PAC-bayesian 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: $\hat{f}_n = \mathcal{A}(S_n) \in \mathcal{F}$.

Our goal is to bound the risk difference:

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

PAC-bayesian bounds

Bounds for deterministic $\mathcal{A}$:

- **Finite $\mathcal{F}$:**

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

- **At most countable $\mathcal{F}$ (McAllester, 1998)**[1]**:**

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

where P is a distribution over $\mathcal{F}$ (**prior**).

[1] Preliminary theorem 2 in `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.1745&rep=rep1&type=pdf`

PAC-bayesian bounds

Consider stochastic learning algorithm: $\hat{f}_n = \mathcal{A}(S_n) \sim Q|S_n$.
Define $R(Q) := \mathbb{E}_{f \sim Q} R(f), \quad \hat{R}_n(Q) := \mathbb{E}_{f \sim Q} \hat{R}_n(f)$.

Corresponding bound:

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

PAC-bayesian bound (McAllester, 1999)[2]:

$$R(Q|S_n) - \hat{R}_n(Q|S_n) \leq \sqrt{\frac{1}{2n-1}\left(\log\frac{4n}{\delta} + KL(Q|S_n \parallel P)\right)} \quad \text{w.p. } \geq 1 - \delta$$

for any distribution P on $\mathcal{F}$.

[2] Theorem 2 in http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.1908&rep=rep1&type=pdf

PAC-bayesian bounds

Define: $\Delta_n(f) := |R(f) - \hat{R}_n(f)|$.

Lemma (McAllester, 1999)[3]**:**

$$\mathbb{E}_{f\sim P} e^{(2n-1)\Delta_n(f)^2} \leq \frac{4n}{\delta} \quad \text{w.p. } \geq 1-\delta \text{ over } S_n$$

for any distribution P on $\mathcal{F}$.

Lemma (Donsker & Varadhan):
Let P and Q be distributions on X. Then:

$$KL(P \,\|\, Q) = \sup_{h\colon X \to \mathbb{R}} \left(\mathbb{E}_{x\sim P} h(x) - \log \mathbb{E}_{x\sim Q} e^{h(x)} \right).$$

[3] Lemma 17 in `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.1908&rep=rep1&type=pdf`

PAC-bayesian bounds

Lemma (Langford & Seeger, 2001)[4]:

$$\mathbb{E}_{f\sim P} e^{(n-1)KL(\hat{R}_n(f)\,\|\,R(f))} \le \frac{2n}{\delta} \quad \text{w.p. } \ge 1-\delta \text{ over } S_n$$

for any distribution P on $\mathcal{F}$.

Theorem (Langford & Seeger, 2001)[5]:

$$KL(\hat{R}_n(Q|S_n)\,\|\,R(Q|S_n)) \le \frac{1}{n-1}\left(\log\frac{2n}{\delta} + KL(Q|S_n\,\|\,P)\right) \quad \text{w.p. } \ge 1-\delta$$

for any distribution P on $\mathcal{F}$.

[4] Lemma 2 in `http://hunch.net/~jl/projects/prediction_bounds/averaging/averaging_tech.pdf`

[5] Theorem 3 there.

PAC-bayesian bounds

Let $X_{1:n}$ be i.i.d., $X_i \sim \mathcal{B}(p)\ \forall i$.

Hoeffding's inequality:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p+\epsilon\right) \leq e^{-2n\epsilon^2}; \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq p-\epsilon\right) \leq e^{-2n\epsilon^2}.$$

Chernoff-Hoeffding's inequality:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p+\epsilon\right) \leq e^{-nKL(p+\epsilon \,\|\, p)};$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq p-\epsilon\right) \leq e^{-nKL(p-\epsilon \,\|\, p)}.$$

PAC-bayesian bounds

PAC-bayesian bound (McAllester, 1999):

$$R(Q|S_n) - \hat{R}_n(Q|S_n) \leq \sqrt{\frac{1}{2n-1}\left(\log\frac{4n}{\delta} + KL(Q|S_n \,\|\, P)\right)} \quad \text{w.p. } \geq 1-\delta$$

for any distribution P on $\mathcal{F}$.

- **Pros:** Depends on learned predictor $\hat{f}_n$.
- **Cons:** Vacuous if $P(A) = 0 \not\Rightarrow Q(A) = 0$. For example, if $P(\{\hat{f}_n\}) = 0$ for $Q|S_n = \delta_{\hat{f}_n}$ we have $KL(Q|S_n \,\|\, P) = +\infty$.

