Why do neural nets learn and generalize?

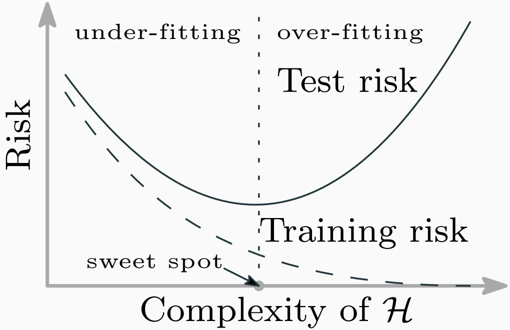
Eugene Golikov

October 4, 2019

Neural Systems and Deep Learning Lab., MIPT

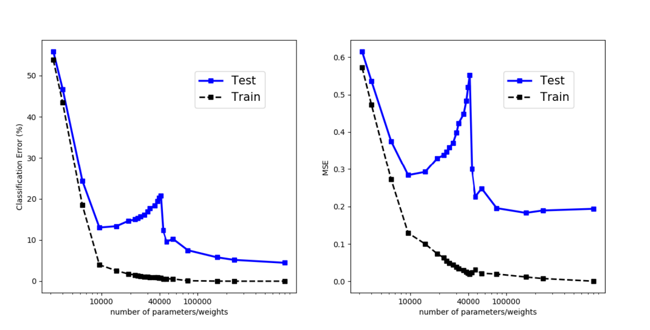
Complexity-risk curve

Classic "bias-variance trade-off" curve:



Complexity-risk curve

Extended curve for neural networks:

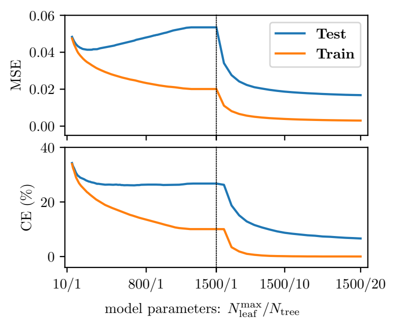


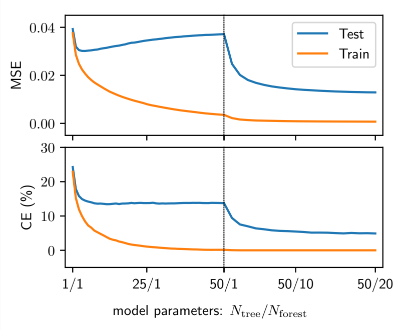
The figure is borrowed from Belkin et al. (2018)[1].

[1] https://arxiv.org/abs/1812.11118

Complexity-risk curve

Similar curves for random forest and boosting:



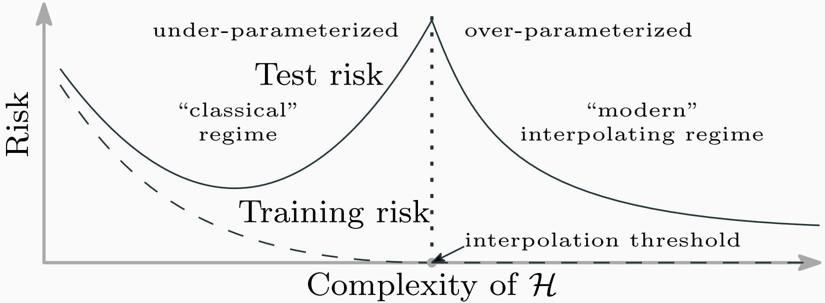


Left: random forest, right: boosting on decision trees.

Figures are borrowed from Belkin et al. (2018).

Complexity-risk curve

General "double descent" curve:



The figure is borrowed from Belkin et al. (2018).

