Institute of Bioinformatics
Johannes Kepler University Linz

Unit 3

Statistical Learning Theory

Motivation

- So far, we have not discussed where a model $g(.; \boldsymbol{w})$ may come from.
- Most of the time, we have assumed that the model (represented by its parameter vector $\boldsymbol{w}$) is determined from the training data $\mathbf{Z}$ and that the resulting generalization error is estimated by considering a test set of withheld samples (or by cross-validation).

Questions

- If a model explains the training samples well, does it also generalize well to future examples?
- Do more training data lead to better models?
- How can we ensure that model selection/training minimizes the generalization error and not only the training error?

Empirical Risk Minimization (ERM)

Minimizing the training error is called *empirical risk minimization.*

Given a training set $\mathbf{Z}_l$, empirical risk minimization is concerned with finding a parameter setting $\boldsymbol{w}$ such that the *empirical risk*

$$R_{\text{emp}}(g(.;\boldsymbol{w}), \mathbf{Z}_l) = \frac{1}{l} \cdot \sum_{i=1}^{l} L(y^i, g(\boldsymbol{x}^i; \boldsymbol{w}))$$

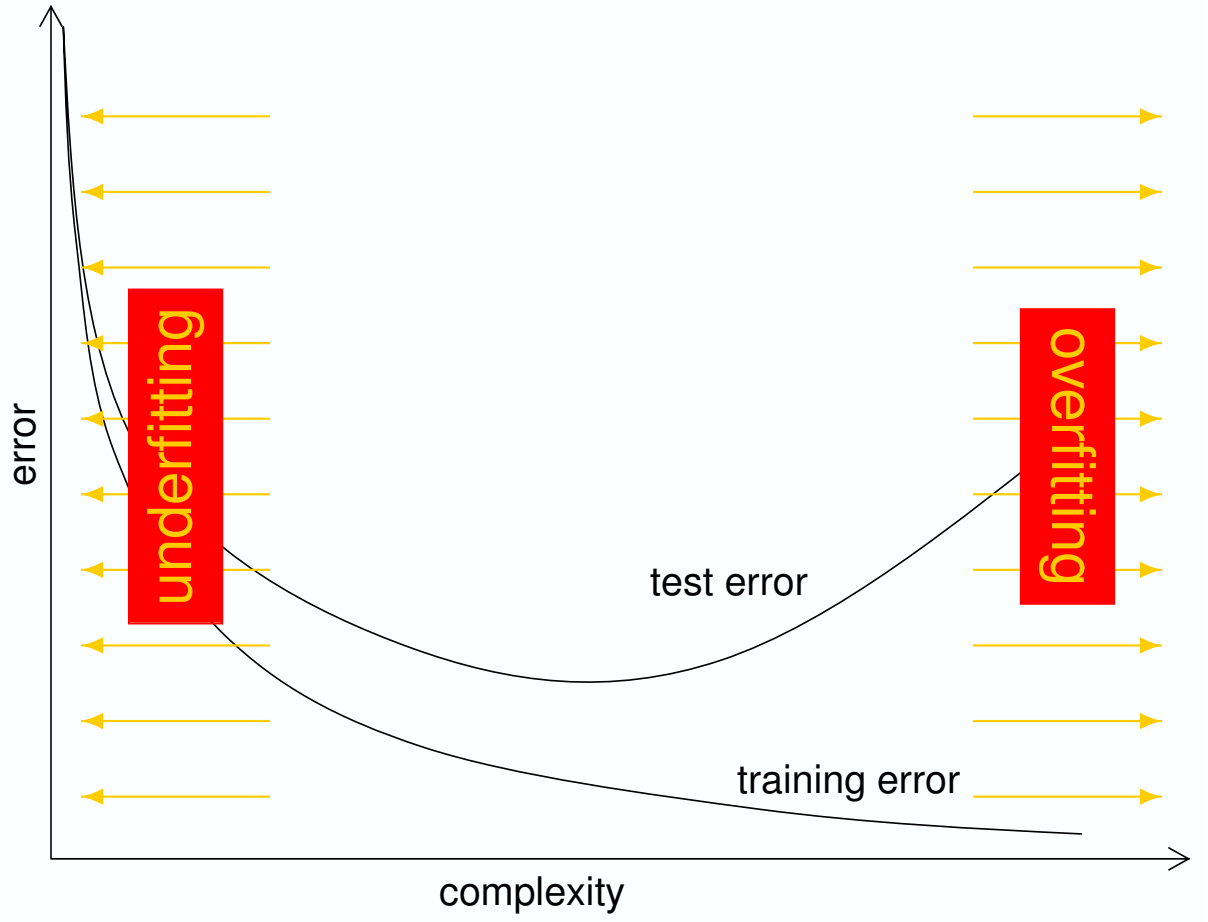
is minimal (or at least as small as possible).

Underfitting and Overfitting

Underfitting: our model is too coarse to fit the data (neither training nor test data); this is usually the result of too restrictive model assumptions (i.e. *too low complexity of model*).

Overfitting: our model works very well on training data, but generalizes poorly to future/test data; this is usually the result of *too high model complexity*.

Notorious Situation in Practice



Questions/Issues

- What are the sources/reasons for under-/overfitting and how to they relate to complexity?
- It is somehow clear what “complexity” means intuitively (flexibility of model class, number of degrees of freedom, etc), but how can we actually define it more formally and use it to deduce useful results?
- It is somehow clear that empirical risk minimization is problematic if we do not take model complexity into account, but how?

Bias-Variance Decomposition for Quadratic Loss (1/3)

We are interested in the expected prediction error for a given $\boldsymbol{x}_0 \in X$ (assuming that the size of the training set is fixed to l examples):

$$\begin{aligned}\mathrm{EPE}(\boldsymbol{x}_0) &= \mathrm{E}_{y|\boldsymbol{x}_0,\mathbf{Z}_l}\left(L_{\mathbf{q}}(y, g(\boldsymbol{x}_0; \boldsymbol{w}(\mathbf{Z}_l)))\right) \\ &= \mathrm{E}_{y|\boldsymbol{x}_0,\mathbf{Z}_l}\left((y - g(\boldsymbol{x}_0; \boldsymbol{w}(\mathbf{Z}_l)))^2\right)\end{aligned}$$

Since $y \mid \boldsymbol{x}_0$ and the selection of training samples are independent (or at least this should be assumed to be the case), we can infer the following:

$$\mathrm{EPE}(\boldsymbol{x}_0) = \mathrm{E}_{y|\boldsymbol{x}_0}\left(\mathrm{E}_{\mathbf{Z}_l}\left((y - g(\boldsymbol{x}_0; \boldsymbol{w}(\mathbf{Z}_l)))^2\right)\right)$$

Bias-Variance Decomposition for Quadratic Loss (2/3)

Using basic properties of expected values, we can infer the following representation:

$$\begin{aligned}\mathrm{EPE}(\boldsymbol{x}_0) =& \mathrm{Var}(y \mid \boldsymbol{x}_0) \\ &+ \Big(\mathrm{E}(y \mid \boldsymbol{x}_0) - \mathrm{E}_{\mathbf{Z}_l}\big(g(\boldsymbol{x}_0; \boldsymbol{w}(\mathbf{Z}_l))\big)\Big)^2 \\ &+ \mathrm{E}_{\mathbf{Z}_l}\Big(\big(g(\boldsymbol{x}_0; \boldsymbol{w}(\mathbf{Z}_l)) - \mathrm{E}_{\mathbf{Z}_l}(g(\boldsymbol{x}_0; \boldsymbol{w}(\mathbf{Z}_l)))\big)^2\Big)\end{aligned}$$

Bias-Variance Decomposition for Quadratic Loss (3/3)

