Supplementary Materials for “Dropout: Explicit Forms and Capacity Control”

A. Auxiliary Results

Lemma 1 (Khintchine-Kahane inequality)**.** *Let $\{\epsilon_i\}_{i=1}^n$ be i.i.d. Rademacher random variables, and $\{\mathrm{x}\}_{i=1}^n \subset \mathbb{R}^d$. Then there exist a pair of universal constants $c_1, c_2 > 0$ such that*

$$c_1\sqrt{\sum_{i=1}^n \|\mathrm{x}_i\|^2} \leq \mathbb{E}\|\sum_{i=1}^n \epsilon_i \mathrm{x}_i\| \leq c_2\sqrt{\sum_{i=1}^n \|\mathrm{x}_i\|^2}.$$

Theorem 4 (Hoeffding’s inequality: Theorem 2.6.2 (Vershynin, 2018))**.** *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-Gaussian random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\widehat{\mathbb{E}}_i X_i\right| \geq t\right) \leq 2e^{-\frac{ct^2N^2}{\sum_{i=1}^N \|X_i\|_{\psi_2}^2}}$$

Theorem 5 (Theorem 3.1 of Mohri et al. (2018))**.** *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over a sample $\mathcal{S} = \{z_1, \ldots, z_n\}$, the following holds for all $g \in \mathcal{G}$*

$$E[g(z)] \leq \frac{1}{n}\sum_{i=1}^n g(z_i) + 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

$$E[g(z)] \leq \frac{1}{n}\sum_{i=1}^n g(z_i) + 2\mathfrak{R}_{\mathcal{S}}(\mathcal{G}) + 3\sqrt{\frac{\log(1/\delta)}{2n}}$$

Theorem 6 (Theorem 10.3 of Mohri et al. (2018))**.** *Assume that $\|h - f\|_\infty \leq M$ for all $h \in \mathcal{H}$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over a sample $\mathcal{S} = \{(\mathrm{x}_i, \mathrm{y}_i),\ i \in [n]\}$ of size n, the following inequalities holds uniformly for all $h \in \mathcal{H}$.*

$$\mathbb{E}[|h(\mathrm{x}) - f(\mathrm{x})|^2] \leq \widehat{\mathbb{E}}_i|h(\mathrm{x}_i) - f(\mathrm{x}_i)|^2 + 4M\mathfrak{R}_n(\mathcal{H}) + M^2\sqrt{\frac{\log(2/\delta)}{2n}}$$

$$\mathbb{E}[|h(\mathrm{x}) - f(\mathrm{x})|^2] \leq \widehat{\mathbb{E}}_i|h(\mathrm{x}_i) - f(\mathrm{x}_i)|^2 + 4M\mathfrak{R}_{\mathcal{S}}(\mathcal{H}) + 3M^2\sqrt{\frac{\log(2/\delta)}{2n}}$$

Theorem 7 (Based on Theorem 1 in Srebro et al. (2010))**.** *Let $\mathcal{X}$ and $\mathcal{Y} = [-1,1]$ denote the input space and the label space, respectively. Let $\mathcal{H} \subseteq \{f : \mathcal{X} \to \mathcal{Y}\}$ be the target function class. For any $f \in \mathcal{H}$, and any $(\mathrm{x}, y) \in \mathcal{X} \times \mathcal{Y}$, let $\ell(f, \mathrm{x}, y) := (f(\mathrm{x}) - y)^2$ be the squared loss. Let $L(f) = \mathbb{E}_{\mathcal{D}}[\ell(f, \mathrm{x}, y)]$ be the population risk with respect to the joint distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. For any $\delta > 0$, with probability at least $1-\delta$ over a sample of size n, we have for any $f \in \mathcal{H}$:*

$$L(f) \leq L_* + K\left(\sqrt{L_*}\left(\sqrt{2}\log(n)^{1.5}\mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{4\log\frac{1}{\delta}}{n}}\right) + 2\log(n)^3\mathfrak{R}_n^2(\mathcal{H}) + \frac{4\log\frac{1}{\delta}}{n}\right)$$

where $L_* := \min_{f\in\mathcal{H}} L(f)$, and K is a numeric constant derived from Srebro et al. (2010).

Theorem 8 (Theorem 3.3 in Mianjy et al. (2018))**.** *For any pair of matrices $U \in \mathbb{R}^{d_2\times d_1}$, $V \in \mathbb{R}^{d_0\times d_1}$, there exist a rotation matrix $Q \in SO(d_1)$ such that rotated matrices $\tilde{U} := UQ, \tilde{V} := VQ$ satisfy $\|\tilde{u}_i\|\|\tilde{v}_i\| = \frac{1}{d_1}\|UV^\top\|_*$, for all $i \in [d_1]$.*

Theorem 9 (Theorem 1 in Foygel et al. (2011))**.** *Assume that $p(i)q(j) \geq \frac{\log(d_2)}{n\sqrt{d_2d_0}}$ for all $i \in [d_2], j \in [d_0]$. For any $\alpha > 0$, let $\mathcal{M}_\alpha := \{M \in \mathbb{R}^{d_2\times d_1} :\ \|\operatorname{diag}(\sqrt{p})M\operatorname{diag}(\sqrt{q})\|_*^2 \leq \alpha\}$ be the class of linear transformations with weighted trace-norm bounded with $\sqrt{\alpha}$. Then the expected Rademacher complexity of $\mathcal{M}_\alpha$ is bounded as follows:*

$$\mathfrak{R}_n(\mathcal{M}_\alpha) \leq O\left(\sqrt{\frac{\alpha d_2\log(d_2)}{n}}\right)$$

B. Matrix Sensing

Proposition 2 (Dropout regularizer in matrix sensing)**.** *The following holds for any $p \in [0,1)$:*

$$\widehat{L}_{drop}(U,V) = \widehat{L}(U,V) + \lambda \widehat{R}(U,V), \tag{6}$$

where $\widehat{R}(U,V) = \sum_{i=1}^{d_1} \widehat{\mathbb{E}}_j (u_i^\top A^{(j)} v_i)^2$ and $\lambda = \frac{p}{1-p}$ is the regularization parameter.

Proof of Proposition 2. Similar statements and proofs can be found in several previous works (Srivastava et al., 2014; Wang & Manning, 2013; Cavazza et al., 2018; Mianjy et al., 2018). For completeness, we include a proof here. The following equality follows from the definition of variance:

$$\mathbb{E}_\mathrm{b}[(y_i - \langle \mathrm{UBV}^\top, \mathrm{A}^{(i)}\rangle)^2] = \Big(\mathbb{E}_\mathrm{b}[y_i - \langle \mathrm{UBV}^\top, \mathrm{A}^{(i)}\rangle]\Big)^2 + \mathrm{Var}(y_i - \langle \mathrm{UBV}^\top, \mathrm{A}^{(i)}\rangle) \tag{7}$$

Recall that for a Bernoulli random variable B_{ii}, we have $\mathbb{E}[\mathrm{B}_{ii}] = 1$ and $\mathrm{Var}(\mathrm{B}_{ii}) = \frac{p}{1-p}$. Thus, the first term on right hand side is equal to $(y_i - \langle \mathrm{UV}^\top, \mathrm{A}^{(i)}\rangle)^2$. For the second term we have

$$\mathrm{Var}(y_i - \langle \mathrm{UBV}^\top, \mathrm{A}^{(i)}\rangle) = \mathrm{Var}(\sum_{j=1}^{d_1} \mathrm{B}_{jj} \mathrm{u}_j^\top \mathrm{A}^{(i)} \mathrm{v}_j) = \sum_{j=1}^{d_1} (\mathrm{u}_j^\top \mathrm{A}^{(i)} \mathrm{v}_j)^2 \,\mathrm{Var}(\mathrm{B}_{jj}) = \frac{p}{1-p} \sum_{j=1}^{d_1} (\mathrm{u}_j^\top \mathrm{A}^{(i)} \mathrm{v}_j)^2$$

Plugging the above into Equation (7) and averaging over samples we get

$$\begin{aligned}
\widehat{L}_{\text{drop}}(\mathrm{U},\mathrm{V}) &= \widehat{\mathbb{E}}_i \mathbb{E}_\mathrm{b}[(y_i - \langle \mathrm{UBV}^\top, \mathrm{A}^{(i)}\rangle)^2] \\
&= \widehat{\mathbb{E}}_i (y_i - \langle \mathrm{UV}^\top, \mathrm{A}^{(i)}\rangle)^2 + \widehat{\mathbb{E}}_i \frac{p}{1-p} \sum_{j=1}^{d_1} (\mathrm{u}_j^\top \mathrm{A}^{(i)} \mathrm{v}_j)^2 \\
&= \widehat{L}(\mathrm{U},\mathrm{V}) + \frac{p}{1-p} \widehat{R}(\mathrm{U},\mathrm{V}).
\end{aligned}$$

which completes the proof. □

Lemma 2 (Concentration in matrix completion)**.** *For $\ell \in [n]$, let $A^{(\ell)}$ be an indicator matrix whose (i,j)-th element is selected according to some distribution. Assume U, V is such that $\|U^\top\|_{2,\infty} \|V\|_{\infty,\infty} \leq \gamma$. Then, with probability at least $1-\delta$ over a sample of size n, we have that*

$$|R(U,V) - \widehat{R}(U,V)| \leq \frac{C\gamma^2 \sqrt{\log(2/\delta)}}{\sqrt{n}}.$$

