

Statistical Learning Theory

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Pattern Recognition and Machine Learning

Lecture 07 Statistical Learning Theory

**Chizhi Chris ZHANG**

zhangchizhi@ciomp.ac.cn

Advanced Computing and Digital Technology Research Center

University of Chinese Academy of Sciences

Autumn 2026

Lecture plan and time allocation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Topic	Minutes
Learning framework	22.5
Risk and empirical risk	22.5
Test error estimation	22.5
Regularization	22.5
Bias and variance	22.5
Generalization theory	22.5

A fitted rule and future observations

Separate observed risk from population risk; examine stability and selection.

The numerical steps are shown on the slides.



A practice winner can lose on new questions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Two classifiers answer the same 20 practice questions. Then they face 20 new questions from the same source. These are illustrative results.

Which one should we trust next?

Rule	Practice	New
A: memorize	0/20	7/20
B: shared rule	2/20	3/20

A wins on the records used to fit it. B makes fewer errors on these new records.

What changed?

The practice answers helped choose the rule. They therefore give an optimistic view of that rule.

The new answers check performance after fitting. If we use them to select B, a separate final test is still needed.

A future-risk claim also needs a stated sampling assumption and model family.

Specify the learning problem before fitting

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Parameters w are fitted. Degree m is selected. An assumption of independent, identically distributed pairs links the observed sample to the target population.

Object	Running regression example
$\mathcal{X}, \mathcal{Y}$	Input $x \in [0, 1]$;responses $y \in \mathbb{R}$
$P(X, Y)$	Unknown joint law producing future input–response pairs
$\mathcal{H}$	Polynomial $f_w(x) = \sum_{j=0}^m w_j x^j$
$L(f(x), y)$	Square error $(f(x) - y)^2$
Learner	From the sample D Fitted coefficient $\hat{w}_D$ fitting process



Loss makes different mistakes comparable

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Squared loss makes a threefold larger residual nine times more costly. The unit becomes the square of the response unit.

Prediction / truth	Absolute loss	Squared loss
$\hat{y} = 8, y = 10$	2	4
$\hat{y} = 4, y = 10$	6	36

$$L(\hat{y}, y) = \mathbf{1}\{\hat{y} \neq y\}$$



The best average prediction need not be an observed outcome

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

At the same departure time x , a trip takes 8 or 12 minutes, each with probability $1/2$. Which constant prediction minimizes expected squared error?

Compare two predictions

Predict a	Squared losses	Mean loss
8	0, 16	8
10	4, 4	4

$$m = E[Y | x] = (8 + 12)/2 = 10.$$

$$\text{Var}(Y | x) = (4 + 4)/2 = 4.$$

Predicting 10 minimizes average loss, although neither trip actually takes 10 minutes.

Why the mean wins

Write $a - Y = (a - m) + (m - Y)$.

$$\begin{aligned} E[(a - Y)^2 | x] &= (a - m)^2 \\ &\quad + 2(a - m) \underbrace{E[m - Y | x]}_0 \\ &\quad + E[(m - Y)^2 | x] \\ &= (a - m)^2 + \text{Var}(Y | x). \end{aligned}$$

Only $(a - m)^2$ changes with a . It is smallest at $a = m$.

Expected risk and empirical risk answer different questions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Empirical risk averages observed losses. Expected risk averages over the target population. The population distribution is usually unknown, so these quantities require different evidence.

Key idea

Squared residuals 1, 4, 0, 9 give empirical risk $14/4 = 3.5$. Another sample could give a different average for the same fixed rule. Large samples reduce sampling fluctuations only when sampling matches the target population.

$$R(f) = \mathbb{E}_{(X,Y) \sim P}[L(f(X), Y)], \quad \hat{R}_D(f) = \frac{1}{n} \sum_{i=1}^n L(f(x_i), y_i).$$

$$14/4 = 3.5$$

Empirical risk minimization searches within a chosen family

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Empirical risk minimization chooses the smallest average training loss within a specified hypothesis set $\mathcal{H}$.

