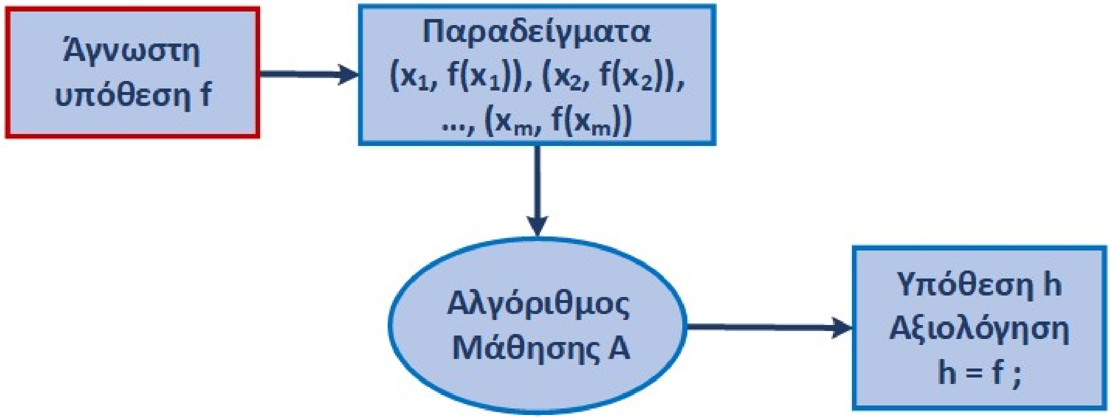
Βασικές Έννοιες

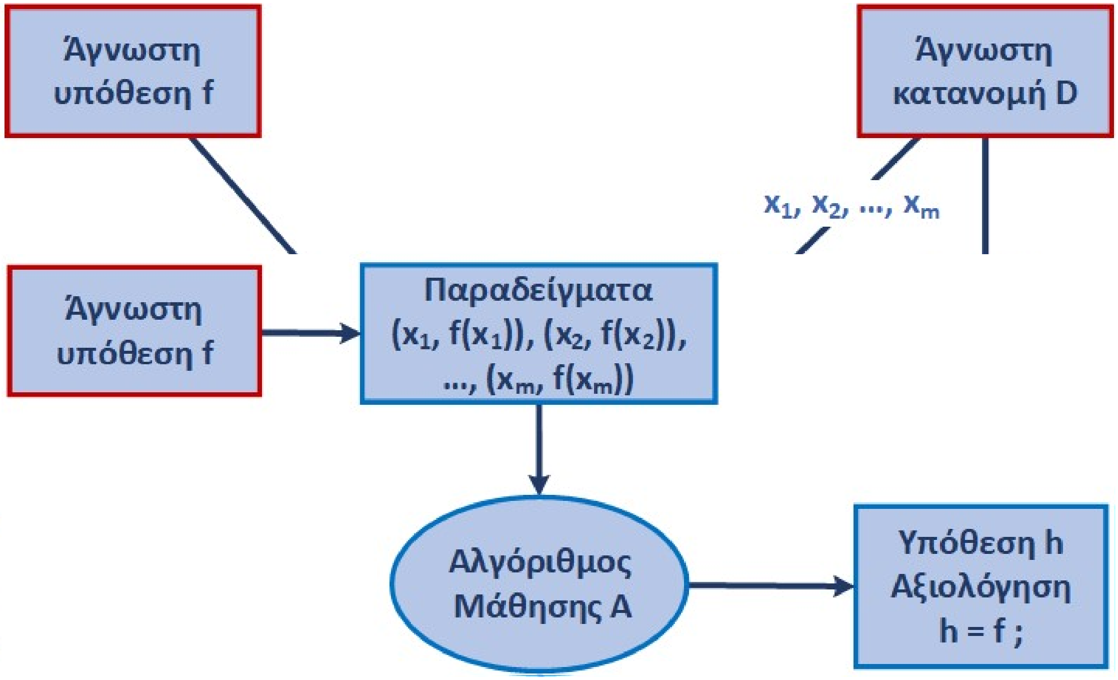
- **Σύμπαν** (domain) **X**: σύνολο (παρα)δειγμάτων για κατηγοριοποίηση, όπως περιγράφονται με βάση χαρακτηριστικά.
  - Έστω όλα τα μήλα: X = Βάρος x Όγκος x Περίμετρος x Χρώμα
- **Κατηγορίες** (labels) **Y** (εστιάζουμε σε |Y| = 2, π.χ. Y = {-1, +1} ή Y = {0, 1})
  - Για μήλα: Y = { Άνοστο, Νόστιμο }
- **Υπόθεση** (hypothesis, concept, classifier) **h : X → Y**
  - Κατηγοριοποιεί μήλο («παράδειγμα») 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 → 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.**