Proof of Lemma 2. Define $X_\ell := \sum_{w=1}^{d_1} (\mathrm{u}_w^\top \mathrm{A}^{(\ell)} \mathrm{v}_w)^2$ and observe that

$$\begin{aligned}
X_\ell &= \sum_{w=1}^{d_1} \left(\sum_{i,j} \mathrm{U}_{iw} \mathrm{V}_{jw} \mathrm{A}_{ij}^{(\ell)} \right)^2 = \sum_{w=1}^{d_1} \sum_{i,i',j,j'} \mathrm{U}_{iw} \mathrm{U}_{i'w} \mathrm{V}_{jw} \mathrm{V}_{j'w} \mathrm{A}_{ij}^{(\ell)} \mathrm{A}_{i'j'}^{(\ell)} \\
&= \sum_{w=1}^{d_1} \sum_{i,j} \mathrm{U}_{iw}^2 \mathrm{V}_{jw}^2 \mathrm{A}_{ij}^{(\ell)} \leq \max_{i,j} \sum_{w=1}^{d_1} \mathrm{U}_{iw}^2 \mathrm{V}_{jw}^2 \\
&\leq \max_{i,j} \|\mathrm{U}(i,:)\|^2 \|\mathrm{V}(j,:)\|_\infty^2 = \|\mathrm{U}^\top\|_{2,\infty}^2 \|\mathrm{V}\|_{\infty,\infty}^2 \leq \gamma^2
\end{aligned}$$

where the third equality follows because for an indicator matrix $\mathrm{A}^{(\ell)}$, it holds that $\mathrm{A}_{ij}^{(\ell)} \mathrm{A}_{i'j'}^{(\ell)} = 0$ if $(i,j) \neq (i',j')$. Thus, $X_{w,\ell}$ is a sub-Gaussian (more strongly, bounded) random variable with mean $\mathbb{E}[X_\ell] = R(\mathrm{U},\mathrm{V})$ and sub-Gaussian norm $\|X_\ell\|_{\psi_2} \leq \gamma^2/\ln(2)$. Furthermore, $\|X_\ell - R(\mathrm{U},\mathrm{V})\|_{\psi_2} \leq C'\|X_\ell\|_{\psi_2} \leq C\gamma^2$, for some absolute constants C', C (Lemma 2.6.8 of Vershynin (2018)). Using Theorem 4, for $t = Cd_1\sqrt{\frac{\log 2/\delta}{n}}$ we get that:

$$\mathbb{P}\left(\left| \widehat{R}(\mathrm{U},\mathrm{V}) - R(\mathrm{U},\mathrm{V}) \right| \geq t \right) = \mathbb{P}\left(\left| \frac{1}{n} \sum_{\ell=1}^{n} X_\ell - R(\mathrm{U},\mathrm{V}) \right| \geq C\gamma^2 \sqrt{\frac{\log 2/\delta}{n}} \right) \leq \delta$$

which completes the proof. □

Proposition 3. *[Induced regularizer] For $j \in [n]$, let $A^{(j)}$ be an indicator matrix whose (i,k)-th element is selected randomly with probability $p(i)q(k)$, where $p(i)$ and $q(k)$ denote the probability of choosing the i-th row and the k-th column. Then $\Theta(M) = \frac{1}{d_1}\|\operatorname{diag}(\sqrt{p})UV^\top \operatorname{diag}(\sqrt{q})\|_*^2$.*

Proof of Proposition 3. For any pair of factors (U,V) it holds that

$$R(\mathrm{U},\mathrm{V}) = \sum_{i=1}^{d_1}\mathbb{E}(\mathrm{u}_i^\top \mathrm{A}\mathrm{v}_i)^2 = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\sum_{k=1}^{d_0} p(j)q(k)(\mathrm{u}_i^\top \mathrm{e}_j\mathrm{e}_k^\top \mathrm{v}_i)^2$$
$$= \sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\sum_{k=1}^{d_0} p(j)q(k)\mathrm{U}(j,i)^2\mathrm{V}(k,i)^2 = \sum_{i=1}^{d_1}\|\operatorname{diag}(\sqrt{p})\mathrm{u}_i\|^2\|\operatorname{diag}(\sqrt{q})\mathrm{v}_i\|^2$$

We can now lower bound the right hand side above as follows:

$$R(\mathrm{U},\mathrm{V}) \geq \frac{1}{d_1}\left(\sum_{i=1}^{d_1}\|\operatorname{diag}(\sqrt{p})\mathrm{u}_i\|\|\operatorname{diag}(\sqrt{q})\mathrm{v}_i\|\right)^2$$
$$= \frac{1}{d_1}\left(\sum_{i=1}^{d_1}\|\operatorname{diag}(\sqrt{p})\mathrm{u}_i\mathrm{v}_i^\top \operatorname{diag}(\sqrt{q})\|_*\right)^2$$
$$\geq \frac{1}{d_1}\left(\|\operatorname{diag}(\sqrt{p})\sum_{i=1}^{d_1}\mathrm{u}_i\mathrm{v}_i^\top \operatorname{diag}(\sqrt{q})\|_*\right)^2 = \frac{1}{d_1}\|\operatorname{diag}(\sqrt{p})\mathrm{U}\mathrm{V}^\top \operatorname{diag}(\sqrt{q})\|_*^2$$

where the first inequality is due to Cauchy-Schwartz and the second inequality follows from the triangle inequality. The equality right after the first inequality follows from the fact that for any two vectors a, b, $\|\mathrm{ab}^\top\|_* = \|\mathrm{ab}^\top\| = \|\mathrm{a}\|\|\mathrm{b}\|$. Since the inequalities hold for any U, V, it implies that

$$\Theta(\mathrm{U}\mathrm{V}^\top) \geq \frac{1}{d_1}\|\operatorname{diag}(\sqrt{p})\mathrm{U}\mathrm{V}^\top \operatorname{diag}(\sqrt{q})\|_*^2.$$

Applying Theorem 8 on $(\operatorname{diag}(\sqrt{p})\mathrm{U}, \operatorname{diag}(\sqrt{p})\mathrm{V})$, there exist a rotation matrix Q such that

$$\|\operatorname{diag}(\sqrt{p})\mathrm{U}\mathrm{q}_i\|\|\operatorname{diag}(\sqrt{q})\mathrm{V}\mathrm{q}_i\| = \frac{1}{d_1}\|\operatorname{diag}(\sqrt{p})\mathrm{U}\mathrm{V}^\top \operatorname{diag}(\sqrt{q})\|_*$$

We evaluate the expected dropout regularizer at UQ, VQ:

$$R(\mathrm{UQ},\mathrm{VQ}) = \sum_{i=1}^{d_1}\|\operatorname{diag}(\sqrt{p})\mathrm{U}\mathrm{q}_i\|^2\|\operatorname{diag}(\sqrt{q})\mathrm{V}\mathrm{q}_i\|^2$$
$$= \sum_{i=1}^{d_1}\frac{1}{d_1^2}\|\operatorname{diag}(\sqrt{p})\mathrm{U}\mathrm{V}^\top \operatorname{diag}(\sqrt{q})\|_*^2 = \frac{1}{d_1}\|\operatorname{diag}(\sqrt{p})\mathrm{U}\mathrm{V}^\top \operatorname{diag}(\sqrt{q})\|_*^2 \leq \Theta(\mathrm{U}\mathrm{V}^\top)$$

which completes the proof of the first part. □

Proof of Theorem 1. We use Theorem 6 to bound the population risk in terms of the Rademacher complexity of the target class. Define the class of predictors with weighted trace-norm bounded by $\sqrt{\alpha}$, i.e.

$$\mathcal{M}_\alpha = \{\mathrm{M}: \ \|\operatorname{diag}(\sqrt{\mathrm{p}})\mathrm{M}\operatorname{diag}(\sqrt{\mathrm{q}})\|_*^2 \leq \alpha\}.$$

In particular dropout empirical risk minimizers U, V belong to this class:

$$\|\operatorname{diag}(\sqrt{p})\mathrm{U}\mathrm{V}^\top \operatorname{diag}(\sqrt{q})\|_*^2 = d_1\Theta(\mathrm{U}\mathrm{V}^\top) \leq d_1 R(\mathrm{U},\mathrm{V}) \leq \alpha$$

where the first inequality holds by definition of the induced regularizer, and the second inequality follows from the assumption of the theorem. Since g is a contraction, by Talagrand's lemma and Theorem 9, we have that $\mathfrak{R}_n(g\circ\mathcal{M}_\alpha) \leq \mathfrak{R}_n(\mathcal{M}_\alpha) \leq$

