

Useful properties of the (empirical) Rademacher complexity

Lemma 1. For $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ let $\mathcal{F}' := \{f + f_0 : f \in \mathcal{F}\}$ be a translated version for some $f_0 : \mathcal{X} \rightarrow \mathbb{R}$. Then, for any m ,

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}') = \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})$$

Proof.

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}') = \mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}'} \left(\frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right) \right] = \mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m \sigma_i (f(z_i) + f_0(z_i)) \right) \right] \quad (1)$$

$$= \mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right) \right] + \underbrace{\mathbb{E}_{\sigma} \frac{1}{m} \sum_{i=1}^m \sigma_i f_0(z_i)}_{= \frac{1}{m} \sum_{i=1}^m [\mathbb{E}_{\sigma} \sigma_i] f_0(z_i) = 0} \quad (2)$$

$$= \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) \quad (3)$$

□

Lemma 2. For $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ let $\mathcal{F}' := \{\lambda f : f \in \mathcal{F}\}$ be scaled by a constant $\lambda \in \mathbb{R}$. Then, for any m ,

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}') = \lambda \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})$$

Proof.

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}') = \mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}'} \left(\frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right) \right] = \mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m \sigma_i \lambda f(z_i) \right) \right] = \lambda \mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right) \right] = \lambda \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) \quad (4)$$

□

Lemma 3. For $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ and $\phi : \mathbb{R} \rightarrow \mathbb{R}$ let $\mathcal{F}' := \{\phi \circ f : f \in \mathcal{F}\}$. If ϕ is L -Lipschitz continuous, i.e. $|\phi(t) - \phi(t')| \leq L|t - t'|$, then for any m ,

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}') \leq L \cdot \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})$$

Proof. We prove it for $m = 1$, the general case works iteratively (see a textbook).

$$\hat{\mathfrak{R}}_{\mathcal{D}_1}(\mathcal{F}') = \mathbb{E}_{\sigma_1} \left[\sup_{f \in \mathcal{F}} (\sigma_1 \phi(f(z_1))) \right] \quad (\text{every function in } \mathcal{F}' \text{ has the form } \phi \circ f \text{ for some } f \in \mathcal{F}) \quad (5)$$

$$= \frac{1}{2} \sup_{f \in \mathcal{F}} (\phi(f(z_1))) + \frac{1}{2} \sup_{f \in \mathcal{F}} (-\phi(f(z_1))) \quad (6)$$

$$= \frac{1}{2} \sup_{f, f' \in \mathcal{F}} (\phi(f(z_1)) - \phi(f'(z_1))) \quad (7)$$

$$\leq \frac{1}{2} \sup_{f, f' \in \mathcal{F}} (L|f(z_1) - f'(z_1)|) \quad (\text{by } L\text{-Lipschitz property of } \phi) \quad (8)$$

$$\leq L \frac{1}{2} \sup_{f, f' \in \mathcal{F}} (f(z_1) - f'(z_1)) \quad (\text{because } \mathbf{sup} \text{ will be where difference is non-negative}) \quad (9)$$

$$= L \frac{1}{2} \sup_{f \in \mathcal{F}} (f(z_1)) + \sup_{f \in \mathcal{F}} (-f'(z_1)) \quad (10)$$

$$= L \mathbb{E}_{\sigma_1} \sup_{f \in \mathcal{F}} (\sigma_1 f(z_1)) \quad (11)$$

$$= L \hat{\mathfrak{R}}_{\mathcal{D}_1}(\mathcal{F}) \quad (12)$$

Rademacher complexity of linear function classes

Lemma 4. Let $\mathcal{F} = \{f = \langle w, z \rangle : \mathcal{Z} \rightarrow \mathbb{R}\}$ be linear functions with $\|w\| \leq B$. Then for any $\mathcal{D}_m = \{z_1, \dots, z_m\}$

$$\hat{\mathfrak{R}}_m(\mathcal{F}) = \frac{B}{m} \sqrt{\sum_i \|z_i\|^2}.$$

Proof.

1. For any fixed $\sigma \in \{\pm 1\}^m$:

$$\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) = \sup_{\|w\| \leq B} \frac{1}{m} \sum_{i=1}^m \sigma_i \langle w, z_i \rangle = \sup_{\|w\| \leq B} \frac{1}{m} \left\langle w, \sum_{i=1}^m \sigma_i z_i \right\rangle \quad (13)$$

$$\stackrel{w \propto \sum_i \sigma_i z_i}{=} \frac{1}{m} \left\langle \frac{B}{\|\sum_i \sigma_i z_i\|} \sum_{i=1}^m \sigma_i z_i, \sum_{i=1}^m \sigma_i z_i \right\rangle = \frac{B}{m} \left\| \sum_{i=1}^m \sigma_i z_i \right\| \quad (14)$$