Μάθηση από Δεδομένα Εκπαίδευσης (Batch Learning)

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



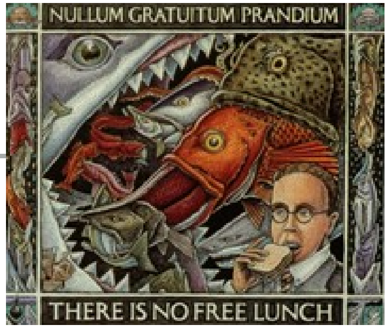
Μάθηση από Δεδομένα Εκπαίδευσης



- Είσοδος **δεδομένα** εκπαίδευσης: $S = \{ (x_1, y_1), \ldots, (x_m, y_m) \}$
- Έξοδος: **υπόθεση** $h : X \to Y$
- **Στόχος**: αν υπάρχει ορθή $f : X \to Y$, μαθαίνουμε f ή άλλη υπόθεση $h \approx f$.
  - Κατανομή D στο X ως μέτρο **«σοβαρότητας» λάθους** κατηγοριοποίησης $h(x) \neq f(x)$.
  - **Σφάλμα $L_{D,f}(h)$** $= \mathrm{Prob}_{x\sim D}[\, f(x) \neq h(x)\,]$ αποτελεί **μέτρο απόκλισης** υπόθεσης h από f.
  - Υπολογισμός υπόθεσης $h : X \to Y$ με αρκετά μικρό σφάλμα $L_{D,f}(h)$.
  - Δεδομένα εκπαίδευσης $S = \{ (x_i, f(x_i)) \}_{i = 1, \ldots, m}$, 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 x Y, σφάλμα **$L_D(h)$** $= \mathrm{Prob}_{(x,y)\sim D}[\, y \neq h(x)\,]$, και προσεγγίζουμε $f = \mathrm{argmin}_h\{ L_D(h) \}$

«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 $\mathbf{L_{D,f}(h) \leq \varepsilon}$, για αρκετά μικρό λάθος $\varepsilon \in (0, 1)$
- «**Μάλλον**» σωστό: αδύνατον $L_{D,f}(h) \leq \varepsilon$ 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 $\mathbf{Prob_S[L_{D,f}(h) > \varepsilon] \leq \delta}$, 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 $\varepsilon, \delta \in (0, 1)$ (είσοδος)
  - Αλγόριθμος «ζητάει» $\mathbf{m(\varepsilon, \delta)}$ ανεξάρτητα δείγματα $(x, f(x))$ από D, και παράγει υπόθεση $h: X \rightarrow Y$ με $\mathbf{Prob_S[L_{D,f}(h) \leq \varepsilon] \geq 1 – \delta}$ (κλάση υποθέσεων;)
  - **Αγνωστική** περίπτωση: υπόθεση $h : X \rightarrow Y$ με $Prob_S[L_D(h) \leq \varepsilon + \min_f L_D(f)] \geq 1 – \delta$



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 f ώστε με πιθανότητα > δ (στην επιλογή m τυχαίων δειγμάτων (x, f(x))), $\mathbf{L_{D,f}(h_A) > \varepsilon}$.
- 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 $\varepsilon, \delta \in (0, 1)$, υπάρχει $\mathbf{m_H(\varepsilon, \delta)}$ και αλγόριθμος A
  - τ.ω. για κάθε κατανομή D στο X και κάθε υπόθεση $f : X \to Y$, με $f \in H$,
  - ο αλγόριθμος A με είσοδο $\mathbf{m_H(\varepsilon, \delta)}$ ανεξάρτητα δείγματα (x, f(x)) από D, υπολογίζει υπόθεση $h \in H$ με $\mathbf{Prob_S[L_{D,f}(h) \leq \varepsilon] \geq 1 - \delta}$ (αγνωστική περίπτωση;)
  - Συνάρτηση $m_H(\varepsilon, \delta)$: **δειγματική πολυπλοκότητα** της κλάσης H.
- Ελαχιστοποίηση εμπειρικού σφάλματος (Empirical Risk Minimization – **ERM**):