$\sqrt{\frac{\alpha d_2 \log(d_2)}{n}}$. To obtain the maximum deviation parameter M in Theorem 6, we note that the assumption $\|\mathrm{M}_*\| \leq 1$ implies that $|\mathrm{M}_*(i,j)| \leq 1$ for all i, j, so that $g(\mathrm{M}_*) = \mathrm{M}_*$. We have that:

$$\max_{\mathrm{A}} |\langle \mathrm{M}_* - g(\mathrm{U}\mathrm{V}^\top), \mathrm{A}\rangle| = \max_{i,j} |\langle \mathrm{M}_* - g(\mathrm{U}\mathrm{V}^\top), \mathrm{e}_i\mathrm{e}_j^\top\rangle| \leq \max_{i,j} |\mathrm{M}_*(i,j)| + \max_{i,j} |\langle \mathrm{U}\mathrm{V}^\top, \mathrm{e}_i\mathrm{e}_j^\top\rangle| \leq \|\mathrm{M}_*\| + 1 \leq 2$$

Let $L(g(\mathrm{U}\mathrm{V}^\top)) := \mathbb{E}(y - \langle g(\mathrm{U}\mathrm{V}^\top), \mathrm{A}\rangle)^2$ and $\widehat{L}(g(\mathrm{U}\mathrm{V}^\top)) := \widehat{\mathbb{E}}_i(y_i - \langle g(\mathrm{U}\mathrm{V}^\top), \mathrm{A}^{(i)}\rangle)^2$ denote the *true risk* and the *empirical risk* of $g(\mathrm{U}\mathrm{V}^\top)$, respectively. Plugging the above results in Theorem 6, we get

$$\begin{aligned} L(g(\mathrm{U},\mathrm{V})) &\leq \widehat{L}(g(\mathrm{U},\mathrm{V})) + 8\mathfrak{R}_n(g \circ \mathcal{M}_\alpha) + 4\sqrt{\frac{\log(2/\delta)}{2n}} \\ &\leq \widehat{L}(\mathrm{U},\mathrm{V}) + 8\sqrt{\frac{\alpha d_2 \log(d_2)}{n}} + 4\sqrt{\frac{\log(2/\delta)}{2n}} \\ &\leq \widehat{L}(\mathrm{U},\mathrm{V}) + 8\sqrt{\frac{2\alpha d_2 \log(d_2) + \frac{1}{4}\log(2/\delta)}{n}} \end{aligned}$$

where the second inequality holds since $\widehat{L}(g(\mathrm{U},\mathrm{V})) \leq \widehat{L}(\mathrm{U},\mathrm{V})$. □

B.1. Optimistic Rates

As we discussed in the main text, under additional assumptions on the value of α, it is possible to give optimistic generalization bounds that decay as $\tilde{O}(\alpha d_2/n)$. This result is given as the following theorem.

Theorem 10. Assume that $d_2 \geq d_0$ and $\|\mathrm{M}_*\| \leq 1$. Furthermore, assume that $\min_{i,k} p(i)q(k) \geq \frac{\log(d_2)}{n\sqrt{d_2 d_0}}$. Let (U,V) be a minimizer of the dropout ERM objective in equation (2). Let α be such that $\max\{R(\mathrm{U},\mathrm{V}), \Theta(\mathrm{M}_*)\} \leq \alpha/d_1$. Then, for any $\delta \in (0,1)$, the following generalization bounds holds with probability at least $1-\delta$ over a sample of size n:

$$L(g(\mathrm{U}\mathrm{V}^\top)) \leq \frac{2K\log(n)^3 \alpha d_2 \log(d_2) + 4K\log(1/\delta)}{n}$$

where K is an absolute constant (Srebro et al., 2010), $g(\mathrm{M})$ thresholds M between $[-1,1]$, and $L(g(\mathrm{U}\mathrm{V}^\top)) := \mathbb{E}(y - \langle g(\mathrm{U}\mathrm{V}^\top), \mathrm{A}\rangle)^2$ is the *true risk* of $g(\mathrm{U}\mathrm{V}^\top)$.

Proof of Theorem 10. We use Theorem 7 to bound the population risk in terms of the Rademacher complexity of the target class. Define the class of predictors with weighted trace-norm bounded by $\sqrt{\alpha}$, i.e.

$$\mathcal{M}_\alpha = \{\mathrm{M} : \ \|\operatorname{diag}(\sqrt{\mathrm{p}})\mathrm{M}\operatorname{diag}(\sqrt{\mathrm{q}})\|_*^2 \leq \alpha\}.$$

In particular dropout empirical risk minimizers U, V belong to this class:

$$\|\operatorname{diag}(\sqrt{p})\mathrm{U}\mathrm{V}^\top\operatorname{diag}(\sqrt{q})\|_*^2 = d_1\Theta(\mathrm{U}\mathrm{V}^\top) \leq d_1 R(\mathrm{U},\mathrm{V}) \leq \alpha$$

where the first inequality holds by definition of the induced regularizer, and the second inequality follows from the assumption of the theorem. Moreover, by assumption $\Theta(\mathrm{M}_*) \leq \alpha$, we have that $\mathrm{M}_* \in \mathcal{M}_\alpha$. With this, we get that

$$L_* := \min_{M \in g \circ \mathcal{M}_\alpha} L(\mathrm{M}) \leq L(g(\mathrm{M}_*)) = L(g(\mathrm{M}_*)) = 0.$$

Since g is a contraction, by Talagrand's lemma and Theorem 9, we have that $\mathfrak{R}_n(g \circ \mathcal{M}_\alpha) \leq \mathfrak{R}_n(\mathcal{M}_\alpha) \leq \sqrt{\frac{\alpha d_2 \log(d_2)}{n}}$. Plugging the above in Theorem 6, we get

$$\begin{aligned} L(g(\mathrm{U},\mathrm{V})) &\leq 2K\log(n)^3\mathfrak{R}_n^2(g \circ \mathcal{M}_\alpha) + \frac{4K\log\frac{1}{\delta}}{n} \\ &\leq \frac{2K\log(n)^3\alpha d_2 \log(d_2) + 4K\log\frac{1}{\delta}}{n} \end{aligned}$$

□

C. Non-linear Neural Networks

Proposition 4 (Dropout regularizer in deep regression)**.**

$$\widehat{L}_{drop}(\mathrm{w}) = \widehat{L}(\mathrm{w}) + \widehat{R}(\mathrm{w}), \;\; \textit{where} \;\; \widehat{R}(\mathrm{w}) = \lambda \sum_{j=1}^{d_1} \|u_j\|^2 \widehat{a}_j^2.$$

where $\widehat{a}_j = \sqrt{\widehat{\mathbb{E}}_i a_j(\mathrm{x}_i)^2}$ *and* $\lambda = \frac{p}{1-p}$ *is the regularization parameter.*

Proof of Proposition 4. Similar statements and proofs can be found in several previous works (Srivastava et al., 2014; Wang & Manning, 2013; Cavazza et al., 2018; Mianjy et al., 2018). Here we include a proof for completeness. Recall that $\mathbb{E}[\mathrm{B}_{ii}] = 1$ and $\mathrm{Var}(\mathrm{B}_{ii}) = \frac{p}{1-p}$. Conditioned on x, y in the current mini-batch, we have that:

$$\mathbb{E}_{\mathrm{B}}\|\mathrm{y} - \mathrm{U}^\top \mathrm{B}\mathrm{a}(\mathrm{x})\|^2 = \sum_{i=1}^{d_2} \left(\mathbb{E}_{\mathrm{B}}[y_i - \mathrm{u}_i^\top \mathrm{B}\mathrm{a}(\mathrm{x})]\right)^2 + \sum_{i=1}^{d_2} \mathrm{Var}(y_i - \mathrm{u}_i^\top \mathrm{B}\mathrm{a}(\mathrm{x}))$$

Since $\mathbb{E}[\mathrm{B}] = \mathrm{I}$, the first term on right hand side is equal to $\|\mathrm{y} - \mathrm{U}^\top \mathrm{a}(\mathrm{x})\|^2$. For the second term we have

$$\sum_{i=1}^{d_2} \mathrm{Var}(y_i - \mathrm{u}_i^\top \mathrm{B}\mathrm{a}(\mathrm{x})) = \sum_{i=1}^{d_2} \mathrm{Var}(\sum_{j=1}^{d_1} \mathrm{U}_{j,i}\mathrm{B}_{jj}a_j(\mathrm{x})) = \sum_{i=1}^{d_2}\sum_{j=1}^{d_1} (\mathrm{U}_{j,i}a_j(\mathrm{x}))^2 \,\mathrm{Var}(\mathrm{B}_{jj}) = \frac{p}{1-p}\sum_{j=1}^{d_1} \|\mathrm{u}_j\|^2 a_j(\mathrm{x})^2$$