1. The first term, $\mathrm{Var}(y \mid \boldsymbol{x}_0)$ is nothing else but the average amount to which the label y varies at $\boldsymbol{x}_0$. This is often termed *unavoidable error.*
2. The second term,

$$\text{bias}^2 = \Big(\mathrm{E}(y \mid \boldsymbol{x}_0) - \mathrm{E}_{\mathbf{Z}_l}\big(g(\boldsymbol{x}_0; \boldsymbol{w}(\mathbf{Z}_l))\big)\Big)^2$$

   measures how close the model in average approximates the average target y at $\boldsymbol{x}_0$; thus, it is nothing else but the *squared bias*.
3. The third term,

$$\text{variance} = \mathrm{E}_{\mathbf{Z}_l}\Big(\big(g(\boldsymbol{x}_0; \boldsymbol{w}(\mathbf{Z}_l)) - \mathrm{E}_{\mathbf{Z}_l}(g(\boldsymbol{x}_0; \boldsymbol{w}(\mathbf{Z}_l)))\big)^2\Big)$$

   is nothing else but the *variance* of models at $\boldsymbol{x}_0$, i.e. $\mathrm{Var}_{\mathbf{Z}_l}(g(\boldsymbol{x}_0; \boldsymbol{w}(\mathbf{Z}_l)))$.

Bias-Variance Decomposition for Quadratic Loss: Simplifications

- Assume that $y(\boldsymbol{x}) = f(\boldsymbol{x}) + \varepsilon$ holds, where f is a deterministic function and ε is a random variable that has mean zero and variance σ_ε^2 and is independent of $\boldsymbol{x}$. Then we can infer the following:

$$\begin{aligned} \mathrm{Var}(y \mid \boldsymbol{x}_0) &= \sigma_\varepsilon^2, \\ \mathrm{E}(y \mid \boldsymbol{x}_0) &= f(\boldsymbol{x}_0), \\ \text{bias}^2 &= \Big(f(\boldsymbol{x}_0) - \mathrm{E}_{\mathbf{Z}_l}\big(g(\boldsymbol{x}_0; \boldsymbol{w}(\mathbf{Z}_l))\big)\Big)^2. \end{aligned}$$

- In the noise-free case ($\sigma_\varepsilon = 0$), consequently, we get $\mathrm{Var}(y \mid \boldsymbol{x}_0) = 0$, i.e. the unavoidable error vanishes and the rest stays the same.

Bias-Variance Decomposition for Binary Classification

Now assume that we are given a binary classification task, i.e. $y \in \{-1, +1\}$ and $g(\boldsymbol{x}; \boldsymbol{w}) \in \{-1, +1\}$. Since $L_{\mathbf{zo}} = \frac{1}{4} L_{\mathbf{q}}$ holds, we can infer the following:

$$\begin{aligned}
\text{EPE}(\boldsymbol{x}_0) &= \mathrm{E}_{y|\boldsymbol{x}_0, \mathbf{Z}_l} \left(L_{\mathbf{zo}}(y, g(\boldsymbol{x}_0; \boldsymbol{w})) \right) \\
&= \frac{1}{4} \cdot \mathrm{E}_{y|\boldsymbol{x}_0} \left(\mathrm{E}_{\mathbf{Z}_l} \left((y - g(\boldsymbol{x}_0; \boldsymbol{w}(\mathbf{Z}_l)))^2 \right) \right) \\
&= \frac{1}{4} \cdot \left(\mathrm{Var}(y \mid \boldsymbol{x}_0) + \text{bias}^2 + \text{variance} \right)
\end{aligned}$$

Note that, in these calculations, g is the final binary classification function and *not* an arbitrary discriminant function. If the latter is the case, the above representation is *not valid!* (see literature)

Bias-Variance Decomposition for Binary Classification (cont'd)

With the notations $p_R = p(y = +1 \mid \boldsymbol{x_0})$ and

$$p_O = p_{\mathbf{Z}_l}(g(\boldsymbol{x_0}; \boldsymbol{w}(\mathbf{Z}_l)) = +1),$$

we can infer further

$$\begin{aligned} \mathrm{Var}(y \mid \boldsymbol{x}_0) &= 4 \cdot p_R \cdot (1 - p_R), \\ \text{bias}^2 &= 4 \cdot (p_R - p_O)^2, \\ \text{variance} &= 4 \cdot p_O \cdot (1 - p_O), \end{aligned}$$

hence, we obtain

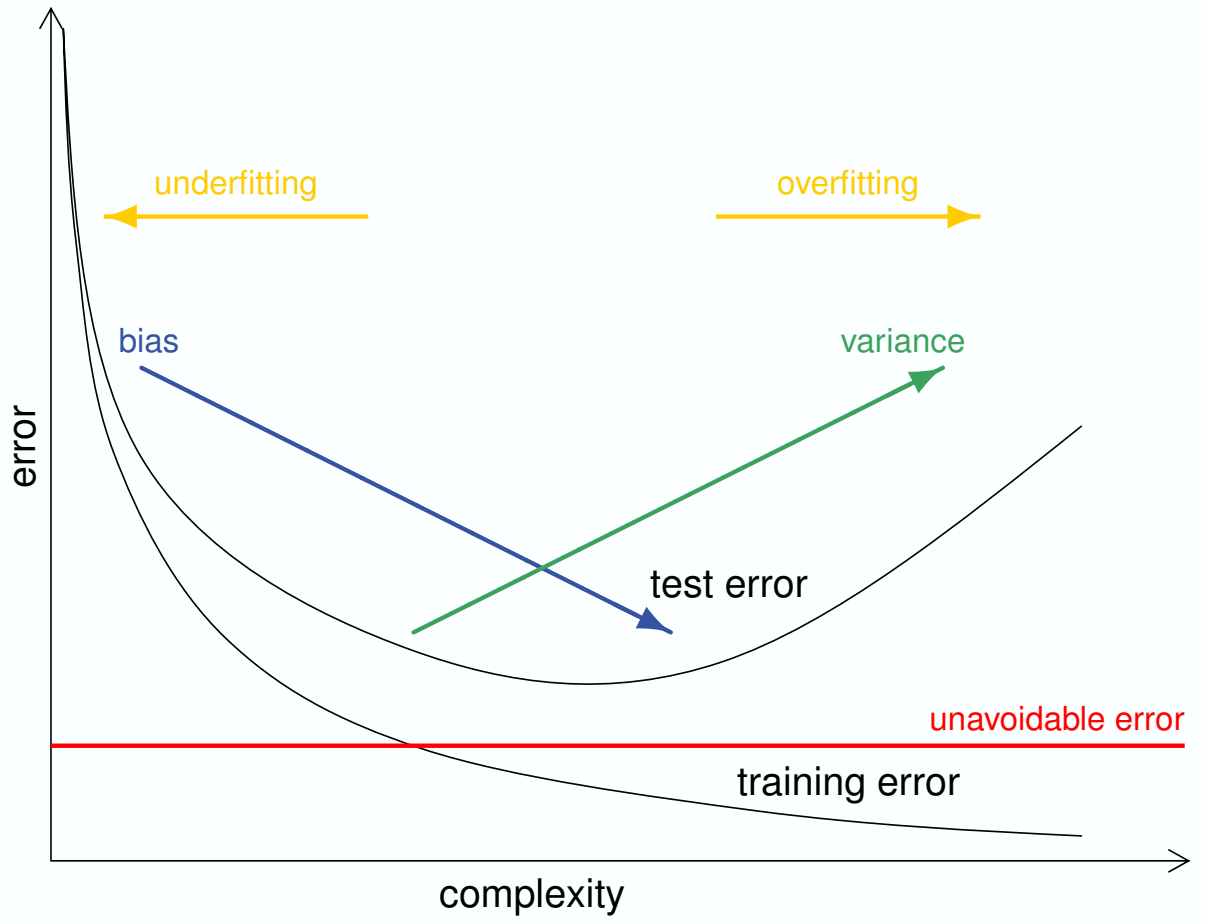
$$\text{EPE}(\boldsymbol{x}_0) = \underbrace{p_R \cdot (1 - p_R)}_{\text{unavoidable error}} + \underbrace{(p_R - p_O)^2}_{\text{squared bias}} + \underbrace{p_O \cdot (1 - p_O)}_{\text{variance}}.$$

The Bias-Variance Trade-Off

- It seems intuitively reasonable that the bias decreases with model complexity.
  Rationale: the more degrees of freedom we allow, the easier we can fit the actual function/relationship.
- It also seems intuitively clear that the variance increases with model complexity.
  Rationale: the more degrees of freedom we allow, the higher the risk to fit to noise.

