Single function

- scaling by ℓ_∞, we obtain:

Proposition: Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. observations of p and f_0 be a fixed predictor. Then, for any $\delta \in (0, 1/2)$, with probability greater than $1 - 2\delta$,

$$\mathcal{R}(f_0) - \hat{\mathcal{R}}(f_0) < \frac{\ell_\infty}{\sqrt{2n}} \sqrt{\log \frac{1}{\delta}} \,,$$

where ℓ_∞ is an upper bound on $\ell(Y_i, f(X_i))$.

From sup to expectation

Example: $F : \mathbb{R}^n \to \mathbb{R}$

$F(z_1, \dots, z_n) = \frac{1}{n}(z_1 + \dots + z_n)$

$F(\dots, z_i, \dots) - F(\dots, z_i', \dots) = \frac{1}{n}(\cancel{\dots} + z_i + \cancel{\dots}) - \frac{1}{n}(\dots + z_i' + \dots) = \frac{z_i - z_i'}{n}$

$|F(\dots, z_i, \dots) - F(\dots, z_i', \dots)| \leq \frac{1}{n}|z_i - z_i'| \leq \frac{2\ell_\infty}{n} =: c$ if $|z_i| \leq \ell_\infty$

- **Problem:** there is often more than one function in $\mathcal{H}$...
- still possible, using for instance:

Proposition (McDiarmid's inequality):[6] Let $Z_1, \dots, Z_n$ be independent random variables and F a function such that $F : \mathbb{R}^n \to \mathbb{R}$

$$\forall i \in [n],\ \forall z_i, z_i' \in \mathbb{R},\quad |F(z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_n) - F(z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_n)| \leq c\,.$$

Then

$$\mathbb{P}\left(|F(Z_1, \dots, Z_n) - \mathbb{E}\left[F(Z_1, \dots, Z_n)\right]| \geq t\right) \leq 2\exp\left(-2t^2/(nc^2)\right)\,.$$

[6] McDiarmid, *On the method of bounded differences*, Survey in Combinatorics, 1989

Application of McDiarmid

$$\sup_{f \in \mathcal{H}} \{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \} \leq \frac{\ell_\infty}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}} + \mathbb{E}\left[\sup_{f \in \mathcal{H}} \{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \} \right]$$

w.h.p.

- set $Z_i := (X_i, Y_i)$, and

$$H(Z_1, \ldots, Z_n) := \sup_{f \in \mathcal{H}} \left\{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right\} .$$

- Mc Diarmid tells us that, with probability higher than $1 - \delta$,

$$\boxed{H(Z_1, \ldots, Z_n) - \mathbb{E}\left[H(Z_1, \ldots, Z_n) \right] \leq \frac{\ell_\infty \sqrt{2}}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}} .}$$

- getting bound on $\mathbb{E}\left[H(Z_1, \ldots, Z_n) \right]$ automatically yields bound on $\sup_{f \in \mathcal{H}} \left\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right\}$
- by symmetry, upper bound on $\sup_{f \in \mathcal{H}} \left| \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right|$

Recall: $H(z_1,\dots,z_n) = \sup_{f \in \mathcal{H}} \left\{ \hat{R}(f) - R(f) \right\}.$

f_0 achieves the sup

$$= \sup_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) - R(f) \right\} = \frac{1}{n} \sum \ell(f_0(x_i), y_i) - R(f_0)$$

$$H(z_1,\dots,z_i,\dots,z_n) - H(z_1,\dots,z_i',\dots,z_n) =$$

$$= \frac{1}{n} \cancel{\sum \ell(f_0(x_j), y_j)} - \cancel{R(f_0)} - \left[\frac{1}{n} \cancel{\sum_{j \neq i} \ell(f_0(x_j), y_j)} + \ell(f_0(x_i'), y_i') - \cancel{R(f_0)} \right]$$

$$= \frac{1}{n} \ell(f_0(x_i), y_i) - \frac{1}{n} \ell(f_0(x_i'), y_i')$$

$$\left| \;\underline{\qquad}\; \right| \leq \frac{2 \ell_\infty}{n} \qquad \square$$

$$\boxed{\Rightarrow H \text{ satisfies the assumptions of McDiarmid inequality}}$$

4.2. Rademacher complexity

Rademacher complexity

- set $Z := (X, Y)$ and $\mathcal{G} := \{(x, y) \mapsto \ell(y, f(x))\}$, with f in some function class $\mathcal{H}$
- **Recall:** we want to bound

$$\mathbb{E} \sup_{f \in \mathcal{H}} \left\{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right\} = \mathbb{E} \sup_{g \in \mathcal{G}} \left\{ \mathbb{E}\left[g(Z)\right] - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \right\} .$$

- set $\mathcal{D} := \{Z_1, \ldots, Z_n\}$ the data

Definition: We call *Rademacher complexity* of the function class $\mathcal{G}$ the quantity

$$R_n(\mathcal{G}) := \mathbb{E}_{\varepsilon, \mathcal{D}} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(Z_i) \right] ,$$

where the ε_is are independent Rademacher random variables (that is, $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$).