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

Πεπερασμένες και Άπειρες Κλάσεις Υποθέσεων

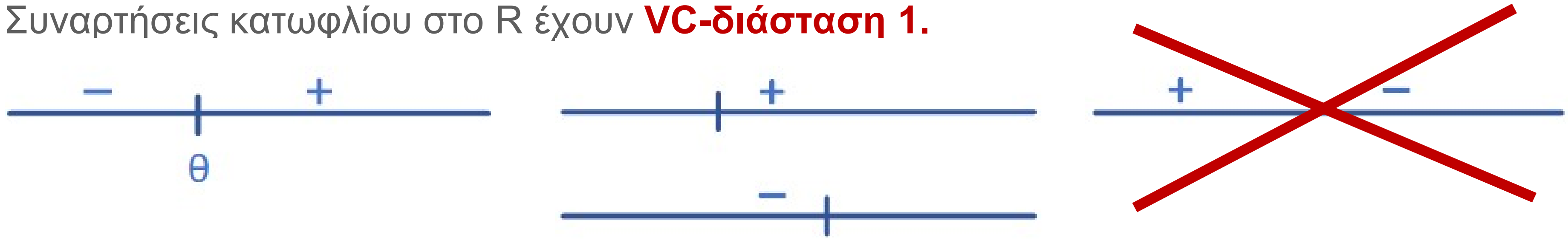
- Κάθε **πεπερασμένη** κλάση H, μαθαίνεται από **ERM** με **m ≥ log(|H|/δ)/ε** δείγματα.
  - $h \in H$ είναι «κακή» αν $h(x_i) = f(x_i)$ για κάθε $(x_i, f(x_i)) \in S$ και $L_{D,f}(h) > \varepsilon$: «κακή» h έχει μεγάλο **πραγματικό** σφάλμα και μηδενικό **εμπειρικό** σφάλμα.

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

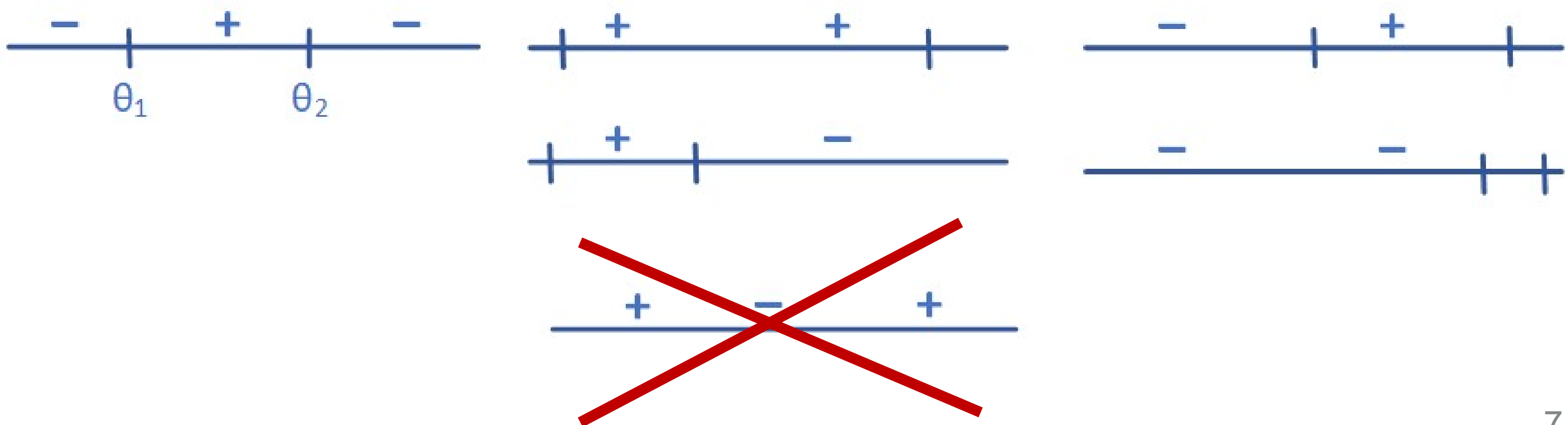
- **Άπειρη** κλάση H: #δειγμάτων ορίζεται από **VC-διάσταση** (Vapnik–Chervonenkis)
  - VC(H) = **μέγιστου** μεγέθους σύνολο δειγμάτων C που μπορεί να κατηγοριοποιηθεί από υποθέσεις στην H με όλους $2^{|C|}$ διαφορετικούς τρόπους.
  - Για κάθε κατηγοριοποίηση C, υπάρχει πλήρως συμβατή υπόθεση στην H.
  - VC(H) = k, αν (ι) **υπάρχει** $C \subseteq X$, |C| = k, που κατηγοριοποιείται από H με 2^k τρόπους, και (ιι) **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 θρυμματίζει C }