$$\hat{f}_D = \arg \min_{f \in \mathcal{H}} \hat{R}_D(f).$$

Four labeled records; two candidate classifiers

Candidate	Observed zero-one losses	Mean loss
Rule A	(0, 1, 0, 1)	$2/4 = .5$
Rule B	(0, 0, 0, 0)	$0/4 = 0$

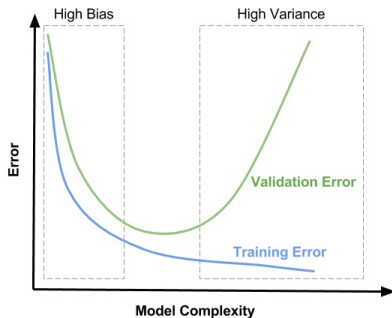


ERM selects B from $\mathcal{H} = \{A, B\}$. It has checked the observed labels. It has not measured B's loss on the future population.

Changing $\mathcal{H}$ changes what can be selected. Training success depends on both the allowed rules and the observed sample.

Read training and validation as complexity changes

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER



The horizontal axis is model complexity. Follow each error curve before comparing them.

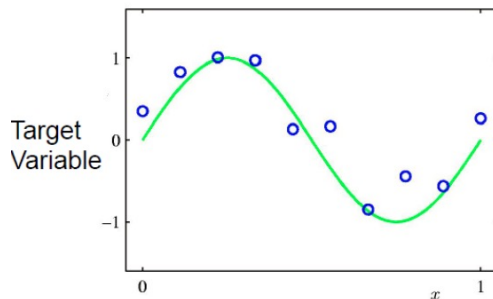


A controlled example separates signal from noise

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

An accurate optimizer can still fit the wrong family. A flexible family can still be estimated poorly. Diagnose the failing step before adding complexity. Successful learning requires structure that connects observed and unobserved inputs. A perfect lookup table supplies no such connection. Green signal is the conditional mean; observed points include random noise.

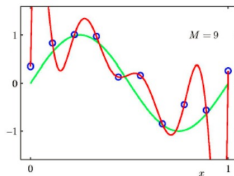
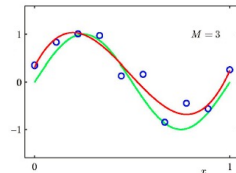
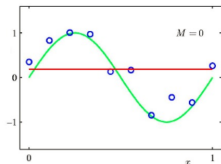
$$y = \sin(2\pi x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.3^2).$$



Compare three fits on the same ten observations

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

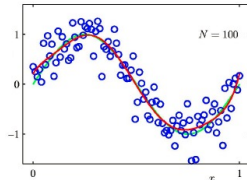
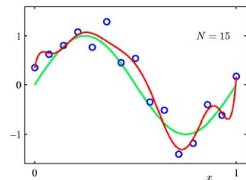
Degree 3 follows the main trend without chasing every noise displacement.



More data changes the usable amount of flexibility

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Degree nine stays fixed. More observations constrain the curve between isolated points and can reduce sample sensitivity.



Cross-validation rotates the held-out observations

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A fitted rule needs evidence beyond its training losses.

For 1,000 independent records, an illustrative split is 600/200/200. A temporal task instead trains on earlier records and tests on later ones. Patient, device, or subject groups must remain in a single side of a split when predicting new groups. Stratification preserves class proportions; it does not remove group dependence.

The same nine development records form three folds. Each model fits two groups and predicts the third. Equal fold sizes allow the three mean losses to be averaged directly.

$$\hat{R}_{CV} = \frac{\sum_{k=1}^3 \sum_{i \in I_k} L(f_{-k}(x_i), y_i)}{\sum_{k=1}^3 |I_k|}.$$
$$(0.20 + 0.32 + 0.26)/3 = 0.26.$$

With unequal fold sizes, weight each mean loss by its validation count.

Check. Development folds select the settings. An independent test set evaluates the selected complete procedure.

