6. Στατιστικές μέθοδοι εκπαίδευσης

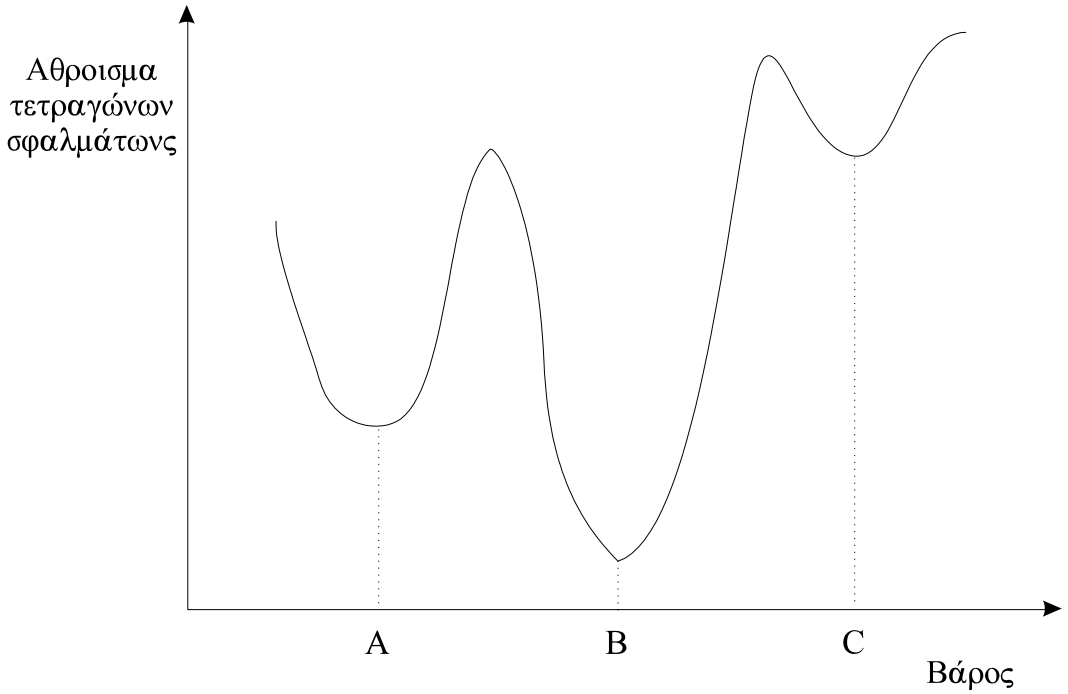
Μία διαφορετική μέθοδος εκπαίδευσης των νευρωνικών δικτύων χρησιμοποιεί ιδέες από την Στατιστική Φυσική για να φέρει τελικά το ίδιο αποτέλεσμα όπως οι άλλες μέθοδοι, πράγμα που ως γνωστό είναι οτι κατά την διάρκεια της εκπαίδευσης πρέπει να αλλάξουν οι τιμές των βαρών w έτσι ώστε το δίκτυο να δίδει το σωστό αποτέλεσμα κάθε φορά που παρουσιάζεται ένα πρότυπο. Η στατιστική φυσική είναι μία περιοχή της φυσικής που χρησιμοποιεί την τυχαιότητα ως την κεντρική ιδέα, αλλά με τέτοιο τρόπο ούτως ώστε στο τέλος το αποτέλεσμα να είναι κοντά στην πραγματικότητα. Αυτό συμβαίνει διότι θεωρεί ένα πολύ μεγάλο αριθμό δειγμάτων, από τα οποία παίρνει μέσες τιμές των ιδιοτήτων (μέσοι όροι), μετά από πολλές πραγματοποιήσεις του φαινομένου που μελετά. Είναι η κατ' εξοχήν περιοχή της φυσικής που χρησιμοποιεί τους υπολογιστές.