This is usually referred to as the *bias-variance trade-off*. sometimes even *bias-variance "dilemma".*

The Bias-Variance Trade-Off (cont'd)



Bias-Variance Decomposition: Summary

- We can state that minimizing the generalization error (learning) is concerned with optimizing bias and variance simultaneously.
- Underfitting $=$ high bias $=$ too simple model
- Overfitting $=$ high variance $=$ too complex model
- It is clear that empirical risk minimization itself does not include any mechanism to assess bias and variance independently (how should it?); more specifically, if we do not care about model complexity (in particular, if we allow highly or even arbitrarily complex models), ERM has a high risk to produce over-fitted models.

Fundamental Question

- Does the empirical risk converge to the real risk if we increase the number of training samples?
- If so, we could be certain that ERM produces better and better models for increasing sizes of the training set, right?
- If not, ERM seems to have a major problem, because it optimizes according to a wrong quality measure (as the “real” generalization risk is our target).
- This question is central in *Vladimir N. Vapnik’s Statistical Learning Theory*, the most common concepts and results of which we will address now.

Complexity vs. Estimating Empirical Risk: A Simple Setting

Assume that our training algorithm has to choose from a finite set of models $\{g_1, \ldots, g_M\}$, i.e. complexity is limited. Further assume $L(.,.) \in [0, c]$ (in the case of zero-one loss, this holds for $c = 1$).

Theorem. For all $g \in \{g_1, \ldots, g_M\}$, the inequality

$$|R(g) - R_{\text{emp}}(g, \mathbf{Z}_l)| \leq \varepsilon(l, M, \delta)$$

holds with a probability of at least $1 - \delta$ over all possible training sets with l elements, where

$$\varepsilon(l, M, \delta) = c \cdot \sqrt{\frac{\ln M + \ln(2/\delta)}{2l}}.$$

Complexity vs. Estimating Empirical Risk: A Simple Setting (cont'd)

Definition. Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of random variables. We say that the sequence $(X_i)_{i\in\mathbb{N}}$ *converges to a value* x *in probability* if

$$\lim_{i\to\infty} p(|X_i - x| > \varepsilon) = 0$$

for all $\varepsilon > 0$. We denote this with $X_i \xrightarrow[i\to\infty]{P} x$.

Corollary. For all $g \in \{g_1, \ldots, g_M\}$, we have

$$R_{\text{emp}}(g, \mathbf{Z}_l) \xrightarrow[l\to\infty]{P} R(g),$$

i.e. the empirical risk uniformly converges to the actual risk in probability.

Finite Sets of Models?

- There are situations in practice, where model classes are finite, e.g. if all input variables are categorical or discretized (e.g. as in some decision tree architectures or other rule-based approaches).
- The more common case, however, is that the model class is *infinite* (in particular, for support vector machines and neural networks).
- The above results state that the empirical risk can only be a good estimate of the generalization error on a finite set of functions, but it cannot necessarily estimate the generalization error for infinitely many functions simultaneously, not even if we can enlarge the training set arbitrarily.

Consistency of ERM (1/4)

We already know that empirical risk minimization is concerned with *minimizing the empirical error*, which means that, for a given training set $\mathbf{Z}_l$, it determines a parameter setting $\hat{\boldsymbol{w}}(\mathbf{Z}_l)$ such that

$$R_{\text{emp}}(g(.;\hat{\boldsymbol{w}}(\mathbf{Z}_l)),\mathbf{Z}_l) = \min_{\boldsymbol{w}} \boxed{R_{\text{emp}}(g(.;\boldsymbol{w}),\mathbf{Z}_l)}.$$

We may also write this in the following way:

$$\hat{\boldsymbol{w}}(\mathbf{Z}_l) = \arg\min_{\boldsymbol{w}} R_{\text{emp}}(g(.;\boldsymbol{w}),\mathbf{Z}_l)$$

But what do we actually want to optimize? Obviously, the objective value we want to minimize is

$$\boxed{R(g(.;\boldsymbol{w}))}.$$

Consistency of ERM (2/4)

Definition. We say that empirical risk minimization is *consistent* for a given model class $g(.;.)$ (and a given learning task) if the following two convergence statements hold:

$$R(g(.;\hat{\boldsymbol{w}}(\mathbf{Z}_l))) \xrightarrow[l\to\infty]{P} \inf_{\boldsymbol{w}} R(g(.;\boldsymbol{w}))$$

$$R_{\text{emp}}(g(.;\hat{\boldsymbol{w}}(\mathbf{Z}_l)),\mathbf{Z}_l) \xrightarrow[l\to\infty]{P} \inf_{\boldsymbol{w}} R(g(.;\boldsymbol{w}))$$

In a weak sense, the first assertion means that our sequence of solutions $\left(\hat{\boldsymbol{w}}(\mathbf{Z}_l)\right)_{l\to\infty}$ actually converges to an optimal solution $\arg\min_{\boldsymbol{w}} R(g(.;\boldsymbol{w}))$. The second one states that the empirical risk observed for $\hat{\boldsymbol{w}}(\mathbf{Z}_l)$ converges to the minimal risk for $l \to \infty$.

Consistency of ERM (3/4)

It is trivial that the consistency of ERM does not hold in general. Consider, for instance, 1-nearest neighbor classification. The empirical risk

