Fondements théoriques du deep learning

Généralisation des réseaux de neurones

Sébastien Gerchinovitz[1,2], François Malgouyres[2], Edouard Pauwels[3] et Nicolas Thome[4]

[1] IRT Saint Exupéry (projet DEEL), Toulouse

[2] Institut de Mathématiques de Toulouse, Université Paul Sabatier

[3] Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier

[4] Centre D'Étude et de Recherche en Informatique et Communications, Conservatoire National des Arts et Métiers

Plan

A short introduction to (classical) generalization bounds

This is a short introduction to generalization bounds, a family of inequalities that are key in statistical learning theory, yet sometimes poorly understood in machine learning practice.

Outline:

1. Problem statement: binary classification with i.i.d. training data
2. Estimating a given probability/risk
3. Estimating the risks of multiple classifiers simultaneously
4. VC-bound for binary classification

The concept of generalization in practice

Example: automatic pedestrian recognition

- Learning task: binary classification (image with/without pedestrians)
- Training set: sequence of labeled images
- Operational domain: within an autonomous vehicle

Broad understanding of generalization

- The ML model/classifier is correct on training images...
- ... yet we want it to 'generalize', that is, to be correct on previously unseen images.

Need for quantitative goals and problem description

- Goal: correct for 'all' or 'most' unseen images? Meaning of 'most'?
- Training data: how was it sampled? How to make sure it is representative of the operational domain?

There are many possible answers corresponding to different practical applications.

Statistical problem: binary classif. with i.i.d. training data

To be able to prove theoretical guarantees, learning problems are usually simplified. We will focus on the simplest problem.

Binary classification with i.i.d. training data

- We observe a sequence $\mathscr{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n) \in \mathscr{X} \times \{0, 1\}$ of input-label examples (the training data).
- Assumption 1: the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are drawn at random and independently from the same probability distribution $P_{X,Y}$.
- Assumption 2: the learned classifier will be used on data (X, Y) drawn at random also from $P_{X,Y}$.

Statistical problem: binary classif. with i.i.d. training data

To be able to prove theoretical guarantees, learning problems are usually simplified. We will focus on the simplest problem.

Binary classification with i.i.d. training data

- We observe a sequence $\mathscr{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n) \in \mathscr{X} \times \{0, 1\}$ of input-label examples (the training data).
- Assumption 1: the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are drawn at random and independently from the same probability distribution $P_{X,Y}$.
- Assumption 2: the learned classifier will be used on data (X, Y) drawn at random also from $P_{X,Y}$.

Learning goal: construct a classifier $\widehat{f}$ depending on the training data $\mathscr{D}_n$ that minimizes the risk (a.k.a. test error or out-of-sample error)

$$R(\widehat{f}) = \mathbb{P}(Y \neq \widehat{f}(X) | \mathscr{D}_n) = \int_{\mathscr{X} \times \{0,1\}} \mathbb{1}_{y \neq \widehat{f}(x)} dP_{X,Y}(x, y)$$

or the expected risk (average of $R(\widehat{f})$ over all possible training data $\mathscr{D}_n$)

$$\mathbb{E}\left[R(\widehat{f})\right] = \mathbb{P}(Y \neq \widehat{f}(X))$$

Statistical problem: binary classif. with i.i.d. training data

The assumptions are strong!

- same distribution across training examples and operational data
- examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are drawn independently (this is a way to increase information as n increases)

There are statistical works where some assumptions are relaxed, e.g.,

- distributional shift, presence of outliers
- weak-dependency in the data

but we should be aware that assumptions are necessary to prove meaningful generalization guarantees with only a finite training sequence $\mathscr{D}_n$.

Estimating a given probability

Toy example 1 (survey): proportion of people under 20 in Toulouse.

- Goal: estimate the proportion p of people under 20 in Toulouse.
- We collect the ages of a (uniform) random subset of people in Toulouse.
- We estimate p with the observed proportion $\widehat{p}$ of people under 20.
- How close is $\widehat{p}$ to p?

Estimating a given probability

Toy example 1 (survey): proportion of people under 20 in Toulouse.

- Goal: estimate the proportion p of people under 20 in Toulouse.
- We collect the ages of a (uniform) random subset of people in Toulouse.
- We estimate p with the observed proportion $\widehat{p}$ of people under 20.
- How close is $\widehat{p}$ to p?