Η διαφορά των στατιστικών μεθόδων από τις άλλες μεθόδους, όπως π.χ. την μέθοδο της οπισθοδιάδοσης, έγκειται στο ότι στις άλλες μεθόδους ακολουθούμε μία αυστηρή διαδικασία, χρησιμοποιώντας μαθηματικές εξισώσεις όπου αλλάζουμε τα βάρη w ανάλογα με τα σφάλματα που παίρνουμε στην έξοδο, ενώ στις στατιστικές μεθόδους αλλάζουμε τα βάρη w τυχαία, με βάση το εξής κριτήριο: Αν η τυχαία αλλαγή κάνει το σφάλμα μικρότερο, τότε την κρατάμε, αν το μεγαλώνει τότε την απορρίπτουμε, και προχωράμε σε μία άλλη τυχαία αλλαγή. Ένας τέτοιος αλγόριθμος περιλαμβάνει τα εξής στάδια:

- Θεωρούμε μία ομάδα προτύπων, τα οποία εισέρχονται στο επίπεδο εισόδου. Με μία μη-γραμμική συνάρτηση βρίσκουμε την έξοδο.
- Συγκρίνουμε έξοδο με στόχο, για κάθε πρότυπο. Βρίσκουμε το άθροισμα των τετραγώνων της διαφοράς, και πρέπει το άθροισμα αυτό να ελαχιστοποιηθεί.
- Διαλέγουμε ένα w τυχαία και το αλλάζουμε επίσης τυχαία, κατά ένα μικρό όμως ποσοστό. Αν η αλλαγή αυτή ελαττώνει το άθροισμα των τετραγώνων, τότε την κρατάμε, εάν όχι τότε το w αυτό δεν αλλάζει καθόλου, αλλά συνεχίζει να έχει την προηγούμενη τιμή του.
- Διαλέγουμε τυχαία ένα άλλο w, και επαναλαμβάνουμε την συνολική διαδικασία των παραπάνω τριών βημάτων τόσες φορές όσες χρειάζεται το δίκτυο για να εκπαιδευθεί.

Η διαδικασία αυτή εκ πρώτης όψης φαίνεται σωστή ότι εξασφαλίζει την λύση που εκπαιδεύει το δίκτυο. Υπάρχει όμως πάντα το πρόβλημα των τοπικών ελαχίστων, που είδαμε στα προηγούμενα κεφάλαια. Συγκεκριμένα, έστω ότι μειώνοντας το σφάλμα το σύστημα βρίσκεται στο σημείο Α. Εάν οι τυχαίες αλλαγές στο w είναι μικρές, τότε κάθε αλλαγή θα απορρίπτεται διότι μεγαλώνει το σφάλμα. Το w θα παραμένει για πάντα στο Α και δεν θα βρει το Β, που είναι το πραγματικό σημείο που θέλουμε. Λέμε ότι το σύστημα παγιδεύεται σε ένα τοπικό ελάχιστο. Εάν, για να αποφύγουμε την δυσκολία

αυτή, κάνουμε τις μεταβολές στα w να είναι μεγάλες, τότε και το Α και το Β θα επισκέπτονται συχνά, αλλά το σύστημα δεν θα κατασταλάξει στο ελάχιστο, γιατί λόγω της μεγάλης μεταβολής θα ξεφεύγει συχνά και θα επισκέπτεται όλα τα σημεία στην καμπύλη του σφάλματος. Ποτέ δεν θα μπορεί να παγιδευθεί σε ένα ελάχιστο, με αποτέλεσμα το δίκτυο να μην μπορεί να εκπαιδευθεί. Ίσως η καλύτερη τακτική είναι να αρχίσουμε με μεγάλα βήματα, τα οποία όμως σιγά-σιγά μικραίνουν, και έτσι το σύστημα απο-παγιδεύεται από τοπικά ελάχιστα, αλλά τελικά παραμένει στο χαμηλότερο ελάχιστο.