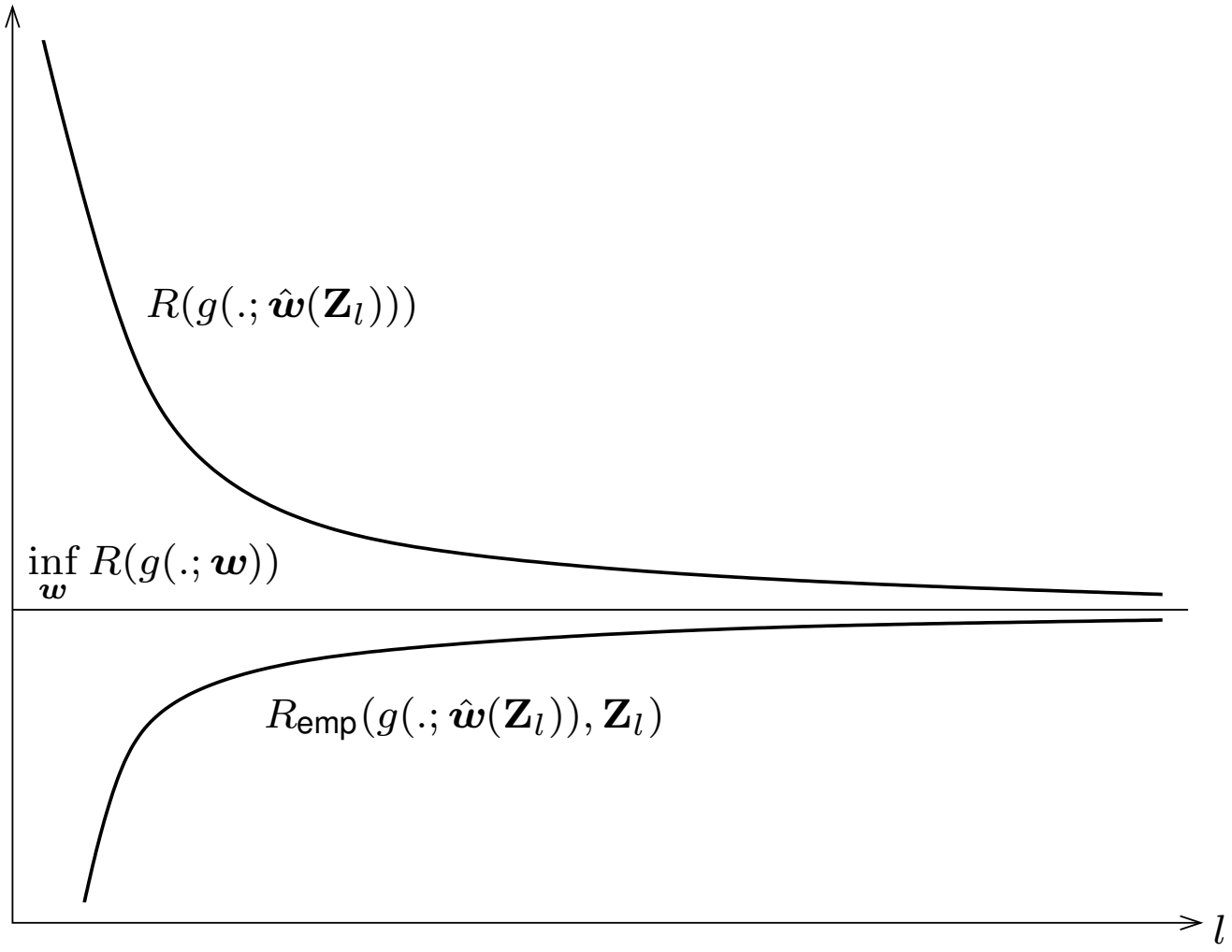
$$R_{\text{emp}}(g_{\text{1-NN}}(.; \mathbf{Z}_l), \mathbf{Z}_l)$$

is always zero, no matter what the values $R(g_{\text{1-NN}}(.; \mathbf{Z}_l))$ are and no matter what

$$\inf_{\substack{\mathbf{Z}_T \subseteq Z \\ \mathbf{Z}_T \text{finite}}} R(g_{\text{1-NN}}(.; \mathbf{Z}_T)),$$

the risk of the best 1-nearest neighbor solution, is.

Consistency of ERM (4/4)



Strict Consistency of ERM

Definition. We say that empirical risk minimization is *strictly consistent* for a given model class $g(.;.)$ (and a given learning task) if the convergence

$$\inf_{\boldsymbol{w}\in\Lambda_g(c)} R_{\text{emp}}(g(.;\boldsymbol{w}),\mathbf{Z}_l) \xrightarrow[l\to\infty]{P} \inf_{\boldsymbol{w}\in\Lambda_g(c)} R(g(.;\boldsymbol{w}))$$

holds for every $c \geq 0$, where

$$\Lambda_g(c) = \{\boldsymbol{w} \mid R(g(.;\boldsymbol{w})) \geq c\}.$$

Lemma. Strict consistency implies consistency.

Agenda

- **Question:** *Under which condition(s) can we guarantee that ERM is (strictly) consistent?*
- ERM is strictly consistent for finite model classes (see above), but that does not help too much.
- We have seen already (even though only intuitively) that the *complexity* of the model class is a major factor.

Shattering Coefficient

Let us assume from here on, that we are dealing with binary classification, i.e. $g(.;.) \in \{-1, +1\}$. For convenience, we will sometimes identify the model class $g(.;.)$ with the set of functions it contains, i.e. $g = \{g(.; \boldsymbol{w}) \mid \boldsymbol{w}\}$.

Definition. Given a model class $g(.;.)$ and a family of l sample inputs $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^l) \in X^l$, the *shattering coefficient* of g for $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^l)$ is defined as

$$\mathcal{N}_g(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^l) = \left|\left\{(g(\boldsymbol{x}^1; \boldsymbol{w}), \ldots, g(\boldsymbol{x}^l; \boldsymbol{w})) \mid \boldsymbol{w}\right\}\right|,$$

i.e. the number of possible labelings of $\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^l\}$ that the model class $g(.;.)$ is able to realize (for any parameter setting $\boldsymbol{w}$).

Shattering Coefficient (cont'd)

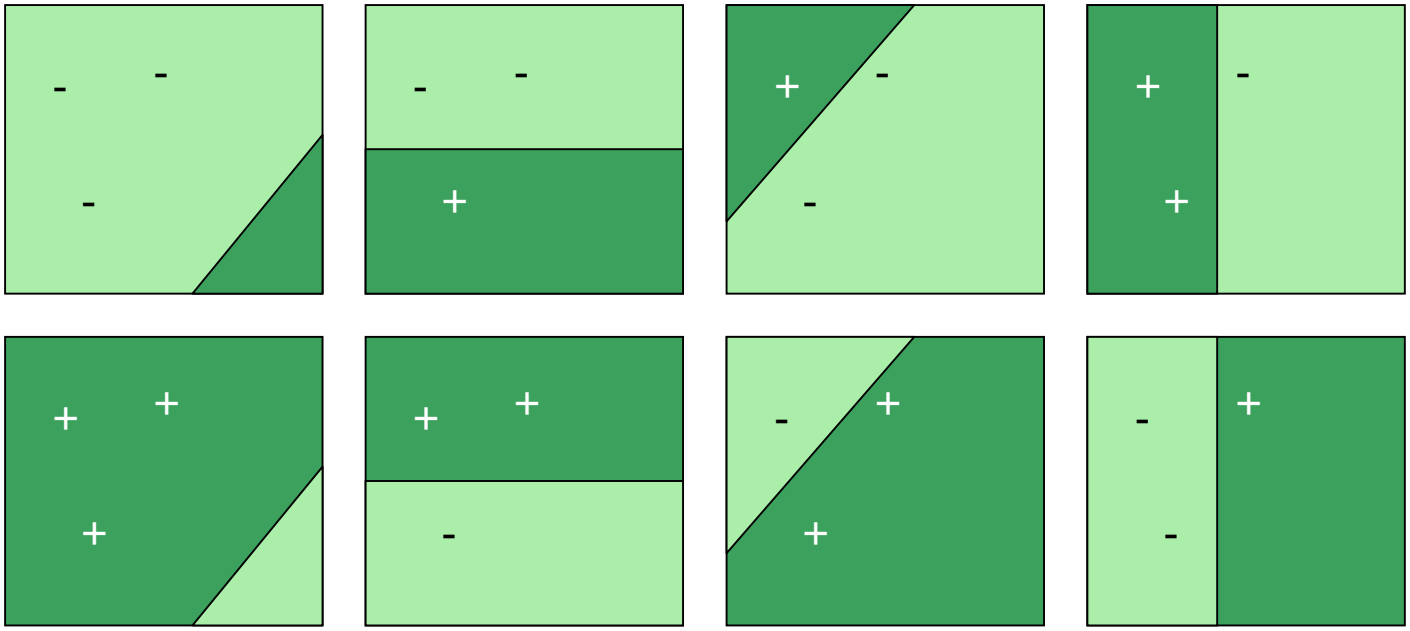
Obviously, if $\mathcal{N}_g(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^l) = 2^l$, $g(.;.)$ can model any labeling of the inputs $\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^l\}$. In this case, we say that $g(.;.)$ *shatters* $\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^l\}$.