Toy example 2 (industrial survey): proportion of defective items in a warehouse.

- Goal: estimate the proportion q of defective items in a warehouse.
- We select items from a spreadsheet uniformly and independently at random, and we observe whether these items are defective or not.
- We estimate q with the observed proportion $\widehat{q}$ of defective items.
- How close is $\widehat{q}$ to q?

Estimating a given probability

Both problems can be modeled (or simplified) as a sequence of independent Bernoulli trials:

- we observe independent random variables $Z_1, \dots, Z_n \in \{0,1\}$ such that $\mathbb{P}(Z_i = 1) = p$ for all $i = 1, \dots, n$;
- we want to estimate p using the observed proportion $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Z_i$

Hoeffding's inequality (Hoeffding, 1963)

Let $Z_1, \dots, Z_n \in \{0,1\}$ be independent $\mathrm{Ber}(p)$ random variables. Then, for any risk level $\delta \in (0,1)$,

$$\mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} Z_i - p \right| \leqslant \sqrt{\frac{\log(2/\delta)}{2n}} \right) \geqslant 1 - \delta \,.$$

Interpretation:

- Unless maybe for a fraction δ of all possible observed sequences $z_1, \dots, z_n$, the observed proportion $\widehat{p}$ is close to the unknown general proportion p, with a difference bounded by $\sqrt{\log(2/\delta)/(2n)}$.
- In practice, we only observe a single sequence $z_1, \dots, z_n$, and we "bet" that it is not among the δ problematic ones.

Estimating a given probability

There are many related mathematical results.

Central Limit Theorem (de Moivre-Laplace) We actually know the exact distribution of a rescaled version of the error $\widehat{p} - p$ when $n \to +\infty$:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - p\right) \xrightarrow{\text{distrib.}} \mathcal{N}\big(0, p(1-p)\big)$$

Refined non-asymptotic concentrations inequalities Hoeffding's inequality only uses information about the range of the random variables. Random variables with additional properties such as small variance may feature improved concentration bounds (e.g., Bernstein's inequality).

Concentration inequalities under relaxed assumptions

- Hoeffding-Azuma or Bernstein inequalities for martingales
- Bernstein's inequality for Markov chains
- etc (though statistical assumptions of some sort are always needed)

Estimating the risk of a classifier

Getting back to the binary classification problem Given a classifier $f : \mathscr{X} \to \{0,1\}$ and a sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent pairs all drawn at random from $P_{X,Y}$, how close is the empirical risk

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \neq f(X_i)}$$

to the risk

$$R(f) = \mathbb{P}(Y \neq f(X)) = \int_{\mathscr{X} \times \{0,1\}} \mathbb{1}_{y \neq f(x)} dP_{X,Y}(x,y) \, ?$$

Answer We can apply, e.g., Hoeffding's inequality with $Z_i = \mathbb{1}_{Y_i \neq f(X_i)}$: for any fixed[1] classifier $f : \mathscr{X} \to \{0,1\}$ and any risk level $\delta \in (0,1)$,

$$\mathbb{P}\left(\left| \widehat{R}(f) - R(f) \right| \leqslant \sqrt{\frac{\log(2/\delta)}{2n}} \right) \geqslant 1 - \delta \, .$$

[1] This means that f is picked before observing the training data.

Estimating the risks of multiple classifiers simultaneously

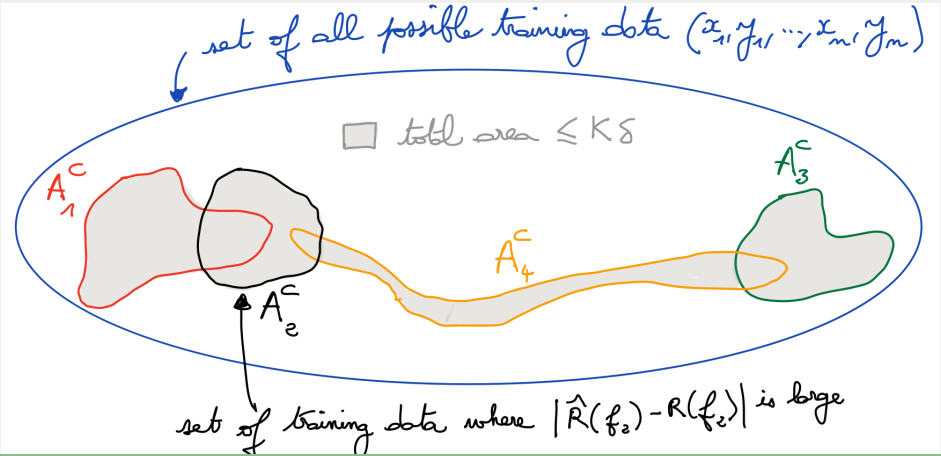
Suppose we want to estimate the risks of K fixed classifiers $f_1, \ldots, f_K$ from the same training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, in order to compare them.