Thus, conditioned on the sample (x, y), we have that

$$\mathbb{E}_{\mathrm{B}}[\|\mathrm{y} - \mathrm{U}^\top \mathrm{B}\mathrm{a}(\mathrm{x})\|^2] = \|\mathrm{y} - \mathrm{U}^\top \mathrm{a}(\mathrm{x})\|^2 + \frac{p}{1-p}\sum_{j=1}^{d_1} \|\mathrm{u}_j\|^2 a_j(\mathrm{x})^2$$

Now taking the empirical average with respect to x, y, we get

$$\widehat{L}_{\mathrm{drop}}(\mathrm{w}) = \widehat{L}(\mathrm{w}) + \frac{p}{1-p}\sum_{j=1}^{d_1} \|\mathrm{u}_j\|^2 \widehat{a}_j^2 = \widehat{L}(\mathrm{w}) + \widehat{R}(\mathrm{w})$$

which completes the proof. □

Proposition 5. *Consider a two layer neural network* $f_{\mathrm{w}}(\cdot)$ *with ReLU activation functions in the hidden layer. Furthermore, assume that the marginal input distribution* $\mathbb{P}_{\mathcal{X}}(\mathrm{x})$ *is symmetric and isotropic, i.e.,* $\mathbb{P}_{\mathcal{X}}(\mathrm{x}) = \mathbb{P}_{\mathcal{X}}(-\mathrm{x})$ *and* $\mathbb{E}[\mathrm{x}\mathrm{x}^\top] = I$*. Then the following holds for the expected explicit regularizer due to dropout:*

$$R(\mathrm{w}) := \mathbb{E}[\widehat{R}(\mathrm{w})] = \frac{\lambda}{2}\sum_{i_0,i_1,i_2=1}^{d_0,d_1,d_2} U(i_1,i_2)^2 V(i_1,i_0)^2, \tag{8}$$

Proof of Proposition 5. Using Proposition 4, we have that:

$$R(\mathrm{w}) = \mathbb{E}[\widehat{R}(\mathrm{w})] = \lambda \sum_{j=1}^{d_1} \|\mathrm{u}_j\|^2 \mathbb{E}[\sigma(\mathrm{V}(j,:)^\top \mathrm{x})^2]$$

It remains to calculate the quantity $\mathbb{E}_{\mathrm{x}}[\sigma(\mathrm{V}(j,:)^\top \mathrm{x})^2]$. By symmetry assumption, we have that $\mathbb{P}_{\mathcal{X}}(\mathrm{x}) = \mathbb{P}_{\mathcal{X}}(-\mathrm{x})$. As a result, for any $\mathrm{v} \in \mathbb{R}^{d_0}$, we have that $\mathbb{P}(\mathrm{v}^\top \mathrm{x}) = \mathbb{P}(-\mathrm{v}^\top \mathrm{x})$ as well. That is, the random variable $z_j := \mathrm{W}_1(j,:)^\top \mathrm{x}$ is also symmetric about the origin. It is easy to see that $\mathbb{E}_z[\sigma(z)^2] = \frac{1}{2}\mathbb{E}_z[z^2]$.

$$\mathbb{E}_z[\sigma(z)^2] = \int_{-\infty}^{\infty} \sigma(z)^2 d\mu(z) = \int_{0}^{\infty} \sigma(z)^2 d\mu(z) = \int_{0}^{\infty} z^2 d\mu(z) = \frac{1}{2}\int_{\infty}^{\infty} z^2 d\mu(z) = \frac{1}{2}\mathbb{E}_z[z^2].$$

Plugging back the above identity in the expression of $R(\mathrm{w})$, we get that

$$R(\mathrm{w}) = \lambda \sum_{j=1}^{d_1} \|\mathrm{u}_j\|^2 \mathbb{E}[(\mathrm{V}(j,:)^\top \mathrm{x})^2] = \frac{\lambda}{2} \sum_{j=1}^{d_1} \|\mathrm{u}_j\|^2 \|\mathrm{V}(j,:)\|^2$$

where the second equality follows from the assumption that the distribution is isotropic. □

Proof of Proposition 1. For $\delta \in (0, \frac{1}{2})$, consider the following random variable:

$$\mathrm{x} = \begin{cases} [1;0] & \text{with probability } \delta \\ [\frac{-\delta}{1-\delta}; \frac{\sqrt{1-2\delta}}{1-\delta}] & \text{with probability } \frac{1-\delta}{2} \\ [\frac{-\delta}{1-\delta}; -\frac{\sqrt{1-2\delta}}{1-\delta}] & \text{with probability } \frac{1-\delta}{2} \end{cases}$$

It is easy to check that the x has zero mean and is supported on the unit sphere. Consider the vector $\mathrm{w} = [\frac{1}{\sqrt{\delta}}; 0]$. It is easy to check that x satisfies $R(\mathrm{w}) = \sqrt{\mathbb{E}\sigma(\mathrm{w}^\top \mathrm{x})^2} = 1$; however, for any given C, it holds that $\|\mathrm{w}\| \geq C$ as long as we let $\delta = C^2$. □

Proof of Theorem 2. For any $j \in [h]$, let $a_j^2 := \mathbb{E}[\sigma(\mathrm{v}_j^\top \mathrm{x})^2]$ denote the average squared activation of the j-th node with respect to the input distribution. Given n i.i.d. samples $\mathcal{S} = \{\mathrm{x}_1, \cdots, \mathrm{x}_n\}$, the empirical Rademahcer complexity is bounded as follows:

$$\begin{aligned}
\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_\alpha) &= \mathbb{E}_\zeta \sup_{f_{\{\mathrm{u},\mathrm{v}\}} \in \mathcal{F}_\alpha} \frac{1}{n} \sum_{j=1}^{h} u_j a_j \sum_{i=1}^{n} \zeta_i \frac{\sigma(\mathrm{v}_j^\top \mathrm{x}_i)}{a_j} \\
&\leq \mathbb{E}_\zeta \sup_{f_{\{\mathrm{u},\mathrm{v}\}} \in \mathcal{F}_\alpha} \frac{1}{n} \sum_{j=1}^{h} |u_j a_j| \, |\sum_{i=1}^{n} \zeta_i \frac{\sigma(\mathrm{v}_j^\top \mathrm{x}_i)}{a_j}| \\
&\leq \mathbb{E}_\zeta \left[\left(\sup_{f_{\{\mathrm{u},\mathrm{v}\}} \in \mathcal{F}_\alpha} \sum_{j=1}^{h} |u_j a_j| \right) \left(\sup_{\mathrm{v}} \max_{j \in [h]} |\frac{1}{n} \sum_{i=1}^{n} \zeta_i \frac{\sigma(\mathrm{v}_j^\top \mathrm{x}_i)}{a_j}| \right) \right]
\end{aligned}$$

where we used the fact that the supremum of product of positive functions is upperbounded by the product of the supremums. By definition of $\mathcal{F}_\alpha$, the first term on the right hand side is bounded by α. To bound the second term in the right hand side, we note that the maximum over rows of $\mathrm{V}^\top$ can be absorbed into the supremum.

$$\begin{aligned}
\frac{1}{n} \mathbb{E}_\zeta \sup_{\mathrm{v}} |\sum_{i=1}^{n} \zeta_i \frac{\sigma(\mathrm{v}^\top \mathrm{x}_i)}{\sqrt{\mathbb{E}[\sigma(\mathrm{v}^\top \mathrm{x})^2]}}| &= \frac{1}{n} \mathbb{E}_\zeta \sup_{\mathbb{E}[\sigma(\mathrm{v}^\top \mathrm{x})^2] \leq 1} |\sum_{i=1}^{n} \zeta_i \sigma(\mathrm{v}^\top \mathrm{x}_i)| \\
&\leq \frac{2}{n} \mathbb{E}_\zeta \sup_{\mathbb{E}[\sigma(\mathrm{v}^\top \mathrm{x})^2] \leq 1} \sum_{i=1}^{n} \zeta_i \sigma(\mathrm{v}^\top \mathrm{x}_i) \\
&\leq \frac{2}{n} \mathbb{E}_\zeta \sup_{\beta \mathbb{E}(\mathrm{v}^\top \mathrm{x})^2 \leq 1} \sum_{i=1}^{n} \zeta_i \sigma(\mathrm{v}^\top \mathrm{x}_i) \qquad (\beta\text{-retentiveness})
\end{aligned}$$

Let $\mathrm{C}^\dagger$ be the pseudo-inverse of C. We perform the following change the variable: $\mathrm{w} \leftarrow \mathrm{C}^{-\dagger/2}\mathrm{v}$.

