A. Section 3 Proofs

We begin by obtaining the decomposition that is instrumental in dividing the excess risk into two pieces that can be then studied separately. Throughout, if W is categorical we fold the sigmoid function σ into the loss for notational convenience.

Proposition A.1 (Proposition 6)**.** *Suppose that f^* is L-Lipschitz relative to $\mathcal{G}$. Then the excess risk $\mathbb{E}[\ell_{\hat{h}}(X,Y) - \ell_{h^*}(X,Y)]$ bounded by,*

$$2LRate_m(\mathcal{G}, P_{X,W}) + Rate_n(\mathcal{F}, \hat{P}).$$

Proof of Proposition 6. Let us split the excess risk into three parts

$$\begin{aligned}\mathbb{E}\Big[\ell_{\hat{h}}(X,Y) - \ell_{h^*}(X,Y)\Big] =& \mathbb{E}\Big[\ell_{\hat{f}(\cdot,\hat{g})}(X,Y) - \ell_{f^*(\cdot,\hat{g})}(X,Y)\Big] \\ &+ \mathbb{E}\Big[\ell_{f^*(\cdot,\hat{g})}(X,Y) - \ell_{f^*(\cdot,g_0)}(X,Y)\Big] + \mathbb{E}\Big[\ell_{f^*(\cdot,g_0)}(X,Y) - \ell_{f^*(\cdot,g^*)}(X,Y)\Big].\end{aligned}$$

By definition, the first term is bounded by $\text{Rate}_n(\mathcal{F}, \hat{P})$. The relative Lipschitzness of f^* delivers the following bound on the second and third terms respectively,

$$\begin{aligned}\mathbb{E}\Big[\ell_{f^*(\cdot,\hat{g})}(X,Y) - \ell_{f^*(\cdot,g_0)}(X,Y)\Big] &\leq L\mathbb{E}_P \ell^{\text{weak}}\Big(A_{\hat{g}}\hat{g}(X), A_{g_0}g_0(X)\Big), \\ \mathbb{E}\Big[\ell_{f^*(\cdot,g_0)}(X,Y) - \ell_{f^*(\cdot,g^*)}(X,Y)\Big] &\leq L\mathbb{E}_P \ell^{\text{weak}}\Big(A_{g_0}g_0(X), A_{g^*}g^*(X)\Big).\end{aligned}$$

Since g^* attains minimal risk, and $W = A_{g_0}g_0(X)$, the sum of these two terms can be bounded by,

$$2L\mathbb{E}_P \ell^{\text{weak}}\Big(A_{\hat{g}}\hat{g}(X), W\Big) \leq 2L\text{Rate}_m(\mathcal{G}, P_{X,W}).$$

Combining this with the bound on the first term yields the claim. □

The next two propositions show, for the two cases of ℓ^{weak} of interest, that the weak central condition is preserved (with a slight weakening in the constant) when replacing the population distribution P by the distribution $\hat{P}$ obtained by replacing the true weak label W by the learned weak estimate $\hat{g}(X)$.

Proposition A.2 (Proposition 7)**.** *Suppose that $\ell^{weak}(w, w') = \mathbb{1}\{w \neq w'\}$ and that ℓ is bounded by $B > 0$, $\mathcal{F}$ is Lipschitz relative to $\mathcal{G}$, and that $(\ell, P, \mathcal{F})$ satisfies the ε-weak central condition. Then $(\ell, \hat{P}, \mathcal{F})$ satisfies the $\varepsilon + \mathcal{O}\Big(Rate_m(\mathcal{G}, P_{X,W})\Big)$-weak central condition with probability at least $1 - \delta$.*

Proof of Proposition 7. Note first that

$$\frac{1}{\eta}\log \mathbb{E}_{\hat{P}} \exp\Big(-\eta(\ell_f - \ell_{f^*})\Big) = \frac{1}{\eta}\log \mathbb{E}_P \exp\Big(-\eta(\ell_{f(\cdot,\hat{g})} - \ell_{f^*(\cdot,\hat{g})})\Big)$$

where we recall that we have overloaded the loss ℓ to include both ℓ_f and ℓ_h. To prove $(\ell, \hat{P}, \mathcal{F})$ satisfies the central condition we therefore need to bound $\frac{1}{\eta}\log \mathbb{E}_P \exp\Big(-\eta(\ell_{f(\cdot,\hat{g})} - \ell_{f^*(\cdot,\hat{g})})\Big)$ above by some constant. We begin bounding (line by line explanations are below),