VC-Διάσταση: Παραδείγματα

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

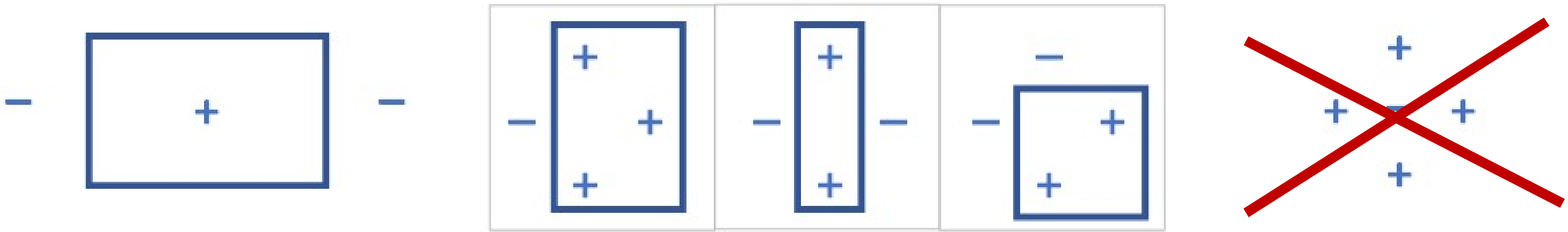


  - Διαστήματα στο R έχουν **VC-διάσταση 2.**



VC-Διάσταση: Παραδείγματα

- VC(H) = k: (ι) υπάρχει $C \subseteq X$, $|C| = k$, που κατηγοριοποιείται από H με 2^k τρόπους, (ιι) κάθε $C' \subseteq X$, $|C'| = k+1$, κατηγοριοποιείται από H με $< 2^{k+1}$ τρόπους.
  - Παραλληλόγραμμα στο R έχουν **VC-διάσταση 4**.



  - Γραμμικοί διαχωριστές στο R^d έχουν **VC-διάσταση d+1**.

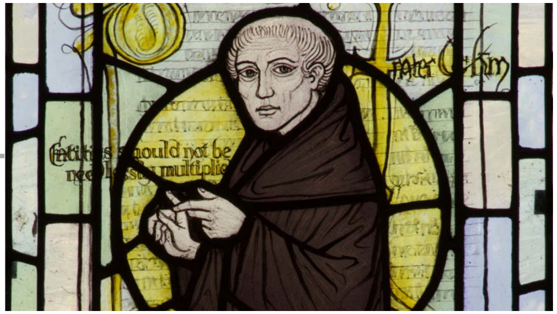
Θεμελιώδες Θεώρημα Στατιστικής Μάθησης

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

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

$$m_H(\varepsilon,\delta) \le \beta' \frac{d + \ln(1/\delta)}{\varepsilon^2}$$

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



Ανομοιομόρφη Μάθηση – Κριτήριο Occam

- Κλάση H μαθαίνεται **ανομοιόμορφα** αν για κάθε υπόθεση $f \in H$, δειγματική πολυπλοκότητα $\mathbf{m_H(\varepsilon, \delta, f)}$ εξασφαλίζει ότι $L_D(h) \leq L_D(f) + \varepsilon$.
  - Κλάση υποθέσεων H μαθαίνεται ανομοιόμορφα ανν 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 $\mathbf{(d\ln(1/\varepsilon)+\ln(1/\delta))/\varepsilon}$
  - Προφανώς μπορεί να **μην** υπάρχει σύντομη υπόθεση με ικανοποιητικό σφάλμα.