PAC-bayesian bounds

Let $\mathcal{F}$ be a set of neural nets of a given architecture.
Denote $f_{\mathbf{w}} \in \mathcal{F}$ a neural net with weights $\mathbf{w} \in \mathcal{W}$.
Consider deterministic learning algorithm $\hat{\mathbf{w}}_n = \mathcal{A}(S_n)$. Then $\hat{f}_n = f_{\hat{\mathbf{w}}_n}$.

PAC-bayesian bound:
Take any distribution P on $\mathcal{W}$. Then, $\forall \delta \in (0,1)$ w.p. $\geq 1-\delta$ over dataset S_n for any distribution $Q|S_n$ on $\mathcal{W}$

$$R(Q|S_n) - \hat{R}_n(Q|S_n) \leq \sqrt{\frac{1}{2n-1}\left(\log\frac{4n}{\delta} + KL(Q|S_n \,\|\, P)\right)}.$$

If we take $Q|S_n = \delta_{\hat{\mathbf{w}}_n}$ and $P:\ \forall \mathbf{w}\ P(\{\mathbf{w}\}) = 0$, we get $KL(Q|S_n \,\|\, P) = \infty$.

PAC-bayesian bounds

If we take $Q|S_n = \delta_{\hat{\mathbf{w}}_n}$ and $P : \forall \mathbf{w}\ P(\{\mathbf{w}\}) = 0$, we get $KL(Q|S_n \,\|\, P) = \infty$.

Ways to deal with it:

- **Stochastization (Dziugaite & Roy, 2017)**[6]:
  With prob. $\geq 1 - \delta$ over S_n for any $Q|S_n$:

  $$R(Q|S_n) \leq \hat{R}_n(Q|S_n) + \mathrm{bound}(KL(Q|S_n \,\|\, P), n, \delta).$$

  Minimize RHS over Q inside some class $\mathcal{Q}$:

  $$\mathrm{RHS} := \hat{R}_n(Q|S_n) + \mathrm{bound}(KL(Q|S_n \,\|\, P), n, \delta) \to \min_{Q \in \mathcal{Q}}.$$

[6] https://arxiv.org/abs/1703.11008

PAC-bayesian bounds

Replace risk R with its differentiable convex surrogate $\mathcal{L}$:

$$\mathrm{RHS} \leq \mathrm{RHS}' := \hat{\mathcal{L}}_n(Q|S_n) + \mathrm{bound}(KL(Q|S_n \parallel P), n, \delta) \to \min_{Q \in \mathcal{Q}}.$$

Instantiating $\mathcal{Q}$ and P:
Take $\mathcal{Q} = \{\mathcal{N}(\mathbf{w}, \operatorname{diag} \exp \mathbf{u}),\ \mathbf{w}, \mathbf{u} \in \mathcal{W}\}$ and $P = \mathcal{N}(\mathbf{w}_*, \exp u_* I)$.
Then,

$$\hat{\mathcal{L}}_n(Q) = \mathbb{E}_{\xi \sim \mathcal{N}(0,I)} \hat{\mathcal{L}}_n(\mathbf{w} + \xi \odot \exp \mathbf{u});$$

$$KL(Q \parallel P) = \frac{1}{2}\left(\frac{1}{\exp u_*}\left(\|\exp \mathbf{u}\|_1 + \|\mathbf{w} - \mathbf{w}_*\|_2^2\right) + \dim \mathcal{W}\,(u_* - 1) - 1 \cdot \mathbf{u}\right).$$

Hence we can optimize RHS' over w and u via GD.
Start optimization from $\mathbf{w}^{(0)} = \hat{\mathbf{w}}_n$, $\mathbf{u}^{(0)} \ll -1$.

PAC-bayesian bounds

$$KL(Q \,\|\, P) = \frac{1}{2}\left(\frac{1}{\exp u_*}\left(\|\exp \mathbf{u}\|_1 + \|\mathbf{w} - \mathbf{w}_*\|_2^2\right) + \dim \mathcal{W}\left(u_* - 1\right) - 1 \cdot \mathbf{u}\right).$$

Choosing $\mathbf{w}_*$:

Let $\mathcal{A}(\cdot)$ be GD starting from $\mathbf{w}_{init}$.