2. Therefore

$$\hat{\mathfrak{R}}_m(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right) \right] = \mathbb{E}_\sigma \frac{B}{m} \left\| \sum_{i=1}^m \sigma_i z_i \right\| = \frac{B}{m} \mathbb{E}_\sigma \sqrt{\sum_{i,j} \sigma_i \sigma_j \langle z_i, z_j \rangle} \quad (15)$$

$$\stackrel{\sqrt{\cdot} \text{ concave}}{\leq} \frac{B}{m} \sqrt{\mathbb{E}_\sigma \sum_{i,j} \sigma_i \sigma_j \langle z_i, z_j \rangle} = \frac{B}{m} \sqrt{\mathbb{E}_\sigma \sum_i \sigma_i \sigma_i \langle z_i, z_i \rangle + \mathbb{E}_\sigma \sum_{i \neq j} \sigma_i \sigma_j \langle z_i, z_j \rangle} \quad (16)$$

$$= \frac{B}{m} \sqrt{\sum_i \langle z_i, z_i \rangle + \sum_{i \neq j} (\mathbb{E}_{\sigma_i} \sigma_i) (\mathbb{E}_{\sigma_j} \sigma_j) \langle z_i, z_j \rangle} \quad (\sigma_i, \sigma_j \text{ independent for } i \neq j) \quad (17)$$

$$= \frac{B}{m} \sqrt{\sum_i \|z_i\|^2} \quad (\mathbb{E} \sigma_i = 0) \quad (18)$$

Rademacher-based generalization bound

Theorem 5. Let $\ell(x, y, h) \leq c$ be a bounded loss function. For a hypothesis set $\mathcal{H} \subset \mathbb{R}^x$ let $\mathcal{F} = \{\ell \circ h : h \in \mathcal{H}\}$, with $(\ell \circ h)(x, y) := \ell(x, y, h)$. Then, with probability at least $1 - \delta$, it holds for all $h \in \mathcal{H}$:

$$\mathcal{R}^\ell(h) \leq \hat{\mathcal{R}}^\ell(h) + 2\mathfrak{R}_m(\mathcal{F}) + c\sqrt{\frac{\log(1/\delta)}{2m}}.$$

Proof. (follow [C. Scott, http://web.eecs.umich.edu/~cscott/past_courses/eecs598w14/index.html])

We drop ℓ from the notation of empirical and expected risk and prove the result in three steps.

Step 1: for every $h \in \mathcal{H}$,

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}_{\mathcal{D}}(h) + \sup_{g \in \mathcal{H}} \left(\mathcal{R}(g) - \hat{\mathcal{R}}_{\mathcal{D}}(g) \right) \quad (19)$$

Step 2: we show that with probability $1 - \delta$ over \mathcal{D} :

$$\sup_{h \in \mathcal{H}} \left(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h) \right) \leq \mathbb{E}_{\mathcal{D}} \sup_{h \in \mathcal{H}} \left(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h) \right) + c\sqrt{\frac{\log \frac{2}{\delta}}{2m}} \quad (20)$$

Step 3: we show

$$\mathbb{E}_{\mathcal{D}} \sup_{h \in \mathcal{H}} \left(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h) \right) \leq 2\mathfrak{R}_m(\mathcal{F}) \quad (21)$$

In combination the first statement of the theorem follows.

Some useful inequalities:

- $\sup_f [A(f) + B(f)] \leq \sup_f A(f) + \sup_f B(f)$ and $\sup_f A(f) - \sup_f B(f) \leq \sup_f [A(f) - B(f)]$
- **Jensen's inequality:** for convex $\phi : \mathbb{R} \rightarrow \mathbb{R}$: $\mathbb{E}_z \phi(A(z)) \geq \phi(\mathbb{E}_z A(z))$.
for concave $\psi : \mathbb{R} \rightarrow \mathbb{R}$: $\mathbb{E}_z \psi(A(z)) \leq \psi(\mathbb{E}_z A(z))$
- \sup is convex, i.e. $\sup_f \mathbb{E}_z(\cdot) \leq \mathbb{E}_z \sup_f(\cdot)$

Step 1: is simple, because h is one element of \mathcal{H} .