Multiple testing issue: From the previous slide, each bound

$$\left|\widehat{R}(f_k) - R(f_k)\right| \leqslant \sqrt{\frac{\log(2/\delta)}{2n}}$$

is correct for a large subset A_k of all possible training sequences $(x_1, y_1), \ldots, (x_n, y_n)$, of probability at least $1 - \delta$.

However, problematic events $A_1^c, \ldots, A_K^c$ may span an event of probability up to $K\delta$ in total.



Estimating the risks of multiple classifiers simultaneously

Resulting **uniform** risk bound: for any $\delta \in (0,1)$,

$$\mathbb{P}\left(\text{for all } k = 1, \dots, K, \quad \left| \widehat{R}(f_k) - R(f_k) \right| \leqslant \sqrt{\frac{\log(2/\delta)}{2n}} \right) \geqslant 1 - K\delta \, .$$

Estimating the risks of multiple classifiers simultaneously

Resulting **uniform** risk bound (Bonferroni correction): for any $\delta \in (0,1)$,

$$\mathbb{P}\left(\text{for all } k = 1, \dots, K, \quad \left| \widehat{R}(f_k) - R(f_k) \right| \leqslant \sqrt{\frac{\log(2K/\delta)}{2n}} \right) \geqslant 1 - \delta \, .$$

Estimating the risks of multiple classifiers simultaneously

Resulting **uniform** risk bound (Bonferroni correction): for any $\delta \in (0,1)$,

$$\mathbb{P}\left(\text{for all } k = 1, \dots, K, \quad \left| \widehat{R}(f_k) - R(f_k) \right| \leqslant \sqrt{\frac{\log(2K/\delta)}{2n}} \right) \geqslant 1 - \delta \, .$$

Consequence: even if we pick a classifier f among $f_1, \dots, f_K$ after seeing the training data, we still have a control on $\widehat{R}(f_k) - R(f_k)$. The price for overfitting here is the extra K in the logarithm.

In particular, if we choose the f_k that best fits the training data, i.e.,

$$\widehat{f} \in \operatorname{argmin}_{f \in \{f_1, \dots, f_K\}} \widehat{R}(f) \, ,$$

then the uniform risk bound above implies that $\widehat{f}$ has a near-optimal risk:

$$\mathbb{P}\left(R(\widehat{f}) \leqslant \min_{f \in \{f_1, \dots, f_K\}} R(f) + 2\sqrt{\frac{\log(2K/\delta)}{2n}} \right) \geqslant 1 - \delta \, .$$

Next: What if K is large or even infinite?

VC-bound for binary classification

Let $\mathscr{F}$ be a set of classifiers $f : \mathscr{X} \to \{0,1\}$. For a given sample size $n \geqslant 1$, the **shattering coefficient** of $\mathscr{F}$ is the quantity

$$\pi_{\mathscr{F}}(n) = \max_{x_1,\ldots,x_n \in \mathscr{X}} \operatorname{card}\left\{\left(f(x_1),\ldots,f(x_n)\right) : f \in \mathscr{F}\right\} .$$

This is the maximum number of sign patterns that classifiers in $\mathscr{F}$ can produce. We always have $\pi_{\mathscr{F}}(n) \leqslant 2^n$.

Vapnik-Chervonenkis dimension

The VC-dimension of $\mathscr{F}$ is the quantity

$$\operatorname{VCdim}(\mathscr{F}) = \sup\left\{n \in \mathbb{N} : \pi_{\mathscr{F}}(n) = 2^n\right\} .$$

This is the largest size of a sequence $(x_1,\ldots,x_n)$ that $\mathscr{F}$ can "shatter", i.e., on which the classifiers in $\mathscr{F}$ can produce all 2^n sign patterns.

VC-bound for binary classification

Example: the perceptron (affine classifiers).

