equality if and only if all the summands inside the summation are equal to each other, and sum up to $\frac{\|\mathrm{W}_{k+1\to 1}\mathrm{C}^{\frac{1}{2}}\|_*}{\prod_{i\in[l]} d_{j_i}}$, i.e. when

$$
\|\mathrm{W}_{k+1\to j_l+1}(:,i_l)\| |\mathrm{W}_{j_l\to j_{l-1}+1}(i_l,i_{l-1})| \cdots |\mathrm{W}_{j_2\to j_1+1}(i_2,i_1)| \|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{j_1\to 1}(i_1,:)\| = \frac{1}{\prod_{i\in[l]} d_{j_i}} \|\mathrm{W}_{k+1\to 1}\mathrm{C}^{\frac{1}{2}}\|_*
$$

for all $(i_l,\dots,i_1) \in [d_{j_l}]\times\cdots\times[d_{j_1}]$. This lowerbound holds for all $l\in[k]$, and for all $(j_l,\dots,j_1)\in\binom{[k]}{l}$. Thus, we get the following lowerbound on the regularizer:

$$
R(\{\mathrm{W}_i\}) \geq \sum_{l\in[k]} \lambda^l \sum_{(j_l,\dots,j_1)\in\binom{[k]}{l}} \frac{1}{\prod_{i\in[l]} d_{j_i}} \|\mathrm{W}_{k+1\to 1}\mathrm{C}^{\frac{1}{2}}\|_*^2 = \nu_{\{\mathrm{d}_i\}} \|\mathrm{W}_{k+1\to 1}\mathrm{C}^{\frac{1}{2}}\|_*^2
$$

which completes the proof. □

Not only $\nu_{\{d_i\}}\|\mathrm{MC}^{\frac{1}{2}}\|_*^2$ is a lowerbound for the induced regularizer, but also is an upperbound for its convex envelope. We prove this result in Lemma 3.2.

Proof of Lemma 3.2. The induced regularizer is non-negative. Hence, the domain of the Fenchel dual of the induced regularizer is the whole $\mathbb{R}^{d_{k+1}\times d_0}$. The Fenchel dual of the induced regularizer $\Theta(\cdot)$ is given by:

$$
\begin{aligned}
\Theta^*(\mathrm{M}) &= \max_{\mathrm{P}} \langle \mathrm{M},\mathrm{P}\rangle - \Theta(\mathrm{P}) \\
&= \max_{\mathrm{P}} \langle \mathrm{M},\mathrm{P}\rangle - \min_{\substack{\{\mathrm{W}_i\} \\ \mathrm{W}_{k+1\to 1}=\mathrm{P}}} R(\{\mathrm{W}_i\}) \\
&= \max_{\{\mathrm{W}_i\}} \langle \mathrm{M}, \mathrm{W}_{k+1\to 1}\rangle - R(\{\mathrm{W}_i\}).
\end{aligned} \tag{16}
$$

Define $\Phi(\{\mathrm{W}_i\}) := \langle \mathrm{M}, \mathrm{W}_{k+1\to 1}\rangle - R(\{\mathrm{W}_i\})$ as the objective in the right hand side of Equation (16). Due to the complicated products of the norms of the weights in the regularizer, maximizing Φ with respect to $\{\mathrm{W}_i\}$ is a daunting task. Here, we find a lower bound on this maximum value. Let $\mathrm{W}_{k+1}^\alpha := \alpha \mathrm{u}_1 1_{d_k}^\top$ and $\mathrm{W}_1^\alpha := 1_{d_1}\mathrm{v}_1^\top \mathrm{C}^{-\frac{1}{2}}$, where $(\mathrm{u}_1,\mathrm{v}_1)$ is

the top singular vectors of $\mathrm{M}\mathrm{C}^{-\frac{1}{2}}$, and 1_d is the d-dimensional vector of all 1s. Furthermore, let $\mathrm{W}_i^\alpha := 1_{d_i} 1_{d_{i-1}}^\top$, for all $i \in \{2, \dots, k\}$. Note that

$$\Theta^*(\mathrm{M}) = \max_{\{\mathrm{W}_i\}} \Phi(\{\mathrm{W}_i\}) \geq \max_{\alpha} \Phi(\{\mathrm{W}_i^\alpha\}).$$