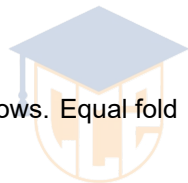
Run three validation folds without moving fitted statistics

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Groups are $A = \{1, 2, 3\}$, $B = \{4, 5, 6\}$, $C = \{7, 8, 9\}$. The final test set is separate.

Fold	Training rows	Validation rows	Validation MSE
1	$B + C$	A	0.20
2	$A + C$	B	0.32
3	$A + B$	C	0.26

Fit the scaler and model on that fold's training rows before predicting validation rows. Equal fold sizes give mean MSE $(0.20 + 0.32 + 0.26)/3 = 0.26$.



Worked example: repair an evaluation design

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Split first. Fit transformations only on the training part; apply those fitted transformations to its held-out part.

A two-point accuracy difference can be smaller than sampling uncertainty. Prefer stable comparisons and evaluate the frozen procedure independently.

$$n = 200 \quad \hat{p} = 0.10$$

$$\sqrt{\hat{p}(1 - \hat{p})/n} = 0.021$$

Keep patient groups together; fit preprocessing inside training folds; tune on development folds; freeze choices before final testing. A smaller honest score is more informative than a leaked high score.



Regularization trades training fit for a controlled rule

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Add a penalty and solve one coefficient completely before returning to several dimensions.

The penalty discourages particular kinds of flexibility; it is not an extra observed prediction error.

$$J(w) = \underbrace{\frac{1}{n} \sum_i (f_w(x_i) - y_i)^2}_{\text{observed fit}} + \lambda \underbrace{\Omega(w)}_{\text{penalty term}}, \quad \lambda \geq 0.$$

$$\Omega(w) = \sum_{j \geq 1} w_j^2$$



Penalty and constraint express related preferences

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Two formulations express a preference for smaller coefficients:

$$\min_w \hat{R}(w) + \lambda \Omega(w), \quad \min_w \hat{R}(w) \text{ subject to } \Omega(w) \leq t.$$

Charge a price: $\lambda = 1$

$$J(w) = (w - 3)^2 + w^2.$$

$$J'(w) = 2(w - 3) + 2w = 0 \Rightarrow w = 1.5.$$

Set a budget: $t = 2.25$

$$\min_w (w - 3)^2 \text{ subject to } w^2 \leq 2.25.$$

Allowed values lie in $[-1.5, 1.5]$. The closest allowed value to 3 is $w = 1.5$.

This matched price and budget give the same solution. The numerical values λ and t have different meanings; arbitrary pairs need not match.

L2 shrinkage can be calculated in one line

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

The fit term and the L2 penalty pull the coefficient in opposite directions. Solving their derivative balance gives $w = 3/(1 + \lambda)$.

$$J(w) = (w - 3)^2 + \lambda w^2,$$

$$J'(w) = \underbrace{2(w - 3)}_{\text{fit term}} + \underbrace{2\lambda w}_{\text{penalty term}}.$$

$$2(w - 3) + 2\lambda w = 0$$

$$\Rightarrow w - 3 + \lambda w = 0$$

$$\Rightarrow (1 + \lambda)w = 3$$

$$\Rightarrow w = \frac{3}{1 + \lambda}.$$

$$J''(w) = 2(1 + \lambda) > 0 \quad \lambda = 1$$

$$w = 1.5$$

For the same fitting term, $\hat{w} = 3/(1 + \lambda)$. Increasing the L2 penalty shrinks this coefficient smoothly. This decoupled example does not describe every path of correlated multivariate coefficients.

λ	0	1	5
$\hat{w}$	3	1.5	0.5

A stronger penalty moves this coefficient closer to zero. In this example, finite λ does not make this coefficient exactly zero.