Take $\mathbf{w}_* = \mathbf{w}_{init}$. Then, bound depends on $\|\mathbf{w} - \mathbf{w}_{init}\|_2$.

Choosing u_*:
Define $u_{*,j} = \log c - j/b$, where $c, b > 0$, $j \in \mathbb{N}$. Take $\delta_j = \frac{6\delta}{\pi^2 j^2}$. Then, w.p. $\geq 1 - \delta$ over S_n for any $j \in \mathbb{N}$ and Q:

$$R(Q) \leq \hat{\mathcal{L}}_n(Q) + \sqrt{\frac{KL(Q \,\|\, \mathcal{N}(\mathbf{w}_*, u_{*,j} I)) + \log(4n) - \log \delta_j}{2n-1}}.$$

Equivalently, w.p. $\geq 1 - \delta$ over S_n **for any** u_* **from a set**, and any Q:

$$R(Q) \leq \hat{\mathcal{L}}_n(Q) + \sqrt{\frac{KL(Q \,\|\, \mathcal{N}(\mathbf{w}_*, u_* I)) + \log \frac{2\pi^2 b^2 n}{3\delta} + \log(\log c - u_*)^2}{2n-1}}.$$

We can optimize RHS' over u_*.

PAC-bayesian bounds

If we take $Q|S_n = \delta_{\hat{\mathbf{w}}_n}$ and $P: \forall \mathbf{w}\ P(\{\mathbf{w}\}) = 0$, we get $KL(Q|S_n \,\|\, P) = \infty$.

Ways to deal with it:

- **Compression & coding (Zhou et al., 2019)**[7]:
  Let $|\mathbf{w}|_c$ — number of bits required to encode $\mathbf{w}$ with coding c.
  Coding-based prior:

$$P_c(\mathbf{w}) = \frac{1}{Z} m(|\mathbf{w}|_c) 2^{-|\mathbf{w}|_c},$$

  where $m(\cdot)$ — some probability measure on $\mathbb{Z}$. Then,

$$KL(\delta_{\hat{\mathbf{w}}_n} \,\|\, P_c) \leq |\hat{\mathbf{w}}_n|_c \log 2 - \log(m(|\hat{\mathbf{w}}_n|_c)).$$

  Need to make $|\hat{\mathbf{w}}_n|_c$ small.

[7] https://openreview.net/forum?id=BJgqqsAct7

PAC-bayesian bounds

Compressing $\hat{\mathbf{w}}_n$:

$$(S, Q, C) := \mathrm{Compress}(\mathbf{w}),$$

where

- $S = \{s_1, \ldots, s_k\} \subset \{1, \ldots, \dim \mathcal{W}\}$ — location of non-zero weights,
- $C = \{c_1, \ldots, c_r\} \subset \mathbb{R}$ — a codebook,
- $Q = \{q_1, \ldots, q_k\}$, $q_i \in \{1, \ldots, r\}$ — quantized values.

Then, compressed weights $\tilde{\mathbf{w}}$ will be:

$$\tilde{\mathbf{w}}_i = c_{q_j} \quad \text{if } i = s_j \text{ else } 0.$$

Hence

$$|\mathrm{Compress}(\hat{\mathbf{w}}_n)|_c = |S|_c + |Q|_c + |C|_c \leq k(\log \dim \mathcal{W} + \log r) + 32r.$$

PAC-bayesian bounds

Good generalization bound if:

1. Solutions found by $\mathcal{A}$ are well-compressible, i.e.

$$|\mathrm{Compress}(\hat{\mathbf{w}}_n)|_c \text{ is small;}$$

2. Compression doesn't lead to performance degradation, i.e.

$$R(\tilde{\hat{\mathbf{w}}}_n) \approx R(\hat{\mathbf{w}}_n).$$

PAC-bayesian bounds

Let $R_\gamma(f) = \mathbb{E}_{x,y\sim\mathcal{D}}[yf(x) < \gamma]$ — γ-margin risk.