ΣΧΗΜΑ 6.1
Τοπικά και ολικό ελάχιστο

Εκπαίδευση Boltzmann

Οι στατιστικές μέθοδοι, από την φύση τους, μπορούν να αποφύγουν το πρόβλημα αυτό, το οποίο όπως είδαμε υπάρχει και στις μεθόδους των προηγουμένων κεφαλαίων. Και το πετυχαίνουν αυτό ακριβώς με το να μεταβάλλουν (να μην κρατούν σταθερό) το μέγεθος της αλλαγής των w. Στην αρχή έχουμε μεγάλες μεταβολές, από τις οποίες κρατάμε μόνο αυτές που ελαττώνουν το σφάλμα. Ακολούθως το μέγεθος του βήματος μικραίνει και τελικά φθάνουμε στο παγκόσμιο ελάχιστο. Η διαδικασία αυτή είναι ανάλογη μιας γνωστής διεργασίας στην φυσική που λέγεται προσομοίωσης ανόπτησης (simulated annealing), και η οποία χρησιμοποιείται σε πολλές διαφορετικές περιπτώσεις. Η πιο συνηθισμένη είναι η κρυστάλλωση, δηλ. η δημιουργία του στερεού κρυστάλλου από τον υγρό κρύσταλλο. Όταν ένα υλικό είναι στην υγρή κατάσταση τα μόρια του έχουν υψηλή σχετικά ενέργεια και κάνουν πολλές βίαιες κινήσεις. Καθ' όσον σιγά-σιγά πέφτει η θερμοκρασία του

συστήματος, πέφτει και η ενέργεια των μορίων, των οποίων οι κινήσεις τώρα είναι πιο μικρές, και αργά το υγρό σύστημα μετατρέπεται σε στερεό. Όπως είναι γνωστό, σε οποιαδήποτε κατάσταση όλα τα μόρια δεν έχουν ακριβώς την ίδια ενέγεια, αλλά άλλα έχουν μεγαλύτερη και άλλα μικρότερη ενέργεια. Η κατανομή των ενεργειών των μορίων υπακούει τον γνωστό νόμο του Boltzmann:

$$P(E) \approx e^{-E/kT} \qquad \textbf{(1)}$$

όπου Ε είναι η ενέργεια, P(E) η πιθανότητα ότι το σύστημα βρίσκεται σε ενέργεια Ε, k η σταθερά Boltzmann, και Τ η θερμοκρασία. Σε υψηλές θερμοκρασίες, κάθε μόριο λέμε ότι έχει μεγάλη θερμική ενέργεια, λόγω της θερμοκρασίας, και έτσι P(E)=1, δηλ. κάθε ενεργειακή κατάσταση είναι πιθανή. Καθ' όσον η θερμοκρασία ελαττώνεται η πιθανότητα να έχουμε μεγάλη Ε πέφτει, και το σύστημα πέφτει σε μικρές ενεργειακές καταστάσεις.

Η εκπαίδευση λοιπόν με στατιστική μέθοδο ακολουθεί την παρακάτω διαδικασία:

- Δίνουμε μία τεχνητή θερμοκρασία στο σύστημα, Τ. Στην αρχή έχουμε μεγάλες τιμές του Τ.
- Βάζουμε στο επίπεδο εισόδου τα πρότυπα, και υπολογίζουμε την έξοδο και το σφάλμα.
- Κάνουμε μία τυχαία αλλαγή στα βάρη w, και υπολογίζουμε πάλι την έξοδο.
- Εάν το σφάλμα μικραίνει κρατάμε την αλλαγή.
- Εάν το σφάλμα μεγαλώνει τότε υπολογίζουμε την πιθανότητα P(c) να δεχθούμε την αλλαγή αυτή.

$$P(c) = e^{-c/kT} \qquad \textbf{(2)}$$

- Διαλέγουμε έναν τυχαίο αριθμό r, από μία ομαλή κατανομή όπου 0<r<1. Εάν P(c)>r, τότε κρατάμε την αλλαγή. Εάν P(c)<r, τότε η αλλαγή απορρίπτεται, και πάμε στον επόμενο υπολογισμό.

Με την διαδικασία αυτή σε μερικές περιπτώσεις το σύστημα πηγαίνει σε κατάσταση υψηλότερης ενέργειας. Το δίκτυο απομακρύνεται περισσότερο από το σημείο εκπαίδευσης, και θα πάρει μεγαλύτερο χρόνο για να βρούμε τελικά τα σωστά w. Αλλά το πλεονέκτημα που έχουμε είναι ότι το δίκτυο μπορεί να ξεφύγει από ένα τοπικό ελάχιστο.