L1 can set a coefficient exactly to zero

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Minimize $J(w) = \frac{1}{2}(w - 3)^2 + \lambda|w|$ for $\lambda \geq 0$. Check positive and negative regions, then the corner at zero.
The subgradient permits zero when $\lambda \geq 3$.

$$w > 0: J'(w) = w - 3 + \lambda = 0 \Rightarrow w = 3 - \lambda \quad (\lambda < 3),$$

$$w < 0: J'(w) = w - 3 - \lambda = 0 \Rightarrow w = 3 + \lambda > 0 \quad \text{contradiction},$$

$$w = 0: 0 \in -3 + \lambda[-1, 1] \iff \lambda \geq 3.$$

Worked calculation

$$\hat{w} = \max(3 - \lambda, 0), \quad \lambda = 1 \Rightarrow \hat{w} = 2, \quad \lambda = 4 \Rightarrow \hat{w} = 0.$$

The subgradient at zero is an interval, allowing a whole range of penalty strengths to select zero.

Scaling and the intercept change what the penalty means

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A unit increase now means one training-set standard deviation. Coefficient penalties compare similar scales.

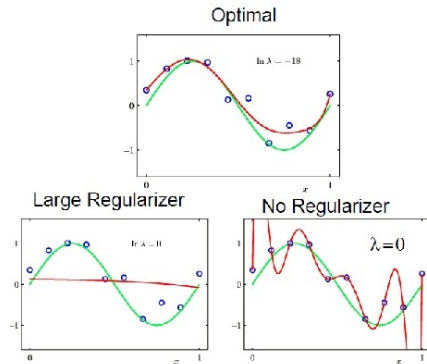
$$z_j = \frac{x_j - \mu_j}{s_j}, \quad f(\mathbf{z}) = b + \sum_j w_j z_j.$$



The same degree-nine family can produce different fits

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Degree stays fixed; only λ changes.
Choose it using held-out loss rather than the visual appeal of the training fit.



Choose a stable validation region rather than the training minimum

The lowest training error favors zero penalty. Validation favors moderate penalties; 0.115 and 0.110 may be indistinguishable given fold variation. Inspect stability, refit the selected setting on development data, then evaluate the fixed procedure on the independent test set.

λ	Train MSE	Mean validation MSE
0	0.000	0.420
10^{-3}	0.018	0.115
10^{-2}	0.041	0.110
10^{-1}	0.120	0.160

A tighter coefficient prior produces stronger shrinkage

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A zero-centered coefficient prior

Assume independent Gaussian observation noise and a Gaussian prior:

$$y_i | w \sim \mathcal{N}(f_w(x_i), \sigma^2), \quad w_j \sim \mathcal{N}(0, \tau^2).$$

For fixed σ, τ , the negative log posterior is

$$\frac{\sum_i (y_i - f_w(x_i))^2}{2\sigma^2} + \frac{\sum_j w_j^2}{2\tau^2} + \text{constant}.$$

Multiplying by $2\sigma^2$ gives an unaveraged squared-error sum plus $\lambda \sum_j w_j^2$, where $\lambda = \sigma^2/\tau^2$.

Degrees, tree depth and early stopping limit fitting in other ways. None repairs a leaking evaluation split.

Reuse the one-coefficient example

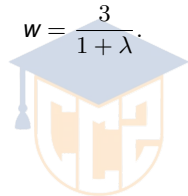
Take the data-fit term $(w - 3)^2$ and $\sigma^2 = 1$.

$$J(w) = (w - 3)^2 + \lambda w^2.$$

$$2(w - 3) + 2\lambda w = 0, \quad w = \frac{3}{1 + \lambda}.$$

τ^2	λ	Fitted w
1	1	1.5
.25	4	.6

A tighter prior gives stronger shrinkage toward zero.



The same machine can receive different predictions after refitting

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The machine setting stays the same.

Refit a prediction rule on four equally likely training sets. Assume the true mean response at this setting is $f^*(x) = 1$ in normalized units.

Training set	1	2	3	4
Rule A prediction	.9	.9	.9	.9
Rule B prediction	.6	.8	1.0	1.2

Both average to .9. A is stable but misses the mean; B has the same average mismatch and also varies across fits.