PAC-bayesian bound:
Take any distribution P on $\mathcal{W}$. Then, $\forall\delta \in (0,1)$ w.p. $\geq 1-\delta$ over dataset S_n for any $\mathbf{w} \in \mathcal{W}$ and any RV $\mathbf{u}$ on $\mathcal{W}$

$$\mathbb{E}_{\mathbf{u}} R_0(f_{\mathbf{w}+\mathbf{u}}) \leq \mathbb{E}_{\mathbf{u}} \hat{R}_{n,0}(f_{\mathbf{w}+\mathbf{u}}) + \sqrt{\frac{KL(\mathbf{w}+\mathbf{u} \,\|\, P) + \log\frac{4n}{\delta}}{2n-1}}.$$

PAC-bayesian bounds

Let $R_\gamma(f) = \mathbb{E}_{x,y\sim\mathcal{D}}[yf(x) < \gamma]$ — γ-margin risk.

Lemma 1 (Neyshabur et al., 2018)[8]:
Take any distribution P on $\mathcal{W}$. Then, $\forall\delta \in (0,1), \gamma > 0$ w.p. $\geq 1-\delta$ over dataset S_n for any $\mathbf{w} \in \mathcal{W}$ and any RV $\mathbf{u}$ on $\mathcal{W}$ s.t.

$$\mathbb{P}_u\left(\max_x |f_{\mathbf{w}+\mathbf{u}}(x) - f_{\mathbf{w}}(x)| < \gamma/2\right) \geq 1/2$$

the following holds:

$$R_0(f_{\mathbf{w}}) \leq \hat{R}_{n,\gamma}(f_{\mathbf{w}}) + \sqrt{\frac{2KL(\mathbf{w}+\mathbf{u} \,\|\, P) + \log\frac{16n}{\delta}}{2n-1}}.$$

[8] `https://openreview.net/forum?id=Skz_WfbCZ`

PAC-bayesian bounds

Let $\mathbf{w} = \{W_l\}_{l=1}^L$, and

$$f_{\mathbf{w}}(x) = W_L \sigma(W_{L-1} \dots \sigma(W_1 x)),$$

where $\sigma(z) = [z]_+$. Define $\mathcal{X}_B := \{x : \ \|x\|_2 < B\}$.

Lemma 2 (Neyshabur et al., 2018):
$\forall B > 0$, $x \in \mathcal{X}_B$, $\mathbf{w} \in \mathcal{W}$, for any perturbation $\mathbf{u} = \{U_l\}_{l=1}^L$ s.t. $\|U_l\|_2 \leq \frac{1}{L}\|W_l\|_2$ the following holds:

$$|f_{\mathbf{w}+\mathbf{u}}(x) - f_{\mathbf{w}}(x)| \leq eB \left(\prod_{l=1}^{L} \|W_l\|_2 \right) \sum_{l=1}^{L} \frac{\|U_l\|_2}{\|W_l\|_2}.$$

PAC-bayesian bounds

Let $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$. Define $d := \max_l d_l$ — maximal width.

Theorem (Neyshabur et al., 2018):
Assume $X_n \in \mathcal{X}_B$ a.s. for some $B > 0$. Then $\forall \delta \in (0,1), \gamma > 0$ w.p. $\geq 1 - \delta$ over dataset S_n for any $\mathbf{w} \in \mathcal{W}$

$$R_0(f_{\mathbf{w}}) \leq \hat{R}_{n,\gamma}(f_{\mathbf{w}}) + \\ + O\left(\sqrt{\frac{B^2 L^2 d \log(Ld) \prod_{l=1}^{L} \|W_l\|_2^2 \sum_{l=1}^{L} \frac{\|W_l\|_F^2}{\|W_l\|_2^2} + \gamma^2 \log \frac{Ln}{\delta}}{\gamma^2 n}}\right).$$

Compression-based bounds

Let $\mathcal{F}, \mathcal{G} \in \mathbb{R}^{\mathcal{X}}$ be sets of predictors on $\mathcal{X}$.

Definitions:

- Let $\hat{f}_n = \mathcal{A}(S_n) \in \mathcal{F}$ — predictor learned on dataset S_n.
- Let $X \subset \mathcal{X}$. f is (γ, X)-compressible via $\mathcal{G}$ if $\exists g \in \mathcal{G}$:

  $$|f(x) - g(x)| \leq \gamma \quad \forall x \in X.$$

  We say "f is (γ, X)-compressible with g".
- For $g \in \mathcal{G}$ let $|g|_c$ be code length of g wrt coding c.

