A. Concentration bounds

In this section we include a series of well known concentration bounds used in the statistical learning literature. In order to prove this bounds we will use the notion of Rademacher complexity.

Definition 7. *Given a sample $z_1, \ldots, z_m \in \mathcal{Z}$ and a class of functions G mapping $\mathcal{Z}$ to $[0,1]$, we define the empirical Rademacher complexity of G as*

$$\Re_m(G) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{g \in G} \sum_{i=1}^m g(z_i)\sigma_i\Big],$$

where σ_i are i.i.d. uniform random varialbes over the set $\{-1, 1\}$.

The Rademacher complexity of a class is closely related to its VC dimension. The following Lemma can be found in (Mohri et al., 2012).

Lemma 3. *Let G be a function class with VC dimension* $\mathrm{VCdim}(h) = d$ *then*

$$\Re(G) \leq \sqrt{2md \log \frac{em}{d}}$$

Lemma 4. *Let L be K-Lipchitz and let $\delta > 0$. Conditioned on the choice of users belonging to the sample the following bound holds with probability at least $1 - \delta$ for for all $h \in H$*

$$\Big|\sum_j \sum_{i=1}^{n_{\tau j}} L(h(x_{ij}), y_{ij}) - \sum_j n_{\tau j}\mathcal{L}_j(h)\Big| \leq 2K\Re_{n_\tau}(H) + \sqrt{\frac{n_\tau \log \frac{1}{\delta}}{2}}$$

Proof. Relabeling the samples we notice that the left hand side of the above inequality is given by

$$\Big|\sum_{i=1}^{n_\tau} L(h(x_i), y_i) - \mathbb{E}[\sum_{i=1}^{n_\tau} L(h(x_i), y_i)]\Big|.$$

Let $H_L = \{(x, y) \mapsto L(h(x), y) | h \in H\}$, using the fact that (x_i, y_i) are independent conditioned on the choice of users and a standard learning theory bound (Mohri et al., 2012) we have with probability at least $1 - \delta$

$$\Big|\sum_{i=1}^{n_\tau} L(h(x_i), y_i) - \mathbb{E}[\sum_{i=1}^{n_\tau} L(h(x_i), y_i)]\Big| \leq \Re_{n_\tau}(H_L) + \sqrt{\frac{n_\tau \log \frac{1}{\delta}}{2}}.$$

Finally by Talagrand's contraction lemma (Mohri et al., 2012) we know that $\Re_{n_\tau}(H_L) \leq K\Re_{n_\tau}(H)$ which concludes the proof. □

Lemma 1. *Conditioned on the outcomes of $\{J_i\}$, with probability at least $1 - \delta$ the following holds uniformly over $h \in H$:*

$$\Big|\mathcal{L}_{\mathcal{S}_\tau}(h) - \sum_j \frac{n_{\tau j}}{n_\tau}\mathcal{L}_j(h)\Big| \leq \sqrt{\frac{2d \log \frac{en}{d}}{\tau_0 n}} + \sqrt{\frac{\log(1/\delta)}{2\tau_0 n}}$$

Proof. The proof follows directly from the previous proposition and a standard bound on the Rademacher complexity by the VC dimension (Mohri et al., 2012). □

Lemma 2. *Fix $\delta > 0$ and let $d = \mathrm{VCdim}(H)$. Then with probability at least $1 - \delta$, the following inequality holds uniformly for h in H.*

$$|\mathcal{L}_{\mathcal{S}_\tau}(h) - \mathcal{L}(h)| \leq \sqrt{\frac{2d \log \frac{en}{d}}{\tau_0 n}} + \sqrt{\frac{\log(2/\delta)}{2\tau_0 n}} + \Big|\sum_j \Big(\frac{n_{\tau j}}{n_\tau} - \frac{n_j}{n}\Big)\mathcal{L}_j(h)\Big| + \sqrt{\frac{\log \frac{4}{\delta}}{2n}}.$$

Proof. We begin by decomposing the loss into three parts.