$$\begin{aligned}\frac{1}{\eta}\log \mathbb{E}_P \exp\Big(-\eta(\ell_{f(\cdot,\hat{g})} - \ell_{f^*(\cdot,\hat{g})})\Big) &= \frac{1}{\eta}\log \mathbb{E}_P\Big[\exp\Big(-\eta(\ell_{f(\cdot,\hat{g})} - \ell_{f^*(\cdot,\hat{g})})\Big)\mathbb{1}\{A_{\hat{g}}\hat{g}(X) = W\}\Big] \\ &\quad + \frac{1}{\eta}\log \mathbb{E}_P\Big[\exp\Big(-\eta(\ell_{f(\cdot,\hat{g})} - \ell_{f^*(\cdot,\hat{g})})\Big)\mathbb{1}\{A_{\hat{g}}\hat{g}(X) \neq W\}\Big] \\ &= \frac{1}{\eta}\log \mathbb{E}_P\Big[\exp\Big(-\eta(\ell_{f(\cdot,g_0)} - \ell_{f^*(\cdot,g_0)})\Big)\mathbb{1}\{A_{\hat{g}}\hat{g}(X) = W\}\Big] \\ &\quad + \frac{1}{\eta}\log \mathbb{E}_P\Big[\exp\Big(-\eta(\ell_{f(\cdot,\hat{g})} - \ell_{f^*(\cdot,\hat{g})})\Big)\mathbb{1}\{A_{\hat{g}}\hat{g}(X) \neq W\}\Big]\end{aligned}$$

where the second line follows from the fact that for any f in the event $\{A_{\hat{g}}\hat{g}(X) = W\}$ we have $\ell_{f(\cdot,\hat{g})} = \ell_{f(\cdot,g_0)}$ and $\ell_{f^*(\cdot,\hat{g})} = \ell_{f^*(\cdot,g_0)}$. This is because $|\ell_{f(\cdot,\hat{g})}(X,Y) - \ell_{f(\cdot,g_0)}(X,Y)| \leq L\ell^{\text{weak}}(A_{\hat{g}}\hat{g}(X), A_{g_0}g_0(X)) = L\ell^{\text{weak}}(W,W) = 0$.

Dropping the indicator $\mathbb{1}\{A_{\hat{g}}\hat{g}(X) = W\}$ from the integrand yields $\frac{1}{\eta}\log \mathbb{E}_P\left[e^{-\eta(\ell_f - \ell_{f^*})}\right]$ which is upper bounded by ε by the weak central condition. We may therefore upper bound the second term by,

$$\begin{aligned}\frac{1}{\eta}\log \mathbb{E}_P\left[\exp\Big(-\eta(\ell_{f(\cdot,\hat{g})} - \ell_{f^*(\cdot,\hat{g})})\Big)\mathbb{1}\{A_{\hat{g}}\hat{g}(X) \neq W\}\right] &\leq \frac{1}{\eta}\log \mathbb{E}_P\left[\exp(\eta B)\mathbb{1}\{A_{\hat{g}}\hat{g}(X) \neq W\}\right] \\ &\leq \frac{\exp(\eta B)}{\eta}\mathbb{P}_P(A_{\hat{g}}\hat{g}(X) \neq W) \\ &= \frac{\exp(\eta B)}{\eta}\text{Rate}(\mathcal{G}, \mathcal{D}_m^{\text{weak}}).\end{aligned}$$

The first inequality uses the fact that ℓ is bounded by B, the second line uses the basic fact $\log x \leq x$, and the final equality holds with probability $1-\delta$ by assumption. Combining this bound with the ε bound on the first term yields the claimed result. □

Proposition A.3 (Proposition 8). *Suppose that $\ell^{weak}(w,w') = \|w - w'\|$ and that ℓ is bounded by $B > 0$, $\mathcal{F}$ is L-Lipschitz relative to $\mathcal{G}$, and that $(\ell, P, \mathcal{F})$ satisfies the ε-weak central condition. Then $(\ell, \hat{P}, \mathcal{F})$ satisfies the $\varepsilon + \mathcal{O}\Big(\sqrt{LRate_m(\mathcal{G}, P_{X,W})}\Big)$-weak central condition with probability at least $1-\delta$.*

Proof of Proposition 8. For any $\delta > 0$ we can split the objective we wish to bound into two pieces as follows,

$$\begin{aligned}\frac{1}{\eta}\log \mathbb{E}_{\hat{P}}\exp\Big(-\eta(\ell_f - \ell_{f^*})\Big) = &\underbrace{\frac{1}{\eta}\log \mathbb{E}_{\hat{P}}\left[\exp\Big(-\eta(\ell_f - \ell_{f^*})\Big)\mathbb{1}\left\{\Big\|A_{\hat{g}}\hat{g}(X) - W\Big\| \leq \frac{\delta}{L}\right\}\right]}_{=:\,\text{I}} \\ &+ \underbrace{\frac{1}{\eta}\log \mathbb{E}_{\hat{P}}\left[\exp\Big(-\eta(\ell_f - \ell_{f^*})\Big)\mathbb{1}\left\{\Big\|A_{\hat{g}}\hat{g}(X) - W\Big\| > \frac{\delta}{L}\right\}\right]}_{=:\,\text{II}}.\end{aligned}$$