Η διαδικασία αυτή γίνεται για όλα τα βάρη w, ένα προς ένα, μέχρις ότου το δίκτυο εκπαιδευθεί. Κατά την διάρκεια της διαδικασίας ελαττώνουμε σιγά-

σιγά την θερμοκρασία T, μέχρις ότου το σφάλμα ελαχιστοποιηθεί. Ακολούθως παρουσιάζουμε τα επόμενα πρότυπα, και ακολουθούμε τα ίδια βήματα.

Το μέγεθος την αλλαγής στο w μπορεί να υπολογισθεί από την κατανομή Gauss:

$$P(w) = e^{-w^2/T^2} \qquad \textbf{(3)}$$

όπου P(w) είναι η πιθανότητα η αλλαγή να έχει μέγεθος w. Βρίσκουμε την συνάρτηση πιθανότητας που αντιστοιχεί στο P(w). Αυτό είναι το ολοκλήρωμα:

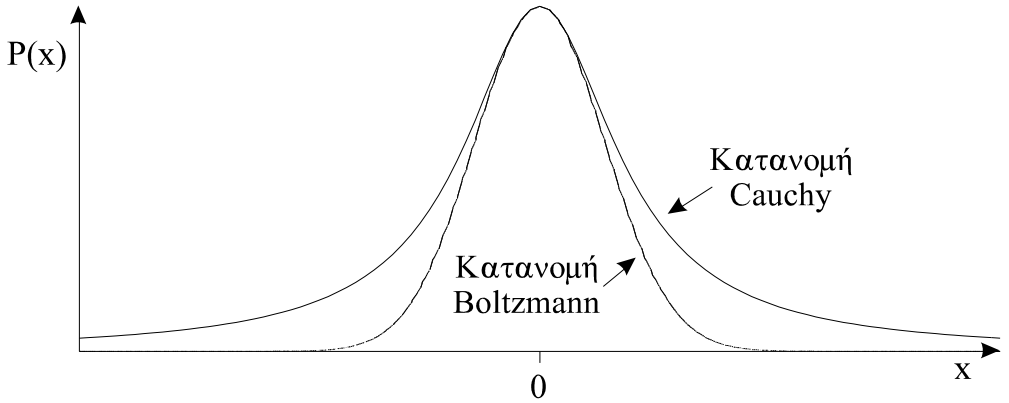
$$\int_0^w P(w) \qquad \textbf{(4)}$$

Το ολοκλήρωμα αυτό υπολογίζεται αριθμητικά. Μετά παίρνουμε έναν τυχαίο αριθμό r, από μία ομαλή κατανομή τυχαίων αριθμών με όρια το διάστημα αυτό. Αυτή είναι η τιμή του P(w), δηλ. P(w)=r, και ακολούθως βρίσκουμε σε ποια τιμή του Δw αντιστοιχεί η τιμή αυτή.

Η μεθοδος εκπαίδευσης αυτής λέγεται μέθοδος Boltzmann, και το δίκτυο λέγεται και μηχανή Boltzamann. Έχουν γίνει πολλές εργασίες πάνω στο μοντέλο αυτό. Έχει αποδειχθεί ότι ο ρυθμός με τον οποίο πρέπει να ελαττώνεται η θερμοκρασία είναι ανάλογος με το αντίστροφο του λογαρίθμου του χρόνου:

$$T(t) = \frac{T_o}{\log(1+t)} \qquad \textbf{(5)}$$

όπου T(t) είναι η θερμοκρασία ως συνάρτηση του χρόνου t, To είναι η αρχική θερμοκρασία. Το αποτέλεσμα βέβαια αυτό σημαίνει ότι το δίκτυο χρειάζεται μεγάλους χρόνους εκπαίδευσης, πράγμα που το καθιστά όχι πολύ χρήσιμο.



ΣΧΗΜΑ 6.2
Οι κατανομές Boltzmann και Cauchy.