Preliminaries

Learning objective:

$$\hat{\mathcal{L}}_n(W) = \mathbb{E}_{x,y \in S_n} \ell(y, f(x; W)) \to \min_W,$$

where

- $S_n = \{x_i, y_i\}_{i=1}^n \sim \mathcal{D}^n$ — dataset of size n;
- $f(x; W) = W_L \sigma(W_{L-1} \ldots \sigma(W_1 x))$ — neural network with weights $W = W_{1:L}$ and non-linearity σ;
- $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$, where d_l — width of l-th layer;
- $\ell(y, \hat{y})$ — convex loss (typically, square).

Learning procedure:

$$W^{(k+1)} = W^{(k)} - \eta \,(\mathrm{stoch})\mathrm{grad}\, \hat{\mathcal{L}}_n(W^{(k)}).$$

Loss landscape analysis

- **Observation:** SGD achieves zero training risk.
- **Natural hypothesis:** All local minima of $\hat{\mathcal{L}}_n$ are global.
- **Motivation:**
  - **Lee et al. (2016)[2]:** If $\hat{\mathcal{L}}_n \in C_l^{2,1}$ GD with $\eta < l^{-1}$ doesn't converge to a strict saddle (or maximum) a.s. wrt random initialization.

[2] https://arxiv.org/abs/1602.04915

Loss landscape analysis

Hypothesis: All local minima of $\hat{\mathcal{L}}_n$ are global.

Cases:

- **Linear regression (square loss and $L = 1$):**
  convex problem $\Rightarrow$ trivial.
- **Deep linear regression (square loss and $\sigma = \mathrm{id}$):**

$$\hat{\mathcal{L}}_{n,deep}(W_{1:L}) = \mathbb{E}_{x,y \in S_n} \ell(y, W_L W_{L-1} \dots W_1 x) \to \min_{W_{1:L}}$$

  is equivalent to:

$$\hat{\mathcal{L}}_{n,shallow}(R) = \mathbb{E}_{x,y \in S_n} \ell(y, Rx) \to \min_{R:\ \mathrm{rk}\, R \le \min d_l}.$$

Loss landscape analysis

Lu & Kawaguchi (2017)[3]:

Theorem 1:
If $W_{1:L}$ is a local minimum of $\hat{\mathcal{L}}_{n,deep}$, than $R = W_L \dots W_1$ is a local minimum of $\hat{\mathcal{L}}_{n,shallow}$.

Theorem 2:
Every local minimum of $\hat{\mathcal{L}}_{n,shallow}$ is global.

Corollary:
Every local minimum of $\hat{\mathcal{L}}_{n,deep}$ is global.

Almost the same result was obtained earlier in Kawaguchi (2016)[4].

[3] `https://arxiv.org/abs/1702.08580`
[4] `http://www.mit.edu/~kawaguch/publications/kawaguchi-nips16.pdf`

Loss landscape analysis

Lu & Kawaguchi (2017):

Theorem 1:
If $W_{1:L}$ is a local minimum of $\hat{\mathcal{L}}_{n,deep}(W_{1:L})$, than $R = W_L \ldots W_1$ is a local minimum of $\hat{\mathcal{L}}_{n,shallow}(R)$.

Proof outline:

- R is a local minimum of $\hat{\mathcal{L}}_{n,shallow} \Leftrightarrow$
  $\forall \delta R \quad \hat{\mathcal{L}}_{n,shallow}(R) \leq \hat{\mathcal{L}}_{n,shallow}(R + \delta R)$;
- $W_{1:L}$ is a local minimum of $\hat{\mathcal{L}}_{n,deep} \Leftrightarrow$
  $\forall \delta W_{1:L} \quad \hat{\mathcal{L}}_{n,deep}(W_{1:L}) \leq \hat{\mathcal{L}}_{n,deep}(W_1 + \delta W_1 \ldots W_L + \delta W_L)$.
- Need to prove that
  $\forall \delta R \quad \exists \delta W_{1:L} : \; R + \delta R = (W_L + \delta W_L) \ldots (W_1 + \delta W_1)$.

