

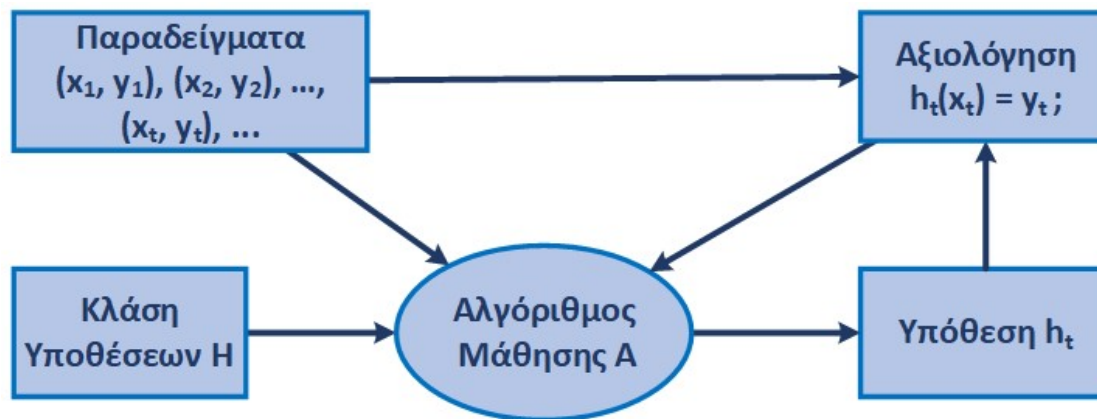
Βασικές Έννοιες



- ▶ **Σύμπαν** (domain) **X**: σύνολο (παρα)δειγμάτων για κατηγοριοποίηση, όπως περιγράφονται με βάση χαρακτηριστικά.
 - ▶ Έστω όλα τα μήλα: $X = \text{Βάρος} \times \text{Όγκος} \times \text{Περίμετρος} \times \text{Χρώμα}$
- ▶ **Κατηγορίες** (labels) **Y** (εστιάζουμε σε $|Y| = 2$, π.χ. $Y = \{-1, +1\}$ ή $Y = \{0, 1\}$)
 - ▶ Για μήλα: $Y = \{\text{Άνοστο}, \text{Νόστιμο}\}$
- ▶ **Υπόθεση** (hypothesis, concept, classifier) $h : X \rightarrow Y$
 - ▶ Κατηγοριοποιεί μήλο («παραδείγμα») ως άνοστο ή νόστιμο με βάση χαρακτηριστικά.
 - ▶ Για $|Y| = 2$, υπόθεση $h \subseteq X$ (σύνολο νόστιμων μήλων).
- ▶ **Στόχος**: δεδομένων ορθά **κατηγοριοποιημένων** παραδειγμάτων, υπολογισμός υπόθεσης $h : X \rightarrow Y$ που κατηγοριοποιεί **ορθά όλα(;) τα παραδείγματα στο X**.
 - ▶ Δοκιμάζοντας **λίγα** μήλα, μαθαίνουμε να αποφεύγουμε **όλα** τα άνοστα!
 - ▶ Στατιστική (λίγα παραδείγματα) και υπολογιστική (ταχύτητα) **αποδοτικότητα**.

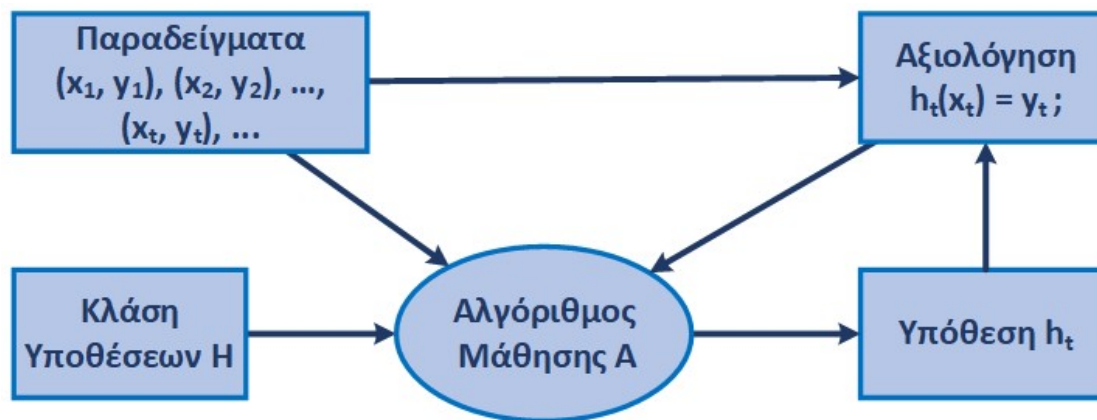
Άμεση Μάθηση (Online Learning)

- ▶ Κάθε χρονική στιγμή $t = 1, 2, \dots$ (επ' άπειρον):
 - ▶ Εμφανίζεται **παράδειγμα** $x_t \in X$ και επιλέγουμε **υπόθεση** $h_t : X \rightarrow Y$ [με βάση παρελθόν και τρέχουσα κατάσταση]
 - ▶ Τοποθετούμε παράδειγμα x_t στην **κατηγορία** $z_t = h_t(x_t)$
 - ▶ Πληροφορούμαστε (ορθή) **κατηγορία** y_t παραδείγματος x_t
 - ▶ Αν $z_t \neq y_t$, έχουμε **λάθος** (κόστος 1), διαφορετικά **σωστό** (κόστος 0)
- ▶ **Στόχος: πεπερασμένο** κόστος (για **άπειρη** ακολουθία παραδειγμάτων)!
 - ▶ «Μικρό» σύμπαν αποτελεί **τετριμμένη** περίπτωση (απομνημόνευση).
 - ▶ Χωρίς γνωστή «δομή», μπορεί **μη εφικτό** για «μεγάλο» (ή **άπειρο**) σύμπαν.