We now simplify $\Phi(\{\mathrm{W}_i^\alpha\})$. First, the following equalities hold for the $\langle \mathrm{M}, \mathrm{W}_{k+1\to 1}^\alpha \rangle$:

$$\begin{aligned}
\langle \mathrm{M}, \mathrm{W}_{k+1\to 1}^\alpha \rangle &= \sum_{(i_{k+1},\dots,i_1)\in[d_{k+1}]\times\cdots\times[d_1]} \langle \mathrm{M}, \mathrm{W}_{k+1}^\alpha(:,i_k) \prod_{j=\{k-1,\dots,1\}} \mathrm{W}_{j+1}^\alpha(i_{j+1},i_j)\mathrm{W}_1^\alpha(i_1,:)^\top \rangle \\
&= \sum_{(i_{k+1},\dots,i_1)\in[d_{k+1}]\times\cdots\times[d_1]} \mathrm{W}_{k+1}^\alpha(:,i_k)^\top \mathrm{M} \mathrm{W}_1^\alpha(i_1,:) \\
&= \sum_{(i_{k+1},\dots,i_1)\in[d_{k+1}]\times\cdots\times[d_1]} \alpha \mathrm{u}_1^\top \mathrm{M}\mathrm{C}^{-\frac{1}{2}} \mathrm{v}_1 \\
&= \sum_{(i_{k+1},\dots,i_1)\in[d_{k+1}]\times\cdots\times[d_1]} \alpha \|\mathrm{M}\mathrm{C}^{-\frac{1}{2}}\|_2 \\
&= \alpha \|\mathrm{M}\mathrm{C}^{-\frac{1}{2}}\|_2 \prod_{j\in[k]} d_j =: \alpha \|\mathrm{M}\mathrm{C}^{-\frac{1}{2}}\|_2 D.
\end{aligned}$$

The following terms show up in the expansion of the regularizer:

$$\mathrm{W}_{j_1\to 1}^\alpha(i_1,:)^\top = \mathrm{W}_{j_1}^\alpha(i_1,:)\mathrm{W}_{j_1-1}^\alpha \cdots \mathrm{W}_2^\alpha \mathrm{W}_1^\alpha = 1_{d_{j_1-1}}^\top 1_{d_{j_1-1}} 1_{d_{j_1-2}}^\top \cdots 1_{d_2} 1_{d_1}^\top 1_{d_1} \mathrm{v}_1^\top = \prod_{i\in[j_1-1]} d_i \mathrm{v}_1^\top \mathrm{C}^{-\frac{1}{2}}$$

$$\begin{aligned}
\mathrm{W}_{j_{p+1}\to j_p+1}^\alpha(i_{p+1},i_p) &= \mathrm{W}_{j_{p+1}}^\alpha(i_{p+1},:)\mathrm{W}_{j_{p+1}-1}^\alpha \cdots \mathrm{W}_{j_p+2}^\alpha \mathrm{W}_{j_p+1}^\alpha(:,i_p) \\
&= 1_{d_{j_{p+1}-1}}^\top 1_{d_{j_{p+1}-1}} 1_{d_{j_{p+1}-2}}^\top \cdots 1_{d_{j_p+2}} 1_{d_{j_p+1}}^\top 1_{d_{j_p+1}} = \prod_{i\in\{j_p+1,\cdots,j_{p+1}-1\}} d_i
\end{aligned}$$

$$\mathrm{W}_{k+1\to j_l+1}^\alpha(:,i_l) = \alpha \mathrm{W}_{k+1}^\alpha \mathrm{W}_k^\alpha \cdots \mathrm{W}_{j_l+2}^\alpha \mathrm{W}_{j_l+1}^\alpha(:,i_l) = \alpha \mathrm{u}_1 1_{d_k}^\top 1_{d_k} 1_{d_{k-1}}^\top \cdots 1_{d_{j_l+2}} 1_{d_{j_l+1}}^\top 1_{d_{j_l+1}} = \alpha \prod_{i\in\{j_l+1,\cdots,k\}} d_i \mathrm{u}_1$$

With the above equalities, the explicit regularizer reduces to:

$$\begin{aligned}
R(\{\mathrm{W}_i^\alpha\}) &= \sum_{l=1}^k \lambda^l \sum_{\substack{(j_l,\dots,j_1)\\ \in\binom{[k]}{l}}} \sum_{\substack{(i_l,\dots,i_1)\\ \in[d_{j_l}]\times\cdots\times[d_{j_1}]}} \|\mathrm{C}^{\frac{1}{2}} \mathrm{W}_{j_1\to 1}^\alpha(i_1,:)\|^2 \prod_{p=1\cdots l-1} \mathrm{W}_{j_{p+1}\to j_p+1}^\alpha(i_{p+1},i_p)^2 \|\mathrm{W}_{k+1\to j_l+1}^\alpha(:,i_l)\|^2 \\
&= \sum_{l=1}^k \lambda^l \sum_{\substack{(j_l,\dots,j_1)\\ \in\binom{[k]}{l}}} \sum_{\substack{(i_l,\dots,i_1)\\ \in[d_{j_l}]\times\cdots\times[d_{j_1}]}} \|\mathrm{C}^{\frac{1}{2}} \mathrm{C}^{-\frac{1}{2}} \mathrm{v}_1 \prod_{i\in[j_1-1]} d_i\|^2 \prod_{p=1\cdots l-1} \prod_{i\in\{j_p+1,\cdots,j_{p+1}-1\}} d_i^2 \|\alpha \mathrm{u}_1 \prod_{i\in\{j_l+1,\cdots,k\}} d_i\|^2 \\
&= \sum_{l=1}^k \lambda^l \sum_{\substack{(j_l,\dots,j_1)\\ \in\binom{[k]}{l}}} \sum_{\substack{(i_l,\dots,i_1)\\ \in[d_{j_l}]\times\cdots\times[d_{j_1}]}} \prod_{i\in[j_1-1]} d_i^2 \prod_{p=1\cdots l-1} \prod_{i\in\{j_p+1,\cdots,j_{p+1}-1\}} d_i^2 \alpha^2 \prod_{i\in\{j_l+1,\cdots,k\}} d_i^2 \\
&= \alpha^2 \sum_{l=1}^k \lambda^l \sum_{(j_l,\dots,j_1)\in\binom{[k]}{l}} \sum_{(i_l,\dots,i_1)\in[d_{j_l}]\times\cdots\times[d_{j_1}]} \frac{\prod_{i\in[k]} d_i^2}{\prod_{i\in[l]} d_{j_i}^2} =: \alpha^2 \rho
\end{aligned}$$