We will bound each term separately. The first term can be rewritten as,

$$\text{I} = \frac{1}{\eta}\log \mathbb{E}_P\left[\exp\Big(-\eta(\ell_{f(\cdot,\hat{g})} - \ell_{f^*(\cdot,\hat{g})})\Big)\mathbb{1}\left\{\Big\|A_{\hat{g}}\hat{g}(X) - W\Big\| \leq \frac{\delta}{L}\right\}\right]$$

Let us focus for a moment specifically on the exponent, which we can break up into three parts,

$$\ell_{f(\cdot,\hat{g})} - \ell_{f^*(\cdot,\hat{g})} = (\ell_{f(\cdot,g_0)} - \ell_{f^*(\cdot,g_0)}) + (\ell_{f(\cdot,\hat{g})} - \ell_{f(\cdot,g_0)}) + (\ell_{f^*(\cdot,g_0)} - \ell_{f^*(\cdot,\hat{g})}).$$

In the event that $\left\{\Big\|A_{\hat{g}}\hat{g}(X) - W\Big\| \leq \frac{\delta}{L}\right\}$ the second and third terms can be bounded using the Lipschitzness of ℓ, and the relative Lipschitzness of $\mathcal{F}$ with respect to $\mathcal{G}$,

$$\begin{aligned}|\ell_{f(\cdot,\hat{g})}(X,Y) - \ell_{f(\cdot,g_0)}(X,Y)| + |\ell_{f^*(\cdot,g_0)}(X,Y) - \ell_{f^*(\cdot,\hat{g})}(X,Y)| &\leq L\Big\|A_{\hat{g}}\hat{g} - A_{g_0}g_0\Big\| + L\Big\|A_{\hat{g}}\hat{g} - A_{g_0}g_0\Big\| \\ &= 2L\Big\|A_{\hat{g}}\hat{g} - W\Big\| \\ &\leq 2\delta.\end{aligned}$$

Plugging this upper bound into the expression for I, we obtain the following bound

$$
\begin{aligned}
\mathrm{I} \leq \frac{1}{\eta} \log \mathbb{E}_P \Bigg[\exp\Big(-\eta(\ell_f - \ell_{f^*}) \Big) \mathbb{1}\Bigg\{ \Big\| A_{\hat{g}} A \hat{g}(X) - W \Big\| \leq \frac{\delta}{L} \Bigg\} \Bigg] \\
+ \frac{1}{\eta} \log \mathbb{E}_P \Bigg[\exp\left(2\eta\delta\right) \mathbb{1}\Bigg\{ \Big\| A_{\hat{g}} \hat{g}(X) - W \Big\| \leq \frac{\delta}{L} \Bigg\} \Bigg] \\
\leq \frac{1}{\eta} \log \mathbb{E}_P \Big[\exp\Big(-\eta(\ell_f - \ell_{f^*}) \Big) \Big] + 2\delta \\
\leq \varepsilon + 2\delta
\end{aligned}
$$

where in the second line we have simply dropped the indicator function from both integrands, and for the third line we have appealed to the ε-weak central condition. Next we proceed to bound the second term (line by line explanations are below) II by,

$$
\begin{aligned}
\frac{1}{\eta} \log \mathbb{E}_{\hat{P}} \Bigg[\exp\Big(-\eta(\ell_f - \ell_{f^*}) \Big) \mathbb{1}\Bigg\{ \Big\| A_{\hat{g}} \hat{g}(X) - W \Big\| > \frac{\delta}{L} \Bigg\} \Bigg] &\leq \frac{1}{\eta} \log \mathbb{E}_{\hat{P}} \Bigg[\exp\left(\eta B\right) \mathbb{1}\Bigg\{ \Big\| A_{\hat{g}} \hat{g}(X) - W \Big\| > \frac{\delta}{L} \Bigg\} \Bigg] \\
&\leq \frac{\exp\left(\eta B\right)}{\eta} \mathbb{E}_{P_{X,W}} \Bigg[\mathbb{1}\Bigg\{ \Big\| A_{\hat{g}} \hat{g}(X) - W \Big\| > \frac{\delta}{L} \Bigg\} \Bigg] \\
&= \frac{\exp\left(\eta B\right)}{\eta} \mathbb{P}_{P_{X,W}} \Bigg(\Big\| A_{\hat{g}} \hat{g}(X) - W \Big\| > \frac{\delta}{L} \Bigg) \\
&\leq \frac{L \exp\left(\eta B\right)}{\delta\eta} \mathbb{E}_P \Big\| A_{\hat{g}} \hat{g}(X) - W \Big\| \\
&\leq \frac{L \exp\left(\eta B\right)}{\delta\eta} \mathrm{Rate}_m(\mathcal{G}, P_{X,W})
\end{aligned}
$$

where the first line follows since ℓ is bounded by B, the second line since $\log x \leq x$, the fourth line is an application of Markov's inequality, and the final inequality holds by definition of $\mathrm{Rate}_m(\mathcal{G}, P_{X,W})$ with probability $1 - \delta$. Collecting these two results together we find that

$$
\frac{1}{\eta} \log \mathbb{E}_{\hat{P}} \exp\Big(-\eta(\ell_f - \ell_{f^*}) \Big) = \mathrm{I} + \mathrm{II} \leq \varepsilon + 2\delta + \frac{L \exp\left(\eta B\right)}{\delta\eta} \mathrm{Rate}_m(\mathcal{G}, P_{X,W}).
$$