Άμεση Μάθηση (Online Learning)

- ▶ Κάθε χρονική στιγμή $t = 1, 2, \dots$ (επ' άπειρον):
 - ▶ Εμφανίζεται παράδειγμα $(\mathbf{x}_t, y_t) \in (X \times Y)$ και επιλέγουμε υπόθεση $h_t : X \rightarrow Y$
 - ▶ Αν $h_t(\mathbf{x}_t) \neq y_t$, έχουμε λάθος (κόστος 1), διαφορετικά σωστό (κόστος 0)
- ▶ **Στόχος: πεπερασμένο** κόστος (για **άπειρη** ακολουθία παραδειγμάτων)!
 - ▶ Χωρίς γνωστή «δομή», μπορεί **μη εφικτό** για «μεγάλο» (ή **άπειρο**) σύμπαν.
- ▶ **Κλάση υποθέσεων** (hypothesis class) $H \subseteq 2^X$ (ή γενικά $H \subseteq Y^X$) και υπάρχει απόλυτα **συνεπής** υπόθεση $f \in H$ ώστε $y_t = f(\mathbf{x}_t)$, για κάθε $\mathbf{x}_t \in X$.
 - ▶ «Ορθή» κατηγοριοποίηση f για X : **πραγματοποιήσιμη** (realizable) περίπτωση.
 - ▶ Απαιτούνται **υποθέσεις** για κλάση H ή/και για μέγεθος περιγραφής f .



Άμεση Μάθηση με Φράγμα Λαθών

- ▶ **Κλάση υποθέσεων H** μαθαίνεται με **M λάθη**, αν υπάρχει αλγόριθμος με $\leq M$ λάθη για κάθε ακολουθία παραδειγμάτων **κατηγοριοποιημένων** από κάποια $f \in H$.
 - ▶ Επιθυμητή **πολυωνυμική** (π.χ., στο $\log_2|H|$) υπολογιστική πολυπλοκότητα!
- ▶ **Παράδειγμα:** εκμάθηση **λογικών διαζεύξεων** σε $\leq n$ μεταβλητές.
 - ▶ Σύμπαν $X = \{0, 1\}^n$, κλάση **$H_{\text{disj-}n}$** περιέχει **όλες λογικές διαζεύξεις** $\leq n$ μεταβλητών. Π.χ., $h_{\{1,3,7\}}(x) = x_1 \vee x_3 \vee x_7$.
$$h_S(x) = \bigvee_{i \in S} x_i, \quad \forall S \subseteq \{1, \dots, n\}$$
 - ▶ Αρχικά $S = \{1, \dots, n\}$ και υπόθεση h_S . Αμετάβλητη: $h_S(x) \geq f(x)$.
 - ▶ Για κάθε **λάθος** $h_S(x) = 1$ και $f(x) = 0$, αφαιρούμε από S κάθε θέση i όπου $x_i = 1$.
 - ▶ Σε κάθε λάθος, $|S|$ **μικραίνει τουλάχιστον κατά 1**: συνολικά $\leq n$ λάθη.
 - ▶ Κλάση **λογικών διαζεύξεων** $H_{\text{disj-}n}$ μαθαίνεται **με n λάθη** (βέλτιστο στη χειρ. περ.).
- ▶ **Συντηρητικός** αλγόριθμος: ενημερώνει κατάσταση του **μόνο όταν κάνει λάθος**.
- ▶ **Perceptron** μαθαίνει υπερεπίπεδο διαχωρισμού με $O(1/\gamma^2)$ λάθη (γ περιθώριο).



Αλγόριθμος Υποδιπλασιασμού (Halving Algorithm)

- ▶ Υποδιπλασιασμός για **πεπερασμένη** κλάση υποθέσεων H :
 - ▶ Αρχικά $S_0 = H$. Διατηρούμε **σύνολο S_t συνεπών** υποθέσεων μέχρι δείγμα t .
 - ▶ Για $t = 1, 2, \dots$, κλάση z_t δείγματος x_t με **πλειοψηφία** σε **συνεπείς** υποθέσεις $h \in S_{t-1}$
 - ▶ $S_t = \{ h \in S_{t-1} \mid h(x_t) = y_t \}$ (υποθέσεις συνεπείς για t πρώτα παραδείγματα).
 - ▶ Συνολικά $\leq \log_2(|H|)$ λάθη: αν έχουμε **λάθος**, $|S_t| \leq |S_{t-1}| / 2$, λόγω πλειοψηφίας.
 - ▶ Αν έχουμε **prior p_h** = πιθανότητα υπόθεση h να είναι απολύτως ορθή.
 - ▶ Πλειοψηφία με βάση πιθανότητες p_h . **Λάθη \leq εντροπία** αντίστοιχης κατανομής.
- ▶ Χρονική **πολυπλοκότητα $O(|H|)$ – εκθετικός** σε μέγεθος περιγραφής $h \in H$!
 - ▶ Αποδοτικές υλοποιήσεις: π.χ., αλγόριθμος **ελλειψοειδούς** μαθαίνει **υπερεπίπεδο** διαχωρισμού σε n -πλέγμα διάστασης d με **$O(d^2 \log(n))$** λάθη, αντί **$O(d \log(n))$** λαθών.
- ▶ Αν κλάση H **δεν** περιέχει **απολύτως ορθή** υπόθεση για X :
αγνωστική (agnostic) περίπτωση.



Επιλέγοντας Συμβουλή Ειδικού

- ▶ Έχουμε $|H|$ ειδικούς, έναν για κάθε $h \in H$, που προβλέπουν αν βρέξει ή όχι.
- ▶ Κάθε πρωί $t = 1, 2, \dots, T$, βλέπουμε **χαρακτηριστικά** καιρού x_t και **επιλέγουμε** h_t
- ▶ Αν $h_t(x_t) = y_t$, έχουμε κόστος **0**, αλλιώς κόστος **1**.