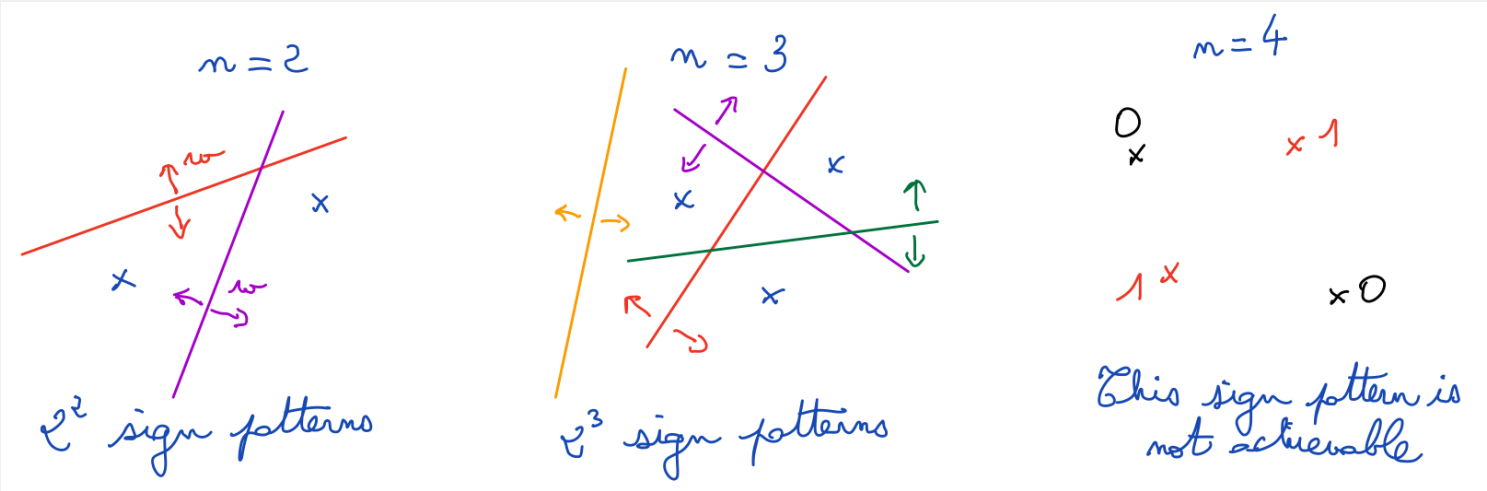
Writing $\mathrm{sign}(t) = \mathbb{1}_{t>0}$ for the sign of t, consider the set of affine classifiers over d real-valued input variables:

$$\mathscr{F} = \left\{ x \in \mathbb{R}^d \mapsto \mathrm{sign}(w \cdot x + b) : w \in \mathbb{R}^d, b \in \mathbb{R} \right\} .$$

VC-dimension of the perceptron

$$\mathrm{VCdim}(\mathscr{F}) = d + 1$$

Example in dimension $d = 2$:



VC-bound for binary classification

We are now ready to state the so-called VC-bound.

VC-bound (cf. Vapnik and Chervonenkis 1971 and subsequent works)

Let $\mathscr{F}$ be a set of classifiers with $V := \mathrm{VCdim}(\mathscr{F}) < +\infty$, and a sequence $(X_1, Y_1), \dots, (X_n, Y_n)$ drawn independently from $P_{X,Y}$. Then,

$$\mathbb{P}\left(\text{for all } f \in \mathscr{F}, \quad \left|\widehat{R}(f) - R(f)\right| \leqslant c\sqrt{\frac{V \max\left\{\log\left(\frac{n}{V\delta}\right), 1\right\}}{n}} \right) \geqslant 1 - \delta \,.$$

Example: feedforward neural networks.

Consider a fixed feedforward ReLU neural network with d inputs, a single (linear) output, L layers, and W weights.

Let G be the set of all functions $g : \mathbb{R}^d \to \mathbb{R}$ that can be represented by this network when varying the weights, and $\mathrm{sign}(G)$ the associated classifiers:

$$\mathrm{sign}(G) = \left\{ x \in \mathbb{R}^d \mapsto \mathrm{sign}(g(x)) : g \in G \right\} \,.$$

Then, by Bartlett et al. (2019, Theorem 6), $\mathrm{VCdim}\big(\mathrm{sign}(G)\big) \lesssim LW\log(W)$.