Συναρτήσεις Σφάλματος – Υπολογιστική Πολυπλοκότητα

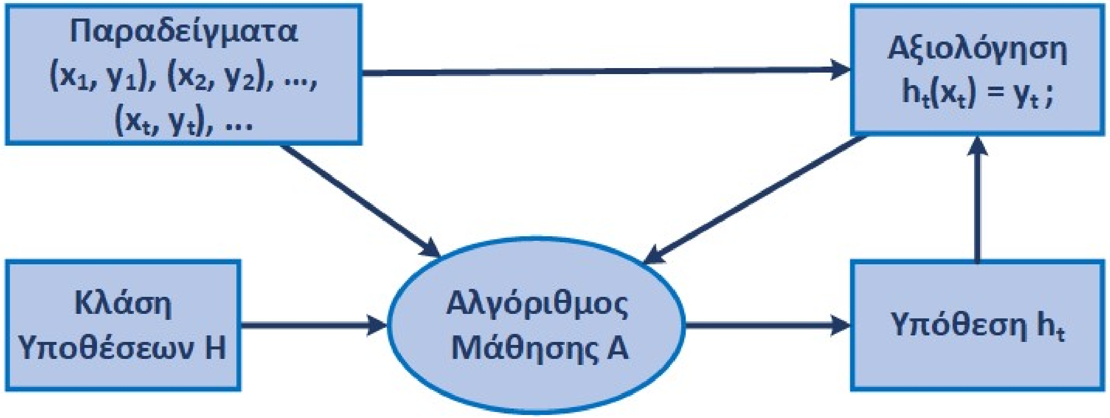
- Συνάρτηση σφάλματος $c: H \times (X \times Y) \to R_{\geq 0}$ αποτιμά **σφάλμα h** σε **δείγμα** (x, y). **Σφάλμα** με βάση συνάρτηση c: $\mathbf{L_D(h) = Exp_{(x,y)\sim D}[c(h, (x, y))]}$
  - 0-1 σφάλμα: c(h, (x, y)) = 0, αν h(x) = y, και 1, αν h(x) ≠ y.
  - Απόλυτη τιμή: $c(h, (x, y)) = |h(x) - y|$
  - 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 #λαθών ≤ M συνεπάγεται PAC-μάθηση με δειγματική πολυπλοκότητα $O((M+\ln(1/\delta))/\varepsilon)$.
  - (Υπολογιστικά «γρήγοροι») **προσεγγιστικοί αλγόριθμοι** για προβλήματα μάθησης: εκπαίδευση δέντρων απόφασης και τυχαίων δασών, νευρωνικά δίκτυα, GANs(;)

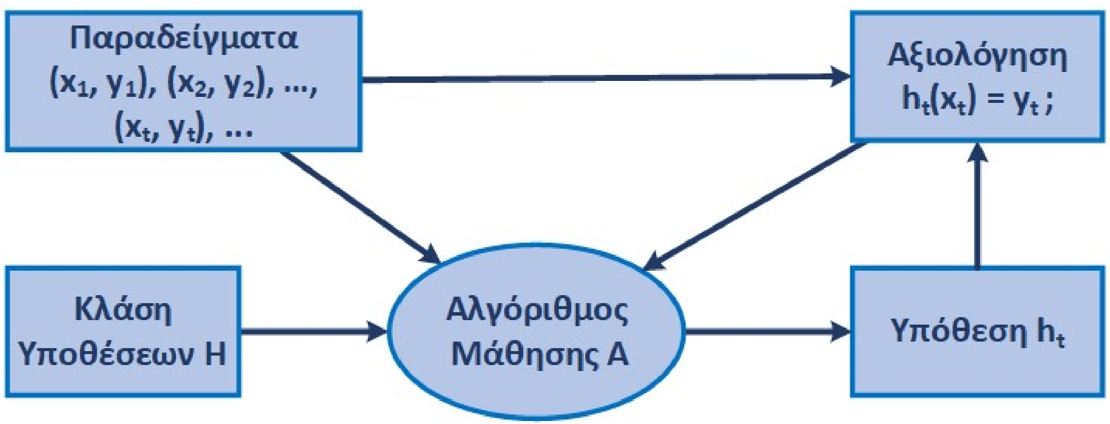
Άμεση Μάθηση (Online Learning)