$$\begin{aligned}
\text{R.H.S.} &\leq \frac{2}{n}\mathbb{E}_\zeta \sup_{\mathbb{E}[(\mathrm{w}^\top \mathrm{C}^{\dagger/2}\mathrm{x})^2]\leq 1/\beta} \sum_{i=1}^n \zeta_i \mathrm{w}^\top \mathrm{C}^{\dagger/2}\mathrm{x}_i \\
&= \frac{2}{n}\mathbb{E}_\zeta \sup_{\|\mathrm{w}\|^2\leq 1/\beta} \langle \mathrm{w}, \sum_{i=1}^n \zeta_i \mathrm{C}^{\dagger/2}\mathrm{x}_i\rangle \\
&= \frac{2}{n\sqrt{\beta}}\mathbb{E}_\zeta \| \sum_{i=1}^n \zeta_i \mathrm{C}^{\dagger/2}\mathrm{x}_i \| \\
&\leq \frac{2}{n\sqrt{\beta}}\sqrt{\mathbb{E}_\zeta \| \sum_{i=1}^n \zeta_i \mathrm{C}^{\dagger/2}\mathrm{x}_i \|^2} = \frac{2}{n\sqrt{\beta}}\sqrt{\sum_{i=1}^n \mathrm{x}_i^\top \mathrm{C}^\dagger \mathrm{x}_i}
\end{aligned}$$

where the last inequality holds due to Jensen's inequality. To bound the expected Rademacher complexity, we take the expected value of both sides with respect to sample $\mathcal{S}$, which gives the following:

$$\mathfrak{R}_n(\mathcal{F}_\alpha) = \mathbb{E}_\mathrm{x}[\mathfrak{R}_\mathcal{S}(\mathcal{F}_\alpha)] \leq \frac{2}{n\sqrt{\beta}}\mathbb{E}_\mathcal{S}\sqrt{\sum_{i=1}^n \mathrm{x}_i^\top \mathrm{C}^\dagger \mathrm{x}_i} \leq \frac{2}{n\sqrt{\beta}}\sqrt{\sum_{i=1}^n \mathbb{E}_{\mathrm{x}_i}[\mathrm{x}_i^\top \mathrm{C}^\dagger \mathrm{x}_i]},$$

where the last inequality holds again due to Jensen's inequality. Finally, we have that $\mathbb{E}_{\mathrm{x}_i}\mathrm{x}_i^\top \mathrm{C}^\dagger \mathrm{x}_i = \mathbb{E}_{\mathrm{x}_i}\langle \mathrm{x}_i\mathrm{x}_i^\top, \mathrm{C}^\dagger\rangle = \langle \mathrm{C}, \mathrm{C}^\dagger\rangle = \mathrm{Rank}(\mathrm{C})$, which completes the proof of the Theorem. □

Proof of Theorem 3. For simplicity, assume that the width of the hidden layer is even. Consider the linear function class:

$$\mathcal{G}_r := \{g_\mathrm{w} : \mathrm{x} \mapsto \mathrm{w}^\top \mathrm{x},\ \mathbb{E}(\mathrm{w}^\top \mathrm{x})^2 \leq d_1 r/2\}.$$

Recall that $\mathcal{H}_r := \{h_\mathrm{w} : \mathrm{x} \mapsto \mathrm{u}^\top \sigma(\mathrm{V}^\top \mathrm{x}),\ R(\mathrm{u}, \mathrm{V}) \leq r\}$. First, we argue that $\mathcal{G}_r \subset \mathcal{H}_r$. Let $g_\mathrm{w} \in \mathcal{G}_r$; we show that there exist u, V such that $g_\mathrm{w} = f_{\mathrm{u},\mathrm{V}}$ and $f_{\mathrm{u},\mathrm{V}} \in \mathcal{H}_r$. Define $\mathrm{u} := \frac{2}{d_1}[1; -1; \cdots 1; -1] \in \mathbb{R}^{d_1}$, and let $\mathrm{V} = \mathrm{w}(\mathrm{e}_1 - \mathrm{e}_2 + \mathrm{e}_3 - \mathrm{e}_4 + \cdots + \mathrm{e}_{d_1-1} - \mathrm{e}_{d_1})^\top$, where $\mathrm{e}_i \in \mathbb{R}^{d_1}$ is the i-th standard basis vector. It's easy to see that

$$\begin{aligned}
f_{\mathrm{u},\mathrm{V}}(\mathrm{x}) = \mathrm{u}^\top \sigma(\mathrm{V}^\top \mathrm{x}) &= \sum_{i=1}^{d_1} u_i \sigma(\mathrm{v}_i^\top \mathrm{x}) \\
&= \sum_{i=1}^{d_1} \frac{2}{d_1}(-1)^{i-1}\sigma(\mathrm{v}_i^\top \mathrm{x}) \\
&= \sum_{i=1}^{d_1/2} \frac{2}{d_1}(\sigma(\mathrm{v}_{2i-1}^\top \mathrm{x}) - \sigma(\mathrm{v}_{2i}^\top \mathrm{x})) \\
&= \sum_{i=1}^{d_1/2} \frac{2}{d_1}(\sigma(\mathrm{w}^\top \mathrm{x}) - \sigma(-\mathrm{w}^\top \mathrm{x})) = \mathrm{w}^\top \mathrm{x} = g_\mathrm{w}.
\end{aligned}$$

Furthermore, it holds for the explicit regularizer that

$$\begin{aligned}
R(\mathrm{u}, \mathrm{V}) = \sum_{i=1}^{d_1} u_i^2 \mathbb{E}\sigma(\mathrm{v}_i^\top \mathrm{x})^2 &= \sum_{i=1}^{d_1/2} \frac{4}{d_1^2}\left(\mathbb{E}\sigma(\mathrm{v}_{2i-1}^\top \mathrm{x})^2 + \mathbb{E}\sigma(\mathrm{v}_{2i}^\top \mathrm{x})^2\right) \\
&= \sum_{i=1}^{d_1/2} \frac{4}{d_1^2}\mathbb{E}[\sigma(\mathrm{w}^\top \mathrm{x})^2 + \sigma(-\mathrm{w}^\top \mathrm{x})^2] \\
&= \frac{2}{d_1}\mathbb{E}(\mathrm{w}^\top \mathrm{x})^2 \leq r
\end{aligned}$$

Thus, we have that $\mathcal{G}_r \subset \mathcal{H}_r$, and the following inequalities follow.

$$\begin{aligned}
\mathfrak{R}_{\mathcal{S}}(\mathcal{H}_r) \geq \mathfrak{R}_{\mathcal{S}}(\mathcal{G}_r) &= \mathbb{E}_{\epsilon_i} \sup_{g_{\mathrm{w}} \in \mathcal{G}_r} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g_{\mathrm{w}}(\mathrm{x}_i) \\
&= \mathbb{E}_{\epsilon_i} \sup_{\mathbb{E}(\mathrm{w}^\top \mathrm{x})^2 \leq d_1 r/2} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathrm{w}^\top \mathrm{x}_i \\
&= \mathbb{E}_{\epsilon_i} \sup_{\mathrm{w}^\top \mathrm{C} \mathrm{w} \leq d_1 r/2} \frac{1}{n} \langle \mathrm{w}, \sum_{i=1}^{n} \epsilon_i \mathrm{x}_i \rangle \\
&= \mathbb{E}_{\epsilon_i} \sup_{\|\mathrm{C}^{1/2}\mathrm{w}\|^2 \leq d_1 r/2} \frac{1}{n} \langle \mathrm{C}^{1/2}\mathrm{w}, \sum_{i=1}^{n} \epsilon_i \mathrm{C}^{-\dagger/2} \mathrm{x}_i \rangle \\
&= \frac{\sqrt{d_1 r}}{\sqrt{2n}} \mathbb{E}_{\epsilon_i} \| \sum_{i=1}^{n} \epsilon_i \mathrm{C}^{\dagger/2} \mathrm{x}_i \| \\
&\geq \frac{c\sqrt{d_1 r}}{\sqrt{2n}} \sqrt{\sum_{i=1}^{n} \|\mathrm{C}^{\dagger/2}\mathrm{x}_i\|^2} = \frac{c\sqrt{d_1 r}\|\mathrm{X}\|_{\mathrm{C}^\dagger}}{\sqrt{2n}}
\end{aligned}$$

where the last inequality follows from Khintchine-Kahane inequality in Lemma 1. □

Next, we define some function classes that will be used frequently in the proofs.