Loss landscape analysis

Hypothesis: All local minima of $\hat{\mathcal{L}}_n$ are global.

Cases:

- **Shallow non-linear regression (square loss and $L = 2$):**
  Theorem (Yu & Chen, 1995[5]):
  If $d_1 \geq n$ and σ is analytic, then all local minima of $\hat{\mathcal{L}}_n$ are global.
- **Deep non-linear regression (square loss and $L \geq 2$):**
  Theorem (Nguyen & Hein, 2017[6]):
  If $\exists l: \; d_l \geq n, \; d_{l+1} \geq \ldots \geq d_L$ and σ is analytic, then all (non-degenerate) local minima of $\hat{\mathcal{L}}_n$ are global.

---

[5] https://ieeexplore.ieee.org/document/410380/
[6] https://arxiv.org/abs/1704.08045

Loss landscape analysis

Theorem (Yu & Chen, 1995):
If $d_1 \geq n$ and σ is analytic, then all local minima of $\hat{\mathcal{L}}_n$ are global.

Proof outline:

- Let $W_{1,2}$ be a local minimum of $\hat{\mathcal{L}}_n$.
- Let $Z = [z_1 \dots z_n] \in \mathbb{R}^{d_1 \times n}$, where $z_i = \sigma(W_1 x_i)$; then $f(x_i; W_{1,2}) = W_2 z_i$.

Lemma: If σ is analytic and $d_1 \geq n$, then the set $\{W_1 : \operatorname{rk} Z < n\}$ has Lebesgue measure zero.

- If $\operatorname{rk} Z = n$, then $\hat{\mathcal{L}}_n(W_{1,2}) = 0$.
- If $\operatorname{rk} Z < n$ and $\hat{\mathcal{L}}_n(W_{1,2}) > 0$, then $\hat{\mathcal{L}}_n(W_{1,2})$ is unstable wrt gradient flow dynamics on W_2: **contradiction**.

Loss landscape analysis

Problem of cross-entropy loss:
$\hat{\mathcal{L}}_n$ can have no minima in $\mathcal{W} = \mathbb{R}^{\dim W}$.

Define:

- **Sublevel set:** $\hat{\mathcal{L}}_n^{-1}((-\infty, \alpha)) \subset \mathcal{W}$;
- **Local valley:** connected component of a sublevel set;
- **Bad local valley:** local valley for which $\inf_{W \in \text{valley}} \hat{\mathcal{L}}_n > \inf_{W \in \mathcal{W}} \hat{\mathcal{L}}_n$.

Loss landscape analysis

More on deep non-linear case:
Let $\ell(y, \cdot)$ be any convex loss, $\sigma(\cdot) \nearrow$, and $\sigma(\mathbb{R}) = \mathbb{R}$.

Theorems (Nguyen, 2019[7]):

1. If $\exists l : d_l \geq n, d_{l+1} > \ldots > d_L$, then $\hat{\mathcal{L}}_n$ has no bad local valleys;
2. If $d_1 \geq 2n$ and $d_2 > \ldots > d_L$, then all sublevel sets of $\hat{\mathcal{L}}_n$ are connected.

Empirical results (Garipov et al., 2018, Draxler et al., 2018[8]):
For realistic networks $\operatorname{Arg\,min} \hat{\mathcal{L}}_n$ is connected.

[7] http://proceedings.mlr.press/v97/nguyen19a/nguyen19a.pdf
[8] https://arxiv.org/abs/1803.00885, https://arxiv.org/abs/1802.10026

Learning dynamics

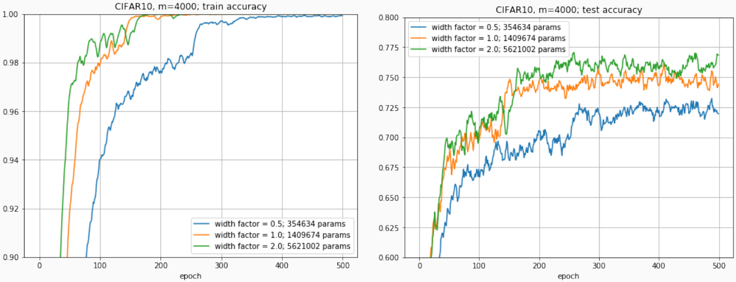


Figure 1: Results of training Conv-small of Miyato et al. (2017)[10] on subset of 4000 samples of CIFAR10. Initial numbers of filters in convolutional layers were multiplied by width factor.