- Κάθε χρονική στιγμή t = 1, 2, … (επ' άπειρον):
  - Επιλέγουμε υπόθεση $h_t : X \rightarrow Y$
  - Εμφανίζεται παράδειγμα $x_t \in X$
  - 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, … (επ’ άπειρον):
  - Επιλέγουμε υπόθεση $h_t : X \rightarrow Y$ και εμφανίζεται παράδειγμα $(x_t, y_t) \in (X \times Y)$
  - Αν $\mathbf{h_t(x_t) \neq y_t}$, 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 $\mathbf{f \in H}$ ώστε $\mathbf{y_t = f(x_t)}$, για κάθε $x_t \in X$.
  - «Ορθή» κατηγοριοποίηση f για X: **πραγματοποιήσιμη** (realizable) περίπτωση.
  - Απαιτούνται **υποθέσεις** για κλάση H ή/και για μέγεθος περιγραφής f.

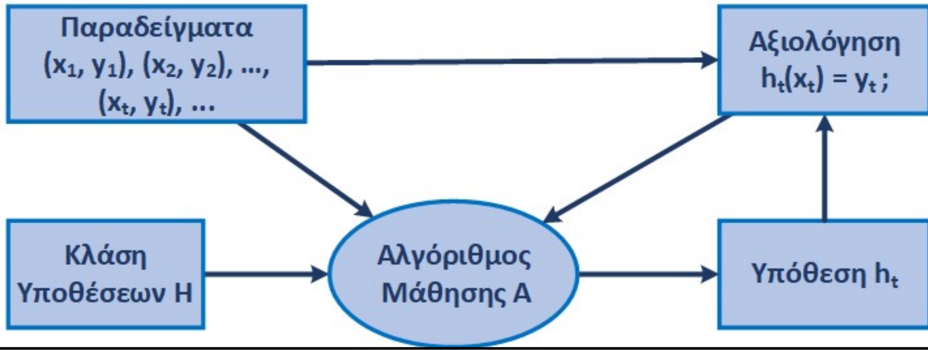


Άμεση Μάθηση με Φράγμα Λαθών

- **Κλάση υποθέσεων H** μαθαίνεται με **M λάθη**, αν υπάρχει αλγόριθμος με ≤ 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 ≤ n μεταβλητές.
  - Σύμπαν $X = \{0, 1\}^n$, κλάση $\mathbf{H_{disj\text{-}n}}$ περιέχει όλες λογικές διαζεύξεις ≤ n μεταβλητών. Π.χ., $h_{\{1,3,7\}}(x) = x_1 \vee x_3 \vee x_7$.

$$h_S(x) = \bigvee_{i \in S} x_i\,, \quad \forall S \subseteq \{1, \ldots, n\}$$

  - Αρχικά S = { 1, …, 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 ≤ n λάθη.
  - Κλάση λογικών διαζεύξεων $H_{disj\text{-}n}$ 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 $\mathbf{S_0 = H}$. Διατηρούμε σύνολο S_t **συνεπών** υποθέσεων μέχρι δείγμα t.
  - Για t = 1, 2, …, κλάση $\mathbf{z_t}$ δείγματος $\mathbf{x_t}$ με **πλειοψηφία** σε **συνεπείς** υποθέσεις $\mathbf{h} \in \mathbf{S_{t-1}}$
  - $\mathbf{S_t = \{ h \in S_{t-1} \mid h(x_t) = y_t \}}$ (υποθέσεις συνεπείς για t πρώτα παραδείγματα).
  - Συνολικά $\mathbf{\leq \log_2(|H|)}$ 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 ≤ **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 $\mathbf{O(d^2 \log(n))}$ 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, …., 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).