Compression-based bounds

Lemma 1:
Let $p(z)$ be pdf on $\mathbb{N}$. Let $\hat{f}_n$ be (γ, X_n)-compressible with $\hat{g}_n \in \mathcal{G}$ w.p. $\geq 1 - \zeta$ over S_n. Then $\forall \delta \in (0,1)$ w.p. $\geq 1 - \zeta - \delta$ over S_n

$$R_0(\hat{g}_n) \leq \hat{R}_{n,\gamma}(\hat{f}_n) + \sqrt{\frac{|\hat{g}_n|_c \log 2 - \log p(|\hat{g}_n|_c) - \log \delta}{2n}}.$$

Corollary:
Let $p(z)$ be pdf on $\mathbb{N}$. Assume $X_n \in \mathcal{X}_B$ a.s. for some $B > 0$. Let $\hat{f}_n$ be $(\gamma, \mathcal{X}_B)$-compressible with $\hat{g}_n \in \mathcal{G}$ a.s. over S_n. Then $\forall \delta \in (0,1)$ w.p. $\geq 1 - \delta$ over S_n

$$R_0(\hat{f}_n) \leq \hat{R}_{n,2\gamma}(\hat{f}_n) + \sqrt{\frac{|\hat{g}_n|_c \log 2 - \log p(|\hat{g}_n|_c) - \log \delta}{2n}}.$$

Compression-based bounds

Instantiating the bound:
Let $\mathcal{F} = \{f_{\mathbf{w}},\ \mathbf{w} \in \mathbb{R}^m\}$.

1. **Discretize weights of $\mathcal{F}$:**
   - Consider only weights with $\|\mathbf{w}\|_\infty \leq w_{max}$.
   - Let $\mathcal{G} = \{f_{\mathbf{w}},\ \mathbf{w} \in A_K^m\}$, where $A_K = \{w_{max}k/K,\ k = -K, \ldots, K\}$.
   - **Proposition 1:** For sufficiently large K $\hat{f}_n$ with $\|\hat{\mathbf{w}}_n\|_\infty \leq w_{max}$ is $(\gamma, \mathcal{X}_B)$-compressible via $\mathcal{G}$ a.s. over S_n.

   Compute code length:

   $$|\hat{g}_n|_c = m \log_2(2K+1).$$

   $|\hat{g}_n|_c \geq m \Rightarrow$ the bound is vacuous.

Compression-based bounds

Instantiating the bound:
Let $\mathbf{w} = \mathrm{vec}(\{W_l\}_{l=1}^L) \in \mathbb{R}^m$, where $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$, and

$$f_{\mathbf{w}}(x) = W_L \sigma(W_{L-1} \ldots \sigma(W_1 x)),$$

where $\sigma(z) = [z]_+$. Define $d = \max_l d_l$.

1. **Reparameterize weights of $\mathcal{F}$:**
   - Substitute W_l with $U_l V_l$ for $U_l \in \mathbb{R}^{d_l \times r_l}$, $V_l \in \mathbb{R}^{r_l \times d_{l-1}}$, $r_l = \mathrm{rk}\, W_l$.
   - Define $\mathcal{F}' = \cup_{r_{1:L}=1}^{d} \{f_{\mathbf{u} \times \mathbf{v}},\ \mathbf{u} = \mathrm{vec}(\{U_l\}_{l=1}^L), \mathbf{v} = \mathrm{vec}(\{V_l\}_{l=1}^L)\}$.
2. **Discretize weights of $\mathcal{F}'$:**
   - Let $\mathcal{G} = \{f_{\mathbf{u} \times \mathbf{v}},\ \mathbf{u} \in A_K^{m_u}, \mathbf{v} \in A_K^{m_v}\}$.
   - **Proposition 1':** For sufficiently large K $\hat{f}_n = f_{\hat{\mathbf{u}}_n \times \hat{\mathbf{v}}_n}$ with $\|\hat{\mathbf{w}}_n\|_\infty \le O(w_{max})$ is $(\gamma, \mathcal{X}_B)$-compressible via $\mathcal{G}$ a.s. over S_n.

$$|\hat{g}_n|_c \le L \log_2 d + 2d \sum_{l=1}^{L} \hat{r}_{n,l} \log_2(2K+1).$$

Non-vacuous if $\hat{r}_{n,l} \ll d$.

