

DSAA3073: Theories in Data Science

Week 4 Validation Sheet · From Uniform Guarantees to Model Selection

Coverage: Week 4

Duration: 30 minutes · **Total:** 130 marks

Answer all questions. For multiple-choice questions, select exactly one answer. Show the essential reasoning requested; unsupported answers may receive no credit.

Part A: Multiple choice

A1 [5 marks]. A class has best-in-class population risk 0.18. Which is an agnostic PAC success statement for $(\epsilon, \delta) = (0.04, 0.05)$?

- (a) $R(A(S)) \leq 0.04$ for every sample.
- (b) $R(A(S)) \leq 0.22$ with probability at least 0.95.
- (c) $\hat{R}_S(A(S)) \leq 0.22$ with probability at least 0.95.
- (d) $\mathbb{E}[R(A(S))] \leq 0.22$.

Solution. (b). Agnostic PAC learning compares population risk with the best population risk available in the declared class. Here the permitted target is $0.18 + 0.04 = 0.22$, and the success probability is $1 - \delta = 0.95$. Option (a) incorrectly treats ϵ as an absolute risk target. Option (c) controls empirical rather than population risk, while (d) replaces the required high-probability statement by an expectation.

Marking. 5 marks for (b), with the best-in-class-plus-tolerance and high-probability interpretation; 0 otherwise.

A2 [5 marks]. Why can the uniform event $\sup_{h \in \mathcal{H}} |\hat{R}_S(h) - R(h)| \leq \alpha$ be applied to an ERM output h_S selected after seeing S ?

- (a) ERM is independent of S .
- (b) The event holds simultaneously for every $h \in \mathcal{H}$ on the same sample.
- (c) ERM always has zero empirical risk.
- (d) A pointwise bound for each fixed h automatically gives the same event.

Solution. (b). The supremum event is simultaneous: once it occurs, its deviation inequality holds for every member of $\mathcal{H}$, including whichever member the data-dependent rule selects. ERM is generally not independent of the sample and need not have zero empirical risk. Separate pointwise high-probability statements do not by themselves provide one shared event for a data-selected hypothesis.

Marking. 5 marks for (b) and the simultaneous-event reason; 0 otherwise.

A3 [5 marks]. In which frame does the Week 4 Fundamental Theorem identify finite VC dimension with learnability?

- (a) Any prediction problem and any loss.
- (b) Distribution-free binary classification with zero-one loss, i.i.d. samples, and suitable measurability.
- (c) Only finite hypothesis classes with deterministic algorithms.
- (d) Regression with squared loss and arbitrary dependent samples.

Solution. (b). The stated equivalence is scoped to distribution-free binary classification under zero-one loss with i.i.d. data and suitable measurability of the risks, suprema, and events. It is not an unrestricted theorem for arbitrary losses, prediction spaces, dependence structures, or nonmeasurable classes.

Marking. 5 marks for (b) together with the complete scope frame; 0 otherwise.

A4 [5 marks]. Which description of a ghost sample S' is correct?

- (a) It is an independent same-distribution sample introduced only in the proof.
- (b) It is a validation set used by the learner to choose a model.
- (c) It is a second training set that the algorithm must receive.
- (d) It is obtained by relabeling the observed training inputs.

Solution. (a). A ghost sample is an independent i.i.d. copy of the training sample, with the same size and distribution, introduced only inside the analysis. It is integrated out in the proof and is not data consumed by the learner. Therefore it is neither an actual validation set nor additional training data, and it is not produced by relabeling observed inputs.

Marking. 5 marks for (a), including the proof-only and independence features; 0 otherwise.

Part B: Definition in use

B1 [20 marks]. Let h_S be an η -approximate ERM in $\mathcal{H}$, so $\hat{R}_S(h_S) \leq \inf_{h \in \mathcal{H}} \hat{R}_S(h) + \eta$. On one sample, the event $\sup_{h \in \mathcal{H}} |\hat{R}_S(h) - R(h)| \leq 0.025$ holds and $\eta = 0.008$.

Starting from the definitions, derive the best-in-class population excess-risk certificate for h_S . State the numerical excess bound, explain why the same sample event protects both hypotheses used in your comparison, and handle the possibility that the population infimum is not attained.

Solution. The target is the excess $R(h_S) - \inf_{h \in \mathcal{H}} R(h)$. Because the population infimum might not be attained, fix any $\rho > 0$ and choose $h_\rho \in \mathcal{H}$ satisfying

$$R(h_\rho) \leq \inf_{h \in \mathcal{H}} R(h) + \rho.$$

This comparator exists by the definition of infimum.