Plugging back the above equalities into the definition of Φ, we arrive at $\Phi(\{\mathrm{W}_i^\alpha\}) = \alpha \|\mathrm{M}\mathrm{C}^{-\frac{1}{2}}\|_2 D - \alpha^2 \rho$. The maximum of $\Phi(\{\mathrm{W}_i^\alpha\})$ with respect to α is achieved when $\alpha^* = \frac{\|\mathrm{M}\mathrm{C}^{-\frac{1}{2}}\|_2 D}{2\rho}$, in which case we have

$$\Theta^*(\mathrm{M}) \geq \Phi(\{\mathrm{W}_i^{\alpha^*}\}) = \frac{D^2}{4\rho} \|\mathrm{M}\mathrm{C}^{-\frac{1}{2}}\|_2^2 =: \Psi(\mathrm{M}).$$

Since Fenchel dual is order reversing, we get

$$\begin{aligned}
\Theta^{**}(\mathrm{M}) &\leq \Psi^*(\mathrm{M}) \\
&= \frac{\rho}{D^2}\|\mathrm{MC}^{\frac{1}{2}}\|_*^2 \\
&= \frac{\sum_{l=1}^k \lambda^l \sum_{(j_l,\ldots,j_1)\in\binom{[k]}{l}} \sum_{(i_l,\ldots,i_1)\in[d_{j_l}]\times\cdots\times[d_{j_1}]} \frac{\prod_{i\in[k]} d_i^2}{\prod_{i\in[l]} d_{j_i}^2}}{\prod_{j\in[k]} d_j^2}\|\mathrm{MC}^{\frac{1}{2}}\|_*^2 \\
&= \sum_{l=1}^k \lambda^l \sum_{(j_l,\ldots,j_1)\in\binom{[k]}{l}} \sum_{(i_l,\ldots,i_1)\in[d_{j_l}]\times\cdots\times[d_{j_1}]} \frac{1}{\prod_{i\in[l]} d_{j_i}^2}\|\mathrm{MC}^{\frac{1}{2}}\|_*^2 \\
&= \sum_{l=1}^k \lambda^l \sum_{(j_l,\ldots,j_1)\in\binom{[k]}{l}} \frac{1}{\prod_{i\in[l]} d_{j_i}}\|\mathrm{MC}^{\frac{1}{2}}\|_*^2 \\
&= \nu_{\{d_i\}}\|\mathrm{MC}^{\frac{1}{2}}\|_*^2
\end{aligned}$$

where the first equality follows from the fact that if $f(\mathrm{M}) = \beta\|\mathrm{MA}\|^2$ and $\mathrm{A} \succ 0$ then $f^*(\mathrm{M}) = \frac{1}{4\beta}\|\mathrm{MA}^{-1}\|_*^2$. This result is standard in the literature, but we prove it here for completeness. Note that

$$\begin{aligned}
\langle \mathrm{Y}, \mathrm{M}\rangle - \beta\|\mathrm{YA}\|^2 &= \langle \mathrm{YA}, \mathrm{MA}^{-1}\rangle - \beta\|\mathrm{YA}\|^2 \\
&\leq \|\mathrm{YA}\|\|\mathrm{MA}^{-1}\|_* - \beta\|\mathrm{YA}\|^2
\end{aligned}$$

where the inequality is due to Holder's identity. The right hand side above is a quadratic in $\|\mathrm{YA}\|$ and is maximized when $\|\mathrm{YA}\| = \frac{1}{2\beta}\|\mathrm{MA}^{-1}\|_*$, in which case we have

$$f^*(\mathrm{M}) = \sup_{\mathrm{Y}} \langle \mathrm{Y}, \mathrm{M}\rangle - \beta\|\mathrm{YA}\|^2 = \frac{1}{2\beta}\|\mathrm{MA}^{-1}\|_*\|\mathrm{MA}^{-1}\|_* - \beta(\frac{1}{2\beta}\|\mathrm{MA}^{-1}\|_*)^2 = \frac{1}{4\beta}\|\mathrm{MA}^{-1}\|_*^2.$$