Since this holds for any $\delta > 0$ we obtain the bound,

$$
\begin{aligned}
\frac{1}{\eta} \log \mathbb{E}_{\hat{P}} \exp\Big(-\eta(\ell_f - \ell_{f^*}) \Big) &\leq \varepsilon + \min_{\delta > 0} \left\{ 2\delta + \frac{L \exp\left(\eta B\right)}{\delta\eta} \mathrm{Rate}_m(\mathcal{G}, P_{X,W}) \right\} \\
&= \varepsilon + 2\sqrt{2} \sqrt{\frac{L \exp\left(\eta B\right)}{\eta}} \sqrt{\mathrm{Rate}_m(\mathcal{G}, P_{X,W})}.
\end{aligned}
$$

The minimization is a simple convex problem that is solved by picking δ to be such that the two terms are balanced. □

The next proposition shows that the weak central condition is sufficient to obtain excess risk bounds. This result generalizes Theorem 1 of (Mehta, 2016), which assumes the strong central condition holds. In contrast, we make only need the weaker assumption that the weak central condition holds.

Proposition A.4 (Proposition 9)**.** *Suppose $(\ell, Q, \mathcal{F})$ satisfies the ε-weak central condition, ℓ is bounded by $B > 0$, each $\mathcal{F}$ is L'-Lipschitz in its parameters in the ℓ_2 norm, $\mathcal{F}$ is contained in the Euclidean ball of radius R, and $\mathcal{Y}$ is compact. Then when $Alg_n(\mathcal{F}, Q)$ is ERM, the excess risk $\mathbb{E}_Q[\ell_{\hat{f}}(U) - \ell_{f^*}(U)]$ is bounded by,*

$$
\mathcal{O}\left(V \frac{d \log \frac{RL'}{\varepsilon} + \log \frac{1}{\delta}}{n} + V\varepsilon \right).
$$

with probability at least $1 - \delta$, where $V = B + \varepsilon$.

Proof of Proposition 9. Before beginning the proof in earnest, let us first introduce a little notation, and explain the high level proof strategy. We use the shorthand $\Delta_f = \ell_f - \ell_{f^*}$. Throughout this proof we are interested in the underlying distribution Q. So, to avoid clutter, throughout the proof we shall write $\mathbb{E}$ and $\mathbb{P}$ as short hand for $\mathbb{E}_{U \sim Q}$ and $\mathbb{P}_{U \sim Q}$.

Our strategy is as follows: we wish to determine an $a > 0$ for which, with high probability, ERM does not select a function $f \in \mathcal{F}$ such that $\mathbb{E}\Delta_f \geq \frac{a}{n}$. Defining $\mathcal{F}_\beta = \{f \in \mathcal{F} : \mathbb{E}\Delta_f \geq \beta\}$ this is equivalent to showing that, with high probability, ERM does not select a function $f \in \mathcal{F}_{\beta_n}$ where $\beta_n = \frac{a}{n}$. In turn this can be re-expressed as showing with high probability that,

$$\frac{1}{n}\sum_{j=1}^{n} \Delta_f(U_j) > 0 \tag{2}$$

for all $f \in \mathcal{F}_{\beta_n}$, where the random variables $\{U_j\}_j$ are i.i.d samples from Q. In order to prove this we shall take a finite cover $\{f_1, f_2, \ldots, f_s\}$ of our function class $\mathcal{F}_{\beta_n}$ and show that, with high probability $\frac{1}{n}\sum_{j=1}^{n} \Delta_f(U_j) > c$ for all f_i for some constant $c > 0$ depending on the radius of the balls. To do this, we use the central condition, and two important tools from probability whose discussion we postpone until Appendix Section B, to bound the probability of selecting each f_i, then apply a simple union bound. This result, combined with the fact that every element of $\mathcal{F}_{\beta_n}$ is close to some such f_i allows us to derive equation (2) for all members of the class $\mathcal{F}_{\beta_n}$.

With the strategy laid out, we are now ready to begin the proof in detail. We start by defining the required covering sets. Specifically, let $\mathcal{F}_{\beta_n,\varepsilon}$ be an optimal proper[1] $\varepsilon/L's$-cover of $\mathcal{F}_{\beta_n}$ in the ℓ_2-norm, where we will pick s later. It is a classical fact (see e.g. (Carl & Stephani, 1990)) that the d-dimensional ℓ_2-ball of radius R has ε-covering number at most $(\frac{4R}{\varepsilon})^d$. Since the cardinality of an optimal proper ε-covering number is at most the $\varepsilon/2$-covering number, and $\mathcal{F}$ is contained in the the d-dimensional ℓ_2-ball of radius R, we have $|\mathcal{F}_{\beta_n,\varepsilon}| \leq (\frac{8RL's}{\varepsilon})^d$. Furthermore, since ℓ is continously differentiable, $\mathcal{Y}$ is compact and f is Lipschitz in its parameter vector, we have that $f \mapsto \ell_f$ is $L's$-Lipschitz in the ℓ_2 norm in the domain and ℓ_∞-norm in the range (for some s, which we have now fixed). Therefore the proper $\varepsilon/L's$-cover of $\mathcal{F}_{\beta_n}$ pushes forward to a proper ε-cover of $\{\ell_f : f \in \mathcal{F}_{\beta_n}\}$ in the ℓ_∞-norm.