Εκπαίδευση Cauchy

Η μέθοδος αυτή χρησιμοποιεί την κατανομή Cauchy αντί της κατανομής Boltzmann που είδαμε προηγουμένως. Η κατανομή C έχει τον τύπο:

$$P(x) = \frac{T(t)}{T^2(t) + x^2} \qquad \textbf{(6)}$$

όπου P(x) είναι η πιθανότητα για ένα βήμα μεγέθους x. Η κατανομή C πέφτει πιο αργά από του B. Αυτό έχει ως αποτέλεσμα να έχουμε πιθανότητες για μεγαλύτερα βήματα. Στην περίπτωση C έχουμε ότι:

$$T(t) = \frac{T_0}{1+t} \quad \textbf{(7)}$$

αντί του λογαρίθμου που είχαμε στον παρανομαστή στην περίπτωση B. Αυτό μικραίνει το χρόνο εκπαίδευσης. Ολοκληρώνουμε το P(x), και λύνοντας ως προς x έχουμε:

$$x_c = \rho\{T(t)\tan[P(x)]\} \quad \textbf{(8)}$$

όπου ρ είναι ο ρυθμός εκπαίδευσης (σταθερά), και x_c είναι η αλλαγή βάρους. Για να βρούμε το x, διαλέγουμε έναν τυχαίο αριθμό r από μία ομαλή κατανομή στο διάστημα $-\pi/2<r<\pi/2$, διότι αυτό είναι το διάστημα ορισμού της εφαπτομένης. Αντικαθιστούμε με την τιμή του P(x), και υπολογίζουμε το μέγεθος του βήματος x.

Μέθοδος ειδικής θερμότητας.

Η ειδική θερμότητα (specific heat) C στην θερμοδυναμική ορίζεται ως η παράγωγος της ενέργειας ως προς την θερμοκρασία:

$$C = \frac{dE}{dT} \qquad \textbf{(9)}$$

Δηλαδή, το C δηλώνει κατά πόσο αλλάζει η ενέργεια του συστήματος με οποιαδήποτε αλλαγή της θερμοκρασίας. Το C αλλάζει απότομα όταν το σύστημα αλλάζει φάση (π.χ. από υγρό σε στερεό). Η αλλαγή φάσης γίνεται σε μία θερμοκρασία που λέγεται κρίσιμη θερμοκρασία Tc. Στα νευρωνικά δίκτυα η απότομη αλλαγή υποδηλώνει ότι το σύστημα βρέθηκε ξαφνικά σε ένα τοπικό ελάχιστο. Εδώ το ανάλογο του C είναι η μέση αλλαγή της θερμοκρασίας ως προς την αλλαγή του σφάλματος. Όταν η θερμοκρασία είναι πολύ χαμηλή, η ειδική θερμότητα είναι σχεδόν σταθερή, και έτσι η θερμοκρασία μπορεί να αλλάζει με μεγάλο ρυθμό χωρίς πρόβλημα.

Στην κρίσιμη θερμοκρασία, Tc, μία μικρή μεταβολή θερμοκρασίας προκαλεί μεγάλη μεταβολή στην μέση τιμή του σφάλματος. Στο σημείο αυτό χρειάζεται προσοχή. Το σημείο αυτό είναι κρίσιμο, γιατί το δίκτυο μπορεί να πάει από το σημείο Α στο σημείο Β, αλλά δεν μπορεί να πάει από το Β στο Α, δηλ. ακριβώς αυτό που θέλουμε. Το σύστημα μπορεί να ξεφύγει από το τοπικό ελάχιστο, αλλά εάν βρεθεί στο παγκόσμιο ελάχιστο, τότε δεν μπορεί να ξεφύγει. Στο σημείο αυτό πρέπει να αλλάζουμε την θερμοκρασία με πολύ αργό ρυθμό. Σημασία έχει να αναγνωρίσουμε πότε είμαστε κοντά όταν έχουμε απότομη ελάττωση στο C, δηλ. έχουμε απότομη αλλαγή στο ρυθμό αλλαγής θερμοκρασίας ως προς το σφάλμα. Μόλις όμως φθάσουμε κοντά στο σημείο αυτό της κρίσιμης θερμοκρασίας, από εδώ και πέρα θα αλλάζουμε την θερμοκρασία σιγά-σιγά, ώστε να βρούμε το παγκόσμιο ελάχιστο. Σε θερμοκρασίες μακριά από την Tc, μπορούμε να χρησιμοποιούμε μεγαλύτερους ρυθμούς ελάττωσης T, πράγμα που επιταχύνει την διαδικασία εκπαίδευσης του δικτύου.