Separate two questions

$$f^*(x) = E[Y | x], \quad \bar{f}(x) = E_D[\hat{f}_D(x)].$$

Average mismatch: $.9 - 1 = -.1$ for both rules.

Fit variation: A has zero variance; B has positive variance.

The next two slides derive the decomposition and calculate all three terms for B.

Derive the squared-error decomposition

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Fix one input x .

The training set D changes the fitted prediction $\hat{f}_D$. Let $m = E_D[\hat{f}_D]$, $f^* = E[Y | x]$, and $Y = f^* + \epsilon$.

$$Y - \hat{f}_D = \underbrace{\epsilon}_{\text{test noise}} - \underbrace{(\hat{f}_D - m)}_{\text{fit variation}} - \underbrace{(m - f^*)}_{\text{systematic bias}}.$$

Assume zero-mean test noise, independent of the training set. Also $E_D[\hat{f}_D - m] = 0$.

$$\begin{aligned} E[(Y - \hat{f}_D)^2] &= E[\epsilon^2] + E_D[(\hat{f}_D - m)^2] + (m - f^*)^2 \\ &\quad - 2E[\epsilon(\hat{f}_D - m)] - 2(m - f^*)E[\epsilon] \\ &\quad + 2(m - f^*)E_D[\hat{f}_D - m] \\ &= \underbrace{\sigma_\epsilon^2}_{\text{noise}} + \underbrace{\text{Var}_D(\hat{f}_D)}_{\text{variance}} + \underbrace{(m - f^*)^2}_{\text{squared bias}}. \end{aligned}$$

The three cross terms vanish under the stated conditions. This is an average over possible training sets and a new target, not a decomposition of one observed residual.



A numerical decomposition makes the three terms tangible

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

At a fixed input, suppose $f^*(x) = 1$, test-noise variance is 0.09, and four equally likely fitted predictions are 0.6, 0.8, 1.0, 1.2.

$$f^*(x) = 1$$

$$\bar{f} = 0.9, \quad \text{Bias}^2 = (0.9 - 1)^2 = 0.01,$$

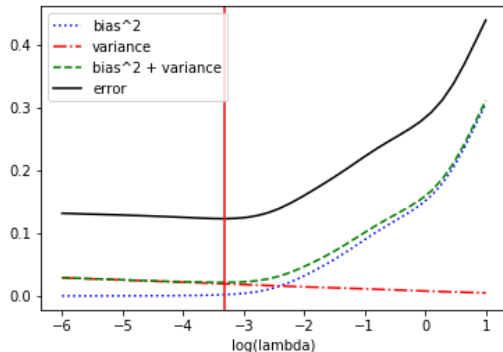
$$\text{Variance} = \frac{(-0.3)^2 + (-0.1)^2 + (0.1)^2 + (0.3)^2}{4} = 0.05.$$

$$\mathbb{E}[(\hat{f} - Y)^2] = 0.01 + 0.05 + 0.09 = 0.15.$$



Track bias, variance and total prediction error

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER



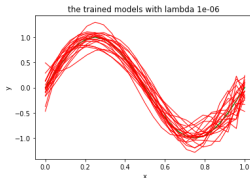
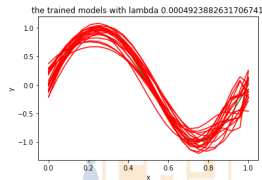
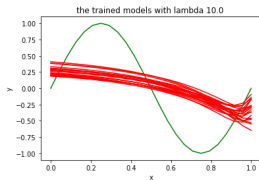
The horizontal axis is log penalty strength. Compare squared bias, variance and their sum. The preferred region balances the components in this example rather than minimizing either one alone.



Repeated fits reveal stability and systematic mismatch

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Closely clustered curves can still miss the true signal: low variance, high bias.