We now tackle the key step in the proof, which is to upper bound the probability that ERM selects an element of $\mathcal{F}_{\beta_n,\varepsilon}$. To this end, fix an $f \in \mathcal{F}_{\beta_n,\varepsilon}$. Since $(\ell, \hat{P}, \mathcal{F})$ satisfies the ε-weak central condition, we have $\mathbb{E}\Big[e^{-\eta\Delta_f}\Big] \leq e^{\eta\varepsilon}$. Rearranging yields,

$$\mathbb{E}\Big[\exp\Big(-\eta(\Delta_f + \varepsilon)\Big)\Big] \leq 1.$$

Lemma B.1 implies that for any $0 < \gamma < a$ there exists a modification $\widetilde{\Delta}_f + \varepsilon$ of $\Delta_f + \varepsilon$, and an $\eta \leq \eta_f \leq 2\eta$ such that $\widetilde{\Delta}_f \leq \Delta_f$, almost surely, and,

$$\mathbb{E}\Big[\exp\Big(-\eta_f(\widetilde{\Delta}_f + \varepsilon)\Big)\Big] = 1 \quad \text{and} \quad \mathbb{E}\widetilde{\Delta}_f \geq \frac{a-\gamma}{n}. \tag{3}$$

Since $\widetilde{\Delta}_f + \varepsilon$ belongs to the shifted interval $[-V, V]$ where $V = B + \varepsilon$, Corollaries 7.4 and 7.5 from (van Erven et al., 2015) imply[2] that,

$$\log \mathbb{E}\Big[\exp\Big(-\eta_f/2(\widetilde{\Delta}_f + \varepsilon)\Big)\Big] \leq -\frac{0.18}{(V \vee 1/\eta_f)}\left(\frac{a-\gamma}{n} + \varepsilon\right) \leq -\frac{0.18(a-\gamma)}{(V \vee 1/\eta_f)n}.$$

where we define $a' = a - \gamma$. By Cramér-Chernoff (Lemma B.2) with $t = ca'\varepsilon$ (where c will also be chosen later) and the η in the lemma being $\eta_f/2$, we obtain

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n}\Big(\widetilde{\Delta}_f(U_j) + \varepsilon\Big) \leq ca'\varepsilon\right) &\leq \exp\left(-\frac{0.18}{V \vee 1/\eta_f}a' + \frac{n\eta_f ca'\varepsilon}{2}\right) \\
&\leq \exp\left(-\frac{0.18}{V \vee 1/\eta}a' + n\eta ca'\varepsilon\right) \\
&= \exp(-Ca')
\end{aligned}$$

[1] For a metric space (M, ρ), let $S \subseteq M$. A set $E \subseteq M$ is an ε-cover for S, if for every $s \in S$ there is an $e \in E$ such that $\rho(s, e) \leq \varepsilon$. An ε-cover is optimal if it has minimal cardinality out of all ε-covers. E is known as a proper cover if $E \subseteq S$.

[2] Note that although the Corollaries in (van Erven et al., 2015) are stated specifically for Δ_f, the claims hold for *any* random variable satisfying the hypotheses, including our case of $\Delta_f + \varepsilon$.

where $C := \frac{0.18}{B \vee 1/\eta} - n\eta c\varepsilon$, and the second inequality follows since $\eta \leq \eta_f \leq 2\eta$. Let us now pick c so as to make C bigger than zero, and in particular so that $C = \frac{0.09}{B \vee 1/\eta}$. That is, let $c = \frac{1}{n\varepsilon}\frac{0.09}{V\eta \vee 1}$. Using the fact that $a' - 2/c \leq a'$, and a union bound over $f \in \mathcal{F}_{\beta_n,\varepsilon}$ we obtain a probability bound on all of $\mathcal{F}_{\beta_n,\varepsilon}$,

$$\mathbb{P}\left(\exists f \in \mathcal{F}_{\beta_n,\varepsilon} : \frac{1}{n}\sum_{j=1}^{n}\widetilde{\Delta}_f(U_j) \leq (ca' - 1)\varepsilon\right) \leq \left(\frac{8RL's}{\varepsilon}\right)^d \exp\left(-\frac{0.09}{B \vee \frac{1}{\eta}}(a' - 2/c)\right).$$