Μη-γραμμικά προβλήματα βελτιστοποίησης (optimization)

Η βελτιστοποίηση είναι μία τεχνική η οποία λύνει διάφορα προβλήματα, στα οποία το ερώτημα που τίθεται έχει την μορφή: Ποιός είναι ο καλύτερος τρόπος για να γίνει μια διεργασία, η οποία υπόκειται σε κάποιους συγκεκριμένους όρους; Προφανώς, σε τέτοια προβλήματα υπάρχουν πολλές λύσεις, αλλά μία μόνο λύση είναι η βέλτιστη, και ικανοποιεί πλήρως τους όρους που τίθενται στο πρόβλημα που εξετάζουμε. Μερικές άλλες λύσεις είναι πολύ καλές, δηλ. χωρίς να είναι ίσες με την βέλτιστη, είναι όμως πολύ κοντά σε αυτήν. Ένα τέτοιο από μαθηματικό πρόβλημα συνάρτησης y=f(x), είναι να βρούμε για ποια τιμή του x έχουμε το μέγιστο (ή το ελάχιστο) στην συνάρτηση y. Η βελτιοστοποίηση θα γίνει στην συνάρτηση y, και η λύση είναι να βρούμε την κατάλληλη τιμή της ανεξάρτητης μεταβλητής x.

Υπάρχουν γνωστές μέθοδοι στον διαφορικό λογισμό που λύνουν το πρόβλημα αυτό, και αποτελούν "κλασσικές" λύσεις, που είναι γνωστές στα μαθηματικά από πολύ καιρό. Πάντοτε όμως υπάρχουν περιπτώσεις που αναζητούμε μια νέα τεχνική για ειδικές περιπτώσεις. Εάν ο αριθμός των μεταβλητών είναι πολύ μεγάλος τότε και η "κλασσική" λύση δεν είναι εύκολη. Άλλες φορές δεν έχουμε την μορφή της συνάρτησης με ακριβή τρόπο, αλλά μόνον ποιοτικά. Σε τέτοιες περιπτώσεις τα νευρωνικά δίκτυα μπορεί να δώσουν πιο ικανοποιητικές λύσεις, με την στατιστική μέθοδο της κατανομής Caushy. Η διαδικασία που ακολουθείται είναι ανάλογη της γνωστής μας εκπαίδευσης ενός νευρωνικού δικτύου και είναι η εξής:

- Βρίσκουμε ορισμένα πρότυπα, τα οποία αποτελούνται από ζεύγη εισόδου-εξόδου.
- Το νευρωνικό δίκτυο εκπαιδεύεται στα ζεύγη αυτά με την γνωστή διαδικασία της αλλαγής των βαρών w. Ουσιαστικά το νευρωνικό δίκτυο

δημιουργεί μία εσωτερική δομή ενός αγνώστου συστήματος. Όσο μεγαλύτερος είναι ο αριθμός των προτύπων που παρουσιάζουμε τόσο καλύτερη και σωστότερη θα είναι η δομή που θα βρεί το νευρωνικό δίκτυο. Τώρα, εάν παρουσιάσουμε ένα άγνωστο πρότυπο στο επίπεδο εισόδου, το νευρωνικό δίκτυο θα πρέπει να δώσει την ίδια απάντηση όπως θα έδινε το σύστημα, του οποίου βρήκαμε την δομή.

- Τέλος προσπαθούμε να βρούμε μία συνάρτηση που αναπαριστά πόσο επιτυχής είναι η δομή που βρήκαμε, και ακολούθως η συνάρτηση αυτή μεγιστοποιείται. Οι τιμές των εισόδων τώρα μεταβάλλονται, όπως προηγουμένως μεταβάλλονταν τα βάρη, με τον ίδιο ακριβώς αλγόριθμο. Η εκπαίδευση λοιπόν συνίσταται, αντί να βρούμε τα βάρη που ελαχιστοποιούν το σφάλμα, στο να βρούμε τις εισόδους που μεγιστοποιούν την συνάρτηση δομής.