On the stated uniform event, the deviation inequality holds for every hypothesis on this one sample. It therefore applies both to the random, data-selected h_S and to h_ρ . The complete comparison is

$$\begin{aligned} R(h_S) &\leq \hat{R}_S(h_S) + 0.025 \\ &\leq \hat{R}_S(h_\rho) + 0.008 + 0.025 \\ &\leq R(h_\rho) + 2(0.025) + 0.008 \\ &\leq \inf_{h \in \mathcal{H}} R(h) + \rho + 0.058. \end{aligned}$$

The first and third inequalities use the two directions of the same simultaneous deviation event; the middle inequality is the approximate-ERM guarantee. Since the bound holds for every $\rho > 0$, let ρ decrease to zero. Thus

$$R(h_S) - \inf_{h \in \mathcal{H}} R(h) \leq 0.058.$$

A bound only for a fixed comparator would not protect the first move because h_S was selected from the data. The uniform event closes exactly that gap.

Marking. 4 marks for introducing a valid ρ -near population comparator; 4 for applying the event to the selected h_S ; 4 for the approximate-ERM comparison; 4 for applying the event to the comparator; 2 for obtaining 0.058 and letting ρ approach zero; 2 for explaining simultaneous protection. Preserve follow-through credit after an arithmetic slip when the comparison chain is valid.

Part C: Basic skills

C1 [7 marks]. On a fixed sample of size $m = 2$, a loss class has restriction vectors $\mathcal{V} = \{(0, 0), (1, 0), (0, 1)\}$. Compute

$$\mathbb{E}_\sigma \left[\max_{v \in \mathcal{V}} \frac{\sigma_1 v_1 + \sigma_2 v_2}{2} \right]$$

by considering all four equally likely sign vectors.

Solution. The sample and restriction vectors are fixed; only $\sigma \in \{(+1, +1), (+1, -1), (-1, +1), (-1, -1)\}$ is random. The maxima are

$$\frac{1}{2}, \quad \frac{1}{2}, \quad \frac{1}{2}, \quad 0,$$

respectively. For example, under $(+1, -1)$ the three signed averages are $0, \frac{1}{2}, -\frac{1}{2}$, so the maximum is $\frac{1}{2}$; under $(-1, -1)$ the zero vector prevents the maximum from being negative. Averaging after maximizing gives

$$\frac{\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 0}{4} = \frac{3}{8}.$$

Therefore the empirical random-sign complexity of this displayed restriction set is $\frac{3}{8}$.

Marking. 2 marks for enumerating the four signs; 3 for the four correct maxima; 2 for averaging to $\frac{3}{8}$.

C2 [8 marks]. Three active SRM classes use weights $(w_1, w_2, w_3) = (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$ and total failure budget $\delta = 0.06$. Their empirical risks are $(0.14, 0.09, 0.04)$ and their valid penalties at the allocated confidences are $(0.04, 0.07, 0.14)$. Compute the three δ_k , the three certified objectives, and the selected class.

Solution. SRM allocates $\delta_k = w_k \delta$, so

$$(\delta_1, \delta_2, \delta_3) = (0.03, 0.02, 0.01).$$

The weights are positive and sum to one, so the union of the three failure events has probability at most 0.06. Adding empirical risk and penalty gives

$$(0.14 + 0.04, 0.09 + 0.07, 0.04 + 0.14) = (0.18, 0.16, 0.18).$$

Therefore SRM selects class $\mathcal{H}_2$, whose certified objective 0.16 is smallest. Empirical risk alone would instead select $\mathcal{H}_3$ and would ignore its larger penalty.

Marking. 3 marks for $(0.03, 0.02, 0.01)$; 3 for the objectives $(0.18, 0.16, 0.18)$; 2 for selecting $\mathcal{H}_2$ with the certificate interpretation.

Part D: Repair and construction

D1 [17 marks]. A report claims:

“For any prediction problem and any loss, a class is learnable exactly when its VC dimension is finite; moreover one particular ERM rule must succeed.”

Rewrite this as a correct Week 4 Fundamental-Theorem statement. Your repair must give the full scope frame and the correct ERM quantifier, and must explain why the original statement cannot be used outside that frame.

Solution. A correct statement is:

In distribution-free binary classification with zero-one loss, i.i.d. samples, and suitable measurability of the risks, suprema, and probability events, the following class properties are equivalent: uniform convergence; success of every ERM rule as an agnostic PAC learner; agnostic PAC learnability; realizable PAC learnability; success of every ERM rule in the realizable setting; and finite VC dimension.

Two repairs are essential. First, the theorem is framed around binary zero-one classification and distribution-free i.i.d. sampling, with the measurability qualification. VC dimension is a combinatorial notion for binary classes and this stated equivalence does not automatically extend to arbitrary real-valued prediction losses or dependent observations. Those settings need their own complexity and concentration conditions.

Second, the ERM clauses quantify over **every** ERM rule, not merely the existence of one favorable tie-breaking rule. The underlying equivalence is a property of the class in the stated frame. Replacing “every ERM rule” by “one particular ERM rule” weakens the claim and no longer reproduces the Fundamental-Theorem equivalence taught in Week 4.