Definition 1. *For any closed subset $[a, b] \subset \mathbb{R}$, let $\Pi_{[a,b]}(y) := \max\{a, \min\{b, y\}\}$. For $z := (\mathrm{x}, y)$ and $f : \mathcal{X} \to \mathcal{Y}$, define the squared loss $\ell_2(f, z) := (1 - yf(\mathrm{x}))^2$. For a given value $\alpha > 0$, consider the following classes*

$$\begin{aligned}
\mathcal{W}_\alpha &:= \{\mathrm{w} = (u, V) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_0 \times d_1},\ \sum_{i=1}^{d_1} |u_i| \sqrt{\mathbb{E}\sigma(v_i^\top \mathrm{x})^2} \leq \alpha\} \\
\mathcal{F}_\alpha &:= \{f_{\mathrm{w}} : \mathrm{x} \mapsto u^\top \sigma(V^\top \mathrm{x}),\ \mathrm{w} \in \mathcal{W}_\alpha\}, \\
\mathcal{G}_\alpha &:= \Pi_{[-1,1]} \circ \mathcal{F}_\alpha = \{g_{\mathrm{w}} = \Pi_{[-1,-1]} \circ f_{\mathrm{w}},\ f_{\mathrm{w}} \in \mathcal{F}_\alpha\} \\
\mathcal{L}_\alpha &:= \{\ell_2 : (g_{\mathrm{w}}, z) \mapsto (y - g_{\mathrm{w}}(\mathrm{x}))^2,\ g_{\mathrm{w}} \in \mathcal{G}_\alpha\}
\end{aligned}$$

Lemma 3. *Let $\mathcal{W}_\alpha, \mathcal{F}_\alpha, \mathcal{G}_\alpha, \mathcal{L}_\alpha$ be as defined in Definition 1. Then the following holds true:*

1. $\mathfrak{R}_{\mathcal{S}}(\mathcal{G}_\alpha) \leq \mathfrak{R}_{\mathcal{S}}(\mathcal{F}_\alpha)$.
2. *If $\mathcal{Y} = \{-1, +1\}$ (binary classification), then it holds that $\mathfrak{R}_{\mathcal{S}}(\mathcal{L}_\alpha) \leq 2\mathfrak{R}_{\mathcal{S}}(\mathcal{G}_\alpha)$.*

Proof. Since $\Pi_{[-1,-1]}(\cdot)$ is 1-Lipschitz, by Talagrand's contraction lemma, we have that $\mathfrak{R}_{\mathcal{S}}(\mathcal{G}_\alpha) \leq \mathfrak{R}_{\mathcal{S}}(\mathcal{F}_\alpha)$. The second claim follows from

$$\begin{aligned}
\mathfrak{R}_{\mathcal{S}}(\mathcal{L}_\alpha) &= \mathbb{E}_\zeta \sup_{\mathrm{w} \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \zeta_i (y_i - g_{\mathrm{w}}(\mathrm{x}_i))^2 \\
&= \mathbb{E}_\zeta \sup_{\mathrm{w} \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \zeta_i (1 - y_i g_{\mathrm{w}}(\mathrm{x}_i))^2 && (y_i \in \{-1, +1\}) \\
&\leq 2\mathbb{E}_\zeta \sup_{\mathrm{w} \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \zeta_i y_i g_{\mathrm{w}}(\mathrm{x}_i) \\
&= 2\mathbb{E}_\zeta \sup_{\mathrm{w} \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \zeta_i g_{\mathrm{w}}(\mathrm{x}_i) = 2\mathfrak{R}_{\mathcal{S}}(\mathcal{G}_\alpha)
\end{aligned}$$

where the first inequality follows from Talagrand's contraction lemma due to the fact that $h(z) = (1 - z)^2$ is 2-Lipschitz for $z \in [-1, 1]$, and the penultimate holds true since for any fixed $(y_i)_{i=1}^n \in \{-1, +1\}^n$, the distribution of $(\zeta_1 y_1, \ldots, \zeta_n y_n)$ is the same as that of $(\zeta_1, \ldots, \zeta_n)$. □

Proof of Corollary 1. We use the standard generalization bound in Theorem 6 for class $\mathcal{G}_\alpha$:

$$\begin{aligned} L_{\mathcal{D}}(g_{\mathrm{w}}) &\le \widehat{L}_{\mathcal{S}}(g_{\mathrm{w}}) + 4M\mathfrak{R}_{\mathcal{S}}(\mathcal{G}_\alpha) + 3M^2\sqrt{\frac{\log(2/\delta)}{2n}} \\ &\le \widehat{L}_{\mathcal{S}}(g_{\mathrm{w}}) + 8\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_\alpha) + 12\sqrt{\frac{\log(2/\delta)}{2n}} && \text{(Lemma 3)} \\ &\le \widehat{L}_{\mathcal{S}}(g_{\mathrm{w}}) + \frac{16\alpha\|\mathrm{X}\|_{\mathrm{C}^\dagger}}{\sqrt{\beta}n} + 12\sqrt{\frac{\log(2/\delta)}{2n}} && \text{(Theorem 2)} \end{aligned}$$

where second inequality follows because the maximum deviation parameter M in Theorem 6 is bounded as

$$M = \sup_{\mathrm{w}\in\mathcal{W}} \sup_{(\mathrm{x},y)\in\mathcal{X}\times\mathcal{Y}} |y - g_{\mathrm{w}}(\mathrm{x})| \le \sup_{\mathrm{w}\in\mathcal{W}} \sup_{(\mathrm{x},y)\in\mathcal{X}\times\mathcal{Y}} |y| + |g_{\mathrm{w}}(\mathrm{x})| \le 2.$$

□

Proof of Corollary 2. Recall that the input is jointly distributed as $(\mathrm{x}, y) \sim \mathcal{D}$. For $\mathcal{X} \subseteq \mathbb{R}_+^{d_0}$, let $\mathcal{X}' = \mathcal{X} \cup -\mathcal{X}$ be the *symmetrized input domain*. Let ζ be a Rademacher random variable. Denote the *symmetrized input* by $\mathrm{x}' = \zeta\mathrm{x}$, and the joint distribution of (x', y) by $\mathcal{D}'$. By construction, $\mathcal{D}'$ is centrally symmetric w.r.t. x', i.e., it holds for all $(\mathrm{x}, y) \in \mathcal{X} \times \mathcal{Y}$ that $\mathcal{D}'(\mathrm{x}, y) = \mathcal{D}'(-\mathrm{x}, y) = \frac{1}{2}\mathcal{D}(\mathrm{x}, y)$. As a result, population risk with respect to the original distribution $\mathcal{D}$ can be bounded in terms of the population risk with respect to the *symmetrized distribution* $\mathcal{D}'$ as follows:

$$\begin{aligned} L_{\mathcal{D}}(f) &:= \mathbb{E}_{\mathcal{D}}[\ell(f(\mathrm{x}), y)] \\ &\le \mathbb{E}_{\mathcal{D}}[\ell(f(\mathrm{x}), y) + \ell(f(-\mathrm{x}), y)] \\ &= 2\mathbb{E}_{\mathcal{D}}[\frac{1}{2}\ell(f(\mathrm{x}), y) + \frac{1}{2}\ell(f(-\mathrm{x}), y)] \\ &= 2\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\zeta}[\ell(f(\zeta\mathrm{x}), y) \mid \mathrm{x}, y] \\ &= 2\mathbb{E}_{\mathcal{D}'}[\ell(f(\mathrm{x}'), y)] = 2L_{\mathcal{D}'}(f) \end{aligned} \tag{9}$$

Moreover, since $\mathcal{D}'$ is centrally symmetric, Assumption 1 holds with $\beta = \frac{1}{2}$. The proof of Corollary 2 follows by doubling the right hand side of inequalities in Corollary 1, and substituting $\beta = \frac{1}{2}$. □

C.1. Classification

Although in the main text we only focus on the task of regression with squared loss, it is not hard to extend the results to binary classification. In particular, the following two Corollaries bound the miss-classification error in terms of the training error and the Rademacher complexity of the target class, with and without symmetrization.

Corollary 3. *Consider a binary classification setting where $\mathcal{Y} = \{-1, +1\}$. For any $\mathrm{w} \in \mathcal{F}_\alpha$, for any $\delta \in (0, 1)$, the following generalization bound holds with probability at least $1 - \delta$ over $\mathcal{S} = \{(\mathrm{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$:*

$$\mathbb{P}\{yf_{\mathrm{w}}(\mathrm{x}) < 0\} \le \widehat{L}_{\mathcal{S}}(g_{\mathrm{w}}) + \frac{8\alpha\|\mathrm{X}\|_{\mathrm{C}^\dagger}}{\sqrt{\beta}n} + 4\sqrt{\frac{\log(1/\delta)}{2n}}$$

where $g_{\mathrm{w}}(\cdot) = \max\{-1, \min\{1, f_{\mathrm{w}}(\cdot)\}\}$ projects the network output onto the range $[-1, 1]$.

