# Worst-case bounds

## Theoretical Deep Learning #2: generalization ability

Eugene Golikov

MIPT, fall 2019

Neural Networks and Deep Learning Lab., MIPT

# Notation and goal

- Data distribution: $\mathcal{D}$;
- Dataset: $S_n = \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$, where all $y_i \in \{-1, 1\}$, all $x_i \in X$;
- Model: $f : X \to \mathbb{R}$;
- Loss function $l(y, f(x))$;
- Risk: $R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}} l(y, f(x))$;
- Empirical risk: $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n l(y_i, f(x_i))$;
- Result of learning on dataset $S_n$: $\hat{f}_n \in \mathcal{F}$.

**Our goal is to bound the risk difference:**

$$R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) < \mathrm{bound}(N(\hat{f}_n), n, \delta) \quad \text{w.p. } \geq 1 - \delta \text{ over } S_n.$$

**Worst-case bound:**

$$R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) \leq \sup_{f \in \mathcal{F}} (R(f) - \hat{R}_n(f)).$$

**Zero-one loss case:**

Assume $l(y, f(x)) = l_{0/1}(y, f(x)) = [yf(x) < 0]$; then

$$\mathbb{P}\left\{ \sup_{f \in \mathcal{F}} (R(f) - \hat{R}_n(f)) > \epsilon \right\} \leq 2\Pi(\mathcal{F}; 2n) e^{-\epsilon^2 n/8},$$

where $\Pi(\mathcal{F}, k) = \sup_{X_k} \#\{\text{labelings } \mathcal{F} \text{ induces on } X_k\}$.

# Worst-case bound: 0/1 loss

**The bound:**

$$R(\hat{f}_n) \leq \hat{R}_n(\hat{f}_n) + \sqrt{\frac{8}{n}\left(\log \Pi(\mathcal{F}; 2n) + \log\left(\frac{2}{\delta}\right)\right)} \quad \text{w.p. } \geq 1-\delta \text{ over } S_n.$$

**Bounding growth function:**

$$\mathrm{VC}(\mathcal{F}) := \max_k \{k : \ \Pi(\mathcal{F}; k) = 2^k\}.$$

Or, in other words,

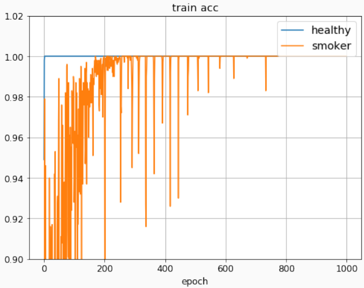
$$\mathrm{VC}(\mathcal{F}) = \max_k \{k : \ \mathcal{F} \text{ shatters some } X_k\}.$$

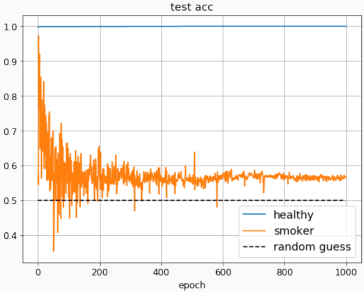
From Sauer lemma:

$$\log \Pi(\mathcal{F}; k) = \mathrm{VC}(\mathcal{F})(1 + \log(k/\mathrm{VC}(\mathcal{F}))) \quad \text{if } k > \mathrm{VC}(\mathcal{F}), \text{ else } k.$$

# Worst-case bound: 0/1 loss

**Two nets, good one and bad one:**



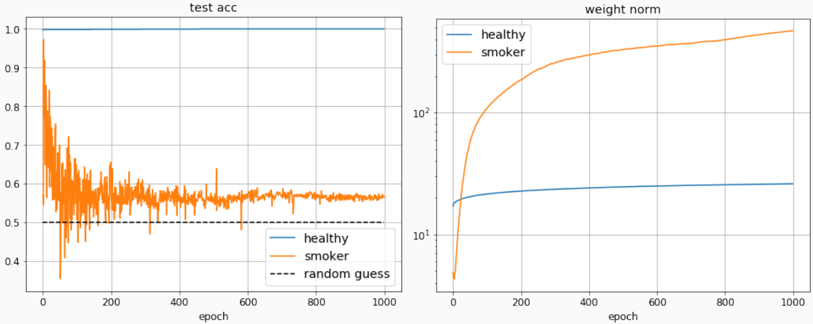

We see here $\Pi(\mathcal{F}; n) = 2^n \Rightarrow \mathrm{VC}(\mathcal{F}) \geq n$. Thus **the bound is vacuous**.

Generally, $\mathrm{VC}(\text{fc-net}) = O(WL \log W)$ (Bartlett et al., 2017a)[1].