Define the right hand side to equal $0 < \delta < 1$. Note that we are allowed to do this thanks to the fact $C > 0$, which implies that the right hand side goes to zero as $a' \to \infty$. This makes it possible to pick a sufficiently large a' for which the right hand side is less than 1. Solving for $a = a' + \gamma$ we choose,

$$a = \frac{V \vee 1/\eta}{0.09}\left(d \log \frac{8RL's}{\varepsilon} + \log \frac{1}{\delta}\right) + 2/c + \gamma.$$

Therefore, with probability at least $1 - \delta$ we have for all $f \in \mathcal{F}_{\beta_n,\varepsilon}$ that $\frac{1}{n}\sum_{j=1}^{n}\widetilde{\Delta}_f(U_j) > (ca' - 1)\varepsilon$. Therefore, for any $f' \in \mathcal{F}_{\beta_n}$ we can find $f \in \mathcal{F}_{\beta_n,\varepsilon}$ such that $\|\ell_f - \ell_{f'}\|_\infty \leq \varepsilon$.

Finally, since $ca \geq 2$ for sufficiently small ε by construction, and $\Delta_f \geq \widetilde{\Delta}_f$ almost surely, we find that $\frac{1}{n}\sum_{j=1}^{n}\Delta_{f'}(U_j) \geq \frac{1}{n}\sum_{j=1}^{n}\Delta_f(U_j) - \varepsilon \geq \frac{1}{n}\sum_{j=1}^{n}\widetilde{\Delta}_f(U_j) - \varepsilon \geq (ca - 1)\varepsilon - \varepsilon > 0$. We have proven that with probability at least $1 - \delta$ that $\frac{1}{n}\sum_{j=1}^{n}\Delta_{f'}(U_j) > 0$ for all $f' \in \mathcal{F}_{\beta_n}$. Therefore, with high probability, ERM will not select any element of $\mathcal{F}_{\beta_n}$. Finally, the bound described in the theorem comes from substituting in the choice of c, and rounding up the numerical constants, recognizing that since the claim holds for all $\gamma > 0$, we may take the limit as $\gamma \to 0^+$ to obtain,

$$a \leq 12(V \vee 1/\eta)\left(d \log \frac{8RL's}{\varepsilon} + \log \frac{1}{\delta}\right) + 12(V\eta \vee 1)n\varepsilon + 1.$$

□

The heavy lifting has now been done by the previous propositions and theorems. In order to obtain the main result, all that remains now is to apply each result in sequence.

Theorem A.5 (Theorem 10)**.** *Suppose that $(\ell, P, \mathcal{F})$ satisfies the central condition and that $Rate_m(\mathcal{G}, P_{X,W}) = \mathcal{O}(1/m^\alpha)$. Then when $Alg_n(\mathcal{F}, \hat{P})$ is ERM we obtain excess risk $\mathbb{E}_P[\ell_{\hat{h}}(X,Y) - \ell_{h^*}(X,Y)]$ that is bounded by,*

$$\mathcal{O}\left(\frac{d\alpha\beta \log RL'n + \log\frac{1}{\delta}}{n} + \frac{L}{n^{\alpha\beta}}\right)$$

with probability at least $1 - \delta$, if either of the following conditions hold,

1. *$m = \Omega(n^\beta)$ and $\ell^{weak}(w, w') = \mathbb{1}\{w \neq w'\}$ (discrete $\mathcal{W}$-space).*
2. *$m = \Omega(n^{2\beta})$ and $\ell^{weak}(w, w') = \|w - w'\|$ (continuous $\mathcal{W}$-space).*

Proof of Theorem 10. **Case 1:** We have $m = \Omega(n^\beta)$, and $\text{Rate}_m(\mathcal{G}, P_{X,W}) = \mathcal{O}(1/m^\alpha)$, together impling that $\text{Rate}(\mathcal{G}, \mathcal{D}_m^{\text{weak}}) = \mathcal{O}(1/n^{\alpha\beta})$. We apply Proposition 7 to conclude that $(\ell, \hat{P}, \mathcal{F})$ satisfies the $\mathcal{O}(1/n^{\alpha\beta})$-weak central condition with probability at least $1 - \delta$.

Proposition 9 therefore implies that $\text{Rate}_n(\mathcal{F}, \hat{P}) = \mathcal{O}\left(\frac{d\alpha\beta \log 8RL'n + \log\frac{1}{\delta}}{n} + \frac{1}{n^{\alpha\beta}}\right)$.

Combining these two bounds using Proposition 6 we conclude that

$$\mathbb{E}[\ell_{\hat{h}}(Z) - \ell_{h^*}(Z)] \leq \mathcal{O}\left(\frac{d\alpha\beta \log 8RL'n + \log\frac{1}{\delta}}{n} + \frac{L}{n^{\alpha\beta}}\right).$$

Case 2: The second case is proved almost identically, however note that since in this case we have $m = \Omega(n^{2\beta})$, that now $\text{Rate}_m(\mathcal{G}, P_{X,W}) = \mathcal{O}(1/n^{2\alpha\beta})$. The factor of two is cancelled our by the extra square root factor in Proposition 8. The rest of the proof is exactly the same as case 1. □

B. Probabilistic Tools

In this section we present two technical lemmas that are key tools used to prove Proposition 9. The first allows us to take a random variable Δ such that $\mathbb{E}e^{-\eta\Delta} \leq 1$ and perturb downwards it slightly to some $\widetilde{\Delta} \leq \Delta$ so that the inequality becomes an equality (for a slightly different η) and yet the expected value changes by an arbitrarily small amount.