Marking. 6 marks for binary classification, zero-one loss, distribution-free i.i.d. samples, and suitable measurability; 5 for the correct finite-VC/learnability equivalence; 3 for the “every ERM rule” quantifier; 3 for explaining why arbitrary losses or dependent/nonmeasurable settings are not covered.

D2 [18 marks]. A proposed countable SRM scheme activates classes $\mathcal{H}_1, \mathcal{H}_2, \dots$ and sets $\delta_k = w_k \delta$ with

$$w_k = \frac{3}{2^{k+1}}, \quad k = 1, 2, \dots$$

It then advertises simultaneous confidence $1 - \delta$ and gives an active extra class weight zero.

Diagnose both defects. Compute the actual union-bound budget, then construct a valid replacement allocation for all countably many active classes and state how a zero-weight class must be treated.

Solution. For $k \geq 1$, the geometric series satisfies

$$\sum_{k=1}^{\infty} 2^{-k} = 1.$$

Since $\frac{3}{2^{k+1}} = \left(\frac{3}{2}\right)2^{-k}$, the proposed weights have total

$$\sum_{k=1}^{\infty} w_k = \frac{3}{2}.$$

Therefore the union bound controls the probability of at least one class failure only by

$$\sum_k \delta_k = \delta \sum_k w_k = 3\frac{\delta}{2},$$

not by δ . The advertised simultaneous confidence $1 - \delta$ is not certified.

One valid repair is

$$\tilde{w}_k = 2^{-k}, \quad k \geq 1.$$

Every weight is strictly positive and the weights sum to one, so taking $\tilde{\delta}_k = \tilde{w}_k \delta$ yields $\sum_k \tilde{\delta}_k = \delta$. A union bound then gives simultaneous success probability at least $1 - \delta$, provided each class penalty was proved at its own allocated confidence.

A class with weight zero receives no positive failure budget, so it cannot remain an active finite-penalty competitor under this SRM certificate. It must be excluded from the active family, equivalently assigned infinite penalty. Thus the repaired scheme uses positive summable weights for every active class and removes the claimed zero-weight active class.

Marking. 5 marks for summing the proposed weights to $\frac{3}{2}$; 4 for deriving the actual failure budget $3\frac{\delta}{2}$ and rejecting the advertised confidence; 5 for a valid positive allocation summing to one (such as $2^{(-k)}$); 4 for excluding a zero-weight class or assigning it infinite penalty.

Part E: Proof and derivation

E1 [25 marks]. A measurable binary zero-one loss class is studied on one i.i.d. sample. On an event that holds simultaneously for every $h \in \mathcal{H}$, the available one-sided certificates are

$$R(h) \leq \hat{R}_S(h) + u(h) \quad \text{and} \quad \hat{R}_S(h) \leq R(h) + \ell(h),$$

where $u(h), \ell(h) \geq 0$ need not be equal. The learner returns an η -approximate empirical minimizer $\hat{h}$, and for every $\rho > 0$ there is $h_\rho \in \mathcal{H}$ with $R(h_\rho) \leq \inf_{h \in \mathcal{H}} R(h) + \rho$.

At the two relevant hypotheses the radii are

$$u(\hat{h}) = 0.040, \quad \ell(\hat{h}) = 0.010, \quad u(h_\rho) = 0.020, \quad \ell(h_\rho) = 0.030,$$

and $\eta = 0.010$. Determine the strongest best-in-class excess-risk certificate justified by this information. Your derivation must make clear which deviation direction is needed at each hypothesis, why simultaneity is essential, and how the unattained-infimum case is handled.

Solution. The decisive issue is direction. We begin at the population risk of the data-selected output and must move to its empirical risk. Hence the needed certificate at $\hat{h}$ is the **upper population** direction

$$R(\hat{h}) \leq \hat{R}_S(\hat{h}) + u(\hat{h}).$$

The smaller number $\ell(\hat{h})$ controls the reverse comparison and cannot replace $u(\hat{h})$ in this step.

Approximate empirical minimization compares the selected hypothesis with every empirical competitor, in particular h_ρ :

$$\hat{R}_S(\hat{h}) \leq \hat{R}_S(h_\rho) + \eta.$$

We must next move from the comparator's empirical risk to its population risk. This requires the **upper empirical** direction

$$\hat{R}_S(h_\rho) \leq R(h_\rho) + \ell(h_\rho).$$

Thus $u(h_\rho)$ is irrelevant to this comparison even though it is smaller.