□

A.3. Characterization of the global optima of the dropout objective

Proof of Proposition 3.3. When the network map has rank equal to one, it can be expressed as $\mathrm{uv}^\top$, where $\mathrm{u} \in \mathbb{R}^{d_{k+1}}$ and $\mathrm{v} \in \mathbb{R}^{d_0}$. We show that for any architecture $\{d_i\}$ and any network mapping $\mathrm{uv}^\top \in \mathbb{R}^{d_{k+1}\times d_0}$, it is always possible to represent $\mathrm{uv}^\top = \mathrm{W}_{k+1}\cdots\mathrm{W}_1$ such that the resulting network is equalized. One such factorization is when $\mathrm{W}_1 = \frac{1_{d_1}\mathrm{v}^\top}{\sqrt{d_1}}$, $\mathrm{W}_{k+1} = \frac{\mathrm{u}1_{d_k}^\top}{\sqrt{d_k}}$, and $\mathrm{W}_i = \frac{1_{d_i}1_{d_{i-1}}^\top}{\sqrt{d_i d_{i-1}}}$ for $i \in \{2,\ldots,k\}$. For these weight parameters, we have that

$$\begin{aligned}
\mathrm{W}_{j_1\to 1}(i_1,:)^\top &= \mathrm{W}_{j_1}(i_1,:)^\top \mathrm{W}_{j_1-1}\cdots\mathrm{W}_2\mathrm{W}_1 = \frac{1_{d_{j_1-1}}^\top}{\sqrt{d_{j_1}d_{j_1-1}}}\frac{1_{d_{j_1-1}}1_{d_{j_1-2}}^\top}{\sqrt{d_{j_1-1}d_{j_1-2}}}\cdots\frac{1_{d_2}1_{d_1}^\top}{\sqrt{d_2 d_1}}\frac{1_{d_1}\mathrm{v}^\top}{\sqrt{d_1}} = \frac{\mathrm{v}^\top}{\sqrt{d_{j_1}}} \\
\mathrm{W}_{j_{p+1}\to j_p+1}(i_{p+1},i_p) &= \mathrm{W}_{j_{p+1}}(i_{p+1},:)^\top \mathrm{W}_{j_{p+1}-1}\cdots\mathrm{W}_{j_p+2}\mathrm{W}_{j_p+1}(:,i_p) \\
&= \frac{1_{d_{j_{p+1}-1}}^\top}{\sqrt{d_{j_{p+1}}d_{j_{p+1}-1}}}\frac{1_{d_{j_{p+1}-1}}1_{d_{j_{p+1}-2}}^\top}{\sqrt{d_{j_{p+1}-1}d_{j_{p+1}-2}}}\cdots\frac{1_{d_{j_p+2}}1_{d_{j_p+1}}^\top}{\sqrt{d_{j_p+2}d_{j_p+1}}}\frac{1_{d_{j_p+1}}}{\sqrt{d_{j_p+1}d_{j_p}}} = \frac{1}{\sqrt{d_{j_{p+1}}d_{j_p}}} \\
\mathrm{W}_{k+1\to j_l+1}(:,i_l) &= \mathrm{W}_{k+1}\mathrm{W}_k\cdots\mathrm{W}_{j_l+2}\mathrm{W}_{j_l+1}(:,i_l) \\
&= \frac{\mathrm{u}1_{d_k}^\top}{\sqrt{d_k}}\frac{1_{d_k}1_{d_{k-1}}^\top}{\sqrt{d_k d_{k-1}}}\cdots\frac{1_{d_{j_l+2}}1_{d_{j_l+1}}^\top}{\sqrt{d_{j_l+2}d_{j_l+1}}}\frac{1_{d_{j_l+1}}}{\sqrt{d_{j_l+1}d_{j_l}}} = \frac{\mathrm{u}}{\sqrt{d_{j_l}}}
\end{aligned}$$

With the above equalities, the regularizer reduces to:

$$
\begin{aligned}
R(\{\mathrm{W}_i\}) &= \sum_{l=1}^{k} \lambda^l \sum_{\substack{(j_l,\dots,j_1)\\ \in\binom{[k]}{l}}} \sum_{\substack{(i_l,\dots,i_1)\\ \in[d_{j_l}]\times\cdots\times[d_{j_1}]}} \|\mathrm{C}^{\frac{1}{2}}\mathrm{W}_{j_1\to 1}(i_1,:)\|^2 \prod_{p=1\cdots l-1} \mathrm{W}_{j_{p+1}\to j_p+1}(i_{p+1},i_p)^2 \|\mathrm{W}_{k+1\to j_l+1}(:,i_l)\|^2 \\
&= \sum_{l=1}^{k} \lambda^l \sum_{\substack{(j_l,\dots,j_1)\\ \in\binom{[k]}{l}}} \sum_{\substack{(i_l,\dots,i_1)\\ \in[d_{j_l}]\times\cdots\times[d_{j_1}]}} \|\mathrm{C}^{\frac{1}{2}}\frac{\mathrm{v}}{\sqrt{d_{j_1}}}\|^2 \prod_{p=1\cdots l-1} \frac{1}{d_{j_{p+1}}d_{j_p}} \|\frac{\mathrm{u}}{\sqrt{d_{j_l}}}\|^2 \\
&= \sum_{l=1}^{k} \lambda^l \sum_{\substack{(j_l,\dots,j_1)\\ \in\binom{[k]}{l}}} \sum_{\substack{(i_l,\dots,i_1)\\ \in[d_{j_l}]\times\cdots\times[d_{j_1}]}} \frac{\|\mathrm{C}^{\frac{1}{2}}\mathrm{v}\|^2\|\mathrm{u}\|^2}{\prod_{p\in[l]} d_{j_p}^2} \\
&= \sum_{l=1}^{k} \sum_{\substack{(j_l,\dots,j_1)\\ \in\binom{[k]}{l}}} \frac{\lambda^l}{\prod_{p\in[l]} d_{j_p}} \|\mathrm{u}\mathrm{v}^\top\mathrm{C}^{\frac{1}{2}}\|_*^2 = \nu_{\{d_i\}}\|\mathrm{u}\mathrm{v}^\top\mathrm{C}^{\frac{1}{2}}\|_*^2
\end{aligned}
$$

where we used the fact that $\|\mathrm{u}\|\|\mathrm{C}^{\frac{1}{2}}\mathrm{v}\| = \|\mathrm{u}\mathrm{v}^\top\mathrm{C}^{\frac{1}{2}}\|_*$. Moreover, note that the network specified by the above weight matrices is equalized, since

$$
|\alpha_{j_1,i_1}| \prod_{p=1\cdots l-1} |\beta_p||\gamma_{j_l,i_l}| = \sqrt{\frac{\|\mathrm{u}\mathrm{v}^\top\mathrm{C}^{\frac{1}{2}}\|_*^2}{\prod_{p\in[l]} d_{j_p}^2}} = \frac{1}{\prod_{p\in[l]} d_{j_p}} \|\mathrm{u}\mathrm{v}^\top\mathrm{C}^{\frac{1}{2}}\|_*.
$$

□

Lemma A.3. *For any integer r, and for any $\nu \in \mathbb{R}_+$, it holds that*

$$
(I_r + \nu I_r I_r^\top)^{-1} = I_r - \frac{\nu}{1+r\nu} I_r I_r^\top.
$$

Lemma A.3 is an instance of the Woodbury's matrix identity. Here, we include a proof for completeness.

Proof of Lemma A.3. The proof simply follows from the following set of equations.

$$
\begin{aligned}
(\mathrm{I}_r + \nu 1_r 1_r^\top)(\mathrm{I}_r - \frac{\nu}{1+r\nu} 1_r 1_r^\top) &= \mathrm{I}_r + \nu 1_r 1_r^\top - \frac{\nu}{1+r\nu} 1_r 1_r^\top - \frac{\nu^2}{1+r\nu} 1_r 1_r^\top 1_r 1_r^\top \\
&= \mathrm{I}_r + \left(\nu - \frac{\nu}{1+r\nu} - \frac{\nu^2 r}{1+r\nu}\right) 1_r 1_r^\top = \mathrm{I}_r
\end{aligned}
$$

□

Lemma A.4. *Consider the following optimization problem where the induced regularizer in Problem 3 is replaced with its convex envelope:*

$$
\min_{\mathrm{W}\in\mathbb{R}^{d_{k+1}\times d_0}} \mathbb{E}[\|\mathrm{y}-\mathrm{W}\mathrm{x}\|^2] + \Theta^{**}(\mathrm{W}), \quad \mathrm{rank}\,(\mathrm{W}) \le \min_{i\in[k+1]} d_i =: r \tag{17}
$$

Define the "model" $\bar{\mathrm{M}} := \mathrm{C}_{\mathrm{yx}}\mathrm{C}^{-\frac{1}{2}}$. The global optimum of problem 17 is given as $\mathrm{M}^* = \mathcal{S}_{\alpha_\rho}(\bar{\mathrm{M}})\mathrm{C}^{-\frac{1}{2}}$, where $\alpha_\rho := \frac{\nu_{\{d_i\}}\sum_{j=1}^{\rho}\sigma_j(\bar{\mathrm{M}})}{1+\rho\nu_{\{d_i\}}}$, and $\rho \in [\min\{r, \mathrm{rank}\,(\bar{\mathrm{M}})\}]$ is the largest integer such that for all $i \in [\rho]$, it holds that $\sigma_i(\bar{\mathrm{M}}) > \alpha_\rho$.