$$|\mathcal{L}_{\mathcal{S}_\tau}(h) - \mathcal{L}(h)| \leq \Big|\mathcal{L}_{\mathcal{S}_\tau}(h) - \sum_j \frac{n_{\tau j}}{n_\tau}\mathcal{L}_j(h)\Big| \tag{7}$$

$$+ \Big|\sum_j \Big(\frac{n_{\tau j}}{n_\tau} - \frac{n_j}{n}\Big)\mathcal{L}_j(h)\Big| \tag{8}$$

$$+ \Big|\sum_j \Big(\frac{n_j}{n} - p_j\Big)\mathcal{L}_j(h)\Big|. \tag{9}$$

Eq. (7) is the generalization error of our empirical loss, conditioned on the outcomes of $\{J_i\}$. We bound it by applying Lemma 1 with $\frac{\delta}{2}$.

Eq. (8) is the error attributable to differences between the original dataset $\mathcal{S}$ and the thresholded data set $\mathcal{S}_\tau$; it appears directly in the bound.

Finally, Eq. (9) is the finite sample error due to the randomness in $\{J_i\}$. Observe that

$$\Big|\sum_j \Big(\frac{n_j}{n} - p_j\Big)\mathcal{L}_j(h)\Big| = \Big|\frac{1}{n}\sum_{i=1}^n L_{J_i}(h) - \sum_j p_j \mathcal{L}_j(h)\Big|,$$

which is just the difference between the sample mean of n i.i.d. random variables bounded in $[0,1]$ and their true mean. Hoeffding's inequality thus bounds (9) by $\sqrt{\frac{\log \frac{4}{\delta}}{2n}}$ with probability $1 - \frac{\delta}{2}$.

Combining these results under a union bound completes the proof. □

B. Bias bounds

Proposition 2. *Let r_j for $j \in \mathbb{N}$ be such that $r_j \geq 0$ and $\sum_{j=1}^n r_j = 1$. Let $0 \leq q_j \leq r_j$, $Q = \sum_j q_j$. Finally let $q'_j = \frac{q_j}{Q}$. If $|L(h, z)| \leq 1$, then the following bound holds for all hypotheses h.*

$$\Big|\sum_j \left(q'_j - r_j\right)\mathcal{L}_j(h)\Big| \leq \sqrt{\frac{1}{2}\log\left(\frac{1}{Q}\right)}$$

Proof. Using the fact that $\mathcal{L}_j(h) \leq 1$ we have

$$\Big|\sum_j (q'_j - r_j)\mathcal{L}_j(h)\Big| \leq \sum_j \Big|q'_j - r_j\Big| \tag{10}$$

Let $\mathbf{r}$ and $\mathbf{q}'$ denote the distributions induced by r_j and q'_j respectively. By Pinsker's inequality we know

$$\sum_{j=1} \Big|q'_j - r_j\Big| \leq \sqrt{\frac{1}{2}\mathrm{KL}(\mathbf{r}||\mathbf{q}')}\ ,$$

where $\mathrm{KL}(\mathbf{r}||\mathbf{q}')$ denotes the Kullback-Leibler divergence between the two distributions. We can bound this divergence as follows:

$$\mathrm{KL}(\mathbf{r}||\mathbf{q}') = \frac{1}{Q}\sum_j q_j \log\left(\frac{q_j}{Qr_j}\right) \leq \frac{1}{Q}\sum_j q_j \log\left(\frac{1}{Q}\right) = \log\left(\frac{1}{Q}\right),$$

where we have used the fact that $q_j < r_j$ for the first inequality. Substituting this bound back in (10) yields the statement of the proposition. ☐

We now define a more general version of the variance term introduced in Section 6.

Definition 8. *Given a distribution $\mathbf{r}$ over $\mathbb{N}$ and a hypothesis $h \in H$ we define the variance of h with respect to $\mathbf{r}$ as*

$$\mathrm{Var}(h, \mathbf{r}) = \sum_j r_j(\mathcal{L}_j(h) - \mathcal{L}_h)^2.$$

Proposition 3. *Under the notation and assumptions of Proposition 2, the following bound holds for every h:*

$$\Big|\sum_j (q'_j - r_j)\mathcal{L}_j(h)\Big| \leq \sqrt{\frac{2\mathrm{Var}(h, \mathbf{r})}{Q}}$$