(22)

Step 2: We will use

Lemma 6 (McDiarmid's inequality). Let $f : \mathcal{Z}^m \rightarrow \mathbb{R}$ be a function of m variables for which a bounded difference inequality holds, namely that there exists a $C > 0$ such that for all $i = 1, \dots, m$ and for all $z_1, \dots, z_m, z'_i \in \mathcal{Z}$:

$$\left| f(z_1, \dots, z_m) - f(z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_m) \right| \leq C.$$

Let Z_1, \dots, Z_m be m independent random variables with values in \mathcal{Z} . Then, with probability at least $1 - \delta$ the following inequality holds:

$$\left| f(Z_1, \dots, Z_m) - \mathbb{E}[f(Z_1, \dots, Z_m)] \right| \leq C\sqrt{\frac{m \log\left(\frac{2}{\delta}\right)}{2}}.$$

Remember $\ell(\cdot) \leq c$. Then $\sup_{h \in \mathcal{H}} (\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h))$ fulfills the bounded difference conditions with $C = \frac{c}{m}$ as a function of the m samples, $z_i = (x_i, y_i)$: for all $i = 1, \dots, m$ and $(x_1, y_1), \dots, (x_m, y_m), (x'_i, y'_i)$ we have

$$\sup_{h \in \mathcal{H}} \left(\mathcal{R}(h) - \frac{1}{m} \sum_i \ell(x_i, y_i, h) \right) - \sup_{h \in \mathcal{H}} \left(\mathcal{R}(h) - \frac{1}{m} \sum_i \ell(x'_i, y'_i, h) \right) \quad (23)$$

$$\leq \sup_{h \in \mathcal{H}} \left(\mathcal{R}(h) - \frac{1}{m} \sum_i \ell(x_i, y_i, h) - \mathcal{R}(h) + \frac{1}{m} \sum_i \ell(x'_i, y'_i, h) \right) \quad (24)$$

$$\leq \sup_{h \in \mathcal{H}} \left(\frac{1}{m} \ell(x_i, y_i, h) - \frac{1}{m} \ell(x'_i, y'_i, h) \right) \leq \frac{1}{m} \sup_{h \in \mathcal{H}} \left(\ell(x_i, y_i, h) \right) \leq \frac{c}{m} \quad (25)$$

and in the same way we can bound the negative of the left hand also by $\frac{c}{m}$.

Therefore, with probability at least $1 - \delta$ over $\mathcal{D}_m \sim p$:

$$\sup_{h \in \mathcal{H}} \left(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h) \right) \leq \mathbb{E}_{\mathcal{D}} \sup_{h \in \mathcal{H}} \left(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h) \right) + c \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \quad (26)$$

Step 3: For any $h \in \mathcal{H}$ and $\mathcal{D}' \sim p$ (of size m) we have $\mathcal{R}^\ell(h) = \mathbb{E}_{\mathcal{D}'} \mathcal{R}_{\mathcal{D}'}^\ell(h)$.

$$\sup_{h \in \mathcal{H}} \left(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h) \right) = \sup_{h \in \mathcal{H}} \left(\mathbb{E}_{\mathcal{D}'} \hat{\mathcal{R}}_{\mathcal{D}'}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h) \right) \quad (27)$$

$$= \sup_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}'} \left(\hat{\mathcal{R}}_{\mathcal{D}'}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h) \right) \quad (28)$$

$$\leq \mathbb{E}_{\mathcal{D}'} \sup_{h \in \mathcal{H}} \left(\hat{\mathcal{R}}_{\mathcal{D}'}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h) \right) \quad (29)$$

Taking expectation over \mathcal{D} on both sides:

$$\mathbb{E}_{\mathcal{D}} \sup_{h \in \mathcal{H}} \left(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h) \right) \leq \mathbb{E}_{\mathcal{D}} \mathbb{E}_{\mathcal{D}'} \sup_{h \in \mathcal{H}} \left(\hat{\mathcal{R}}_{\mathcal{D}'}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h) \right) = \mathbb{E}_{\mathcal{D}, \mathcal{D}'} \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m (f(z_i) - f(z'_i)) \right) \quad (30)$$

for $z = (x, y)$ and $f = \ell \circ h$, because $\hat{\mathcal{R}}_{\mathcal{D}} = \frac{1}{m} \sum_i \ell(x_i, y_i, h) = \frac{1}{m} \sum_i (\ell \circ h)(x_i, y_i)$.

$$= \frac{1}{m} \mathbb{E}_{\mathcal{D}, \mathcal{D}'} \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^m (f(z_i) - f(z'_i)) \right) \quad (31)$$

For any $j = 1, \dots, m$, e.g. $j = 1$, z_j and z'_j are i.i.d., so the expected value doesn't notice if we swap $z_j \leftrightarrow z'_j$

$$\mathbb{E}_{\mathcal{D}, \mathcal{D}'} \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^m (f(z_i) - f(z'_i)) \right) = \mathbb{E}_{\mathcal{D}, \mathcal{D}'} \sup_{f \in \mathcal{F}} \left((f(z'_j) - f(z_j) + \sum_{i \neq j} (f(z_i) - f(z'_i))) \right) \quad (32)$$

$$= \mathbb{E}_{\mathcal{D}, \mathcal{D}'} \sup_{f \in \mathcal{F}} \left(-(f(z_j) - f(z'_j) + \sum_{i \neq j} (f(z_i) - f(z'_i))) \right) \quad (33)$$