Need for less conservative bounds

The generalization bounds presented before are from classical statistical learning theory textbooks. See, e.g., Shalev-Shwartz and Ben-David (2014); Mohri et al. (2018); Anthony and Bartlett (2009).

Though nearly optimal in the worst case, the VC-bound for deep learning is non informative (bound larger than 1) in many practical deep learning settings.

Research about tighter generalization bounds that can explain good empirical performances is ongoing.

1. A short introduction to (classical) generalization bounds

2. Overparameterized neural networks: a statistical paradox?

3. An (almost) ready-to-use approach: post-processing risk control

Overparameterized neural networks: a statistical paradox?

Statistical paradox

Overparameterized neural networks can perfectly fit training data, while generalizing well to unseen data (Zhang et al., 2021).

- This contradicts statistical wisdom against overfitting or overparamerization!
- For overparameterized networks, VC-type bounds from statistical learning theory only provide vacuous guarantees (e.g., a misclassification risk smaller than 1).
- Active mathematical research to solve this paradox, and understand the role of overparameterization and implicit bias (Belkin 2021; Bartlett et al. 2021; Grohs and Kutyniok 2022).

Towards tighter generalization bounds

Let us mention two attempts aimed at tighter generalization bounds.

- "Size-independent bounds" by Golowich et al. (2018).
  Assume the matrices of a depth-L ReLU feedforward neural network are constrained, say, in Frobenius norm by $M_F(j)$ for layer j.

  Then, the Rademacher complexity of $F = \{f_{\mathbf{w}} : w \in \mathbb{R}^W\}$ (which controls a *uniform* generalization gap) is bounded with high probability roughly by

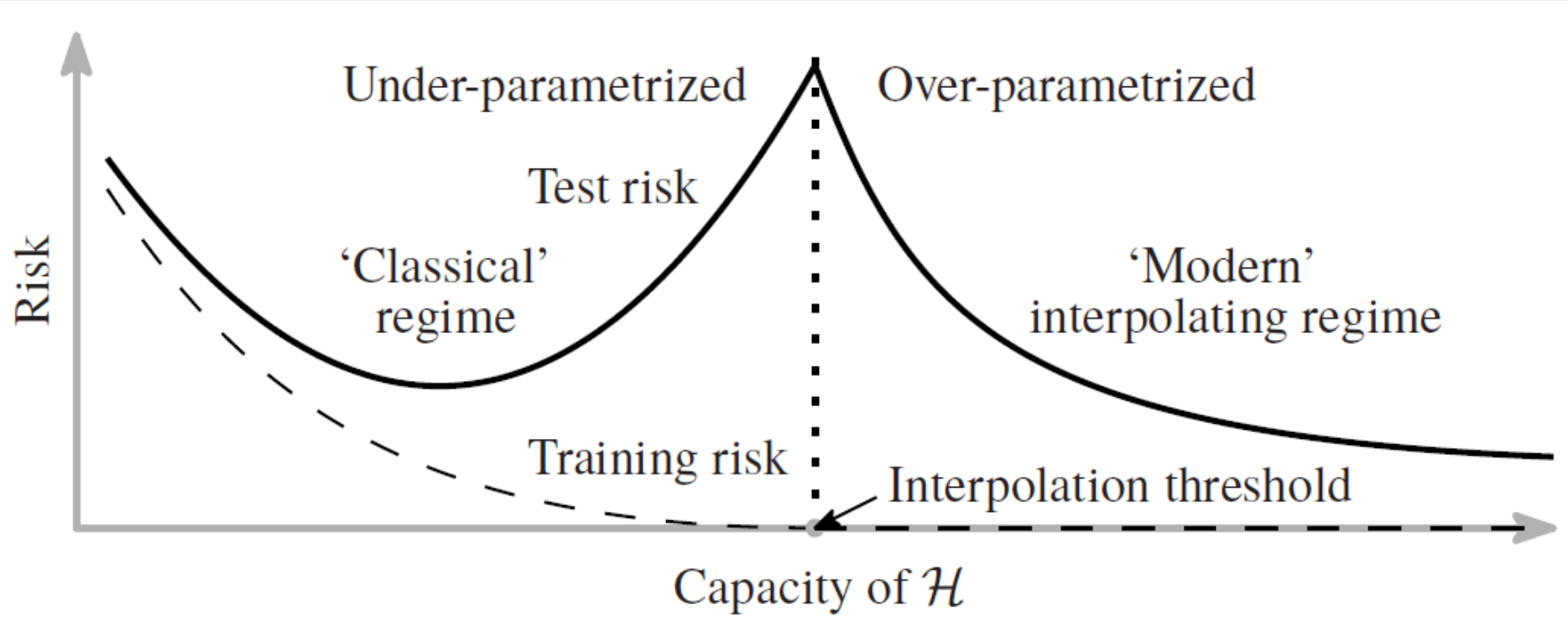
$$\widehat{\mathscr{R}}_n(F) \lesssim \frac{B\sqrt{L}\prod_{j=1}^{L} M_F(j)}{\sqrt{n}}$$