[9] https://arxiv.org/abs/1704.03976
[10] https://arxiv.org/abs/1704.03976

Learning dynamics

- **Observation:** optimization becomes easier as number of parameters grows.
- **Hypothesis:** $\mathcal{L}(W^{(k)}) \leq (1-\beta)^k \mathcal{L}(W^{(0)})$ whp over initialization $W^{(0)}$, and β grows with $\dim W$.
- **Problems:** In general, more params $\Rightarrow$ harder to optimize:
  - **Theorem (Jin et al., 2017)[11]:**
    Suppose $\mathcal{L} \in C^{2,2}(\mathbb{R}^{\dim W})$ — general function to minimize.
    Then $\forall\, \epsilon > 0$ for appropriate choice of hyperparameters (perturbed) GD achieves an ϵ-2nd-order stationary point of $\mathcal{L}$ in

$$K_\epsilon = O(\log^4(\dim W)/\epsilon^2) \quad \text{iterations whp.}$$

[11] https://arxiv.org/abs/1703.00887

Learning dynamics

Hypothesis: $\mathcal{L}(W^{(k)}) \leq (1-\beta)^k \mathcal{L}(W^{(0)})$ whp over initialization $W^{(0)}$, and β grows with $\dim W$.

Consider a 2-layer non-linear net with square loss:

$$f(W, a, x) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(w_r^T x), \quad x \in \mathbb{R}^d,\ w_r \in \mathbb{R}^d,\ a_r \in \mathbb{R}.$$

$$\mathcal{L}(W, a) = \frac{1}{2} \sum_{i=1}^{n} (f(W, a, x_i) - y_i)^2.$$

Continuous-time GD:

$$\frac{dw_r(t)}{dt} = -\frac{\partial \mathcal{L}(W(t), a(t))}{\partial w_r}; \qquad \frac{da(t)}{dt} = -\frac{\partial \mathcal{L}(W(t), a(t))}{\partial a}.$$

Learning dynamics

Consider training the output layer only:

$$\frac{da(t)}{dt} = -\frac{\partial \mathcal{L}(W, a(t))}{\partial a}.$$

Denote: $z_i = \sigma(Wx_i), \quad Z = [z_1 \dots z_n] \in \mathbb{R}^{m \times n}$.

$$\frac{d\mathcal{L}(W, a(t))}{dt} \leq -2\lambda_{min}(H^{out})\mathcal{L}(W, a(t)),$$

where

$$H^{out} = \frac{1}{m} Z^T Z \in \mathbb{R}^{n \times n}.$$

From lemma of Yu & Chen (1995):
If σ is analytic and $m \geq n$, then H^{out} is full rank a.s. wrt $W \sim \mathcal{N}(0, I)$.

Hence $\lambda_{min}(H^{out}) > 0$, and:

$$\mathcal{L}(W, a(t)) \leq e^{-2\lambda_{min}(H^{out})t} \mathcal{L}(W, a(0)).$$

Learning dynamics

Consider training the input layer only:

$$\frac{dw_r(t)}{dt} = -\frac{\partial \mathcal{L}(W(t), a)}{\partial w_r}.$$

Loss dynamics:

$$\frac{d\mathcal{L}(W(t), a)}{dt} \leq -2\lambda_{min}(H(t))\mathcal{L}(W(t), a),$$

where $H(t) = H(W(t))$, and

$$H_{ij}(W) := \frac{1}{m}\sum_{r=1}^{m}\left(\frac{\partial f(W, a, x_i)}{\partial w_r}\right)^T \frac{\partial f(W, a, x_j)}{\partial w_r} \quad \forall i, j = 1, \ldots, n.$$