[1] https://arxiv.org/abs/1703.02930

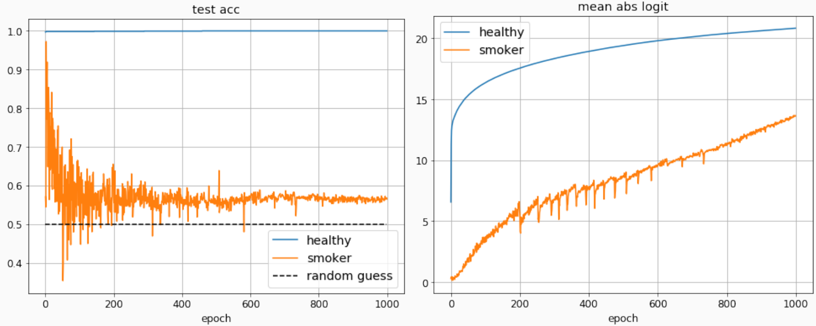
# Worst-case bound: 0/1 loss

**Symptoms:**



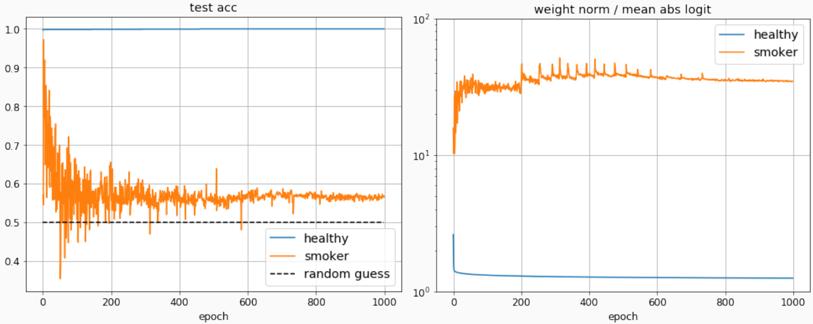
# Worst-case bound: 0/1 loss

**Symptoms:**



# Worst-case bound: 0/1 loss

## Symptoms:



**Worst-case bound:**

$$R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) \leq \sup_{f \in \mathcal{F}} (R(f) - \hat{R}_n(f)).$$

**$\gamma$-margin loss case (Bartlett, 1998)[2]:**

Assume $l(y, f(x)) = l_\gamma(y, f(x)) = [yf(x) < \gamma]$; then

$$\mathbb{P}\left\{ \sup_{f \in \mathcal{F}} (R(f) - \hat{R}_{n,\gamma}(f)) > \epsilon \right\} \leq 2\mathcal{N}_\infty(\pi_\gamma(\mathcal{F}), \gamma/2, 2n) e^{-\epsilon^2 n/8},$$

where

$$\mathcal{N}_\infty(\mathcal{H}, \epsilon, k) = \sup_{S_k} \inf_{\bar{\mathcal{H}} \subset \mathcal{H}} \{ |\bar{\mathcal{H}}| : \ \forall h \in \mathcal{H} \ \exists \bar{h} \in \bar{\mathcal{H}} : \max_{z \in S_k} |h(z) - \bar{h}(z)| < \epsilon \}.$$

---

[2] `https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=661502`

# Worst-case bound: $\gamma$-margin loss

**The bound:**

$$R(\hat{f}_n) \leq \hat{R}_{n,\gamma}(\hat{f}_n) + \sqrt{\frac{8}{n}\left(\log \mathcal{N}_\infty(\pi_\gamma(\mathcal{F}), \gamma/2, 2n) + \log\left(\frac{2}{\delta}\right)\right)} \quad \text{w.p. } \geq 1-\delta.$$

**Bounding growth function:**

$$d := \mathrm{fat}_\gamma(\mathcal{F}) := \max_k \{k : \ \mathcal{F} \text{ shatters some } X_k \text{ with confidence } \gamma\}.$$

It is possible to show:

$\log \mathcal{N}_\infty(\pi_\gamma(\mathcal{F}), \gamma/2, 2n) = O(d \log(n/d) \log n)$ if $n > O(d \log(n/d))$, else $O(n)$.