- "Path-norms". Consider the vector $\Phi(\mathbf{w})$ made of (roughly) all weight products along all possible paths in a feedforward network. Gonon et al. (2023) prove a generalization bound proportional to the ℓ^1-norm $\|\Phi(\mathbf{w})\|_1$.

But: "Uniform convergence may be unable to explain generalization in deep learning" (Nagarajan and Kolter, 2019).

The double-descent phenomenon

For various DL experiments, the classical *U*-shaped risk curve was shown to be replaced with a **double-descent** generalization curve (source: Belkin et al. 2019).



The double-descent phenomenon

The double descent phenomenon and that of "benign overfitting" in the overparameterized regime have been *demonstrated* in simpler settings.

Simplest example: linear regression with isotropic covariates (Hastie et al., 2022), where we observe $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$ drawn i.i.d. with, e.g.,

- $y_i = x_i^T \beta^* + \varepsilon_i$ for all $i = 1, \ldots, n$ ($\beta^* \in \mathbb{R}^p$ is unknown)
- $x_{i,j} \sim \mathcal{N}(0,1)$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
- all $x_{i,j}$ and ε_i independent

We consider:

- (Excess) Risk: $R(\beta) = \mathbb{E}\left[(x_{n+1}^T \beta - x_{n+1}^T \beta^*)^2\right]$
- Minimum ℓ^2-norm least-squares estimator:

$$\widehat{\beta} = \operatorname{argmin}\left\{\|b\|_2 : b \text{ minimizes } \|y - Xb\|_2\right\}$$

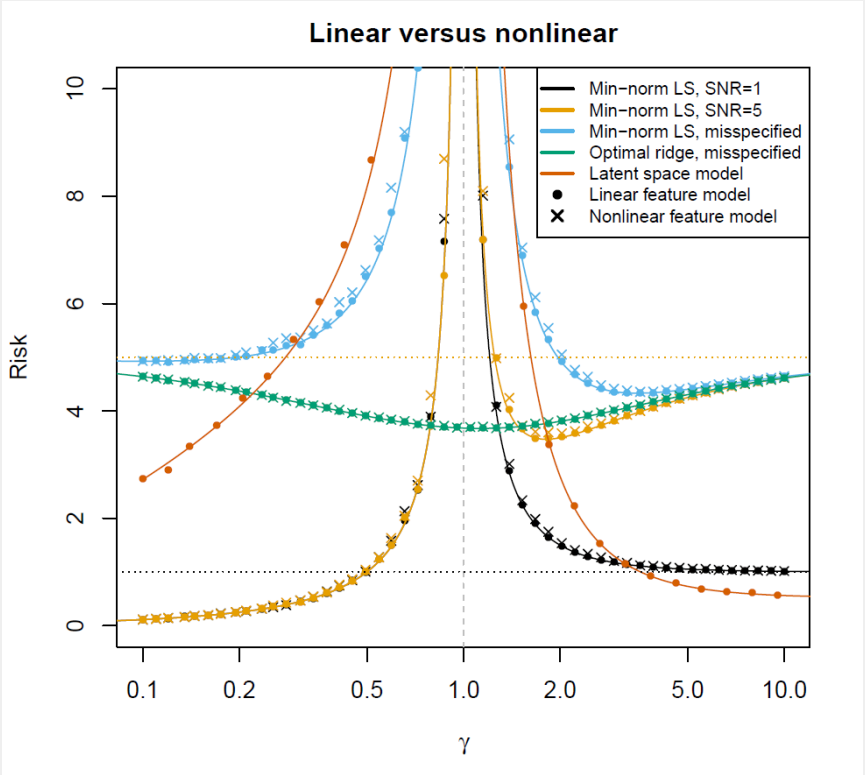
The double-descent phenomenon

We consider the asymptotic regime when $n, p \to +\infty$ with $p/n \to \gamma > 0$.