Learning dynamics

Loss dynamics:

$$\frac{d\mathcal{L}(W(t), a)}{dt} \leq -2\lambda_{min}(H(t))\mathcal{L}(W(t), a),$$

Let $W(0) \sim \mathcal{N}(0, I)$ and $\lambda_0 := \lambda_{min}(\mathbb{E}_{W(0)} H(0))$.

Lemma (Du et al., 2019[12]):
If $\forall i, j\ x_i \nparallel x_j$, then $\lambda_0 > 0$.

Assume $\forall t \geq 0 \quad \lambda_{min}(H(t)) \geq \kappa\lambda_0 > 0$. Then,

$$\mathcal{L}(W(t), a) \leq e^{-2\kappa\lambda_0 t}\mathcal{L}(W(0), a).$$

Similar for discrete-time GD with step η:

$$\mathcal{L}^{(k)} \leq (1 - \alpha\kappa\lambda_0)^k\, \mathcal{L}^{(0)} \quad \forall k \geq 0 \quad \text{for sufficiently small } \eta.$$

[12] https://openreview.net/forum?id=S1eK3i09YQ

Learning dynamics

Theorem (Du et al., 2019):
Assume $\|x_i\| = 1, |y_i| < C \quad \forall i = 1 \dots n$, and

$$w_r(0) \sim \mathcal{N}(0, I),\ a_r \sim U(\{-1, 1\}) \quad \forall r = 1, \dots, m.$$

Let $\delta \in (0, 1)$ and $m = \Omega\left(\frac{n^6}{\lambda_0^4 \delta^3}\right)$; then w.p. $\geq 1 - \delta$ over initialization

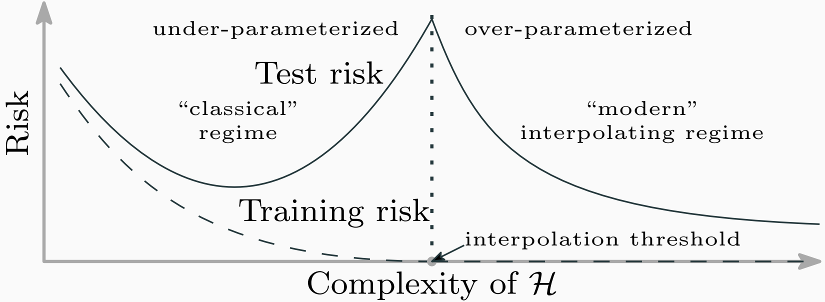
$$\lambda_{min}(H(t)) \geq \frac{\lambda_0}{2} \quad \forall t \geq 0.$$

Theorem (Song & Yang, 2019)[13]:
The same holds for $m = \Omega\left(\frac{n^4}{\lambda_0^4} \log^3\left(\frac{n}{\delta}\right)\right)$.

[13] `https://arxiv.org/abs/1906.03593`

Complexity-risk curve

General "double descent" curve:



Generalization ability

- **Observation:** Test risk of networks found by SGD decreases as width grows.
- **Hypothesis:** There is a network complexity measure with following properties:
    1. It correlates with test risk;
    2. It is implicitly minimized by SGD.

Generalization bounds

Our goal is to bound the risk difference:

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

where

- $R(f)$ — risk of predictor f,
- $\hat{R}_n(f)$ — empirical risk of predictor f on dataset S_n,
- $\hat{f}_n = \mathcal{A}(S_n) \in \mathcal{F}$ — solution found by algorithm $\mathcal{A}$ (e.g. SGD) on S_n,
- $N(f)$ — complexity measure of predictor f.

