IST Austria: Statistical Machine Learning 2015/16

Christoph Lampert <chl@ist.ac.at>

Lecture 8 – Notes

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 $f_0 : \mathcal{X} \to \mathbb{R}$. Then, for any m,*

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

Proof.

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

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

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

□

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.

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

□

Lemma 3. *For $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ and $\phi : \mathbb{R} \to \mathbb{R}$ 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).

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

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

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

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

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

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

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

$$= L\,\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) \tag{12}$$

□

Rademacher complexity of linear function classes

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

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

If $\langle \cdot, \cdot \rangle$ is given by a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then

$$\hat{\mathfrak{R}}_m(\mathcal{F}) = \frac{B}{m}\sqrt{\text{trace}(K)}.$$

where $K \in \mathbb{R}^{m\times m}$ is the kernel matrix with entries $k_{ij} = k(z_i, z_j) = \langle z_i, z_j \rangle$.

Proof. The second statement follows from the first, since $\text{trace}(K) = \sum_i \langle z_i, z_i\rangle = \|z_i\|^2$.

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

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

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

2. Therefore

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

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

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

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

3. Note: we could have written kernel evaluations everywhere instead of inner products

Rademacher-based generalization bound

Theorem 5. *Let $\ell(x,y,h) \leq c$ be a bounded loss function. For a hypothesis set $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ 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 $h \in \mathcal{H}$:*

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

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

We drop p, ℓ and m from the notation and prove the result in three steps: first, for every $h \in \mathcal{H}$,

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}_{\mathcal{D}}(h) + \sup_{h \in \mathcal{H}} \Big(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h) \Big) \tag{19}$$

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

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

Finally we show

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

In combination the first statement of the theorem follows.

Some useful inequalities:

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

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

$$\tag{22}$$

Step 2: We will use

Lemma 6 (McDiarmid's inequality)**.** *Let $f : \mathcal{Z}^m \to \mathbb{R}$ 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 $z_1, \dots, z_m, z'_i \in \mathcal{Z}$:*

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

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:

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

Remember $\ell(\cdot) \leq c$. Then $\mathbf{sup}_{h\in\mathcal{H}}\Big(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h)\Big)$ 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

$$\mathbf{sup}_{h\in\mathcal{H}}\Big(\mathcal{R}(h) - \frac{1}{m}\sum_i \ell(x_i, y_i, h)\Big) \;-\; \mathbf{sup}_{h\in\mathcal{H}}\Big(\mathcal{R}(h) - \frac{1}{m}\sum_i \ell(x_i', y_i', h)\Big) \tag{23}$$

$$\leq \mathbf{sup}_{h\in\mathcal{H}}\Big(\mathcal{R}(h) - \frac{1}{m}\sum_i \ell(x_i, y_i, h) - \mathcal{R}(h) + \frac{1}{m}\sum_i \ell(x_i', y_i', h)\Big) \tag{24}$$

$$\leq \mathbf{sup}_{h\in\mathcal{H}}\Big(\frac{1}{m}\ell(x_i, y_i, h) - \frac{1}{m}\ell(x_i', y_i', h)\Big) \leq \frac{1}{m}\mathbf{sup}_{h\in\mathcal{H}}\Big(\ell(x_i, y_i, h)\Big) \leq \frac{c}{m} \tag{25}$$

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 $\mathcal{D}_m \sim p$:

$$\mathbf{sup}_{h\in\mathcal{H}}\Big(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h)\Big) \leq \mathop{\mathbb{E}}_{\mathcal{D}}\mathbf{sup}_{h\in\mathcal{H}}\Big(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h)\Big) + c\sqrt{\frac{\log\frac{2}{\delta}}{2m}} \tag{26}$$

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

$$\mathbf{sup}_{h\in\mathcal{H}}\Big(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h)\Big) = \mathbf{sup}_{h\in\mathcal{H}}\Big(\mathop{\mathbb{E}}_{\mathcal{D}'}\hat{\mathcal{R}}_{\mathcal{D}'}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h)\Big) \tag{27}$$

$$= \mathbf{sup}_{h\in\mathcal{H}}\mathop{\mathbb{E}}_{\mathcal{D}'}\Big(\hat{\mathcal{R}}_{\mathcal{D}'}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h)\Big) \tag{28}$$

$$\leq \mathop{\mathbb{E}}_{\mathcal{D}'}\mathbf{sup}_{h\in\mathcal{H}}\Big(\hat{\mathcal{R}}_{\mathcal{D}'}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h)\Big) \tag{29}$$