Example: Consider $X = \mathbb{R}^2$ and

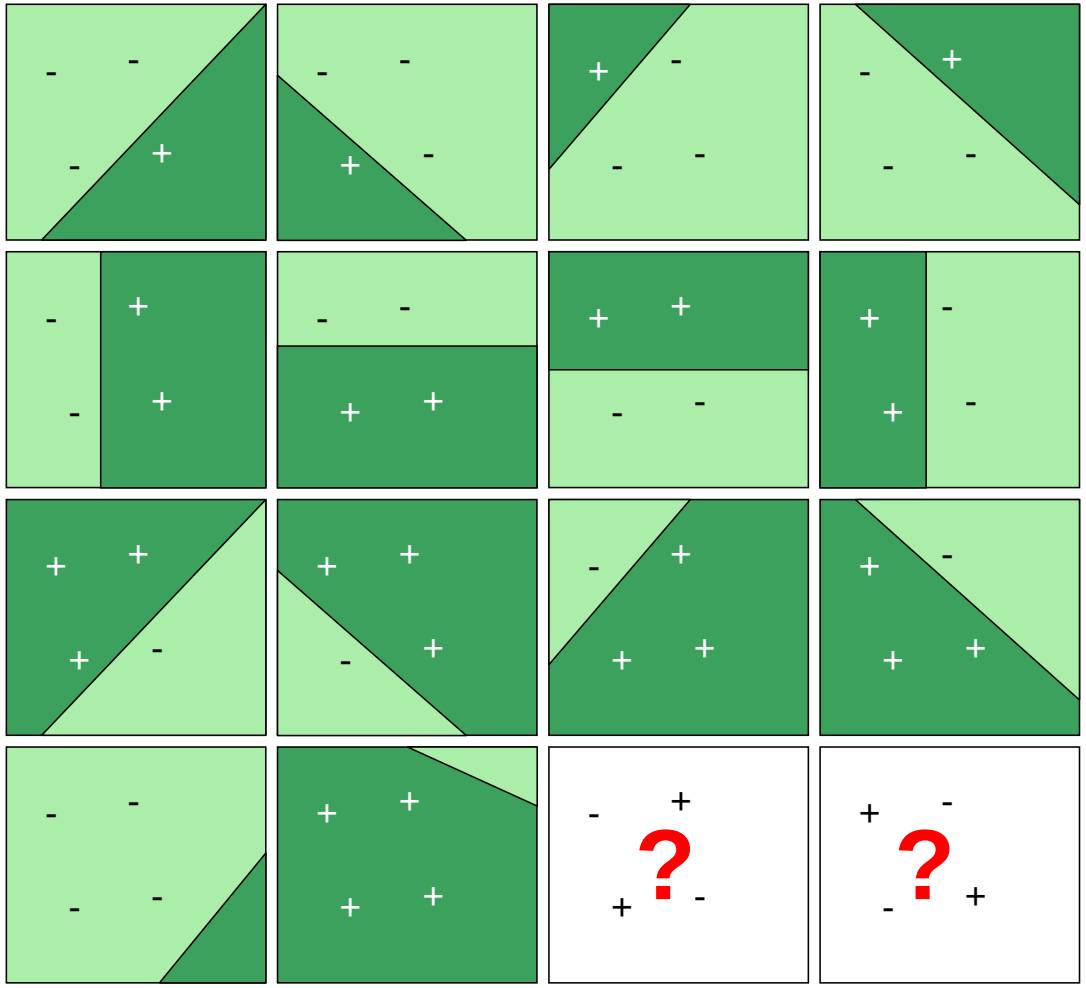
$$g_{\text{lin}}((x_1, x_2); (w_1, w_2, b)) = \begin{cases} 1 & \text{if } w_1 x_1 + w_2 x_2 \geq b, \\ -1 & \text{otherwise,} \end{cases}$$

i.e. linear separation. Then, for any three points $\boldsymbol{x}^1, \boldsymbol{x}^2, \boldsymbol{x}^3$ that are not collinear, we have $\mathcal{N}_{g_{\text{lin}}}(\boldsymbol{x}^1, \boldsymbol{x}^2, \boldsymbol{x}^3) = 8$. For four points $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^4$ arranged as a general tetragon, we obtain $\mathcal{N}_{g_{\text{lin}}}(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^4) = 14$.

Shattering Coefficient Example #1



Shattering Coefficient Example #2



Entropy and Annealed Entropy

Definition. Given a model class $g(.;.)$ and an input distribution $p(\boldsymbol{x})$, the *entropy* of the model class $g(.;.)$ for a given number of points l is defined as

$$\mathcal{H}_g(l) = \mathrm{E}_{(\boldsymbol{x}^1,\ldots,\boldsymbol{x}^l)}\big(\ln \mathcal{N}_g(\boldsymbol{x}^1,\ldots,\boldsymbol{x}^l)\big).$$

The *annealed entropy* of the model class $g(.;.)$ for a given number of points l is defined as

$$\mathcal{H}_g^{\mathrm{ann}}(l) = \ln \mathrm{E}_{(\boldsymbol{x}^1,\ldots,\boldsymbol{x}^l)}\big(\mathcal{N}_g(\boldsymbol{x}^1,\ldots,\boldsymbol{x}^l)\big).$$

Growth Function

Definition. Given a model class $g(.;.)$, the *growth function* of the model class $g(.;.)$ for a given number of points l is defined as

$$\mathcal{G}_g(l) = \ln \max_{(\boldsymbol{x}^1,\ldots,\boldsymbol{x}^l)} \mathcal{N}_g(\boldsymbol{x}^1,\ldots,\boldsymbol{x}^l).$$

Proposition. The following inequalities always hold:

$$\mathcal{H}_g(l) \leq \mathcal{H}_g^{\text{ann}}(l) \leq \mathcal{G}_g(l)$$

Examples

Consider again $X = \mathbb{R}^2$. If we assume that $p(\boldsymbol{x})$ is some uniform distribution over a non-degenerate area $\Omega \subset \mathbb{R}^2$, we can infer

$$\mathcal{H}_{g_{\text{lin}}}(3) = \mathcal{H}^{\text{ann}}_{g_{\text{lin}}}(3) = \mathcal{G}_{g_{\text{lin}}}(3) = \ln 8 = 3 \ln 2.$$

Note that the equalities $\mathcal{H}_{g_{\text{lin}}}(3) = \mathcal{H}^{\text{ann}}_{g_{\text{lin}}}(3) = 3 \ln 2$ hold even though there are configurations (when the three points are collinear or if two or more points coincide), for which $\mathcal{N}_{g_{\text{lin}}}(\boldsymbol{x}^1, \boldsymbol{x}^2, \boldsymbol{x}^3) < 8$ holds. Since these are a null set (having probability zero) in the set of possible triples $\{(\boldsymbol{x}^1, \boldsymbol{x}^2, \boldsymbol{x}^3)\}$, they do not influence the expectations in the definitions of the entropy and annealed entropy.

Examples (cont'd)

With the same assumptions as above, we obtain the following:

$$\mathcal{H}_{g_{\text{lin}}}(4) \overset{!}{=} \mathcal{H}^{\text{ann}}_{g_{\text{lin}}}(4) \overset{!}{=} \mathcal{G}_{g_{\text{lin}}}(4) = \underbrace{\ln 14}_{\approx 2.64} < \underbrace{4 \ln 2}_{\approx 2.77}$$

If the model class consists of a finite set of classification functions $\{g(.;1), \ldots, g(.;M)\}$, then we trivially have that the inequalities

$$\left.\begin{aligned} \mathcal{H}_g(l) \\ \mathcal{H}^{\text{ann}}_g(l) \\ \mathcal{G}_g(l) \end{aligned}\right\} \leq \min(\ln M, l \ln 2) \leq \ln M$$

hold for all $l \in \mathbb{N}$.