Combining the three compatible arrows on the same simultaneous event gives

$$\begin{aligned} R(\hat{h}) &\leq \hat{R}_S(\hat{h}) + u(\hat{h}) \\ &\leq \hat{R}_S(h_\rho) + \eta + u(\hat{h}) \\ &\leq R(h_\rho) + \ell(h_\rho) + \eta + u(\hat{h}) \\ &\leq \inf_{h \in \mathcal{H}} R(h) + \rho + 0.030 + 0.010 + 0.040. \end{aligned}$$

Therefore, for every $\rho > 0$,

$$R(\hat{h}) - \inf_{h \in \mathcal{H}} R(h) \leq \rho + 0.080.$$

Letting ρ approach zero yields the strongest supported certificate

$$R(\hat{h}) - \inf_{h \in \mathcal{H}} R(h) \leq 0.080.$$

Simultaneity is essential because $\hat{h}$ is selected after seeing the sample and the proof must use certificates for both $\hat{h}$ and h_ρ on that same sample. Separate pointwise high-probability statements do not by themselves supply this shared event. The tempting sum $\ell(\hat{h}) + u(h_\rho) + \eta = 0.040$ is invalid: both radii point opposite to the inequalities required by the comparison chain. Finally, using h_ρ and then letting ρ decrease to zero avoids assuming that the population infimum is attained.

Marking. 4 marks for identifying the needed direction $R(\hat{h}) \leq \hat{R}_S(\hat{h}) + u(\hat{h})$; 5 for inserting the approximate-ERM comparison; 5 for independently identifying the needed comparator direction $\hat{R}_S(h_\rho) \leq R(h_\rho) + \ell(h_\rho)$; 5 for a coherent chain yielding $\rho + 0.080$; 2 for letting ρ approach zero; 2 for explaining simultaneity for the data-selected output and comparator; 2 for rejecting the numerically smaller but directionally invalid 0.040 combination. Award method and follow-through credit for a correct directional chain after one arithmetic error.

Part F: Synthesis and limitations

F1 [15 marks]. Before collecting data, a team declares a fixed library $\mathcal{H}_0$ of 500 binary classifiers and an infinite binary class $\mathcal{H}_\infty$ with finite VC dimension d . Both use zero-one loss and satisfy the Week 4 measurability conditions. From one i.i.d. sample S , the team may select its final classifier from $\mathcal{H}_0 \cup \mathcal{H}_\infty$.

The team observes only 12 distinct loss vectors from $\mathcal{H}_\infty$ on S . It therefore proposes to apply fixed-classifier concentration followed by a union bound over $500 + 12$ observed vectors, claiming a simultaneous population-risk certificate for the final data-selected classifier.

Diagnose this proposal and replace it with a sound Week 4 proof strategy. State precisely which part of the finite-union argument remains valid, why the count 12 does not certify the infinite-class search, and how the final selection can instead receive one simultaneous high-probability guarantee.

Solution. The finite-union argument is valid for $\mathcal{H}_0$. Its 500 classifiers were fixed before S was drawn. A fixed-hypothesis concentration inequality can be applied to each of them, and a union bound then produces one event that controls all 500 population-versus-empirical deviations simultaneously. Consequently, selecting a member of $\mathcal{H}_0$ after seeing S is protected on that event.

The 12 observed loss vectors do not turn $\mathcal{H}_\infty$ into 12 fixed classifiers. The set of restrictions was obtained from the same random sample whose deviations are being controlled. Moreover, two hypotheses that agree on every point in S have the same observed loss vector but may make different predictions outside S and therefore have different population risks. Choosing one observed representative per vector cannot justify a union bound over all population-deviation events in the infinite class.

The repair is to derive a genuine class-wide event for $\mathcal{H}_\infty$. Introduce an independent same-distribution copy only inside the proof to replace the population expectation by a paired empirical process. Exchangeability permits random signs, and splitting the signed supremum leads to Rademacher-complexity control. Because the class is binary with finite VC dimension d , Sauer-Shelah bounds its restriction growth and Massart's lemma converts that count into a random-sign bound. Concentration then yields, under the stated i.i.d. and measurability conditions, a high-probability event controlling

$$\sup_{h \in \mathcal{H}_\infty} |R(h) - \hat{R}_S(h)|.$$

Finally, split the total failure budget between the finite-library event and the infinite-class event, for example assigning $\frac{\delta}{2}$ to each. A union bound over these **two valid simultaneous events** then controls every member of $\mathcal{H}_0 \cup \mathcal{H}_\infty$ on the same sample. The team's final data-selected classifier is protected regardless of which of the two declared classes supplies it.

Marking. 4 marks for locating the legal finite union on the 500 classifiers fixed before sampling; 5 for explaining both sample-dependence of the 12 restrictions and the possibility of different population risks outside S ; 5 for giving the finite-VC route through an independent paired proof process, random-sign control, Sauer-Shelah/Massart, concentration, and the i.i.d./measurability conditions; 1 mark for combining the two valid simultaneous events with an allocated failure budget.