Proof of Lemma A.4. Denote the objective in the optimization problem (17) as $\mathcal{E}_{\nu_{\{d_i\}}}(\mathrm{W}) := \mathbb{E}[\|\mathrm{y}-\mathrm{W}\mathrm{x}\|^2] + \nu_{\{d_i\}}\|\mathrm{W}\mathrm{C}^{\frac{1}{2}}\|_*^2$.

Let $\mathrm{C_y} := \mathbb{E}[\mathrm{yy}^\top]$ and $\mathrm{C_{xy}} := \mathbb{E}[\mathrm{xy}^\top]$. Note that

$$\begin{aligned}\min_{\mathrm{rank(W)}\le r} \mathcal{E}_{\nu_{\{d_i\}}}(\mathrm{W}) &= \min_{\mathrm{rank(W)}\le r} \mathbb{E}[\|\mathrm{y}\|^2] + \mathbb{E}[\|\mathrm{Wx}\|^2] - 2\mathbb{E}[\langle \mathrm{y}, \mathrm{Wx}\rangle] + \nu_{\{d_i\}}\|\mathrm{WC}^{\frac{1}{2}}\|_*^2 \\ &\equiv \min_{\mathrm{rank(W)}\le r} \mathrm{Tr}\left(\mathbb{E}[\mathrm{Wxx}^\top\mathrm{W}^\top]\right) - 2\mathrm{Tr}\left(\mathbb{E}[\mathrm{Wxy}^\top]\right) + \nu_{\{d_i\}}\|\mathrm{WC}^{\frac{1}{2}}\|_*^2 \\ &= \min_{\mathrm{rank(W)}\le r} \mathrm{Tr}\left(\mathrm{WCW}^\top\right) - 2\mathrm{Tr}\left(\mathrm{WC_{xy}}\right) + \nu_{\{d_i\}}\|\mathrm{WC}^{\frac{1}{2}}\|_*^2\end{aligned}$$

Make the change of variable $\bar{\mathrm{W}} \leftarrow \mathrm{WC}^{\frac{1}{2}}$ and denote $\bar{\mathrm{M}} := \mathrm{C_{yx}C}^{-\frac{1}{2}}$, the goal is to solve the following problem

$$\min_{\mathrm{rank(W)}\le r} \mathrm{Tr}\left(\bar{\mathrm{W}}\bar{\mathrm{W}}^\top\right) - 2\langle \bar{\mathrm{W}}, \bar{\mathrm{M}}\rangle + \nu_{\{d_i\}}\|\bar{\mathrm{W}}\|_*^2 \equiv \min_{\mathrm{rank}\left(\bar{\mathrm{W}}\right)\le r} \|\bar{\mathrm{M}} - \bar{\mathrm{W}}\|_F^2 + \nu_{\{d_i\}}\|\bar{\mathrm{W}}\|_*^2 \tag{18}$$

If $\bar{\mathrm{W}}$ is a solution to the above problem, then a solution to the original problem in Equation (17) is given as $\bar{\mathrm{W}}\mathrm{C}^{-\frac{1}{2}}$. Following (Cavazza et al., 2018; Mianjy et al., 2018), we show that the global optimum of Problem 18 is given in terms of an appropriate shrinkage-thresholding on the spectrum of $\bar{\mathrm{M}}$. Define $r' := \max\{\mathrm{rank}\left(\bar{\mathrm{M}}\right), r\}$. Let $\bar{\mathrm{M}} = \mathrm{U}_{\bar{\mathrm{M}}}\Sigma_{\bar{\mathrm{M}}}\mathrm{V}_{\bar{\mathrm{M}}}^\top$ and $\bar{\mathrm{W}} = \mathrm{U}_{\bar{\mathrm{W}}}\Sigma_{\bar{\mathrm{W}}}\mathrm{V}_{\bar{\mathrm{W}}}^\top$ be rank-r' SVDs of $\bar{\mathrm{M}}$ and $\bar{\mathrm{W}}$ respectively, such that $\sigma_i(\bar{\mathrm{M}}) \ge \sigma_{i+1}(\bar{\mathrm{M}})$ and $\sigma_i(\bar{\mathrm{W}}) \ge \sigma_{i+1}(\bar{\mathrm{W}})$ for all $i \in [r'-1]$. Rewriting objective of Problem 18 in terms of these decompositions gives:

$$\begin{aligned}&\|\bar{\mathrm{M}} - \bar{\mathrm{W}}\|_F^2 + \nu_{\{d_i\}}\|\bar{\mathrm{W}}\|_*^2 = \|\mathrm{U}_{\bar{\mathrm{M}}}\Sigma_{\bar{\mathrm{M}}}\mathrm{V}_{\bar{\mathrm{M}}}^\top - \mathrm{U}_{\bar{\mathrm{W}}}\Sigma_{\bar{\mathrm{W}}}\mathrm{V}_{\bar{\mathrm{W}}}^\top\|_F^2 + \nu_{\{d_i\}}\|\mathrm{U}_{\bar{\mathrm{W}}}\Sigma_{\bar{\mathrm{W}}}\mathrm{V}_{\bar{\mathrm{W}}}^\top\|_*^2 \\ &= \|\Sigma_{\bar{\mathrm{M}}} - \mathrm{U}'\Sigma_{\bar{\mathrm{W}}}\mathrm{V}'^\top\|_F^2 + \nu_{\{d_i\}}\|\Sigma_{\bar{\mathrm{W}}}\|_*^2 \\ &= \|\Sigma_{\bar{\mathrm{M}}}\|_F^2 + \|\Sigma_{\bar{\mathrm{W}}}\|_F^2 - 2\langle \Sigma_{\bar{\mathrm{M}}}, \mathrm{U}'\Sigma_{\bar{\mathrm{W}}}\mathrm{V}'^\top\rangle + \nu_{\{d_i\}}\|\Sigma_{\bar{\mathrm{W}}}\|_*^2\end{aligned}$$

where $\mathrm{U}' = \mathrm{U}_{\bar{\mathrm{M}}}^\top\mathrm{U}_{\bar{\mathrm{W}}}$ and $\mathrm{V}' = \mathrm{V}_{\bar{\mathrm{M}}}^\top\mathrm{V}_{\bar{\mathrm{W}}}$. By Von Neumann's trace inequality, for a fixed $\Sigma_{\bar{\mathrm{W}}}$ we have that

$$\langle \Sigma_{\bar{\mathrm{M}}}, \mathrm{U}'\Sigma_{\bar{\mathrm{W}}}\mathrm{V}'^\top\rangle \le \sum_{i=1}^{r'} \sigma_i(\bar{\mathrm{M}})\sigma_i(\bar{\mathrm{W}}),$$

where the equality is achieved when $\mathrm{U}_{\bar{\mathrm{M}}} = \mathrm{U}_{\bar{\mathrm{W}}}$ and $\mathrm{V}_{\bar{\mathrm{M}}} = \mathrm{V}_{\bar{\mathrm{W}}}$. Hence, problem 18 is reduced to

$$\min_{\substack{\|\Sigma_{\bar{\mathrm{W}}}\|_0\le r,\\ \Sigma_{\bar{\mathrm{W}}}\ge 0}} \|\Sigma_{\bar{\mathrm{M}}} - \Sigma_{\bar{\mathrm{W}}}\|_F^2 + \nu_{\{d_i\}}\left(\mathrm{Trace}(\Sigma_{\bar{\mathrm{W}}})\right)^2 = \min_{\bar{\sigma}\in\mathbb{R}_+^r} \sum_{i=1}^{r}\left(\lambda_i(\bar{\mathrm{M}}) - \bar{\sigma}_i\right)^2 + \nu_{\{d_i\}}\left(\sum_{i=1}^{r}\bar{\sigma}_i\right)^2$$

The Lagrangian is given by

$$L(\bar{\sigma}, \alpha) = \sum_{i=1}^{r}\left(\lambda_i(\bar{\mathrm{M}}) - \bar{\sigma}_i\right)^2 + \nu_{\{d_i\}}\left(\sum_{i=1}^{r}\bar{\sigma}_i\right)^2 - \sum_{i=1}^{r}\alpha_i\bar{\sigma}_i$$

The KKT conditions ensures that at the optima it holds for all $i \in [r]$ that

$$\bar{\sigma}_i \ge 0,\ \alpha_i \ge 0,\ \bar{\sigma}_i\alpha_i = 0, \quad 2(\bar{\sigma}_i - \sigma_i(\bar{\mathrm{M}})) + 2\nu_{\{d_i\}}\sum_{j=1}^{r}\bar{\sigma}_j - \alpha_i = 0$$

Let $\rho = |i : \bar{\sigma}_i > 0| \le r$ be the number of nonzero $\bar{\sigma}_i$, i.e. rank of the global optimum $\bar{\mathrm{W}}$. For $i \in [\rho]$, we have $\alpha_i = 0$. Therefore, we have that:

$$\begin{aligned}\bar{\sigma}_i + \nu_{\{d_i\}}\sum_{j=1}^{r}\bar{\sigma}_j = \sigma_i(\bar{\mathrm{M}}) &\implies (\mathrm{I}_\rho + \nu_{\{d_i\}}1_\rho 1_\rho^\top)\bar{\sigma}_{1:\rho} = \sigma_{1:\rho}(\bar{\mathrm{M}}) \\ &\implies \bar{\sigma}_{1:\rho} = (\mathrm{I}_\rho - \frac{\nu_{\{d_i\}}}{1+\rho\nu_{\{d_i\}}}1_\rho 1_\rho^\top)\sigma_{1:\rho}(\bar{\mathrm{M}}) = \sigma_{1:\rho}(\bar{\mathrm{M}}) - \frac{\nu_{\{d_i\}}\rho\kappa_\rho}{1+\rho\nu_{\{d_i\}}}1_\rho\end{aligned}$$