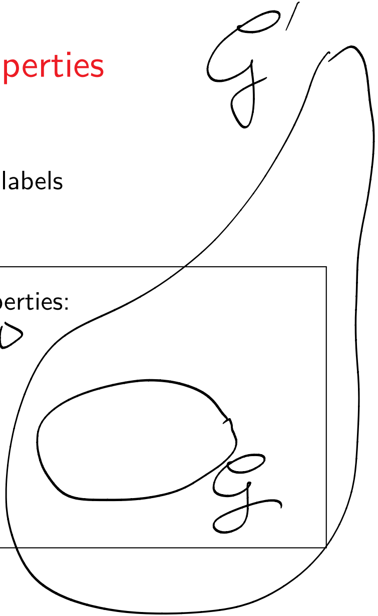
$$R_n(\mathcal{G}) := \mathbb{E}\Big[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum \varepsilon_i g(Z_i)\Big]$$

Rademacher complexity, first properties

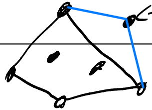
- **Intuition:** expectation of maximal dot-product with random labels
- measures the *capacity* of the set $\mathcal{G}$

Properties: Rademacher complexity satisfies the following properties:

- if $\mathcal{G} \subset \mathcal{G}'$, then $R_n(\mathcal{G}) \leq R_n(\mathcal{G}')$;
- $R_n(\mathcal{G} + \mathcal{G}') = R_n(\mathcal{G}) + R_n(\mathcal{G}')$;
- $R_n(\alpha\mathcal{G}) = |\alpha|\, R_n(\mathcal{G})$;
- if g_0 is a function, $R_n(\mathcal{G} + \{g_0\}) = R_n(\mathcal{G})$;
- $R_n(\mathcal{G}) = R_n(\mathrm{conv}(\mathcal{G}))$.



convex hull



outside!

Symmetrization (argument)

- **Question:** why is it useful?
- Rademacher complexity directly controls expected uniform deviation

~~uniformly~~ in $\mathcal{G}$.

Proposition (symmetrization): With the previous notation,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\underbrace{\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}\left[g(Z)\right]}_{\hat{R}(g)\,-\,R(g)}\right\}\right] \leq 2R_n(\mathcal{G})\,, = 2\,\mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum \varepsilon_i g(Z_i)\right]$$

and

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\mathbb{E}\left[g(Z)\right] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right\}\right] \leq 2R_n(\mathcal{G})\,.$$

~~$Z_1' = Z_1$~~

Symmetrization, proof

$\mathcal{D} = \{Z_1, \dots, Z_n\}$

- let $\mathcal{D}' := \{Z_1', \dots, Z_n'\}$ be an independent copy of $\mathcal{D}'$
- in particular, one has $\mathbb{E}\left[g(Z_i') \mid \mathcal{D}\right] = \mathbb{E}\left[g(Z_i)\right]$
- we write

$\hat{\mathbb{E}}[g(Z_i')]$

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\mathbb{E}\left[g(Z)\right] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right\}\right] = \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\mathbb{E}\left[g(Z_i') \mid \mathcal{D}\right] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right\}\right]$$

$$= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[g(Z_i') - g(Z_i) \mid \mathcal{D}\right]\right\}\right].$$

$\mathbb{E}[g(Z_i) \mid \mathcal{D}]$ since $g(Z_i)$ is $\mathcal{D}$-mes.

$X_1 \le \max(X_1, X_2)$ / $X_2 \le \max$

$\mathbb{E}X_1 \le \mathbb{E}\max(X_1, X_2)$ $\mathbb{E}X_2 \le \mathbb{E}\max$

$\max(\mathbb{E}X_1, \mathbb{E}X_2) \le \mathbb{E}\max$. D

$\max(\mathbb{E}[X_1], \mathbb{E}[X_2]) \le \mathbb{E}[\max(X_1, X_2)]$

Symmetrization, proof ctd.

- since the sup of expectation is $\le$ than expectation of the sup, (max)

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\cancel{\mathbb{E}\left[g(Z)\right] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)}\right\}\right] \le \mathbb{E}\left[\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_i') - g(Z_i))\right\} \mid \mathcal{D}\right]\right]$$

$\sup_{g} \mathbb{E}[g(Z_i') - g(Z_i) \mid \mathcal{D}]$

$$= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_i') - g(Z_i))\right\}\right]$$

by the tower property. (Z_i indep. copy of Z_i').

- we notice that

$$g(Z_i') - g(Z_i) \quad \text{and} \quad \varepsilon_i(g(Z_i') - g(Z_i)) \text{ have the same distribution}$$

(this is what we call symmetrization)

Remember: $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = \frac{1}{2}$

Symmetrization proof, ctd.

$W_i \sim \mathcal{N}(0,1)$, $\varepsilon_i W_i \overset{d}{=} W_i$

- thus

$$
\begin{aligned}
\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_i') - g(Z_i))\right\}\right] &= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(g(Z_i') - g(Z_i))\right\}\right] \\
&\leq \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i)\right\}\right] + \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}-\varepsilon_i g(Z_i)\right\}\right] \\
&= 2R_n(\mathcal{G})
\end{aligned}
$$

since ε and $-\varepsilon$ have the same distribution. $\square$