Consistency of ERM vs. Complexity (1/3)

Theorem (Sufficient Condition for Consistency of ERM). Consider a given learning task and a model class $g(.;.)$. If

$$\lim_{l \to \infty} \frac{\mathcal{H}_g(l)}{l} = 0$$

holds, then ERM is consistent.

Corollary. The ERM procedure to select the best classification function from a finite set of functions according to the empirical error is consistent.

Consistency of ERM vs. Complexity (2/3)

Definition. We say that ERM procedure has a *fast (exponential) rate of convergence* if there exist positive real constants b, c such that, for all $\varepsilon > 0$, there exists an $l_0 \in \mathbb{N}$ such that the following holds for all $l \geq l_0$:

$$p(\sup_{\boldsymbol{w}} |R(g(.;\boldsymbol{w})) - R_{\text{emp}}(g(.;\boldsymbol{w}), \mathbf{Z}_l)| > \varepsilon) \leq b \exp(-c\varepsilon^2 l).$$

Theorem (Sufficient Condition for Consistency of ERM and Fast Rate of Convergence for Given Distribution). Consider a given learning task and a model class $g(.;.)$. If

$$\lim_{l \to \infty} \frac{\mathcal{H}_g^{\text{ann}}(l)}{l} = 0$$

holds, then ERM is consistent and has a fast rate of convergence.

Consistency of ERM vs. Complexity (3/3)

Theorem (Characterization of Consistency of ERM and Fast Rate of Convergence for Any Distribution). Consider a given learning task and a model class $g(.;.)$. Then ERM is consistent and has a fast rate of convergence *if and only if* the following holds:

$$\lim_{l \to \infty} \frac{\mathcal{G}_g(l)}{l} = 0$$

Note that, in the above theorem, the distribution $p(\boldsymbol{x})$ does *not* occur. Hence, the growth function allows us to draw conclusions about the consistency and fast convergence of ERM for a given model class *independently of the learning task.*

The Vapnik-Chervonenkis (VC) Dimension

Definition. The *Vapnik-Chervonenkis dimension (VC dimension)* of a model class $g(.;.)$ is defined as

$$d_{\mathsf{vc}}(g) = \sup\left\{l \in \mathbb{N} \mid \mathcal{G}_g(l) = l \ln 2\right\}.$$

From the definition of the growth function, it is easy to see that

$$d_{\mathsf{vc}}(g) = \sup\{l \in \mathbb{N} \mid \exists((\boldsymbol{x}^1, \ldots, \boldsymbol{x}^l) \in X^l)\, \mathcal{N}_g(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^l) = 2^l\},$$

i.e. the VC dimension is the largest number l for which a configuration of l samples can be found that can be shattered by a model from the model class g. If this works for all l, the VC dimension is ∞.

VC Dimension: Examples

- For $X = \mathbb{R}^2$, we have $d_{\text{VC}}(g_{\text{lin}}) = 3$.
- For $X = \mathbb{R}^d$, we have $d_{\text{VC}}(g_{\text{lin}}) = d + 1$ (where g_{lin} is generalized to the p-dimensional case in the obvious way).
- For $X = \mathbb{R}$ and any model class that contains only non-decreasing functions, we have $d_{\text{VC}}(g) = 1$, regardless of how many parameters are necessary to parametrize g.
- For $X = \mathbb{R}$ and $g_{\sin}(x, w) = \text{sign}(\sin(wx))$, we obtain $d_{\text{VC}}(g_{\sin}) = \infty$, although $g_{\sin}$ only depends on one parameter.

We conclude that there is not necessarily a dependency between the VC dimension and the number of parameters which describe a model class.

The VC Dimension Bounds the Growth Function

Theorem. The following holds for a given model class $g(.;.)$:

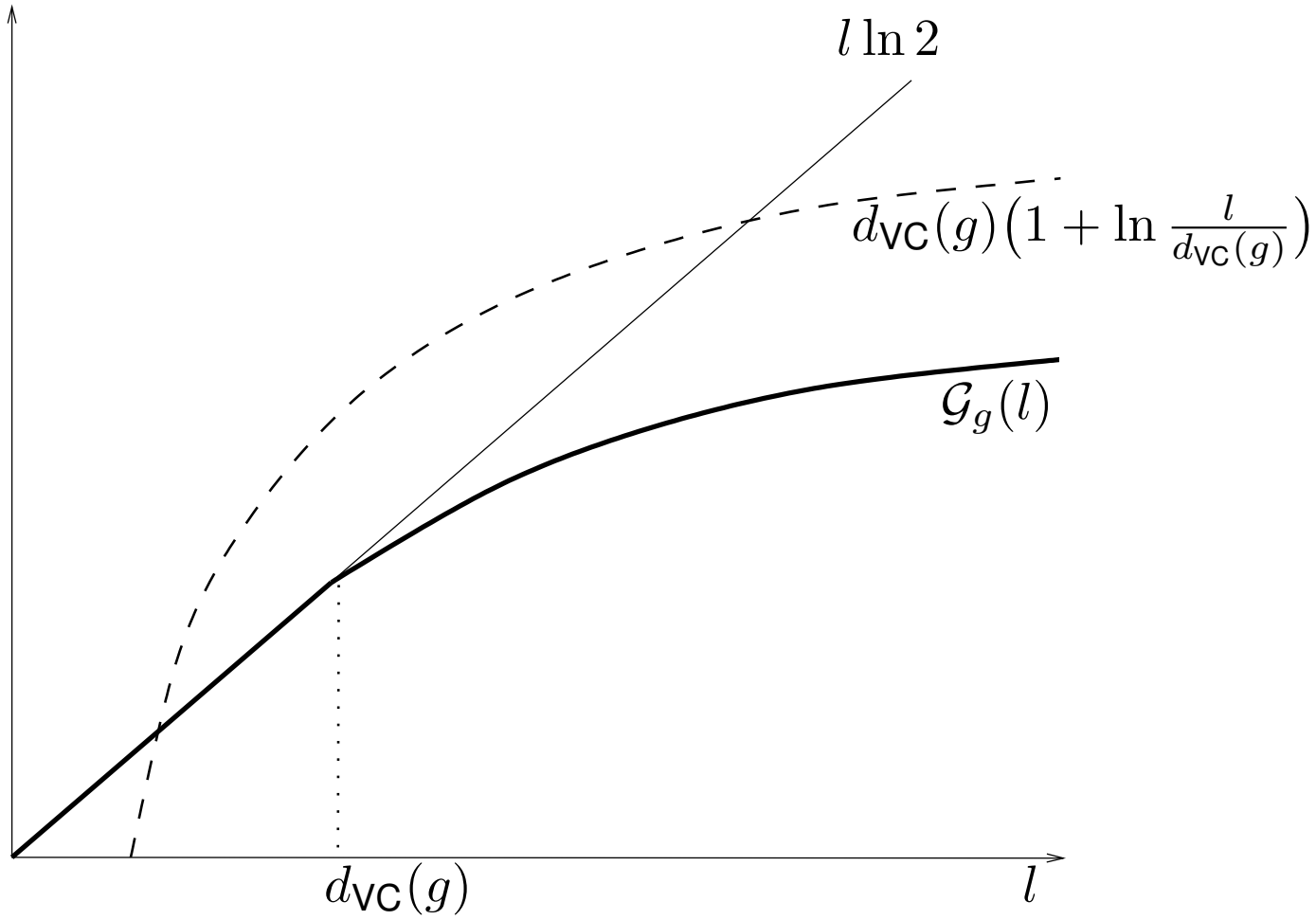
$$\mathcal{G}_g(l) \begin{cases} = l \ln 2 & \text{if } l \leq d_{\mathsf{VC}}(g) \\ \leq \min\left(l \ln 2, d_{\mathsf{VC}}(g)\left(1 + \ln \frac{l}{d_{\mathsf{VC}}(g)}\right)\right) & \text{if } l > d_{\mathsf{VC}}(g) \end{cases}$$

Corollary. If a given model class $g(.;.)$ has a finite VC dimension $d_{\mathsf{VC}}(g) < \infty$, then

$$\lim_{l \to \infty} \frac{\mathcal{G}_g(l)}{l} = 0$$

holds, hence, ERM is consistent and has a fast rate of convergence.