- ▶ **Στόχος:** κόστος συγκρίσιμο με κόστος **καλύτερου ειδικού** (εκ των υστέρων).

$$\text{regret}(T) = \sum_{t=1}^T (h_t(x_t) \neq y_t) - \min_{h \in H} \sum_{t=1}^T (h(x_t) \neq y_t)$$

- ▶ Αν τέλειος ειδικός, **πλειοψηφία αλάνθαστων** μέχρι στιγμής εγγυάται **$\leq \log_2 |H|$ λάθη**
- ▶ Αν **όχι**, φάσεις με **πλειοψηφία αλάνθαστων με επανεκκίνηση** όταν όλοι ≥ 1 λάθος.
- ▶ Αν $L = \#$ λαθών καλύτερου ειδικού, έχουμε **$\leq L+1$ φάσεις**, και **$\leq \log_2 |H|$ λάθη/φάση**.
- ▶ Συνολικός $\#$ λαθών **$\leq \log_2 |H| (L + 1)$** .
- ▶ «Ξεχνάμε» εντελώς προηγούμενες φάσεις: περιθώριο **σημαντικής βελτίωσης!**

Πλειοψηφία με Βάρη Εμπιστοσύνης (WMA)

- ▶ **Λάθος δεν αποκλείει** κάποιον ειδικό h , αλλά **μειώνει βάρος** εμπιστοσύνης $w(h)$.
 - ▶ Αρχικά $w_1(h) = 1$ για κάθε $h \in H$. Σε κάθε βήμα $t = 1, 2, \dots, T$:
 - ▶ Για παρατήρηση x_t , υιοθετούμε **πρόταση z_t** που συγκεντρώνει **βεβαρημένη πλειοψηφία**.

$$\sum_{h \in H: h(x_t) = z_t} w_t(h) \geq W_t(H)/2$$
 - ▶ Για κάθε $h \in H$ με $h(x_t) \neq z_t$, $w_{t+1}(h) = w_t(h)/2$ (γενικότερα $w_{t+1}(h) = (1 - \epsilon)w_t(h)$).
- ▶ $L = \#$ λαθών καλύτερου ειδικού, $M = \#$ λαθών αλγόριθμου, $W =$ συνολικό βάρος.
 - ▶ Αρχικά $W_1 = |H|$, **μειώνεται κατά 25%** σε κάθε λάθος: $W_{t+1} \leq 3 W_t / 4$

$$(1/2)^L \leq |H|(3/4)^M$$

$$-L \leq \log(|H|) - M \log(4/3)$$
 - ▶ Βάρος καλύτερου ειδικού \leq συνολικό βάρος

$$M \leq 2.41(L + \log(|H|))$$
 - ▶ $\#$ λαθών $\leq 2.41(L + \log_2|H|)$, αντί του $\log_2|H| (L + 1)$ για πλειοψηφία με φάσεις.
 - ▶ **$2.5 \log_2|H|$ λάθη** αρχικά, και μετά **2.5 λάθη** αλγόριθμου για **κάθε αναπόφευκτο** λάθος!
 - ▶ Όχι ικανοποιητικό για «δύσκολες» προβλέψεις με «μεγάλο» L . Μπορούμε καλύτερα;

Αναλογικότητα με Βάρη Εμπιστοσύνης (RWMA)

- ▶ Αντίστοιχα, αλλά χρησιμοποιούμε τα **βάρη** ως **πιθανότητες!**
 - ▶ Αρχικά $\mathbf{w}_1(\mathbf{h}) = \mathbf{1}$ για κάθε $h \in H$. Σε κάθε βήμα $t = 1, 2, \dots, T$:
 - ▶ Για παρατήρηση x_t , υιοθετούμε **πρόταση** \mathbf{z}_t με **πιθανότητα ανάλογη βάρους** υποστήριξης.

$$\frac{\sum_{h \in H: h(x_t) = z_t} w_t(h)}{W_t(H)}$$
 - ▶ Για κάθε $h \in H$ με $h(x_t) \neq y_t$, $\mathbf{w}_{t+1}(\mathbf{h}) = (1 - \varepsilon)\mathbf{w}_t(\mathbf{h})$.
- ▶ $L = \#$ λαθών καλύτερου ειδικού, $M = \mathbf{αναμενόμενο} \#$ λαθών αλγόριθμου, $F_t = \mathbf{πιθανότητα λάθους}$ σε βήμα t , και $M = F_1 + F_2 + \dots + F_t + \dots$
 - ▶ $W_{t+1} = W_t[\sigma\omega\sigma\tau\acute{o}] + (1 - \varepsilon)W_t[\lambda\acute{\alpha}\theta\omicron\varsigma]$

$$(1 - \varepsilon)^L \leq |H| \prod_t (1 - \varepsilon F_t)$$
 - ▶ Αφαιρούμε εF_t από συνολικό βάρος W_t σε κάθε βήμα: $\mathbf{W}_{t+1} = \mathbf{W}_t(1 - \varepsilon F_t)$

$$L \ln(1 - \varepsilon) \leq \ln(|H|) - \varepsilon \sum_t F_t$$
 - ▶ $\#$ λαθών $\leq (1 + \varepsilon/2)L + \log_2 |H|/\varepsilon, \forall \varepsilon > 0$.

$$-L(\varepsilon + \varepsilon^2/2) \leq \ln(|H|) - \varepsilon M$$
 - ▶ **Regret** = $\#$ λαθών – βέλτιστος $\#$ λαθών