Theorem (Hastie et al. 2022)

Assume $n, p \to +\infty$ with $p/n \to \gamma > 0$, and $\|\beta\|_2^2 = r^2$ for all n, p. Then, almost surely,

$$R(\widehat{\beta}) \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{if } \gamma < 1 \text{ (underparameterized regime)} \\ r^2\left(1 - \frac{1}{\gamma}\right) + \sigma^2 \frac{1}{\gamma - 1} & \text{if } \gamma > 1 \text{ (overparameterized regime)} \end{cases}$$

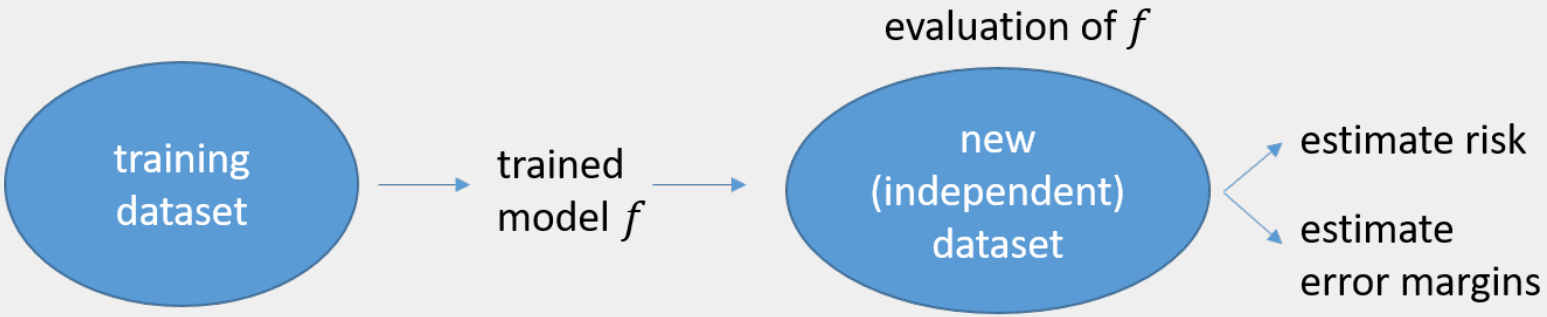


1. A short introduction to (classical) generalization bounds
2. Overparameterized neural networks: a statistical paradox?
3. An (almost) ready-to-use approach: post-processing risk control

A more practical approach: post-processing risk control

As seen previously, VC bounds (or their equivalent for regression) are typically too conservative in practice. This is because they hold for all data distributions and **control the generalization gap uniformly over all predictors in a class**.

Simple alternative: **post-processing** methods.



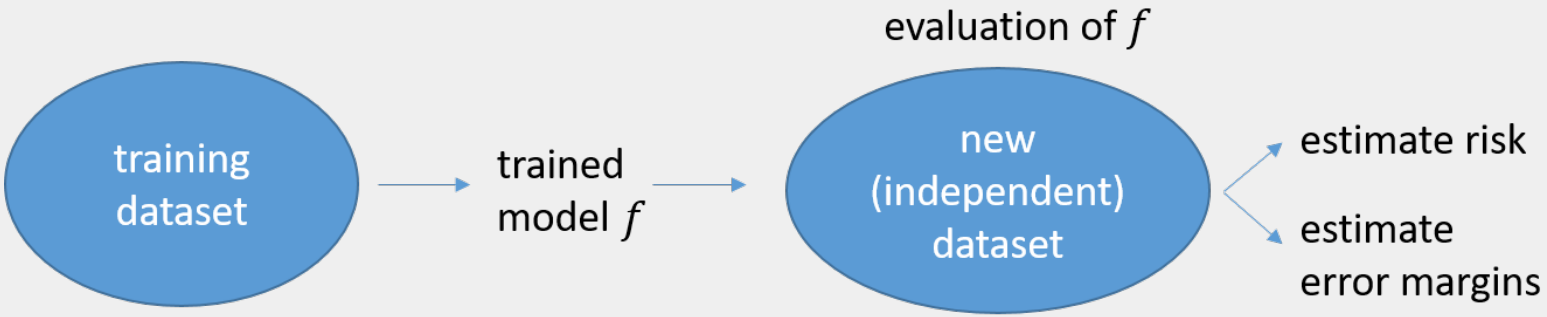
Simple alternative approach #1: **evaluate risk after training**