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

$$\mathop{\mathbb{E}}_{\mathcal{D}}\mathbf{sup}_{h\in\mathcal{H}}\Big(\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h)\Big) \leq \mathop{\mathbb{E}}_{\mathcal{D}}\mathop{\mathbb{E}}_{\mathcal{D}'}\mathbf{sup}_{h\in\mathcal{H}}\Big(\hat{\mathcal{R}}_{\mathcal{D}'}(h) - \hat{\mathcal{R}}_{\mathcal{D}}(h)\Big) = \mathop{\mathbb{E}}_{\mathcal{D},\mathcal{D}'}\mathbf{sup}_{f\in\mathcal{F}}\Big(\frac{1}{m}\sum_{i=1}^{m}\big(f(z_i) - f(z_i')\big)\Big) \tag{30}$$

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

$$= \frac{1}{m}\mathop{\mathbb{E}}_{\mathcal{D},\mathcal{D}'}\mathbf{sup}_{f\in\mathcal{F}}\Big(\sum_{i=1}^{m}\big(f(z_i) - f(z_i')\big)\Big) \tag{31}$$

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 $z_j \leftrightarrow z_j'$

$$\mathop{\mathbb{E}}_{\mathcal{D},\mathcal{D}'}\mathbf{sup}_{f\in\mathcal{F}}\Big(\sum_{i=1}^{m}\big(f(z_i) - f(z_i')\big)\Big) = \mathop{\mathbb{E}}_{\mathcal{D},\mathcal{D}'}\mathbf{sup}_{f\in\mathcal{F}}\Big((f(z_j') - f(z_j) + \sum_{i\neq j}\big(f(z_i) - f(z_i')\big)\Big) \tag{32}$$

$$= \mathop{\mathbb{E}}_{\mathcal{D},\mathcal{D}'}\mathbf{sup}_{f\in\mathcal{F}}\Big(-(f(z_j) - f(z_j') + \sum_{i\neq j}\big(f(z_i) - f(z_i')\big)\Big) \tag{33}$$

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

$$\mathop{\mathbb{E}}_{\mathcal{D},\mathcal{D}'}\mathbf{sup}_{f\in\mathcal{F}}\sum_{i=1}^{m}\big(f(z_i) - f(z_i')\big) = \mathop{\mathbb{E}}_{\mathcal{D},\mathcal{D}'}\mathbf{sup}_{f\in\mathcal{F}}\sum_i \sigma_i\big(f(z_i) - f(z_i')\big) \tag{34}$$

$$\leq \mathop{\mathbb{E}}_{\mathcal{D},\mathcal{D}'}\Big(\mathbf{sup}_{f\in\mathcal{F}}\sum_i \sigma_i f(z_i) + \mathbf{sup}_{f'\in\mathcal{F}}\sum_i -\sigma_i f(z_i')\Big) \tag{35}$$

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

$$\mathop{\mathbb{E}}_{\mathcal{D},\mathcal{D}'}\mathbf{sup}_{f\in\mathcal{F}}\Big(\sum_{i=1}^{m}\big(f(z_i) - f(z_i')\big)\Big) \leq \mathop{\mathbb{E}}_{\mathcal{D},\sigma}\mathbf{sup}_{f\in\mathcal{F}}\sum_i \sigma_i f(z_i) \;+\; \mathop{\mathbb{E}}_{\mathcal{D}',\sigma}\mathbf{sup}_{f'\in\mathcal{F}}\sum_i \sigma_i f(z_i')) \tag{36}$$

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

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

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: $\ell(x, y, h) = \mathbf{min}\{\ \mathbf{max}\{0, 1 - y\langle w, x \rangle\},\ 1\ \} \in [0, 1]$
- 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 $i = 1, \ldots, m$: $\quad \hat{\mathcal{R}}^{\ell}_{\mathcal{D}_m}(h) = 0$
- 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. $1 - \delta$: $\qquad \Pr\{h_S(x) \neq y\} \leq \dfrac{2R\|w^*\|}{\sqrt{m}} + \sqrt{\dfrac{\log(1/\delta)}{2m}}$

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: $\ell(x, y, h) = \mathbf{min}\{\ \mathbf{max}\{0, 1 - y\langle w, x \rangle\},\ 1\ \} \in [0, 1]$
- 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 $w \in \mathcal{H}$:

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

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

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