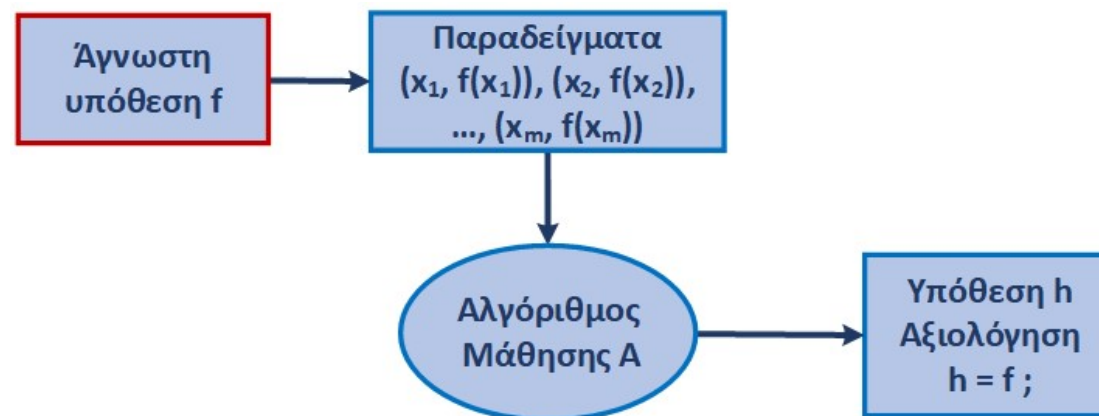
$$M \leq (1 + \varepsilon/2)L + \ln(|H|)/\varepsilon$$
 - ▶ $\text{Regret} / L \rightarrow 0$, καθώς L (ή T) μεγαλώνει. $\varepsilon = \sqrt{\log(|H|)/L} \Rightarrow M \leq L + 2\sqrt{L \log(|H|)}$
 - ▶ **No-regret** αλγόριθμοι: πρακτικά **βέλτιστος** μέσος $\#$ λαθών, καθώς T μεγαλώνει!

Γενικεύσεις – Εφαρμογές

- ▶ Γενικεύσεις – Παραλλαγές:
 - ▶ Πλαίσιο **πολλαπλασιαστικής ενημέρωσης βαρών** (multiplicative weight updates).
 - ▶ Fictitious play, winnow (**αυξομείωση** βαρών), Hedge όπου $w_{t+1}(h) = w_t(h)\exp(-\eta \text{cost})$, AdaBoost για συνδυασμό ταξινομητών, ...
 - ▶ **Bandits**: επιλογή μόνο ενός h_t , για κάθε t , και μαθαίνουμε μόνο για $h_t(x_t) = y_t$
 - ▶ **Εκθετικά μεγάλο $|H|$** : μείωση διάστασης, πλαίσιο **Follow the (Regularized) Leader**.
- ▶ Εφαρμογές:
 - ▶ Μηχανική μάθηση (γραμμικός διαχωρισμός, boosting, ...).
 - ▶ **Θεωρία παιγνίων**: 2-person 0-sum games, coarse correlated equilibrium
 - ▶ **Εξελικτική** θεωρία παιγνίων: replicator dynamics.
 - ▶ **Βελτιστοποίηση**: stochastic gradient descent, άμεση κυρτή βελτιστοποίηση
 - ▶ Προσεγγιστική επίλυση packing / covering γραμμικών προγραμμάτων.
 - ▶ Άμεσοι και προσεγγιστικοί αλγόριθμοι, αλγοριθμικός σχεδιασμός μηχανισμών
 - ▶ (Arora, Hazan, Kale, ToC 2012, <https://theoryofcomputing.org/articles/v008a006/v008a006.pdf>)

Μάθηση από Δεδομένα Εκπαίδευσης (Batch Learning)

- ▶ **Σύμπαν** (domain) X : σύνολο παραδειγμάτων για κατηγοριοποίηση.
- ▶ **Κατηγορίες** (labels) Y (εστιάζουμε σε $|Y| = 2$).
- ▶ **Υπόθεση** (hypothesis, concept, classifier) $h : X \rightarrow Y$
- ▶ Είσοδος **δεδομένα εκπαίδευσης**: $S = \{ (x_1, y_1), \dots, (x_m, y_m) \} \in (X \times Y)^m$
- ▶ Έξοδος: **υπόθεση** $h : X \rightarrow Y$ (ορθή στο μεγαλύτερο μέρος του X)
- ▶ **Στόχος**: αν υπάρχει ορθή $f : X \rightarrow Y$, μαθαίνουμε f ή άλλη υπόθεση $h \approx f$.