Compare:

$$\log \Pi(\mathcal{F}, 2n) = O(d_{VC} \log(n/d_{VC})) \quad \text{if } n > O(d_{VC}), \text{ else } 2n.$$

# Comparing dimensions

**Consider the following class of predictors:**

$$\mathcal{F}_A = \left\{ \sum_{j=1}^{m} w_j f_j : \; m \in \mathbb{N}, \; f_j : X \to [-1, 1], \; \sum_{j=1}^{m} |w_j| \le A \right\}.$$

Due to Cybenko theorem (Cybenko, 1989)[3]:

$$\mathrm{VC}(\mathcal{F}_A) = \infty.$$

However, (Bartlett, 1998):

$$\mathrm{fat}_\gamma(\mathcal{F}_A) = O\left( \frac{A^2}{\gamma^2} \log^2(A/\gamma) \right).$$

[3] https://web.archive.org/web/20151010204407/http://deeplearning.cs.cmu.edu/pdfs/Cybenko.pdf

# Worst-case bound: another approach

**Worst-case bound:**

$$R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) \le \sup_{f \in \mathcal{F}} (R(f) - \hat{R}_n(f)).$$

**From McDiarmid's inequality:**

$$\sup_{f \in \mathcal{F}} (R(f) - \hat{R}_n(f)) \le \mathbb{E} \sup_{f \in \mathcal{F}} (R(f) - \hat{R}_n(f)) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} \quad \text{w.p. } \ge 1 - \delta.$$

# Worst-case bound: another approach

$$\sup_{f\in\mathcal{F}}(R(f) - \hat{R}_n(f)) \leq \mathbb{E}\sup_{f\in\mathcal{F}}(R(f) - \hat{R}_n(f)) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}} \quad \text{w.p. } \geq 1-\delta.$$

**Symmetrization:**

$$\mathbb{E}\sup_{f\in\mathcal{F}}(R(f) - \hat{R}_n(f)) \leq 2\mathbb{E}\,\mathrm{Rad}(l \odot \mathcal{F} \mid S_n),$$

where we have introduced **Rademacher complexity**:

$$\mathrm{Rad}(\mathcal{H} \mid S_n) := \mathbb{E}_{\Sigma_n}\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(z_i)\right|.$$

# Worst-case bound: another approach

Need to bound $\mathrm{Rad}(l \odot \mathcal{F} \mid S_n)$.

**Zero-one loss:**

Using Hoeffding's lemma:

$$\mathrm{Rad}(l \odot \mathcal{F} \mid S_n) \leq \sqrt{\frac{2}{n} \log(2\Pi(\mathcal{F}, n))}.$$

The corresponding bound:

$$R(\hat{f}_n) \leq \hat{R}_n(\hat{f}_n) + \sqrt{\frac{8}{n} \log 2\Pi(\mathcal{F}; n)} + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} \quad \text{w.p. } \geq 1 - \delta \text{ over } S_n.$$

# Worst-case bound: another approach

Need to bound $\mathrm{Rad}(l \odot \mathcal{F} \mid S_n)$.

$\gamma$**-margin loss:**
From Dudley's entropy integral:

$$\mathrm{Rad}(\tilde{l}_\gamma \odot \mathcal{F} \mid S_n) \leq \frac{4\epsilon}{\sqrt{n}} + \frac{12}{n} \int_\epsilon^{\sqrt{n}/2} \sqrt{\log \mathcal{N}_2(\tilde{l}_\gamma \odot \mathcal{F}, t, S_n)}\, dt \quad \forall \epsilon > 0,$$

where

$$\mathcal{N}_2(\mathcal{H}, t, S_n) = \inf_{\bar{\mathcal{H}} \subset \mathcal{H}} \left\{ |\bar{\mathcal{H}}| :\ \forall h \in \mathcal{H}\ \exists \bar{h} \in \bar{\mathcal{H}} : \sum_{i=1}^n (h(z_i) - \bar{h}(z_i))^2 < t^2 \right\}.$$

# Worst-case bound: another approach

Now, we need to bound $\log \mathcal{N}_2(\tilde{l}_\gamma \odot \mathcal{F}, t, S_n)$.

Obviously,

$$\log \mathcal{N}_2(\tilde{l}_\gamma \odot \mathcal{F}, t, S_n) \leq \log \mathcal{N}_2(\mathcal{F}(X_n), \gamma t).$$

**Consider multilayer fc-net with weights $\mathcal{A}$:**

$$f_{\mathcal{A}}(x) = a_L^T \sigma(A_{L-1}\sigma(\ldots A_1 x)).$$

Let

$$\mathcal{F}_{s,b} = \{f_{\mathcal{A}} : \ \|A_l\|_2 \leq s_l, \|A_l^T\|_{2,1} \leq b_l\}.$$

Need to bound $\log \mathcal{N}_2(\mathcal{F}_{s,b}(X_n), \gamma t)$.

# Worst-case bound: another approach

Need to bound $\log \mathcal{N}_2(\mathcal{F}_{s,b}(X_n), \gamma t)$.