Proof of Corollary 3. We use the standard generalization bound in Theorem 5. Recall that $g_w = \Pi_{[-1,1]}(f_{\mathrm{w}})$, where $\Pi_{[-1,1]}(y) = \max\{-1, \max\{1, y\}\}$ projects onto the range $[-1, 1]$. It is easy to bound the classification error of f_{w} in terms of the ℓ_2-loss of g_{w}:

$$\mathbb{P}\{\mathrm{sgn}(f_{\mathrm{w}}(\mathrm{x})) \ne y\} = \mathbb{P}\{yf_{\mathrm{w}}(\mathrm{x}) < 0\} = \mathbb{E}[1_{yf_{\mathrm{w}}(\mathrm{x})<0}] = \mathbb{E}[1_{yg_{\mathrm{w}}(\mathrm{x})<0}] \le \mathbb{E}(1 - yg_{\mathrm{w}}(\mathrm{x}))^2 = L_{\mathcal{D}}(g_{\mathrm{w}}). \tag{10}$$

We use Theorem 5 for class $\frac{1}{4}\mathcal{L}_\alpha$ to get the generalization bound as follows:

$$
\begin{aligned}
&\frac{1}{4}L_{\mathcal{D}}(g_{\mathrm{w}}) \leq \frac{1}{4}\widehat{L}_{\mathcal{S}}(g_{\mathrm{w}}) + 2\mathfrak{R}_{\mathcal{S}}(\frac{1}{4}\mathcal{L}_\alpha) + \sqrt{\frac{\log(1/\delta)}{2n}} \\
\implies &L_{\mathcal{D}}(g_{\mathrm{w}}) \leq \widehat{L}_{\mathcal{S}}(g_{\mathrm{w}}) + 4\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_\alpha) + 4\sqrt{\frac{\log(1/\delta)}{2n}} && \text{(by Lemma 3)} \\
\implies &L_{\mathcal{D}}(g_{\mathrm{w}}) \leq \widehat{L}_{\mathcal{S}}(g_{\mathrm{w}}) + \frac{8\alpha\|\mathrm{X}\|_{\mathrm{C}^\dagger}}{\sqrt{\beta}n} + 4\sqrt{\frac{\log(1/\delta)}{2n}} && \text{(by Theorem 2)}
\end{aligned}
$$

□

Corollary 4. *Consider a binary classification setting where $\mathcal{Y} = \{-1, +1\}$. For any $\mathrm{w} \in \mathcal{F}'_\alpha$, for any $\delta \in (0, 1)$, the following generalization bound holds with probability at least $1 - \delta$ over a sample of size n and the randomization in symmetrization*

$$
\mathbb{P}\{y g_{\mathrm{w}}(\mathrm{x}) < 0\} \leq 2\widehat{L}_{\mathcal{S}'}(g_{\mathrm{w}}) + \frac{23\alpha'\|\mathrm{X}\|_{\mathrm{C}^\dagger}}{n} + 8\sqrt{\frac{\log(1/\delta)}{2n}}
$$

Proof of Corollary 4. Akin to proof of Corollary 2, we have that $L_{\mathcal{D}}(f) \leq 2L_{\mathcal{D}'}(f)$, and the marginal distribution is $\frac{1}{2}$-retentive. Proof of Corollary 4 follows by doubling the right hand side of inequalities in Corollary 3, and substituting $\beta = \frac{1}{2}$. □

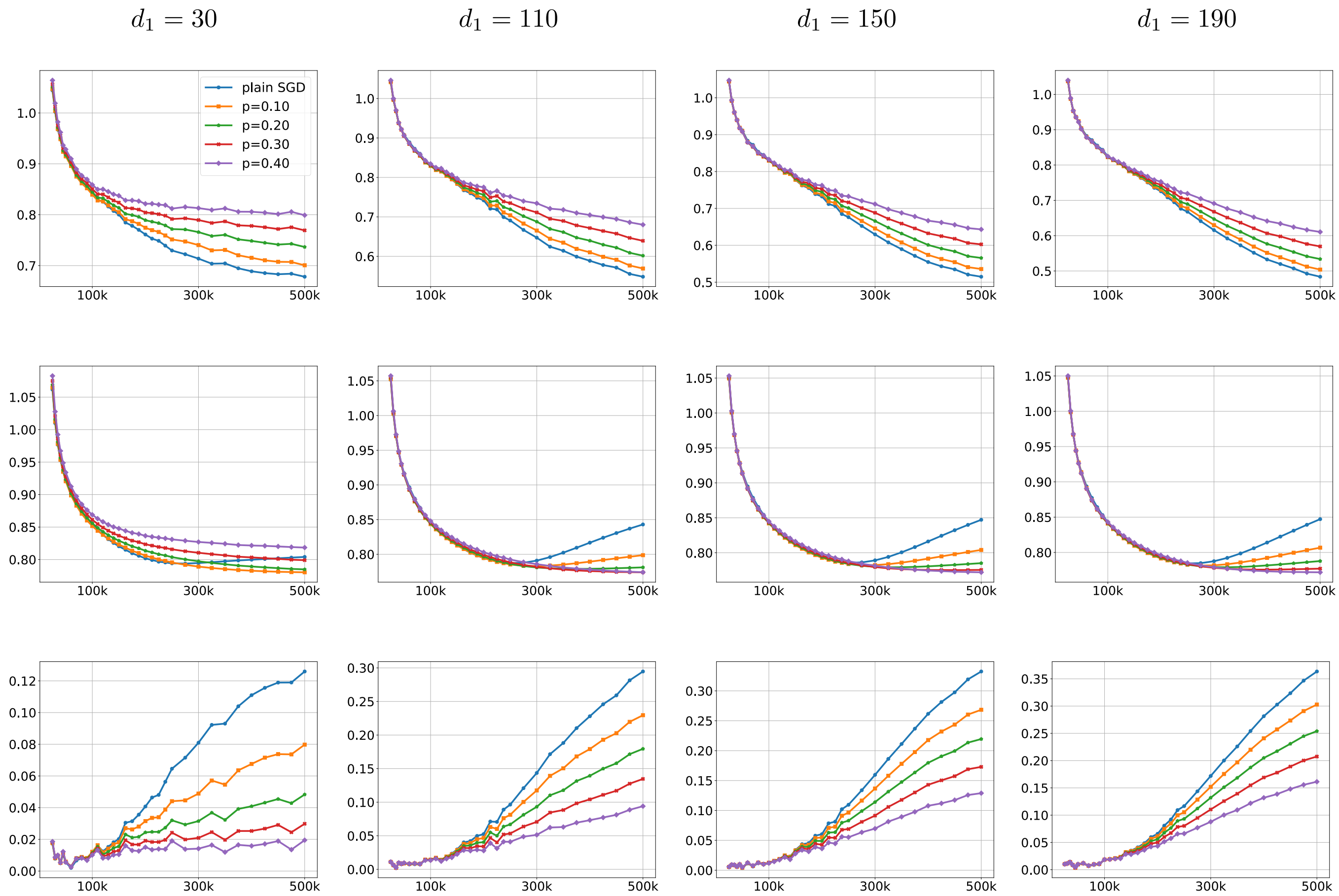

Figure 1. MovieLens dataset: the training error (**top**), the test error (**middle**), and the generalization gap for plain SGD as well as dropout with $p \in \{0.25, 0.50, 0.75\}$ as a function of the number of iterates, for different factorization sizes $d_1 = 30$ (first column), $d_1 = 110$ (second column), $d_1 = 150$ (third column), and $d_1 = 190$ (forth column).

D. Additional Experiments

In this section, we include additional plots which was not reported in the main paper due to the space limitations.

D.1. Matrix Completion

Figure 1 in the main paper shows comparisons between plain SGD and the dropout algorithm on the MovieLens dataset for a factorization size of $d_1 = 70$. The observation that we make with regard to those plots is not at all limited to the specific choice of the factorization size. In Figure 1 here, we report similar experiments with factorization sizes $d_1 \in \{30, 110, 150, 190\}$. It can be seen that the overall behaviour of plain SGD and dropout are very similar in all experiments. In particular, plain SGD always achieves the best training error but it has the largest generalization gap. Furthermore, increasing the dropout rate increases the training error but results in a tighter generalization gap.

It can be seen that an appropriate choice of the dropout rate *always* perform better than the plain SGD in terms of the test error. For instance, a dropout rate of $p = 0.2$ seems to always outperform plain SGD. Moreover, as the factorization size increases, the function class becomes more complex, and a larger value of the dropout rate is more helpful. For example, when $d_1 = 30$, the dropout with rates $p = 0.3, 0.4$ fail to achieve a good test performance, where as for larger factorization sizes ($d_1 \in \{110, 150, 190\}$), they consistently outperform plain SGD as well as other dropout rates.