The bias–variance tradeoff is a guide, not an automatic law

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The true conditional mean and all possible training samples are unavailable. Diagnose using observable training/validation behavior; do not label a single numerical gap as an exact variance estimate.

$$\text{Expected squared error} = \text{bias}^2 + \text{variance} + \text{noise} .$$



Learning curves vary sample size, not model degree

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

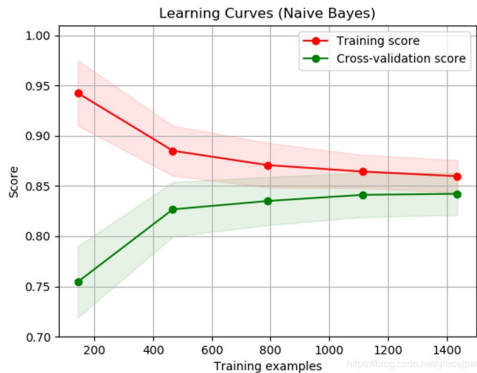
Fit the same model and hyperparameters on increasing training subsets; evaluate each fitted model on held-out observations.

Training size	Train MSE	Valid MSE	Gap
20	0.02	0.40	0.38
80	0.06	0.22	0.16
320	0.09	0.13	0.04



Read scores as training sample size increases

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER



The horizontal axis is training sample size. These curves show scores, so higher is better. Training and validation scores approach each other as evidence increases. Shading shows variation across evaluations.



Worked diagnosis: choose the next intervention

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

These are hypotheses to test, not diagnoses established by a single graph. First confirm optimization is adequate and the held-out observations were truly excluded.

A suggests a restrictive representation if optimization has converged. B suggests unstable fitting; more data help. C is the current validation choice.

Candidate	Train MSE	Valid MSE	With more data
A: degree 1	.42	.44	Little change
B: degree 9, no penalty	0	.39	Validation error decreases to .15
C: degree 9, penalty	.06	.12	stabilizes at .11 Near.

Check. Compare on the same observations and test C on untouched data. More iterations alone do not expand A's linear family.

Convert a probability bound into an error tolerance

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A test error is an average of bounded, independent losses for a fixed rule.

Key idea

For a classifier fixed independently of an i.i.d. sample, let $Z_i = \mathbf{1}\{f(x_i) \neq y_i\} \in [0, 1]$.

$$Z_i = \mathbf{1}\{f(x_i) \neq y_i\} \in [0, 1]$$

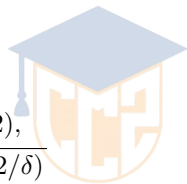
$$\hat{R}(f) = \frac{1}{n} \sum_i Z_i, \quad R(f) = \mathbb{E}[Z_i].$$

$$P(|\hat{R}(f) - R(f)| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}.$$

With $n = 200$ and $\delta = 0.05$, $\varepsilon \approx 0.096$. The procedure $\hat{R} + .096$ covers $R(f)$ on at least 95 percent of repeated independent test samples. For observed error .10, its computed upper limit is .196.

$$\begin{aligned} 2e^{-2n\varepsilon^2} &= \delta, \\ -2n\varepsilon^2 &= \log(\delta/2), \\ \varepsilon &= \sqrt{\frac{\log(2/\delta)}{2n}}. \end{aligned}$$

$n = 200, \delta = 0.05$, and
 $\varepsilon = \sqrt{\log 40/400} \approx 0.0960$. With observed error .10, the risk upper bound is
 $0.10 + 0.0960 = 0.1960$.



Choosing a rule after observing the sample changes the problem

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Control the deviation for every candidate simultaneously: $\sup_{f \in \mathcal{H}} |\hat{R}(f) - R(f)| \leq \varepsilon$.

$$\sup_{f \in \mathcal{H}} |\hat{R}(f) - R(f)| \leq \varepsilon$$



Selecting among models requires simultaneous control

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Let $E_j = \{|\hat{R}(f_j) - R(f_j)| > \epsilon\}$. For each of M predeclared candidates, $P(E_j) \leq 2e^{-2n\epsilon^2}$.