Compression-based bounds

Instantiating the bound (Arora et al., 2018)[9]:

1. **Compress weights of $\mathcal{F}$:**
   - Define W_l^α as W_l with sing. values $< \alpha \|W_l\|_2$ substituted with zero.
   - **Proposition 2:** For sufficiently small α $\hat{f}_n = f_{\hat{\mathbf{w}}_n}$ is $(\gamma, \mathcal{X}_B)$-compressible with $\hat{f}_n^\alpha = f_{\hat{\mathbf{w}}_n^\alpha}$ a.s. over S_n.
   - Denote $\hat{r}_{n,l}^\alpha = \mathrm{rk}\, \hat{W}_{n,l}^\alpha$.
2. **Reparameterize weights of $\mathcal{F}$:**
   - Define $\mathcal{F}' = \cup_{r_{1:L}=1}^d \{f_{\mathbf{u}\times\mathbf{v}},\ \mathbf{u} = \mathrm{vec}(\{U_l\}_{l=1}^L), \mathbf{v} = \mathrm{vec}(\{V_l\}_{l=1}^L)\}$.
3. **Discretize weights of $\mathcal{F}'$.** Compute code length:

$$|\hat{g}_n|_c \leq L \log_2 d + 2d \sum_{l=1}^{L} \hat{r}_{n,l}^\alpha \log_2(2K+1).$$

[9] http://proceedings.mlr.press/v80/arora18b.html

Compression-based bounds

Compress weights of $\mathcal{F}$:

- Define W_l^α as W_l with sing. values $< \alpha \| W_l \|_2$ substituted with zero.
- **Lemma 2 (Arora et al., 2018)**[10]**:**

$$\| W_l^\alpha - W_l \|_2 \leq \alpha \| W_l \|_2, \qquad \operatorname{rk} W_l^\alpha \leq \frac{\| W_l \|_F^2}{\alpha^2 \| W_l \|_2^2}.$$

- **Proposition 2:** For $\alpha = \gamma (eBL \prod_{l=1}^L \| \hat{W}_{n,l} \|_2)^{-1}$ $\hat{f}_n = f_{\hat{\mathbf{w}}_n}$ is $(\gamma, \mathcal{X}_B)$-compressible with $\hat{f}_n^\alpha = f_{\hat{\mathbf{w}}_n^\alpha}$ a.s. over S_n.

$$\hat{r}_{n,l}^\alpha = \operatorname{rk} \hat{W}_{n,l}^\alpha \leq e^2 B^2 L^2 \gamma^{-2} \left(\prod_{l=1}^L \| \hat{W}_{n,l} \|_2^2 \right) \frac{\| W_l \|_F^2}{\| W_l \|_2^2}.$$

[10] Lemma 1 in `http://proceedings.mlr.press/v80/arora18b.html`

Compression-based bounds

Discrete weights of $\mathcal{F}$:

- Consider only weights with $\|\mathbf{u}\|_\infty \leq w_{max}$ and $\|\mathbf{v}\|_\infty \leq w_{max}$.
- Let $\mathcal{G} = \{f_{\mathbf{u}\times\mathbf{v}},\ \mathbf{u} \in A_K^{m_u}, \mathbf{v} \in A_K^{m_v}\}$.
- **Proposition 1':** For sufficiently large K $\hat{f}_n = f_{\hat{\mathbf{u}}_n \times \hat{\mathbf{v}}_n}$ with $\|\hat{\mathbf{w}}_n\|_\infty \leq O(w_{max})$ is $(\gamma, \mathcal{X}_B)$-compressible via $\mathcal{G}$ a.s. over S_n.

Compute code length:

$$|\hat{g}_n|_c \leq L \log_2 d + 2d \sum_{l=1}^{L} \hat{r}_{n,l} \log_2(2K+1) =$$

$$= L \log_2 d + 2de^2 B^2 L^2 \gamma^{-2} \log_2(2K+1) \left(\prod_{l=1}^{L} \|\hat{W}_{n,l}\|_2^2 \right) \sum_{l=1}^{L} \frac{\|W_l\|_F^2}{\|W_l\|_2^2} =$$

$$= O\left(dB^2 L^2 \gamma^{-2} \log_2(2K+1) \left(\prod_{l=1}^{L} \|\hat{W}_{n,l}\|_2^2 \right) \sum_{l=1}^{L} \frac{\|W_l\|_F^2}{\|W_l\|_2^2} \right).$$