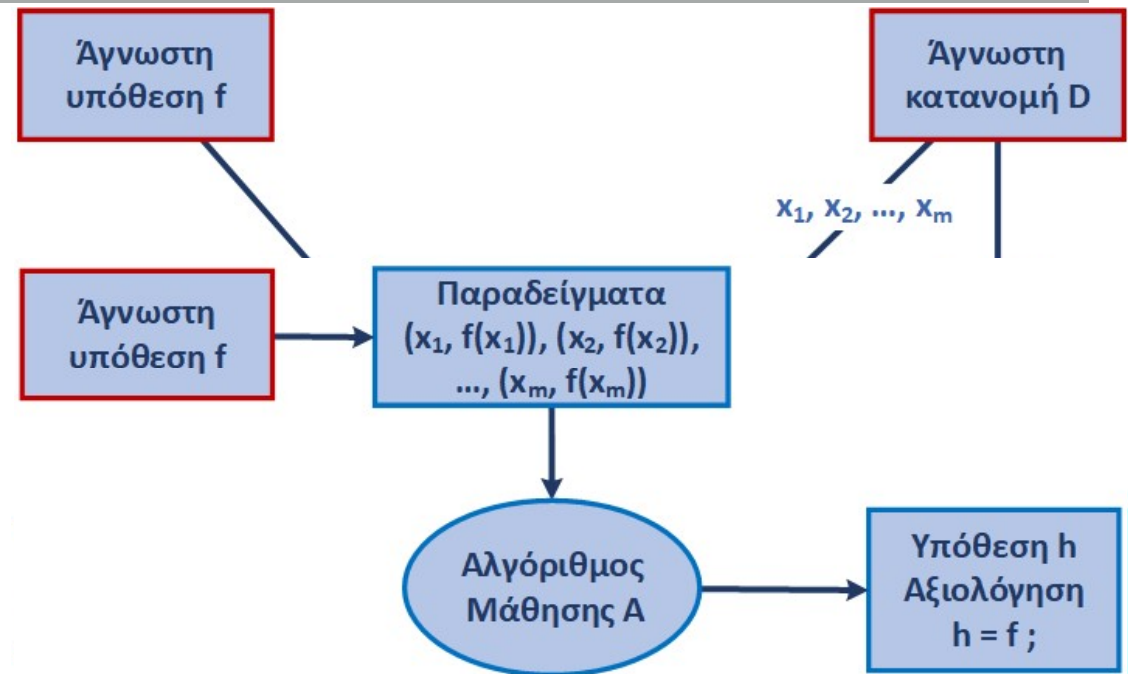
Μάθηση από Δεδομένα Εκπαίδευσης

- ▶ Είσοδος **δεδομένα** εκπαίδευσης:
 $S = \{ (x_1, y_1), \dots, (x_m, y_m) \}$

- ▶ Έξοδος: **υπόθεση** $h : X \rightarrow Y$

- ▶ **Στόχος**: αν υπάρχει ορθή $f : X \rightarrow Y$, μαθαίνουμε f ή άλλη υπόθεση $h \approx f$.

- ▶ Κατανομή D στο X ως μέτρο «**σοβαρότητας**» **λάθους** κατηγοριοποίησης $h(x) \neq f(x)$.
- ▶ **Σφάλμα** $L_{D,f}(h) = \text{Prob}_{x \sim D}[f(x) \neq h(x)]$ αποτελεί **μέτρο απόκλισης** υπόθεσης h από f .
- ▶ Υπολογισμός υπόθεσης $h : X \rightarrow Y$ με αρκετά **μικρό σφάλμα** $L_{D,f}(h)$.
- ▶ Δεδομένα εκπαίδευσης $S = \{ (x_i, f(x_i)) \}_{i=1, \dots, m}$, με κάθε $x_i \sim D$, ώστε να **αντανακλούν** κατανομή D (ανεξάρτητα δείγματα) και **κατηγοριοποίηση** f .
- ▶ **Αγνωστική** περίπτωση: δεδομένα $S = \{ (x_i, y_i) \}$ από κατανομή D στο $X \times Y$, σφάλμα $L_D(h) = \text{Prob}_{(x,y) \sim D}[y \neq h(x)]$, και προσεγγίζουμε $f = \text{argmin}_h \{ L_D(h) \}$





«Μάλλον Σχεδόν Σωστή» (PAC) μάθηση

- ▶ **«Σχεδόν» σωστό:** αδύνατον $L_{D,f}(h) = 0$ (στη χειρότερη περίπτωση).
 - ▶ Π.χ., $X = \{x_1, x_2\}$ με πιθανότητες $(1 - 1/m^2, 1/m^2)$, οπότε δεν μαθαίνουμε $f(x_2)$.
 - ▶ Απαιτούμε $L_{D,f}(h) \leq \epsilon$, για αρκετά μικρό λάθος $\epsilon \in (0, 1)$
- ▶ **«Μάλλον» σωστό:** αδύνατον $L_{D,f}(h) \leq \epsilon$ με **σιγουριά** (πιθανότητα = 1).
 - ▶ Σύνολο δεδομένων εκπαίδευσης S αποτελείται από **ανεξάρτητα δείγματα**: μικρή, αλλά θετική, πιθανότητα να έχουμε δει **λίγα διαφορετικά στοιχεία** του X .
 - ▶ Απαιτούμε $\text{Prob}_S[L_{D,f}(h) > \epsilon] \leq \delta$, για κάποιο αρκετά μικρό $\delta \in (0, 1)$
- ▶ **«Μάλλον σχεδόν σωστή»** (Probably Approximately Correct – **PAC**) μάθηση.
 - ▶ Για κάθε **κατανομή** D και κάθε $f : X \rightarrow Y$ (άγνωστα), για κάθε $\epsilon, \delta \in (0, 1)$ (είσοδος)
 - ▶ Αλγόριθμος «ζητάει» **$m(\epsilon, \delta)$** ανεξάρτητα δείγματα $(x, f(x))$ από D , και παράγει υπόθεση $h : X \rightarrow Y$ με $\text{Prob}_S[L_{D,f}(h) \leq \epsilon] \geq 1 - \delta$ (κλάση υποθέσεων;)
 - ▶ **Άγνωστική** περίπτωση: υπόθεση $h : X \rightarrow Y$ με $\text{Prob}_S[L_D(h) \leq \epsilon + \min_f L_D(f)] \geq 1 - \delta$