Usual form of bound:

$$\mathrm{bound}(N, n, \delta) = O\left(\sqrt{\frac{N + \log(1/\delta)}{n}}\right).$$

Generalization bounds

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

Worst-case bounds:

$$\mathrm{bound} = \sup_{f \in \mathcal{F}} \left|R(f) - \hat{R}_n(f)\right|.$$

Lead to complexity measures that depend on $\mathcal{F}$ (and do not depend on $\hat{f}_n$ directly).

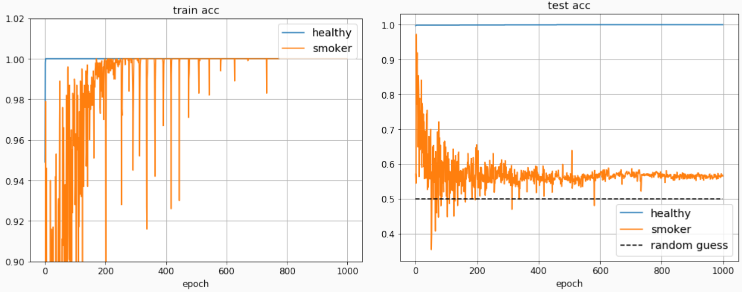
Generalization bounds

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

- Let $\mathcal{F}$ be the set of all nets of the given architecture;
- Let $R(f) = \mathbb{E}_{x,y \sim \mathcal{D}}[yf(x) < 0] = 1 -$ accuracy of f.

Generalization bounds

Bad nets usually exist:



Experiments similar to Zhang et al. (2017)[14].

Hence the bound is vacuous.

[14] https://arxiv.org/abs/1611.03530

Generalization bounds

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

- Let $\mathcal{F}$ be the set of all nets of the given architecture;
- Let $R(f) = \mathbb{E}_{x,y\sim\mathcal{D}}[yf(x) < 0] = 1 -$ accuracy of f.

Leads to vacuous bounds; way to mitigate it:

- Narrow $\mathcal{F}$;
- Use scale-sensitive R, i.e. $R_\gamma(f) = \mathbb{E}_{x,y\sim\mathcal{D}}[yf(x) < \gamma]$.

Generalization bounds

Way to narrow $\mathcal{F}$:

Let $\hat{f}_n$ be network with weights $\left\{\hat{W}_n^{(l)}\right\}_{l=1}^{L}$. Consider

$$\mathcal{F}(\hat{f}_n) = \left\{f : \left\|W^{(l)}\right\| \leq \left\|\hat{W}_n^{(l)}\right\|\right\}.$$

Lead to bounds that depend on $\left\{\left\|\hat{W}_n^{(l)}\right\|\right\}_{l=1}^{L}$.

Examples:

- **Bartlett (1998)[15]:** Tanh-nets, bound depends on l_1-norm of output layer;
- **Bartlett et al. (2017)[16]:** Arbitrary feed-forward nets, bound depends on Lipschitz constant of learned net $\hat{f}_n$.

---

15 https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=661502

16 https://arxiv.org/abs/1706.08498

Generalization bounds

Consider stochastic learning algorithm: $\hat{f}_n = \mathcal{A}(S_n) \sim Q|S_n$.

Corresponding bound:

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

PAC-bayesian bound (McAllester, 1999)[17]**:**

$$N(Q) = KL(Q \,\|\, P),$$

where P denotes prior over predictors f.

- **Pros:** Depends on learned predictor $\hat{f}_n$.
- **Cons:** Vacuous if $P(A) = 0 \not\Rightarrow Q(A) = 0$.

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

Generalization bounds

Usually $\hat{f}_n$ is deterministic and P is continuous $\Rightarrow KL(\delta_{\hat{f}_n} \parallel P) = \infty$.

How to deal with it?