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

  - Αν τέλειος ειδικός, πλειοψηφία αλάνθαστων μέχρι στιγμής εγγυάται **≤ $\log_2$|H|** λάθη
  - Αν **όχι**, φάσεις με πλειοψηφία αλάνθαστων με επανεκκίνηση όταν όλοι ≥ 1 λάθος.
  - Αν L = #λαθών καλύτερου ειδικού, έχουμε **≤ L+1** φάσεις, και **≤ $\log_2$|H|** λάθη/φάση.
  - Συνολικός #λαθών **≤ $\log_2$|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, …, T:
  - Για παρατήρηση x_t, υιοθετούμε πρόταση $\mathbf{z_t}$ 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 $h(x_t) \neq y_t$, **$w_{t+1}(h) = w_t(h)/2$** (γενικότερα $w_{t+1}(h) = (1 - \varepsilon)w_t(h)$).
- L = #λαθών καλύτερου ειδικού, M = #λαθών αλγόριθμου, W = συνολικό βάρος.
  - Αρχικά $W_1 = |H|$, μειώνεται κατά 25% σε κάθε λάθος: $W_{t+1} \leq 3\, W_t / 4$
  - Βάρος καλύτερου ειδικού ≤ συνολικό βάρος

$$(1/2)^L \leq |H|(3/4)^M$$
$$-L \leq \log(|H|) - M \log(4/3)$$
$$M \leq 2.41(L + \log(|H|))$$

  - #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 $\mathbf{w_1(h) = 1}$ για κάθε $h \in H$. Σε κάθε βήμα $t = 1, 2, \ldots, T$:
  - 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 $h(x_t) \neq y_t$, $\mathbf{w_{t+1}(h) = (1 - \varepsilon)w_t(h)}$.
- L = #λαθών καλύτερου ειδικού, M = **αναμενόμενο** #λαθών αλγόριθμου, F_t = πιθανότητα λάθους σε βήμα t, και $M = F_1 + F_2 + \ldots + F_t + \ldots$
  - $W_{t+1} = W_t[\text{σωστό}] + (1 - \varepsilon)W_t[\text{λάθος}]$
  - Αφαιρούμε εF_t από συνολικό βάρος W_t σε κάθε βήμα: $\mathbf{W_{t+1} = W_t(1 - \varepsilon F_t)}$
  - #λαθών $\mathbf{\leq (1+\varepsilon/2)L + \log_2|H|/\varepsilon}$, $\forall \varepsilon > 0$.
  - **Regret** = #λαθών – βέλτιστος #λαθών
    Regret / L $\to$ 0, καθώς L (ή T) μεγαλώνει.
  - **No-regret** αλγόριθμοι: πρακτικά **βέλτιστος** μέσος #λαθών, καθώς T μεγαλώνει!

$$(1-\varepsilon)^L \leq |H| \prod_t (1-\varepsilon F_t)$$

$$-L\ln(1-\varepsilon) \leq \log(|H|) - \varepsilon \sum_t F_t$$

$$-L\ln(1-\varepsilon) \leq \log(|H|) - \varepsilon M$$

$$M \approx (1+\varepsilon/2)L + \log(|H|)/\varepsilon$$

$$\varepsilon = \sqrt{\log(|H|)/L} \Rightarrow M \leq L + 2\sqrt{L\log(|H|)}$$

Γενικεύσεις – Εφαρμογές

- Γενικεύσεις – Παραλλαγές:
  - Πλαίσιο **πολλαπλασιαστικής ενημέρωσης βαρών** (multiplicative weight updates).
  - Fictitious play, winnow (αυξομείωση βαρών), Hedge όπου $w_{t+1}(h) = w_t(h)\exp(-\eta \mathrm{cost})$, 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)

Συμπεράσματα

- Βασικές αρχές (υπολογιστικής) **θεωρίας μάθησης.**
  - **Μοντέλο άμεσης μάθησης** με φράγμα στο #λαθών.
  - Αλγόριθμοι πλειοψηφίας (δυαδική 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.