Προκαθορισμένη Κλάση Υποθέσεων

- ▶ Δεν υπάρχει «καθολικός» αλγόριθμος, ανεξάρτητος από κλάση υποθέσεων H .
- ▶ Για κάθε αλγόριθμο A , domain X και #δειγμάτων $m \leq |X|/2$, υπάρχει (π.χ. τυχαία) f τ.ω. με πιθανότητα $> \delta$ (στην επιλογή m τυχαίων δειγμάτων $(x, f(x))$), $L_{D,f}(h_A) > \epsilon$.
- ▶ Ανάγκη για επιπλέον πληροφορία, με **προκαθορισμένη κλάση** υποθέσεων H .
- ▶ «**Μάλλον σχεδόν σωστή**» (Probably Approximately Correct – **PAC**) μάθηση.
 - ▶ Κλάση υποθέσεων H **μαθαίνεται κατά PAC**, αν για κάθε $\epsilon, \delta \in (0, 1)$, υπάρχει $m_H(\epsilon, \delta)$ και αλγόριθμος A
 - ▶ τ.ω. για κάθε κατανομή D στο X και κάθε υπόθεση $f : X \rightarrow Y$, με $f \in H$,
 - ▶ ο αλγόριθμος A με είσοδο $m_H(\epsilon, \delta)$ ανεξάρτητα δείγματα $(x, f(x))$ από D , υπολογίζει υπόθεση $h \in H$ με $\text{Prob}_S[L_{D,f}(h) \leq \epsilon] \geq 1 - \delta$ (αγνωστική περίπτωση;)
 - ▶ Συνάρτηση $m_H(\epsilon, \delta)$: **δειγματική πολυπλοκότητα** της κλάσης H .
- ▶ Ελαχιστοποίηση εμπειρικού σφάλματος (Empirical Risk Minimization – **ERM**):

$$\text{ERM}_H(S) = \arg \min_{h \in H} \{(x_i, y_i) \in S : h(x_i) \neq y_i\}$$



Πεπερασμένες και Άπειρες Κλάσεις Υποθέσεων

- ▶ Κάθε **πεπερασμένη** κλάση H , μαθαίνεται από **ERM** με $m \geq \log(|H|/\delta)/\varepsilon$ δείγματα.
 - ▶ $h \in H$ είναι «κακή» αν $h(x_i) \neq f(x_i)$ για κάθε $(x_i, f(x_i)) \in S$ και $L_{D,f}(h) > \varepsilon$:
«κακή» h έχει **μεγάλο πραγματικό** σφάλμα και **μηδενικό εμπειρικό** σφάλμα.

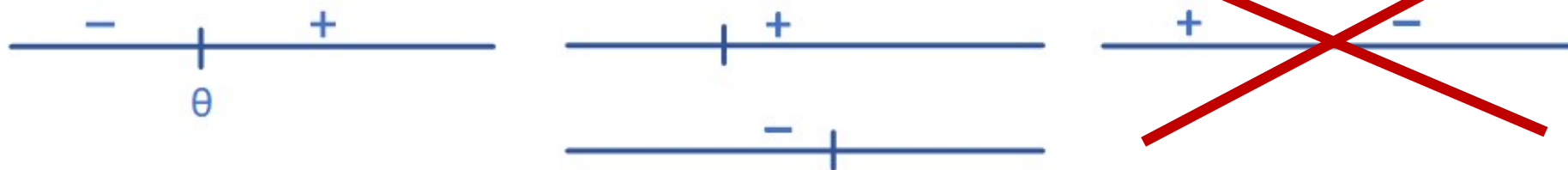
$$\text{Prob}_{S \sim D^m}[h \text{ "bad"}] < (1 - \varepsilon)^m$$

$$\text{Prob}_{S \sim D^m}[\exists h \in H : h \text{ "bad"}] < |H|(1 - \varepsilon)^m \leq \delta$$

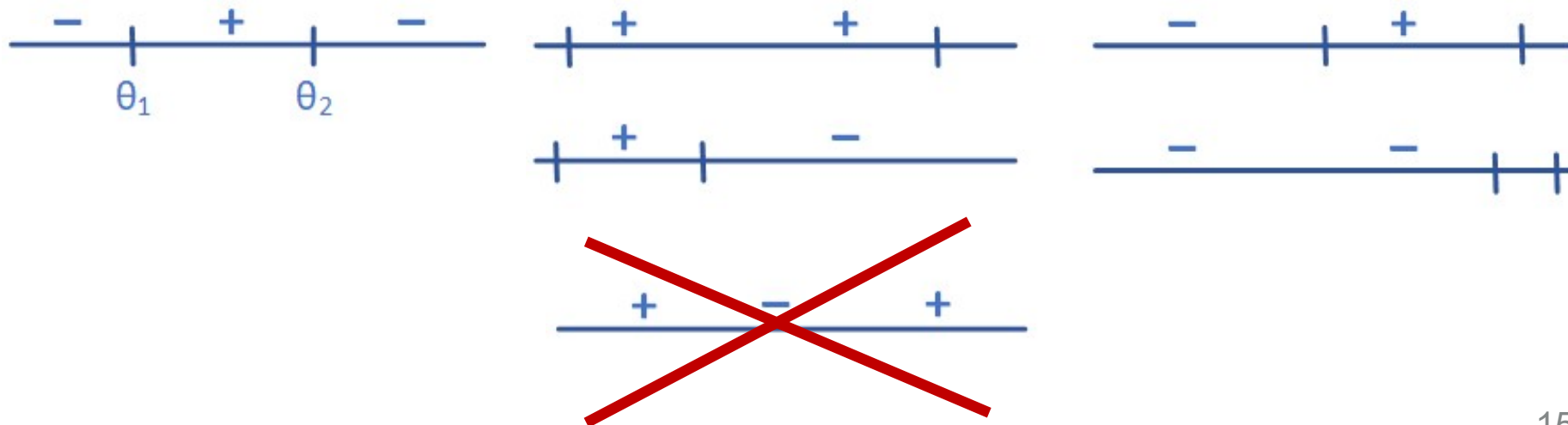
- ▶ **Άπειρη** κλάση H : #δειγμάτων ορίζεται από **VC-διάσταση** (Vapnik–Chervonenkis)
 - ▶ $VC(H) =$ **μέγιστου** μεγέθους **σύνολο δειγμάτων C** που μπορεί να κατηγοριοποιηθεί από υποθέσεις στην H με **όλους** $2^{|C|}$ διαφορετικούς τρόπους.
 - ▶ Για κάθε κατηγοριοποίηση C , υπάρχει **πλήρως συμβατή υπόθεση** στην H .
 - ▶ $VC(H) = k$, αν (i) **υπάρχει** $C \subseteq X$, $|C| = k$, που κατηγοριοποιείται από H με 2^k τρόπους, και (ii) **κάθε** $C' \subseteq X$, $|C'| = k+1$, κατηγοριοποιείται από H με $< 2^{k+1}$ τρόπους.
 - ▶ $H \subseteq 2^X$ ως σύστημα συνόλων. H **θρυμματίζει** σύνολο C αν $|\{h \cap C : h \in H\}| = 2^{|C|}$
 - ▶ $VC(H) = \sup\{|C| : H \text{ θρυμματίζει } C\}$

