Exercise 1.20. Suppose that we are in the multi-class classification setting, with $\mathsf{Y} = \{1, \dots, K\}$. Suppose that we posit the following probabilistic model:

$$p_{w_{1:K}}(y = k \mid x) \propto \exp(\langle w_k, x \rangle),\ k \in \{1, \dots, K\}.$$

Show that the corresponding empirical risk for the population risk

$$L[w] = \mathbb{E}[\mathrm{KL}(p(y \mid x) \parallel p_{w_{1:K}}(y \mid x))]$$

is given by the following *cross-entropy loss*:

$$\hat{L}_n[w] = -\frac{1}{n} \sum_{i=1}^{n} \left[\langle w_{y_i}, x_i \rangle - \log \left[\sum_{j=1}^{K} \exp(\langle w_j, x_i \rangle) \right] \right].$$

That is, show that $\mathbb{E}[\hat{L}_n[w]] = L[w] + C$ for every fixed $w = \{w_k\}_{k=1}^K$, where C is a constant that does not depend on w.

$$p_{w_{1:K}}(y = k \mid x) = \frac{\exp(\langle w_k, x \rangle)}{\sum_{j=1}^{K} \exp(\langle w_j, x \rangle)}$$

Take log on both sides, we obtain

$$\log p_{w_{1:K}}(y = k \mid x) = \langle w_k, x \rangle - \log\left(\sum_{j=1}^{K} \exp(\langle w_j, x \rangle)\right)$$

$$\Rightarrow \hat{L}_n[w] = -\frac{1}{n} \sum_{i=1}^{n} \log p_{w_{1:K}}(y = y_i \mid x_i)$$

Recall <u>Kullback-Leibler Divergence</u>:

Given discrete prob. distr. P and Q defined on the same sample space, $\mathcal{Y}$, the KL divergence of P from Q is defined to be

$$\mathrm{KL}(P \| Q) = \sum_{x \in \mathcal{Y}} P(x) \log \frac{P(x)}{Q(x)}$$

$\Rightarrow KL(p(y|x) \| p_{w_{1:K}}(y|x))$

$$L[w] = E_x\left[\sum_{k=1}^{K} p(y=k|x) \log \frac{p(y=k|x)}{p_{w_{1:K}}(y=k|x)}\right]$$

$$= -C - \underline{E\left[\sum_{k=1}^{K} p(y=k|x) \log p_{w_{1:K}}(y=k|x)\right]}$$

where $C = -E\left[\sum_{k=1}^{K} p(y=k|x) \log p(y=k|x)\right]$

$\downarrow$ independent of w.

$$E[\hat{L}_n[w]] = E_{(x_i, y_i)}\left[-\frac{1}{n}\sum_{i=1}^{n} \log p_{w_{1:K}}(y=y_i | x_i)\right]$$

$$= -E_{(x,y)}[\log p_{w_{1:K}}(y|x)]$$

$$= \underline{-E_x\left[\sum_{k=1}^{K} p(y=k|x) \log p_{w_{1:K}}(y=k|x)\right]}$$

$$= L[w] + C$$

$$w^* = \underset{w}{\operatorname{argmin}}\ L[w]$$

$$= \underset{w}{\operatorname{argmin}}\ L[w] + C$$

$$= \underset{w}{\operatorname{argmin}}\ E[\hat{L}_n[w]]$$

Proposition 1.23. *Suppose that ℓ is the zero-one loss, $\mathscr{F}$ is finite, and there exists $f \in \mathscr{F}$ satisfying $L[f] = 0$. Then, the empirical risk minimizer $\hat{f}_n$ satisfies, with probability at least $1 - \delta$,*

$$L[\hat{f}_n] \leqslant \frac{\log(|\mathscr{F}|/\delta)}{n}.$$

$>$ with prob. $< \delta$

Proof: For any $t > 0$, want to look at $\{L(\hat{f}_n) > t\}$

Define $B(t) = \{f \in \mathcal{F} \mid L(f) > t\}$

$\{L[\hat{f}_n] > t\} = \{\hat{f}_n \in B(t)\}$

Since by assumption, we have an $f \in \mathcal{F}$ s.t. $L(f) = 0$, then the ERM $\hat{f}_n$ will always achieve zero training error, i.e., $\hat{L}_n[\hat{f}_n] = 0$

$\hat{f}_{erm} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \hat{L}_n[f] = \underbrace{\frac{1}{n}\sum_{i=1}^n \ell(f(x_i), y_i)}_{\neq 0}$

$L[f] = \underline{E[\ell(f(x), y)]} = 0$

If $\hat{f}_n \in B(t)$, then $\exists f \in B(t)$ s.t. $\hat{L}_n[f] = 0$

For a fixed $f \in B(t)$

$$\mathbb{P}\{\hat{L}_n[f] = 0\} = \mathbb{P}\{\cap_{i=1}^n \{\operatorname{sgn}(f(x_i)) = y_i\}\}$$

$$= \prod_{i=1}^n (1 - \underbrace{\mathbb{P}\{\operatorname{sgn}(f(x_i)) \neq y_i\}}_{L(f)})$$

$$= (1 - L(f))^n$$

$$\leq (1-t)^n \qquad 1 - x \leq \exp(x)\ \forall x \in \mathbb{R}$$

$$\leq \exp(-tn)$$

$$\mathbb{P}\{\hat{f}_n \in \mathcal{B}(H)\} \le \mathbb{P}\{\exists f \in \mathcal{B}(H),\ \hat{L}_n(f)=0\}$$
$$\le \sum_{f \in \mathcal{B}(H)} \mathbb{P}(\hat{L}_n(f)=0)$$
$$\le |\mathcal{B}(H)| \exp(-tn)$$
$$\le |\mathcal{F}| \exp(-tn) = \delta$$

- Rademacher Random Variable

ε_i is a Rademacher r.v.

$$\mathbb{P}(\varepsilon_i = \pm 1) = \frac{1}{2}$$
$$\mathbb{P}(\varepsilon_i = k) = 0 \quad \forall k \neq \pm 1$$

- Rademacher Complexity

$$R_n(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} n^{-1} \sum_{i=1}^n \varepsilon_i f(x_i)\right]$$

We first look at R.C. of linear function classes
$(x \sim N(0, I))$

$$\mathcal{F}_p := \{x \mapsto \langle x, w\rangle \mid \|w\|_p \le 1\}$$

Can use dual norm $\|\cdot\|_*$

$\|\cdot\| = \|\cdot\|_p$
$\|\cdot\|_* = \|\cdot\|_q$

$$\|x\|_* := \sup_{\|z\| \le 1} \langle z, x\rangle$$

$\frac{1}{p} + \frac{1}{q} = 1$

$$R_n(\mathcal{F}_p) = \mathbb{E}\left[\sup_{\|w\|_p \le 1} n^{-1} \sum_{i=1}^n \varepsilon_i \langle x_i, w\rangle\right]$$
$$= n^{-1}\,\mathbb{E} \sup_{\|w\|_p \le 1} \langle w, \sum_{i=1}^n \varepsilon_i x_i\rangle$$

$$= n^{-1}\,\mathbb{E} \left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_q$$

where $\frac{1}{p} + \frac{1}{q} = 1$