$$E_j = \{|\hat{R}(f_j) - R(f_j)| > \epsilon\}$$

$$P\left(\bigcup_{j=1}^M E_j\right) \leq \sum_{j=1}^M P(E_j) \leq 2Me^{-2n\epsilon^2}.$$

$$2Me^{-2n\epsilon^2} = \delta \Rightarrow \epsilon = \sqrt{\frac{\log(2M/\delta)}{2n}}.$$



A finite class pays a logarithmic selection cost

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For 100 predefined rules, divide the failure probability .05 among the rules. A union bound then covers all their deviations simultaneously.

$$R(f) \leq \hat{R}(f) + \sqrt{\frac{\log(2M/\delta)}{2n}}.$$

$$M = 100, n = 1000, \delta = 0.05 \Rightarrow \varepsilon = \sqrt{\frac{\log 4000}{2000}} \approx 0.0644.$$



Worked calculation

For a selected rule with empirical error .03, the upper bound is $0.03 + 0.0644 = 0.0944$. The statement covers the entire candidate set simultaneously with probability at least .95.

Check. Fix the candidate set before inspecting these data; otherwise this finite-set count does not apply.

Infinite parameter values can produce few label patterns

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Let $f_t(x) = \mathbf{1}\{x \geq t\}$ on three fixed points $x_1 < x_2 < x_3$. Although t varies continuously, the predicted labels change only when it crosses a point.

Threshold position	Predicted labels on three points
--------------------	----------------------------------

$t \leq x_1$	$(1, 1, 1)$
--------------	-------------

$x_1 < t \leq x_2$	$(0, 1, 1)$
--------------------	-------------

$x_2 < t \leq x_3$	$(0, 0, 1)$
--------------------	-------------

$t > x_3$	$(0, 0, 0)$
-----------	-------------

Infinitely many thresholds yield only four patterns on these points; $(1, 0, 1)$ cannot be realized by moving this positive-ray threshold.



One input coordinate can support different model capacities

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Two fixed points: $x_1 = 1, x_2 = 2$.

Shattering requires every labeling of the same points. Compare two model families on the same input axis.

Positive-ray threshold

$$f_t(x) = \mathbf{1}\{x \geq t\}.$$

Move t : labels 00, 01 and 11 are possible.

Labels 10 are impossible.

One point admits both labels, but any ordered pair misses 10. Thus VC dimension is 1.

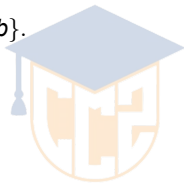
A bounded positive interval

$$f_{a,b}(x) = \mathbf{1}\{a \leq x \leq b\}.$$

Labels at (1, 2)	One interval $[a, b]$
00	$[3, 4]$
01	$[1.5, 2.5]$
10	$[\cdot 5, 1.5]$
11	$[\cdot 5, 2.5]$

Any three ordered points fail labeling 101: an interval containing both outer points also contains the middle. Thus VC dimension is 2.

Both use one input coordinate. Their permitted decision sets, and therefore their capacities, differ.

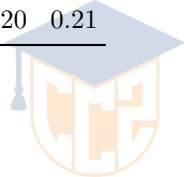


Structural risk minimization compares fit and capacity together

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The illustrative criterion selects $\mathcal{H}_2$. Added flexibility improves fit, but its uncertainty cost can exceed that gain. Theory motivates this comparison; in practice validation often selects degree or penalty because distribution-free bounds are conservative. A penalty criterion is not the same thing as observed test loss.

Class	Empirical risk	Capacity penalty	Sum
$\mathcal{H}_1$	0.18	0.04	0.22
$\mathcal{H}_2$	0.07	0.09	0.16
$\mathcal{H}_3$	0.01	0.20	0.21



What evidence supports a future-risk claim?

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

These principles guide model design. They do not replace independent evaluation or guarantee robustness to a changed population.

A selected rule has empirical error .03. Can we report .03 as its future error?

The observed error is a sample result. Searching many candidates adds selection optimism. The finite-set bound accounts for a predefined candidate count, while validation evaluates the selected procedure on excluded cases.

Scope of a guarantee

The sampling conditions and candidate class are part of the claim. A changed population requires new evidence.