Lemma B.1. *Suppose $\eta > 0$ and Δ is an absolutely continuous random variable on the probability space $(\Omega, \mathbb{P})$ such that Δ is almost surely bounded, and $\mathbb{E}e^{-\eta\Delta} \leq 1$. Then for any $\varepsilon > 0$ there exists an $\eta \leq \eta' \leq 2\eta$ and another random variable $\widetilde{\Delta}$ (called a "modification") such that,*

1. *$\widetilde{\Delta} \leq \Delta$ almost surely,*
2. *$\mathbb{E}e^{-\eta'\widetilde{\Delta}} = 1$, and*
3. *$|\mathbb{E}[\Delta - \widetilde{\Delta}]| \leq \varepsilon$.*

Proof. We may assume that $\mathbb{E}e^{-\eta\Delta} < 1$ since otherwise we can simply take $\widetilde{\Delta} = \Delta$ and $\eta = \eta'$. Due to absolute continuity, for any $\delta > 0$ there is a measurable set $A_\delta \subset \Omega$ such that $\mathbb{P}(A_\delta) = e^{-1/\delta}$. Now define $\widetilde{\Delta} : \Omega \to \mathbb{R}$ by,

$$\widetilde{\Delta}(\omega) = \begin{cases} \Delta(\omega) & \text{if } \omega \notin A_\delta \\ -\frac{1}{2\delta\eta} & \text{if } \omega \in A_\delta \end{cases} \tag{4}$$

We now prove that as long as δ is small enough, all three claimed properties hold.

Property 1: Since Δ is almost surely bounded, there is a $V > 0$ such that $|\Delta| \leq V$ almost surely. Taking δ small enough that $-\frac{1}{2\delta\eta} \leq -V$ we guarantee that $\widetilde{\Delta} \leq \Delta$ almost surely.

Property 2: We can lower bound the 2η case,

$$\mathbb{E}e^{-2\eta\widetilde{\Delta}} \geq e^{-2\eta(-\frac{1}{2\eta\delta})}\mathbb{P}(A_\delta) = e^{1/\delta}\mathbb{P}(A_\delta) = e^{1/\delta}e^{-1/\delta} = 1.$$

We can similarly upper bound the η case,

$$\begin{aligned} \mathbb{E}e^{-\eta\widetilde{\Delta}} &= \int e^{-\eta\widetilde{\Delta}(\omega)}\mathbf{1}\{\omega \in A_\delta\}\mathbb{P}(\mathrm{d}\omega) + \int e^{-\eta\widetilde{\Delta}(\omega)}\mathbf{1}\{\omega \notin A_\delta\}\mathbb{P}(\mathrm{d}\omega) \\ &= e^{1/2\delta}\mathbb{P}(A_\delta) + \int e^{-\eta\Delta(\omega)}\mathbf{1}\{\omega \notin A_\delta\}\mathbb{P}(\mathrm{d}\omega) \\ &\leq e^{-1/2\delta} + \int e^{-\eta\Delta(\omega)}\mathbb{P}(\mathrm{d}\omega) \\ &\leq e^{-1/2\delta} + \mathbb{E}e^{-\eta\Delta}. \end{aligned}$$

Recall that by assumption $\mathbb{E}e^{-\eta\Delta} < 1$, so we may pick δ sufficiently small so that $e^{-1/2\delta} + \mathbb{E}e^{-\eta\Delta} < 1$. Using these two bounds, and observing that boundedness of Δ implies continuity of $\eta \mapsto \mathbb{E}\left[e^{-\eta\Delta}\right]$, we can guarantee that there is an $\eta \leq \eta' \leq 2\eta$ such that $\mathbb{E}\left[e^{-\eta'\widetilde{\Delta}}\right] = 1$.

Property 3: Since Δ and $\widetilde{\Delta}$ only disagree on A_δ,

$$\mathbb{E}|\widetilde{\Delta} - \Delta| = \int |\widetilde{\Delta}(\omega) - \Delta(\omega)|\mathbf{1}\{w \in A_\delta\}\mathbb{P}(\mathrm{d}\omega) \leq \left(\frac{1}{2\delta\eta} + V\right)\mathbb{P}(A_\delta) = \left(\frac{1}{2\delta\eta} + V\right)e^{-1/\delta}$$

which converges to 0 as $\delta \to 0^+$. We may, therefore, make the difference in expectations smaller than ε by taking δ to be sufficiently close to 0. □

The second lemma is a well known Cramér-Chernoff bound that is used to obtain concentration of measure results. A proof was given, for example, given in (van Erven et al., 2015). However, since the proof is short and simple we include it here for completeness.