- **Make stochastic (Dziugaite & Roy, 2017)[18]:**

$$\mathbb{E}_{Q|S_n} \hat{R}_n(\hat{f}_n) + \mathrm{bound}(KL(Q|S_n \parallel P), n, \delta) \to \min_{Q},$$

where Q is initialized with $\delta_{\hat{f}_n}$.

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

Generalization bounds

Usually $\hat{f}_n$ is deterministic and P is continuous $\Rightarrow KL(\delta_{\hat{f}_n} \parallel P) = \infty$.

How to deal with it?

- **Use margin loss (Neyshabur et al., 2018)[19]:**
  Let $Q = \mathcal{N}(\hat{f}_n, \sigma)$. Take $\delta' \in (0, 1)$ and $\gamma > 0$.
  Then take maximal σ:

$$R(\hat{f}_n) - \hat{R}_{n,\gamma}(\hat{f}_n) \leq \mathbb{E}_{f \sim Q|S_n}(R_{\gamma/2}(f) - \hat{R}_{n,\gamma/2}(f)) \quad \text{w.p. } \geq 1 - \delta' \text{ over } S_n.$$

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

Generalization bounds

Usually $\hat{f}_n$ is deterministic and P is continuous $\Rightarrow KL(\delta_{\hat{f}_n} \parallel P) = \infty$.

How to deal with it?

- **Use discrete coding (Zhou et al., 2019)[20]:**
  Let $|f|_c$ — number of bits required to encode f with coding c.
  Coding-based prior:

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

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

$$KL(\delta_{\hat{f}_n} \parallel P_c) \leq |\hat{f}_n|_c \log 2 - \log(m(|\hat{f}_n|_c)).$$

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

Generalization bounds

Sanity checks (Nagarajan & Kolter, 2019)[21]:

1. Non-vacuous (bound < 1);
2. Reflect the same width/depth/batch size dependence as generalization error;
3. Decrease with dataset size;
4. Increase with proportion of randomly flipped labels;
5. Applies directly to original learned network.

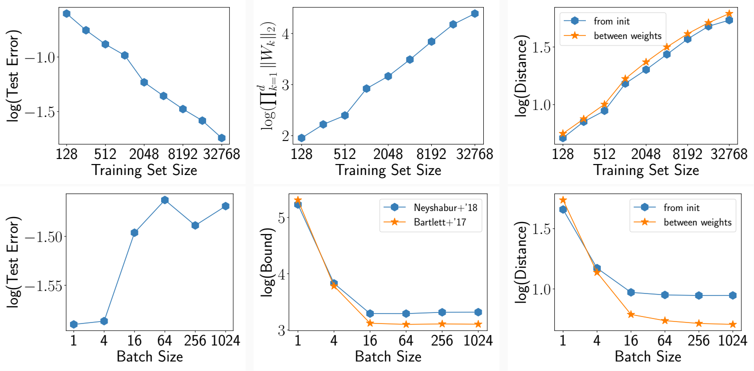
All of the above-mentioned bounds fail at least one of them.

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

Generalization bounds

Most of the bounds depend either on:

1. Lipschitz constant of $\hat{f}_n$ (Bartlett et al., 2017, Neyshabur et al., 2018), **or**
2. l_2 distance from init (Dziugaite & Roy, 2017).



Generalization ability

Observation: Test risk of networks found by SGD decreases as width grows.

Hypothesis: There is a network complexity measure $N(\cdot)$ with following properties:

1. It correlates with test risk;
2. It is implicitly minimized by SGD:

$$\hat{f}_n = \mathrm{SGD}(S_n) \in \operatorname*{Arg\,min}_{f:\, \hat{R}_n(f)=0} N(f).$$

Implicit bias

Results (teaser):

- **Linear regression:** for zero init GD chooses minimum l_2-norm solution.
- **Neural network:** depends on magnitude of init (Chizat et al., 2018)[22]:
  - Large init: SGD finds minimum norm solution in some RKHS;
  - Small init: ??

[22] `https://arxiv.org/abs/1812.07956`