- take a trained model $f : \mathscr{X} \to \mathscr{Y}$ as is;
- evaluate this single model on an independent data set: observe $\widehat{R}(f)$
- bound the true risk by $R(f) \leqslant \widehat{R}(f) + \varepsilon(n, \delta)$ with confidence level $1 - \delta$, where $\varepsilon(n, \delta)$ is given by some (out of many) concentration inequality.

A more practical approach: post-processing risk control

As seen previously, VC bounds (or their equivalent for regression) are typically too conservative in practice. This is because they hold for all data distributions and **control the generalization gap uniformly over all predictors in a class**.

Simple alternative: **post-processing** methods.



Simple alternative approach #2: **learn predictive uncertainties after training**

- instead of estimating the risk of single-point predictions $f(x)$,
- fix a target risk α and learn (from some independent data) the "size" of a prediction set $C(x)$ that contains the unknown label with probability $\geqslant 1-\alpha$. This "size" can be an "error margin" around some classical prediction $f(x)$.

Conformal prediction in a nutshell

We briefly describe approach #2, of a black-box type (**applies to DNNs**).

We want to predict some random $Y \in \mathscr{Y}$ given some random $X \in \mathscr{X}$.

Suppose we have access to predefined nested prediction sets $(C_\lambda(x))_{\lambda \geqslant 0}$.
For instance, with Y in one or two dimensions ($f^\pm$ and f can be pre-trained DNNs):

$$C_\lambda(x) = [f^-(x) - \lambda; f^+(x) + \lambda] \qquad \text{or} \qquad C_\lambda(x) = \{y \in \mathbb{R}^2 : \|y - f(x)\| \leqslant \lambda\}$$

Task: find a size $\lambda \geqslant 0$ (or "predictive uncertainty") such that, on average over a random draw of (X, Y),

$$\mathbb{P}\big(Y \in C_\lambda(X)\big) \geqslant 1 - \alpha\,.$$

We will learn λ using some **calibration set** $(X_1, Y_1), \ldots, (X_n, Y_n)$ independent of the training set.

Conformal prediction in a nutshell (2)

One possible approach consists in computing the smallest[2] valid λ for each (X_i, Y_i):

$$\lambda_i := \min\{\lambda \geqslant 0 : Y_i \in C_\lambda(X_i)\}, \quad 1 \leqslant i \leqslant n$$

Then, for $\alpha \in (1/(n+1), 1)$, we rank the λ_i's in increasing order and pick the one at rank $\lceil (n+1)(1-\alpha) \rceil$. We denote this value by $\widehat{\lambda}_\alpha$.

Theorem (Papadopoulos et al. 2002; Gupta et al. 2022)

If $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are exchangeable and independent from the training set used to build $(C_\lambda(x))_{\lambda \geqslant 0}$, then

$$\mathbb{P}\big(Y_{n+1} \in C_{\widehat{\lambda}_\alpha}(X_{n+1})\big) \geqslant 1 - \alpha,$$

where the probability is w.r.t. all sources of randomness (training, calibration, test).

In fact, the coverage guarantee holds conditionally on the training set. However, it only holds **on average** over the calibration set $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ and the test point (X_{n+1}, Y_{n+1}). See, e.g., Angelopoulos and Bates (2021) and Bates et al. (2021) for details.

[2] We consider cases for which the minimum is always achieved (as in the previous examples).

References I

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. arXiv:2107.07511.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 2009.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30: 87–201, 2021.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 297–299, 2018.

Antoine Gonon, Nicolas Brisebarre, Elisa Riccietti, and Rémi Gribonval. A path-norm toolkit for modern networks: consequences, promises and challenges, 2023. arXiv:2310.01225.

Philipp Grohs and Gitta Kutyniok, editors. *Mathematical Aspects of Deep Learning*. Cambridge University Press, 2022.

Chirag Gupta, Arun K. Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58(301):13–30, 1963.

References II

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, second edition, 2018.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002*, pages 345–356. Springer Berlin Heidelberg, 2002.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.