Proof. The proof relies on the simple fact that:

$$\sum_i\sum_j(\mathcal{L}_j(h) - \mathcal{L}_i(h))r_i q'_j = \sum_j \mathcal{L}_j(h)q'_j - \sum_i \mathcal{L}_i(h)r_i.$$

This is easy to verify using the fact that $\sum r_i = 1$ and $\sum q'_j = 1$. We can now apply the Cauchy-Schwarz inequality as follows:

$$\begin{aligned}
&\Big|\sum_j (q'_j - r_j)\mathcal{L}_j(h)\Big| \\
&= \Big|\sum_i\sum_j(\mathcal{L}_j(h) - \mathcal{L}_i(h))q'_j r_i\Big| \\
&= \Big|\sum_i\sum_j(\mathcal{L}_j(h) - \mathcal{L}_i(h))\sqrt{r_i r_j}\frac{q'_j}{\sqrt{r_j}}\sqrt{r_i}\Big| \\
&\leq \sqrt{\sum_i\sum_j(\mathcal{L}_j(h) - \mathcal{L}_i(h))^2 r_i r_j}\sqrt{\sum_i\sum_j\frac{(q'_j)^2}{r_j}r_i} \\
&= \sqrt{\sum_i\sum_j(\mathcal{L}_j(h) - \mathcal{L}_i(h))^2 r_i r_j}\sqrt{\sum_j\frac{(q'_j)^2}{r_j}}
\end{aligned}$$

A simple calculation shows that the first term in the above expression is in fact equal to $2\mathrm{Var}(h, \mathbf{r})$. Therefore we need only to prove that the second term is bounded by $\frac{1}{Q}$. We have

$$\begin{aligned}
\sum_j \frac{(q'_j)^2}{r_j} &= \frac{1}{Q^2}\sum_j\frac{q_j^2}{r_j} \\
&\leq \frac{1}{Q^2}\sum_j q_j = \frac{1}{Q},
\end{aligned}$$

where we used the fact that $q_j \leq r_j$. ☐

The proof of Proposition 1 is easily derived from Propositions 2 and 3. Indeed, letting $r_j = \frac{n_j}{n}$ and $q_j = \frac{n_{j\tau}}{n}$ we have $q_j \leq r_j$, and thus the result follows.

C. Additional bounds

Proposition 4. *Let $\tau \leq n$ be the cap on user contributions. Then $n_\tau > \tau$.*

Proof. There are only two possibilities: either $n_j < \tau$ for all j or $n_j \geq \tau$ for some j. In the latter case $n_\tau \geq n_j = \tau$ by definition. On the other hand, if $n_j < \tau$ for all j then

$$n_\tau = \sum_j n_{j\tau} = \sum_j n_j = n \geq \tau.$$

☐

Proposition 5. *Let $1 > \tau_0 > 0$ and $\tau = \tau_0 n$. Let $K(\tau_0) = |\{j \mid p_j > \tau_0\}|$ and let $\delta > 0$. With probability at least $1 - \delta$,*

$$\frac{n_\tau}{n} \geq \frac{\tau_0 K(\tau_0)}{4} - \sqrt{\frac{\log(1/\delta)}{2n}}\ .$$

Proof. Recall that J_i is the random variable that denotes the user corresponding to example i. We know that $n_j = \sum_{i=1}^n \mathbb{1}_{J_i=j}$ and $n_\tau = \sum_{i=1}^n \min(n_i, \tau)$. Let $\phi(J_1, \ldots, J_n) = \frac{n_\tau}{n}$. We want to bound the change in ϕ as we perturb a single coordinate:

$$|\phi(J_1, \ldots, J_n) - \phi(J_1', \ldots, J_n)|.$$

If we change only one point in the sample then, clearly, we change the contribution of at most two users i_1 and i_2. Let n_{i_1}' and n_{i_2}' denote the user contributions under the perturbation. Then the above expression is equal to

$$\frac{1}{n}|\min(n_{i_1}, \tau) - \min(n_{i_1}', \tau) + \min(n_{i_2}, \tau) - \min(n_{i_2}', \tau)|. \tag{11}$$