VC-Διάσταση: Παραδείγματα

- ▶ $VC(H) = k$: (i) **υπάρχει** $C \subseteq X$, $|C| = k$, που κατηγοριοποιείται από H με 2^k τρόπους, (ii) **κάθε** $C' \subseteq X$, $|C'| = k+1$, κατηγοριοποιείται από H με $< 2^{k+1}$ τρόπους.
- ▶ Συναρτήσεις κατωφλίου στο \mathbb{R} έχουν **VC-διάσταση 1**.



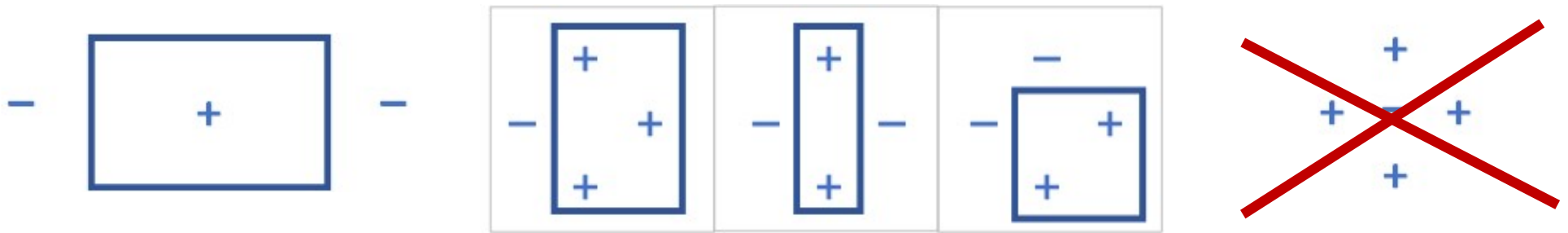
- ▶ Διαστήματα στο \mathbb{R} έχουν **VC-διάσταση 2**.



VC-Διάσταση: Παραδείγματα

- ▶ $VC(H) = k$: (i) υπάρχει $C \subseteq X$, $|C| = k$, που κατηγοριοποιείται από H με 2^k τρόπους, (ii) κάθε $C' \subseteq X$, $|C'| = k+1$, κατηγοριοποιείται από H με $< 2^{k+1}$ τρόπους.

- ▶ Παραλληλόγραμμα στο \mathbb{R}^2 έχουν **VC-διάσταση 4**.



- ▶ Γραμμικοί διαχωριστές στο \mathbb{R}^d έχουν **VC-διάσταση $d+1$** .



Θεμελιώδες Θεώρημα Στατιστικής Μάθησης

- ▶ Κλάση H **μαθαίνεται** αν και μόνο αν έχει **πεπερασμένη** VC-διάσταση $VC(H) = d$.
- ▶ Αλγόριθμος ERM, **δειγματική πολυπλοκότητα** για **πραγματοποιήσιμη** και **αγνωστική** περίπτωση:

$$m_H(\varepsilon, \delta) \leq \beta \frac{d \ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon}$$

$$m_H(\varepsilon, \delta) \leq \beta' \frac{d + \ln(1/\delta)}{\varepsilon^2}$$
- ▶ **Μάθηση με ομοιόμορφη σύγκλιση:**
 - ▶ Σύνολο S δειγμάτων **ε -αντιπροσωπευτικό** για H : για κάθε $h \in H$, $|L_S(h) - L_D(h)| \leq \varepsilon$
 - ▶ Αν σύνολο εκπαίδευσης S είναι **$(\varepsilon/2)$ -αντιπροσωπευτικό** για H , **ERM** υπολογίζει υπόθεση $h_{ERM} \in H$ με $L_D(h_{ERM}) \leq \min_h L_D(h) + \varepsilon$
 - ▶ Κλάση H έχει **ιδιότητα ομοιόμορφης σύγκλισης**: για κάθε κατανομή D , και $\varepsilon, \delta \in (0, 1)$, σύνολο **$m_{H,UC}(\varepsilon, \delta)$** τυχαίων δειγμάτων ε -αντιπροσωπευτικό με πιθανότητα $\geq 1 - \delta$.
 - ▶ Κλάση H έχει **ιδιότητα ομοιόμορφης σύγκλισης** ανν έχει **πεπερασμένη** $VC(H) = d$.
 - ▶ Δειγματική πολυπλοκότητα: $m_H^{UC}(\varepsilon, \delta) \leq \beta'' \frac{d + \ln(1/\delta)}{\varepsilon^2}$



Ανομοιομόρφη Μάθηση – Κριτήριο Occam

- ▶ Κλάση H μαθαίνεται **ανομοιόμορφα** αν για κάθε υπόθεση $f \in H$, δειγματική πολυπλοκότητα **$m_H(\epsilon, \delta, f)$** εξασφαλίζει ότι $L_D(h) \leq L_D(f) + \epsilon$.
 - ▶ Κλάση υποθέσεων H μαθαίνεται **ανομοιόμορφα** αν H αριθμήσιμη **ένωση κλάσεων** με πεπερασμένη VC-διάσταση.
- ▶ **Κριτήριο Occam**
 - ▶ **Συντομότερες** εξηγήσεις (βλ. υποθέσεις) είναι **προτιμότερες** (μεταξύ εξηγήσεων παρόμοιας ακρίβειας).
 - ▶ Συντομότερες υποθέσεις έχουν **μικρότερη δειγματική** πολυπλοκότητα.
 - ▶ Υποθέσεις με περιγραφή **μήκους** d , δειγματική πολυπλοκότητα **$(d \ln(1/\epsilon) + \ln(1/\delta)) / \epsilon$**
 - ▶ Προφανώς μπορεί να **μην** υπάρχει **σύντομη υπόθεση** με ικανοποιητικό σφάλμα.