**Theorem (Bartlett et al., 2017b)[4]:**

$$\log \mathcal{N}_2(\mathcal{F}_{s,b}(X_n), \epsilon) \leq O\left(\frac{\|X\|_F^2}{\epsilon^2} \mathcal{R}_{s,b}^2\right),$$

where we have introduced **spectral complexity**:

$$\mathcal{R}_{s,b} = \left(\prod_{l=1}^{L} s_l\right) \times \left(\sum_{l=1}^{L} (b_l/s_l)^{2/3}\right)^{3/2}.$$

[4] https://arxiv.org/abs/1706.08498

# Worst-case bound: another approach

$$\log \mathcal{N}_2(\mathcal{F}_{s,b}(X_n), \epsilon) \leq O\left(\frac{\|X\|_F^2}{\epsilon^2}\mathcal{R}_{s,b}^2\right).$$

**Plug this into Dudley's integral:**

$$\mathrm{Rad}(\tilde{l}_\gamma \odot \mathcal{F}_{s,b} \mid S_n) \leq \tilde{O}\left(\frac{\|X\|_F}{\gamma n}\mathcal{R}_{s,b}\right).$$

**The bound (Bartlett et al., 2017b):**

$$R(\hat{f}_n) \leq \hat{R}_n(\hat{f}_n) + \tilde{O}\left(\frac{\|X\|_F}{\gamma n}\mathcal{R}_{\mathcal{A}}\right) + 3\sqrt{\frac{1}{2n}\log\frac{2}{\delta}} \quad \text{w.p. } \geq 1-\delta \text{ over } S_n,$$

where

$$\mathcal{R}_{\mathcal{A}} := \mathcal{R}_{s,b} \quad \text{for } s_l = \|A_l\|_2 \text{ and } b_l = \|A_l^T\|_{2,1}.$$

# Failure of worst-case bounds

**All considered bounds are essentially two-sided:**

Compare:

$$R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) \leq 2\mathbb{E}\,\mathrm{Rad}(l \odot \mathcal{F} | S_n) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} \quad \text{w.p. } \geq 1 - \delta \text{ over } S_n$$

and:

$$|R(\hat{f}_n) - \hat{R}_n(\hat{f}_n)| \leq 2\mathbb{E}\,\mathrm{Rad}(l \odot \mathcal{F} | S_n) + \sqrt{\frac{1}{2n} \log \frac{2}{\delta}} \quad \text{w.p. } \geq 1 - \delta \text{ over } S_n.$$

# Failure of worst-case bounds

Recall that $\hat{f}_n = \mathcal{A}(S_n)$ — result of applying learning algorithm $\mathcal{A}$ on dataset $S_n$.

**Generalization error:**

$$\epsilon_{gen}(n, \delta) = \min\left\{\epsilon :\ \mathbb{P}_{S_n}\left(R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) > \epsilon\right) < \delta\right\}.$$

**Best two-sided uniform bound:**

$$\epsilon_{unif}(n, \delta) = \min\left\{\epsilon :\ \mathbb{P}_{S_n}\left(\sup_{f \in \mathcal{F}_\mathcal{A}} |R(f) - \hat{R}_n(f)| > \epsilon\right) < \delta\right\},$$

where $\mathcal{F}_\mathcal{A} = \{\mathcal{A}(S_n)\}_{S_n \subset (X \times Y)^n}$.

# Failure of worst-case bounds

Let $\exists \tau_n \geq 0 : \ \forall S_n \quad \hat{R}_n(\hat{f}_n) \leq \tau_n$.

**Consider dataset negation procedure:**

$$S_n' := \mathrm{neg}(S_n)$$

with following properties:

1. $\forall S_n \quad (S_n')' = S_n$;
2. $\exists \tau_n' \geq 0 : \ \forall S_n \quad \hat{R}_n'(\hat{f}_n) \geq 1 - \tau_n'$;
3. $\forall \mathcal{S}_n \subset (X \times Y)^n \quad \mathbb{P}(\mathcal{S}_n) = \mathbb{P}(\mathcal{S}_n')$.

# Failure of worst-case bounds

**Theorem (Nagarajan & Kolter, 2019)[5]:**
Suppose dataset negation procedure exists.
Let $\delta \in (0, 1/2)$. Then

$$\epsilon_{unif}(n, \delta) \geq 1 - \tau_n - \tau'_n - \epsilon_{gen}(n, \delta).$$

[5] `https://arxiv.org/abs/1902.04742`

# Failure of worst-case bounds

**Dataset-dependent bound:**

$$\epsilon_{\text{data-dep}}(n, \delta) = \min \left\{ \epsilon : \ \mathbb{P}_{S_n} \left( \sup_{f \in \mathcal{F}(S_n)} |R(\hat{f}_n) - \hat{R}(\hat{f}_n)| > \epsilon \right) < \delta \right\}.$$

The same argument holds if

$$\forall S_n \quad \hat{f}'_n \in \mathcal{F}(S_n).$$