For any $\sigma \in \{\pm 1\}^m$, swap $z_j \leftrightarrow z'_j$ whenever $\sigma_j = -1$:

$$\mathbb{E}_{\mathcal{D}, \mathcal{D}'} \sup_{f \in \mathcal{F}} \sum_{i=1}^m (f(z_i) - f(z'_i)) = \mathbb{E}_{\mathcal{D}, \mathcal{D}'} \sup_{f \in \mathcal{F}} \sum_i \sigma_i (f(z_i) - f(z'_i)) \quad (34)$$

$$\leq \mathbb{E}_{\mathcal{D}, \mathcal{D}'} \left(\sup_{f \in \mathcal{F}} \sum_i \sigma_i f(z_i) + \sup_{f' \in \mathcal{F}} \sum_i -\sigma_i f'(z'_i) \right) \quad (35)$$

Taking expectations over σ on both sides and noticing that $-\sigma_i$ has the same distribution as σ_i :

$$\mathbb{E}_{\mathcal{D}, \mathcal{D}'} \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^m (f(z_i) - f(z'_i)) \right) \leq \mathbb{E}_{\mathcal{D}, \sigma} \sup_{f \in \mathcal{F}} \sum_i \sigma_i f(z_i) + \mathbb{E}_{\mathcal{D}', \sigma} \sup_{f' \in \mathcal{F}} \sum_i \sigma_i f'(z'_i) \quad (36)$$

$$\leq 2m \mathbb{E}_{\mathcal{D}, \sigma} \frac{1}{m} \sup_{f \in \mathcal{F}} \sum_i \sigma_i f(z_i) \quad (37)$$

$$= 2m \mathfrak{R}_{\mathcal{D}}(\mathcal{F}) \quad (38)$$

Hard-margin SVM bound

- $\|x\| \leq R$ with probability 1
- $\mathcal{H} = \{h(x) = \langle w, x \rangle : \|w\| \leq B\}$ for B that we'll specify later
- ramp-loss: $\ell(x, y, h) = \min\{\max\{0, 1 - y\langle w, x \rangle\}, 1\} \in [0, 1]$
- ℓ is an upper bounds to the 0/1 error

$$\Pr\{h(x) \neq y\} = \mathcal{R}_p^{0/1}(h) \leq \mathcal{R}_p^\ell(h)$$

- hard-margin h fulfills $y_i \langle w, x_i \rangle \geq 1$ for $i = 1, \dots, m$: $\hat{\mathcal{R}}_{\mathcal{D}_m}^\ell(h) = 0$
- ℓ is 1-Lipschitz, i.e. for $\mathcal{F} = \{\ell \circ h : h \in \mathcal{H}\}$:

$$\mathfrak{R}_m(\mathcal{F}) \leq \mathfrak{R}_m(\mathcal{H}) \leq BR\sqrt{\frac{1}{m}}$$

- $B = \|w^*\|$ ensures that hard-margin SVM $h_S \in \mathcal{H}$.

With prob. $1 - \delta$: $\Pr\{h_S(x) \neq y\} \leq \frac{2R\|w^*\|}{\sqrt{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}$

Soft-margin SVM bounds

- $\|x\| \leq R$ with probability 1
- $\mathcal{H} = \{h(x) = \langle w, x \rangle : \|w\| \leq B\}$ for fixed B
- ramp-loss: $\ell(x, y, h) = \min\{\max\{0, 1 - y\langle w, x \rangle\}, 1\} \in [0, 1]$
- ℓ is 1-Lipschitz, i.e. for $\mathcal{F} = \{\ell \circ h : h \in \mathcal{H}\}$:

$$\mathfrak{R}_m(\mathcal{F}) \leq \mathfrak{R}_m(\mathcal{H}) \leq BR\sqrt{\frac{1}{m}}$$

- ℓ is an upper bounds to the 0/1 error

$$\Pr\{h(x) \neq y\} = \mathcal{R}_p^{0/1}(h) \leq \mathcal{R}_p^\ell(h)$$

With prob. $1 - \delta$ for every $w \in \mathcal{H}$:

$$\mathcal{R}^\ell(w) \leq \hat{\mathcal{R}}_{\mathcal{D}_m}^\ell(w) + \frac{RB}{\sqrt{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

Note: ℓ is upper bound to 0/1-loss, and ℓ is lower bound to Hinge-loss, therefore

$$\Pr\{\text{sign}\langle w, x \rangle \neq y\} \leq \frac{1}{m} \sum_{i=1}^m \max\{0, 1 - y_i \langle w, x_i \rangle\} + \frac{RB}{\sqrt{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}$$