The VC Dimension Bounds the Growth Function (cont'd)



Error Bounds

Theorem. If we consider all possible l-element training sets $\mathbf{Z}_l$, the following holds with a probability of at least $1 - \delta$:

$$R(g(.; \boldsymbol{w}(\mathbf{Z}_l))) \leq R_{\text{emp}}(g(.; \boldsymbol{w}(\mathbf{Z}_l)), \mathbf{Z}_l) + \sqrt{\varepsilon(l, g, \delta)}$$

With a probability of at least $1 - 2\delta$, the following holds:

$$R(g(.; \boldsymbol{w}(\mathbf{Z}_l))) \leq \inf_{\boldsymbol{w}} R(g(.; \boldsymbol{w})) + \sqrt{\varepsilon(l, g, \delta)} + \sqrt{\frac{-\ln \delta}{l}}$$

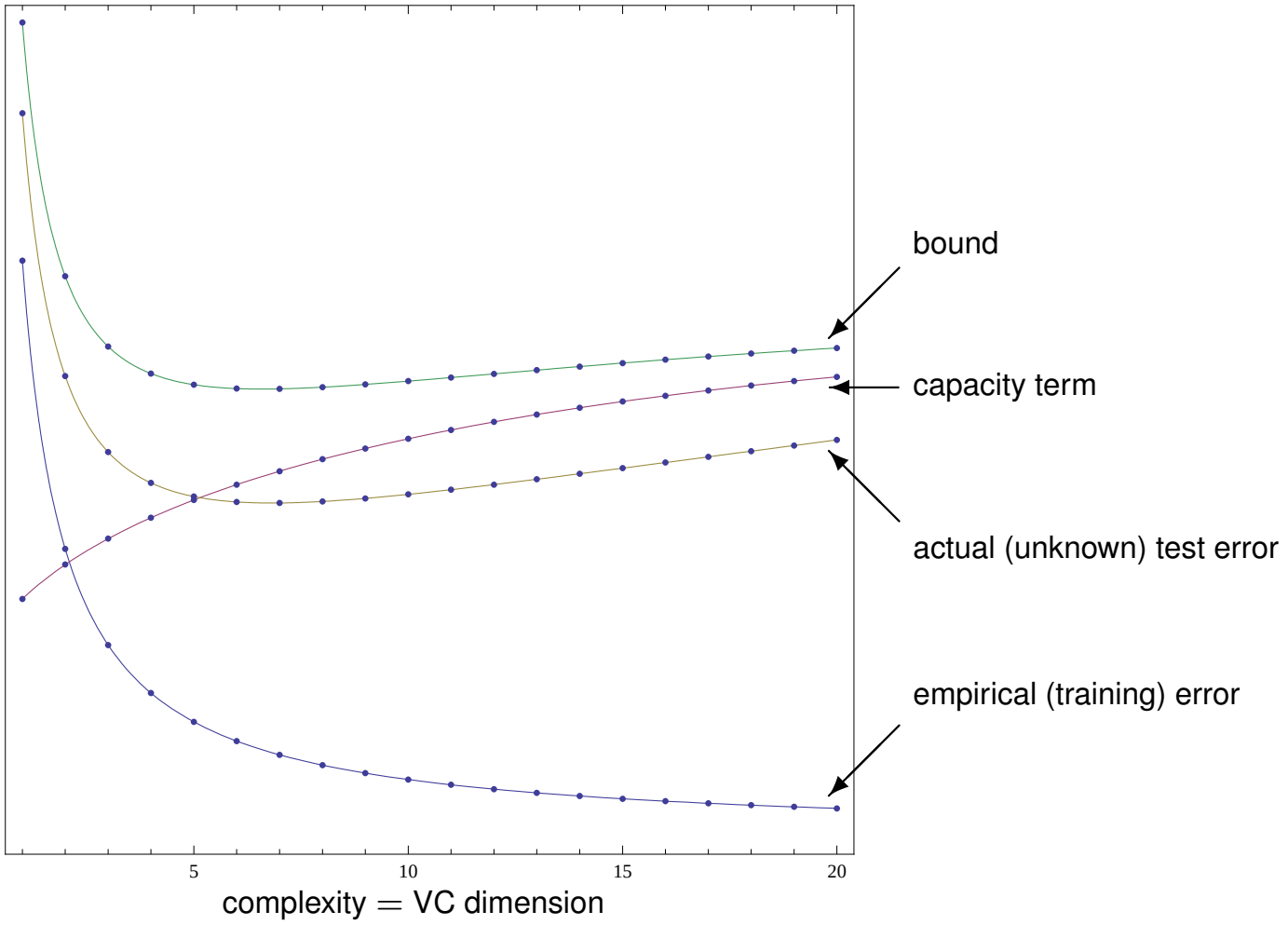
Error Bounds (cont'd)

Theorem (cont'd). In the inequalities above, we can use the following functions $\varepsilon(l, g, \delta)$:

$$\varepsilon(l,g,\delta) = \begin{cases} \frac{8}{l}\left(\mathcal{H}_g^{\text{ann}}(2l) + \ln(4/\delta)\right) & \text{for given } p(\boldsymbol{x}) \\ \frac{8}{l}\left(\mathcal{G}_g(2l) + \ln(4/\delta)\right) & \left.\begin{matrix}\\ \end{matrix}\right\} \text{for any } p(\boldsymbol{x}) \\ \frac{8}{l}\left(d_{\text{VC}}(g)(1 + \ln(2l/d_{\text{VC}}(g))) + \ln(4/\delta)\right) & \end{cases}$$

Note that numerous bounds of this flavor are available in literature, some tighter, some looser, some making special assumptions. All results have in common that they provide bounds for the deviation of the empirical error from the actual risk. Such bounds are often termed *capacity terms* or *VC confidence*.