Lemma B.2 (Cramér-Chernoff (van Erven et al., 2015))**.** *Let $\Delta, \Delta_1, \ldots, \Delta_n$ be i.i.d. and define $\Lambda_\Delta(\eta) = \log \mathbb{E}[e^{-\eta\Delta}]$. Then, for any $\eta > 0$ and $t \in \mathbb{R}$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\Delta_i \leq t\right) \leq \exp\left(\eta nt + n\Lambda_\Delta(\eta)\right).$$

Proof. Note that since $x \mapsto \exp(-\eta x)$ is a bijection, we have,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\Delta_i \leq t\right) = \mathbb{P}\left(\exp\Big(-\eta\sum_{i=1}^{n}\Delta_i\Big) \geq \exp(-\eta n t)\right).$$

Applying Markov's inequality to the right hand side of the equality yields the upper bound,

$$\exp(\eta n t)\mathbb{E}\Big[\exp(-\eta\sum_{i=1}^{n}\Delta_i)\Big] = \exp(\eta n t)\Big[\mathbb{E}\ \exp(-\eta\Delta)\Big]^{n} = \exp\left(\eta n t + n\Lambda_{\Delta}(\eta)\right).$$

□

C. Hyperparameter and Architecture Details

All models were trained using PyTorch (Paszke et al., 2019) and repeated from scratch 4 times (20 for TREC) to give error bars. All layers were initialized using the default uniform initialization.

Architecture For the MNIST experiments we used the ResNet-18 architecture as a deep feature extractor for the weak task (He et al., 2016), followed by a single fully connected layer to the output. For the strong model, we used a two hidden layer fully connected neural network as a feature extractor with ReLU activations. The first hidden layer has 2048 neurons, and the second layer has 1024. This feature vector is then concatenated with the ResNet feature extractor, and passed through a fully connected one hidden layer network with 1024 hidden neurons. For all other datasets (SVHN, CIFAR-10, CIFAR-100) the exact same architecture was used except for replacing the ResNet-18 feature extractor by ResNet-34. We also ran experiments using smaller models for the weak feature map, and obtained similar results. That is, the precise absolute learning rates changed, but the comparison between the learning rates remained the similar. The backbone of the weak model for the TREC language understanding problem was the concatenation of three convolutional networks with filter sizes $2, 3, 4$.

Optimization We used Adam (Kingma & Ba, 2015) with initial learning rate 0.0001, and $\beta_1 = 0.5$, and $\beta_2 = 0.999$. We used batches of size 100, except for MNIST, for which we used 50. We used an exponential learning rate schedule, scaling the learning rate by 0.97 once every two epochs.

Data pre-processing For CIFAR-10, CIFAR-100, and SVHN we used random cropping and horizontal image flipping to augment the training data. We normalized CIFAR-100 color channels by subtracting the dataset mean pixel values $(0.5071, 0.4867, 0.4408)$ and dividing by the standard deviation $(0.2675, 0.2565, 0.2761)$. For CIFAR-10 and SVHN we normalize each pixel to live in the interval $[-1, 1]$ by channel-wise subtracting $(0.5, 0.5, 0.5)$ and dividing by $(0.5, 0.5, 0.5)$. For MNIST the only image processing was to normalize each pixel to the range $[0, 1]$. For the TREC language understanding task used GloVe input word embeddings with a vocabulary size of 25000.

Number of training epochs The weak networks were trained for a number of epochs proportional to $1/m$. For example, for all CIFAR-10 experiments the weak networks were trained for $500000/m$ epochs. This was sufficient to train all models to convergence.

Once the weak network was finished training, we stopped all gradients passing through that module, thereby keeping the weak network weights fixed during strong network training. To train the strong network, we used early stopping to avoid overfitting. Specifically, we tested model accuracy on a holdout dataset once every 5 epochs. The first time the accuracy decreased we stopped training, and measured the final model accuracy using a test dataset.

Dataset size The amount of strong data is clearly labeled on the figures. For the weak data, we used the following method to compute the amount of weak data to use:

$$m_i^{(1)} = c_1 n_i$$
$$m_i^{(2)} = c_2 n_i^2$$

where $m_i^{(1)}$ is the amount of weak data for the linear growth, $m_i^{(2)}$ for quadratic growth, and $n_1, n_2, \ldots, n_7$ are the different strong data amounts. For MNIST we took $(c_1, c_2) = (4, 0.02)$, for SVHN we took $(c_1, c_2) = (4.8, 0.0024)$ and for CIFAR-10, for CIFAR-100 we took $(c_1, c_2) = (4, 0.002)$ and for TREC we took $(c_1, c_2) = (0.7652, 0.0019)$. The main design choice is that in each case we have $m_1^{(1)} = m_1^{(2)}$, i.e. weak and quadratic growth begin with the same amount of weak labels.