where $\kappa_j := \frac{1}{j}\sum_{i=1}^{j}\sigma_i(\bar{\mathrm{M}})$. The equation above tell us that for $i \in [\rho]$, the singular values of $\bar{\mathrm{W}}$ are just a shrinkage of the singular values of $\bar{\mathrm{M}}$. In particular, it means that $\rho \leq \mathrm{rank}\,(\bar{\mathrm{M}})$. Therefore, without loss of generality, we assume that $r \leq \mathrm{rank}\,(\bar{\mathrm{M}})$. Also, since $\bar{\sigma}_i > 0$ for all $i \in [\rho]$, it holds that $\sigma_i(\bar{\mathrm{M}}) > \frac{\nu_{\{d_i\}}\rho\kappa_\rho}{1+\rho\nu_{\{d_i\}}}$ for all $i \in [\rho]$. For $i \in \{\rho+1,\dots,r\}$, on the other hand, $\bar{\sigma}_i = 0$ and we have

$$\frac{1}{2}\alpha_i = \bar{\sigma}_i + \nu_{\{d_i\}}\sum_{j=1}^{r}\bar{\sigma}_j - \sigma_i(\bar{\mathrm{M}}) = 0 + \frac{\nu_{\{d_i\}}}{1+\rho\nu_{\{d_i\}}}\sum_{j=1}^{\rho}\sigma_j(\bar{\mathrm{M}}) - \sigma_i(\bar{\mathrm{M}}) = -\sigma_i(\bar{\mathrm{M}}) + \frac{\nu_{\{d_i\}}\rho\kappa_\rho}{1+\rho\nu_{\{d_i\}}},$$

where we used the fact that

$$\sum_{i=1}^{r}\bar{\sigma}_i = 1_\rho^\top\bar{\sigma}_{1:\rho} = \sum_{i=1}^{\rho}\sigma_i(\bar{\mathrm{M}}) - \frac{\nu_{\{d_i\}}\rho^2\kappa_\rho}{1+\rho\nu_{\{d_i\}}} = (1 - \frac{\nu_{\{d_i\}}\rho}{1+\rho\nu_{\{d_i\}}})\kappa_\rho = \frac{\rho\kappa_\rho}{1+\rho\nu_{\{d_i\}}}.$$

By dual feasibility, we conclude that $\sigma_i(\bar{\mathrm{M}}) \leq \frac{\nu_{\{d_i\}}\rho\kappa_\rho}{1+\rho\nu_{\{d_i\}}}$ for all $i \in \{\rho+1,\dots,r\}$, which completes the proof. □

Proof of Theorem 2.7. Consider W^*, a global optimum of problem 3. If all such global optima can be implemented by equalized networks, then by Theorem 2.6 it holds that $\Theta(\mathrm{W}^*) = \Theta^{**}(\mathrm{W}^*) = \nu_{\{d_i\}}\|\mathrm{W}^*\mathrm{C}^{\frac{1}{2}}\|_*^2$. In this case, the lifted problem in Equation 3 boils down to the following convex problem

$$\min_{\mathrm{W}\in\mathbb{R}^{d_{k+1}\times d_0}} \mathbb{E}[\|\mathrm{y} - \mathrm{Wx}\|^2] + \nu_{\{d_i\}}\|\mathrm{WC}^{\frac{1}{2}}\|_*^2, \quad \mathrm{rank}\,(\mathrm{W}) \leq \min_{i\in[k+1]} d_i =: r. \tag{19}$$

Proposition 3.3, on the other hand, states that any rank-1 network map can be implemented by an equalized network. Therefore, the key idea of the proof is to make sure that the global optimum of problem 19 has rank equal to one. It suffices to notice that under the assumption $\sigma_1(\bar{\mathrm{M}}) - \sigma_2(\bar{\mathrm{M}}) \geq \frac{1}{\nu_{\{d_i\}}}\sigma_2(\bar{\mathrm{M}})$, it holds that $\sigma_1(\bar{\mathrm{M}}) > \frac{\nu_{\{d_i\}}\sigma_1(\bar{\mathrm{M}})}{1+\nu_{\{d_i\}}}$ and $\sigma_j(\bar{\mathrm{M}}) \leq \frac{\nu_{\{d_i\}}\sigma_1(\bar{\mathrm{M}})}{1+\nu_{\{d_i\}}}$ for all $j > 1$. In this case, using Lemma A.4, the solution $\mathcal{S}_{\alpha_1}(\bar{\mathrm{M}})\mathrm{C}^{-\frac{1}{2}}$ has rank equal to one, which completes the proof. □