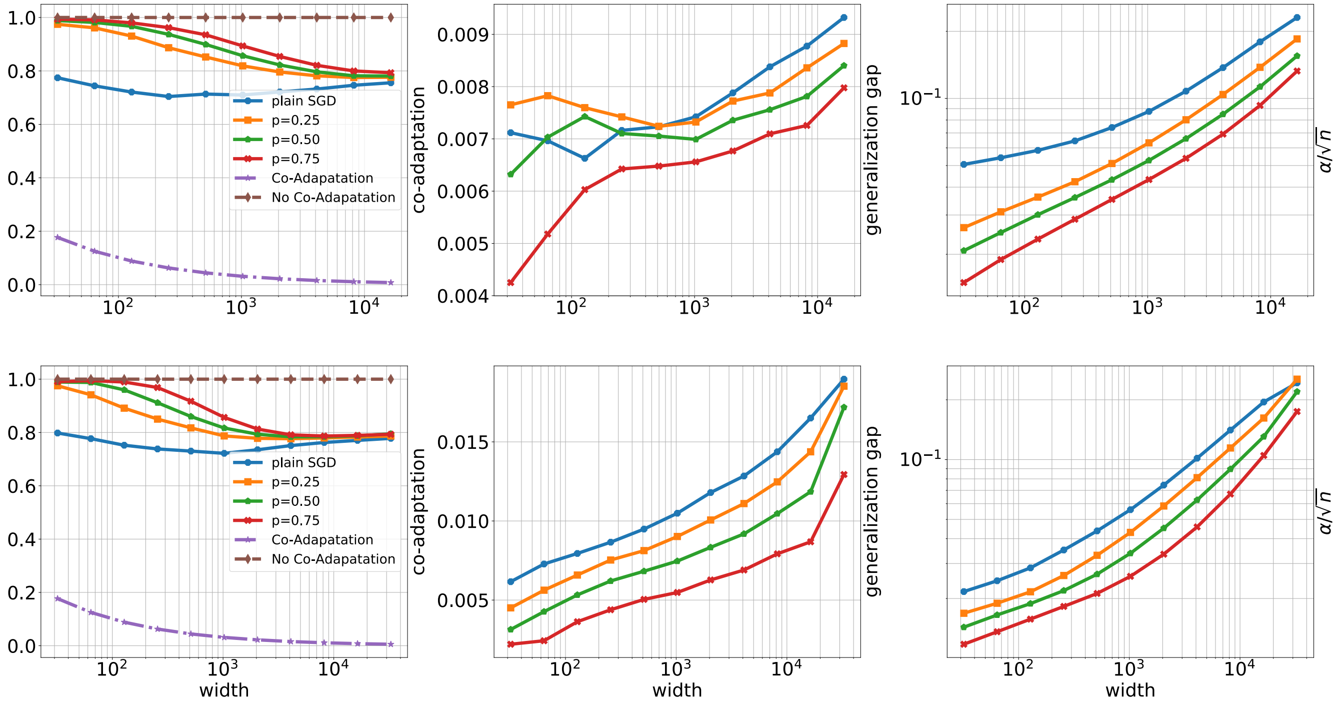

Figure 2. (**left**) *"co-adaptation"*; (**middle**) generalization gap; and (**right**) $\alpha/\sqrt{n}$ (**top**) with symmetrization on FashionMNIST; and (**bottom**) without symmetrization on MNIST. In left column, the dashed brown and dotted purple lines represent minimal and maximal co-adaptations, respectively.

D.2. Shallow Neural Networks

In Figure 2, we plot the co-adaptation measure, the generalization gap, as well as the complexity measure $\alpha/\sqrt{n}$ as a function of width of the network, for FashionMNIST with symmetrization, and for MNIST without symmetrization.

The co-adaptation plot is very similar to Figure 2 in the main text. In particular, 1) increasing the dropout rate results in less co-adaptation; 2) even plain SGD is biased towards networks with less co-adaptation; and 3) as the networks becomes wider, the co-adaptation curves corresponding to plain SGD converge to those of dropout. We also make similar observations for the generalization gap as well as the complexity term $\alpha/\sqrt{n}$. In particular, 1) a higher dropout rate corresponds to a lower generalization gap, uniformly for all widths; 2) the generalization gap is higher for wider networks; and 3) curves with smaller complexity terms in the right plot correspond to curves with smaller generalization gaps in the middle plot.

D.3. Deep Neural Networks

In Section 3, we derived generalization bounds that scale with the explicit regularizer as $O(\sqrt{\frac{\texttt{width}\cdot R(\mathrm{w})}{n}})$. Although our theoretical analysis is limited to two-layer networks; empirically, we show in Figure 3 that the generalization gap correlates well with this measure even for deep neural networks. In particular, we train deep convolutional neural networks with a dropout layer on top of the feature extractor, i.e. the top hidden layer. Let $\texttt{feature}_i$ denote the i-th hidden node in the top hidden layer. Akin to the derivation presented in Proposition 4, it is easy to see that the (expected) explicit regularizer is given by $R(\mathrm{w}) = \frac{p}{1-p}\sum_{i=1}^{\texttt{width}} \|\mathrm{u}_i\|^2 a_i^2$, where $\texttt{width}$ is the width of the top hidden layer, U denotes the top layer weight matrix, and $a_i^2 = \mathbb{E}_{\mathrm{x}}[\texttt{feature}_i(\mathrm{x})^2]$ is the second moment of the i-th node in the top hidden layer.

We train convolutional neural networks with and without dropout, on MNIST, Fashion MNIST, and CIFAR-10. The CIFAR-10 dataset consists of 60K 32 × 32 color images in 10 classes, with 6k images per class, divided into a training set and a test set of sizes 50K and 10K respectively (Krizhevsky et al., 2009). We do not perform symmetrization in these experiments. In contrast with the experiments in the previous section, here we run the experiments on full datasets, representing each of the ten classes as a one-hot target vector.

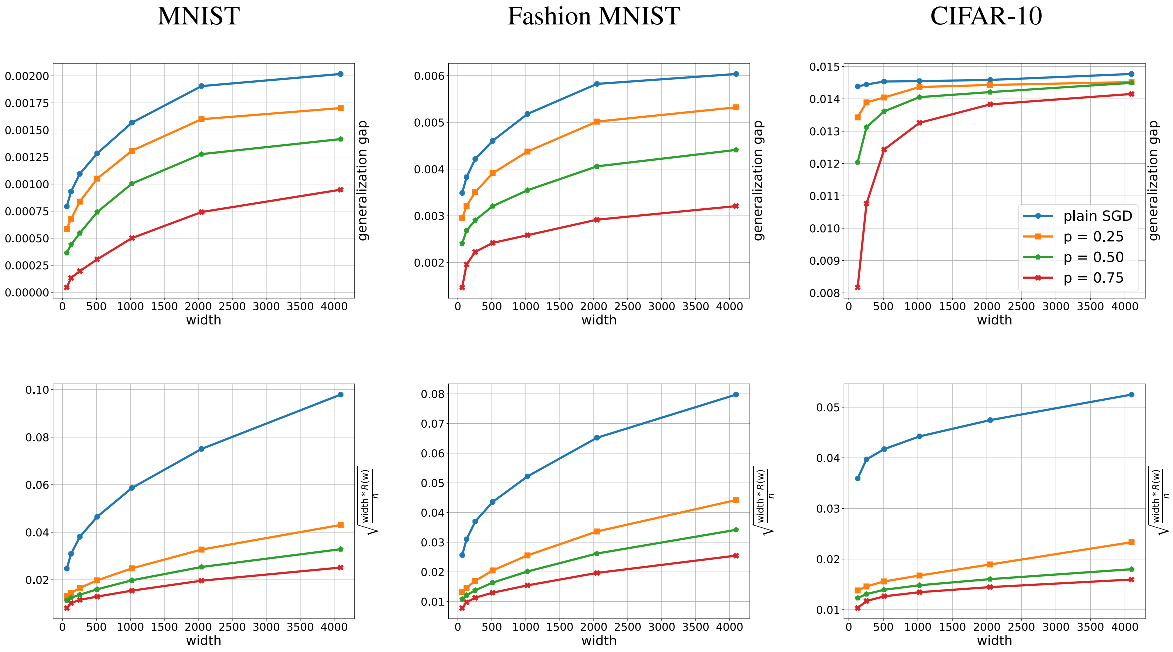

Figure 3. (**top**) generalization gap and (**bottom**) the complexity measure ($\sqrt{\frac{\texttt{width}\cdot R(\mathrm{w})}{n}}$) as a function of the width of the top hidden layer on (**left**) MNIST, (**middle**) Fashion MNIST, and (**right**) CIFAR-10.

For MNIST and Fashion MNIST datasets, we use a convolutional neural network with one convolutional layer and two fully connected layers. The convolutional layer has 16 convolutional filters, padding and stride of 2, and kernel size of 5. We report experiments on networks with the width of the top hidden layer chosen from $\texttt{width} \in \{2^6, 2^7, 2^8, 2^9, 2^{10}, 2^{11}\}$. In all the experiments, a fixed learning rate $\texttt{lr} = 0.5$ and a mini-batch of size 256 is used to perform the updates. We train the models for 30 epochs over the whole training set.

For CIFAR-10, we use an AlexNet (Krizhevsky et al., 2012), where the layers are modified accordingly to match the dataset. The only difference here is that we apply dropout to the top hidden layer, whereas in Krizhevsky et al. (2012), dropout is used on top of the second and the third hidden layers from the top. We report experiments on networks with the width of the top hidden layer chosen from $\texttt{width} \in \{2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}, 2^{11}, 2^{12}\}$. In all the experiments, an initial learning rate $\texttt{lr} = 5$ and a mini-batch of size 256 is used to perform the updates. We train the models for 100 epochs over the whole training set. We decay the learning rate by a factor of 10 every 30 epochs.