Let us assume w.l.o.g. that $n_{i_1} \geq n_{i_1}'$; this implies that $n_{i_2} \leq n_{i_2}'$. Therefore $0 \leq \min(n_{i_1}, \tau) - \min(n_{i_1}', \tau) \leq 1$ and $0 \geq \min(n_{i_2}, \tau) - \min(n_{i_2}', \tau) \geq -1$. This readily implies that (11) is bounded by $\frac{1}{n}$. We can now apply McDiarmid's inequality and see that for any $\eta > 0$

$$P\Big(\frac{n_\tau}{n} \leq \frac{1}{n}\mathbb{E}[n_\tau] - \eta\Big) \leq e^{-2n\eta^2}. \tag{12}$$

Now let $Q(\tau_0) = \sum_{j=1}^n \min(p_j, \tau_0)$. It is easy to see that

$$Q(\tau_0) = \sum_{j: p_j > \tau_0} \tau_0 + \sum_{j: p_j \leq \tau_0} p_j \geq K(\tau_0).$$

Therefore from Corollary 2 we know that

$$\begin{aligned} P\Big(\frac{n_\tau}{n} \leq \frac{\tau_0 K(\tau_0)}{4} - \eta\Big) &\leq P\Big(\frac{n_\tau}{n} \leq \frac{Q(\tau_0)}{4} - \eta\Big) \\ &\leq P\Big(\frac{n_\tau}{n} \leq \frac{1}{n}\mathbb{E}[n_\tau] - \eta\Big) \end{aligned}$$

The result follows from (12) by setting $\delta = e^{-2n\eta^2}$ and solving for η. ☐

Lemma 2. *Let $S_n = \sum_{i=1}^N X_i$ be a sum of i.i.d. Bernoulli random variables with $P(X_i = 1) = p$. Then*

$$\mathbb{E}[\min(S_n, \tau)] \geq \frac{1}{4}\min(pn, \tau) \tag{13}$$

Proof. First let us assume that $\tau < np$ in that case we have:

$$\begin{aligned} \mathbb{E}[\min(S_n, \tau)] &= \mathbb{E}[S_n \mathbb{1}_{S_n < \tau}] + \tau P(S_n > \tau) \\ &\geq \tau P(S_n > \tau) \\ &\geq \tau P(S_n > np) \geq \frac{\tau}{4}, \end{aligned}$$

where we used the fact that $P(S_n > np) > \frac{1}{4}$ (Greenberg & Mohri, 2013; Vapnik, 1998).

On the other hand if $\tau > np$ then

$$\begin{aligned} \mathbb{E}[\min(S_n, \tau)] \geq \mathbb{E}[S_n \mathbb{1}_{S_n < \tau}] &\geq \mathbb{E}[S_n \mathbb{1}_{S_n > np}] \\ &= \int_0^\infty P(S_n \mathbb{1}_{S_n > np} > t)dt \\ &= \int_0^{np} P(S_n > t)dt \\ &\geq \int_0^{np} P(S_n > np)dt \\ &\geq \frac{1}{4}np \end{aligned}$$

Combining the two cases yields the statement of the proposition. ☐

Corollary 2. *Let J_k, $k = 1, \ldots, n$ be a random variable in $\mathbb{N}$ such that $P(J_k = j) = p_j$. Let $n_j = \sum_{i=1}^n \mathbb{1}_{J_k=j}$, $\tau_0 > 0$ and $\tau = \tau_0 n$. Finally, let $n_\tau = \sum_j \min(n_j, \tau)$; then we have*

$$\frac{1}{n}\mathbb{E}[n_\tau] \geq \frac{1}{4}\sum_j \min(p_j, \tau_0)$$

Proof. By Fubini's theorem,

$$\mathbb{E}[n_\tau] = \mathbb{E}[\sum_j \min(n_j, \tau)] = \sum_j \mathbb{E}[\min(n_j, \tau)] .$$

On the other hand, n_j is a sum of independent Bernoulli random variables with probability p_j. So from the previous proposition we have

$$\begin{aligned} \frac{1}{n}\sum_j \mathbb{E}[\min(n_j, \tau)] &\geq \frac{1}{4n}\sum_j \min(p_j n, \tau) \\ &= \frac{1}{4}\sum_j \min(p_j, \tau_0) \end{aligned}$$

☐