Error Bounds Visualized



Structural Risk Minimization (SRM)

- *Structural Risk Minimization (SRM)* is an alternative learning scheme proposed by Vapnik.
- Instead of considering only the empirical error as in ERM, the idea is to minimize an estimate of the test error (given as sum of empirical error and a capacity term).
- We consider a nested family of model classes

$$g_1 \subset g_2 \subset \cdots \subset g_n \subset \ldots$$

such that

$$d_{\mathsf{VC}}(g_1) \leq d_{\mathsf{VC}}(g_2) \leq \cdots \leq d_{\mathsf{VC}}(g_n) \leq \cdots$$

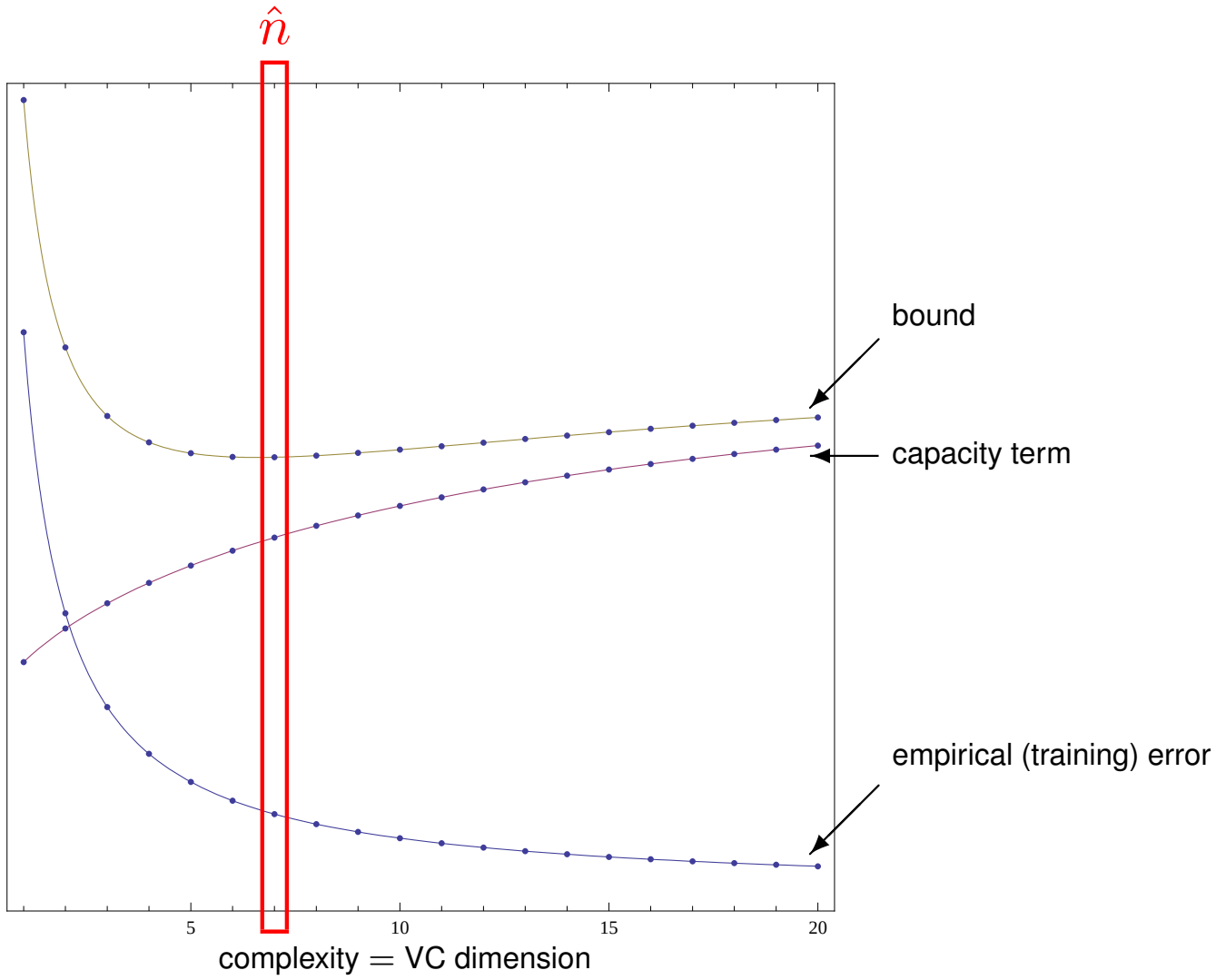
Structural Risk Minimization (SRM) (cont'd)

- For each $n = 1, 2, \ldots,$ we select a model from the model class g_n according to ERM, i.e., we determine solutions such that

$$\hat{\boldsymbol{w}}_n(\mathbf{Z}_l) = \arg\min_{\boldsymbol{w}} R_{\text{emp}}(g_n(.; \boldsymbol{w}), \mathbf{Z}_l).$$

- Finally, we select the level $\hat{n}$ such that the sum of the empirical error $R_{\text{emp}}(g_n(.; \hat{\boldsymbol{w}}_n(\mathbf{Z}_l)), \mathbf{Z}_l)$ and the capacity term for g_n is minimal. The finally selected model is then $g_{\hat{n}}(.; \hat{\boldsymbol{w}}_{\hat{n}}(\mathbf{Z}_l))$.
- In the case of very low (or even zero) empirical error, SRM is concerned with minimizing complexity (compare with *Occam's razor*).

Structural Risk Minimization Visualized



Consistency of Structural Risk Minimization

Theorem. The convergence

$$R(g_{\hat{n}(\mathbf{z}_l)}(.;\hat{\boldsymbol{w}}_{\hat{n}(\mathbf{z}_l)}(\mathbf{Z}_l))) \xrightarrow[l\to\infty]{P} \underbrace{\inf_{g\in\bar{g}} R(g)}_{=R_{\min}}$$

holds, where $\bar{g} = \overline{\bigcup_{n\in\mathbb{N}} g_n}$. The asymptotic rate of convergence is

$$V(l) = \Big|\inf_{g\in g_{\hat{n}(\mathbf{z}_l)}} R(g) - R_{\min}\Big| + \sqrt{\frac{d_{\mathrm{VC}}(g_{\hat{n}(\mathbf{z}_l)})\cdot \ln l}{l}},$$

i.e.

$$p\Big(\lim_{l\to\infty}\sup V^{-1}(l)\Big|R(g_{\hat{n}(\mathbf{z}_l)}(.;\hat{\boldsymbol{w}}_{\hat{n}(\mathbf{z}_l)}(\mathbf{Z}_l))) - R_{\min}\Big| < \infty\Big) = 1$$

holds provided that $\frac{d_{\mathrm{VC}}(g_{\hat{n}(\mathbf{z}_l)})\cdot\ln l}{l} \xrightarrow{l\to\infty} 0$ (even if $\hat{n}(l) \xrightarrow{l\to\infty} \infty$, otherwise this is trivially fulfilled).

Structural Risk Minimization: Remarks and Caveats

- SRM does nothing else but ERM for each model class g_n from a family of model classes $g_1, g_2, \ldots$ (since the VC confidence term is constant if we fix a model class g_n).
- The choice of the complexity level $\hat{n}(\mathbf{Z}_l)$, however, is not based on the risk itself (which we normally do not know), but on a bound that may be rather loose. Thus, for a given training set $\mathbf{Z}_l$, the chosen complexity level $\hat{n}(\mathbf{Z}_l)$ need not be optimal.
- Consistency only tells us that SRM produces solutions that converge to an optimal one.

Statistical Learning Theory: Summary

- We introduced entropies, growth functions, and the VC dimensions as meaningful *measures of complexity* of a given model class.
- We were able to formulate the *consistency of ERM* and sufficient (and necessary) conditions for consistency on the basis of the complexity of the model class considered.
- *Structural Risk Minimization (SRM)* has been introduced as a means to explicitly address the bias-variance trade-off on the basis of the empirical error only.

Statistical Learning Theory: General Caveats

- All the results formulated here are restricted to binary classification. For regression problems, generalizations of the complexity measures have to be used (fat shattering, ε-covers).
- The model class g has always been considered a black box. In particular, we have not addressed the question how powerful/appropriate the model class g is. The reference to which we compared all estimates was $R_{\min} = \inf_{\boldsymbol{w}} R(g(.;\boldsymbol{w}))$, but we have no results that actually tell us how good $R_{\min}$ is, i.e. how well the model class is actually able to fit the data.
- Almost all results hold "in probability". There may always exist data sets for which the worst possible case occurs (*No Free Lunch*).