Συναρτήσεις Σφάλματος – Υπολογιστική Πολυπλοκότητα

- ▶ Συνάρτηση σφάλματος $c: H \times (X \times Y) \rightarrow \mathbb{R}_{\geq 0}$ αποτιμά **σφάλμα h σε δείγμα (x, y)** .
Σφάλμα με βάση συνάρτηση c : $L_D(h) = \mathbb{E}_{(x,y) \sim D}[c(h, (x, y))]$
 - ▶ 0-1 σφάλμα: $c(h, (x, y)) = 0$, αν $h(x) = y$, και 1 , αν $h(x) \neq y$.
 - ▶ Απόλυτη τιμή: $c(h, (x, y)) = |h(x) - y|$
 - ▶ Τετραγωνικό σφάλμα (γραμμική παλινδρόμηση): $c(h, (x, y)) = (h(x) - y)^2$
 - ▶ Hinge σφάλμα (SVM): $c(h, (x, y)) = \max\{1 - y \cdot h(x), 0\}$
 - ▶ Εκθετικό σφάλμα (λογιστική παλινδρόμηση): $c(h, (x, y)) = \ln(1 + e^{-y \cdot h(x)})$
- ▶ ERM μπορεί **υπολογιστικά δύσκολο** πρόβλημα (π.χ., μη-κυρτές συναρτήσεις σφάλματος, σύνθετες κλάσεις υποθέσεων).
 - ▶ Υιοθετούμε (απλές) **κυρτές κλάσεις υποθέσεων** (π.χ., υπερεπίπεδα) και **κυρτές συναρτήσεις σφάλματος** και βελτιστοποιούμε με **(stochastic) gradient descent**.
 - ▶ **Συνολικό σφάλμα**: (σφάλμα περιορισμού σε απλή κλάση H) + (σφάλμα βελτιστοποίησης) + (σφάλμα αντιπροσωπευτικότητας δεδομένων)

Υπολογιστική Πολυπλοκότητα

- ▶ ERM μπορεί **υπολογιστικά δύσκολο** πρόβλημα (π.χ., μη-κυρτές συναρτήσεις σφάλματος, σύνθετες κλάσεις υποθέσεων).
 - ▶ Υιοθετούμε (απλές) **κυρτές κλάσεις υποθέσεων** (π.χ., υπερεπίπεδα) και **κυρτές συναρτήσεις σφάλματος** και βελτιστοποιούμε με **(stochastic) gradient descent**.
 - ▶ **Υλοποίηση** (online projected / stochastic) gradient descent ως **άμεση μάθηση: τρέχουσα υπόθεση** (π.χ., υπερεπίπεδο) ενημερώνεται μετά από **κάθε δείγμα** με χρήση **(sub)gradient σφάλματος** υπόθεσης για δείγμα.
 - ▶ **Perceptron** και **SVM** προκύπτουν ως **ειδικές περιπτώσεις** γενικού πλαισίου.
 - ▶ **Άμεση μάθηση** H με $\# \text{λαθών} \leq M$ συνεπάγεται **PAC-μάθηση** με δειγματική πολυπλοκότητα $O((M + \ln(1/\delta))/\epsilon)$.
 - ▶ (Υπολογιστικά «γρήγοροι») **προσεγγιστικοί αλγόριθμοι** για προβλήματα μάθησης: εκπαίδευση δέντρων απόφασης και τυχαίων δασών, νευρωνικά δίκτυα, GANs(;)

Συμπεράσματα

- ▶ Βασικές αρχές (υπολογιστικής) **θεωρίας μάθησης**.
 - ▶ **Μοντέλο άμεσης μάθησης** με φράγμα στο #λαθών.
 - ▶ **Αλγόριθμοι πλειοψηφίας** (δυναμική αναζήτηση!), έννοια regret, **no-regret** αλγόριθμοι
 - ▶ **Σημαντικές εφαρμογές** σε κυρτή βελτιστοποίηση, υπολογισμό ισορροπιών σε παίγνια, άμεσους και προσεγγιστικούς αλγόριθμους, αλγοριθμικό σχεδιασμό μηχανισμών.
 - ▶ **«Μάλλον σχεδόν σωστή»** (PAC) μάθηση.
 - ▶ **VC-διάσταση**, ελαχιστοποίηση εμπειρικού σφάλματος (**ERM**), ομοιόμορφη σύγκλιση και ανομοιόμορφη μάθηση, υπολογιστική πολυπλοκότητα.
 - ▶ **Στενή σχέση** μεταξύ των δύο μοντέλων και με **άμεση (και κυρτή)** βελτιστοποίηση
 - ▶ Γνωστοί αλγόριθμοι / παραδείγματα (επιβλεπόμενης) μηχανικής μάθησης **εμπίπτουν(;) στο θεωρητικό** πλαίσιο με στόχο βαθύτερη **κατανόηση**.
 - ▶ Κατανόηση θεωρητικού πλαισίου **μπορεί(;) να οδηγήσει** σε νέα παραδείγματα ή/και **νέους αλγόριθμους** μηχανικής μάθησης.