ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΠΟΥΚΙΟΥ ΔΑΝΑΗ

«ΜΕΘΟΔΟΙ ΕΛΑΧΙΣΤΟΠΟΙΗΣΗΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ»

ΕΠΙΒΛΕΠΩΝ:

ΚΩΝΣΤΑΝΤΙΝΟΣ ΧΡΥΣΑΦΙΝΟΣ
ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ

ΚΩΝΣΤΑΝΤΙΝΟΣ ΧΡΥΣΑΦΙΝΟΣ
ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΕΜΜΑΝΟΥΗΛ ΓΕΩΡΓΟΥΛΗΣ
ΚΑΘΗΓΗΤΗΣ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΜΙΧΑΗΛ ΛΟΥΛΑΚΗΣ
ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΑΘΗΝΑ, ΦΕΒΡΟΥΑΡΙΟΣ 2021

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον Αναπληρωτή Καθηγητή Ε.Μ.Π. κ. Κωνσταντίνο Χρυσαφίνο, για την επίβλεψη και τη συνεχή βοήθεια και καθοδήγησή του καθ΄ όλη τη διάρκεια εκπόνησης της παρούσας εργασίας.

Θα ήθελα, επίσης, να ευχαριστήσω την οικογένειά μου, τους φίλους μου και όλους τους ανθρώπους που με βοήθησαν, με στήριξαν και πίστεψαν σε μένα σε όλα τα χρόνια των σπουδών μου.

Ιτέα, Ιανουάριος 2021
Δανάη Μπουκίου

Περίληψη

Η παρούσα εργασία βασίζεται στο αναγνωρισμένο από την Εταιρεία Βιομηχανικών και Εφαρμοσμένων Μαθηματικών (Society for Industrial and Applied Mathematics - SIAM) έργο των L. Bottou, F. E. Curtis και J. Nocedal, «Optimization Methods in Large-Scale Machine Learning» (SIAM Review, 2018).

Ξεκινάμε θέτωντας τα μαθηματικά θεμέλια που χρειάζονται για την κατανόηση του τρόπου λειτουργίας των αλγορίθμων. Στη συνέχεια, δίνονται οι κυριότερες μεθόδοι αριθμητικής βελτιστοποίησης χωρίς περιορισμούς και συγκεκριμένα οι μέθοδοι κλίσης, συζυγούς κλίσης, μέγιστης καθόδου, Newton-Raphson και quasi-Newton. Συνεχίζουμε με την εισαγωγή ορισμών και κανόνων από τη στατιστική και τη θεωρία πιθανοτήτων, απαραίτητων για τη μηχανική μάθηση. Επίσης, γίνεται μια συνοπτική παρουσίαση των μαθηματικών εργαλείων που μοντελοποιούν ένα πρόβλημα μηχανικής μάθησης. Δίνεται ένα παράδειγμα από τη μηχανική μάθηση μεγάλης κλίμακας, η ταξινόμηση κειμένου. Με βάση αυτό το πρόβλημα, δείχνουμε πώς μια εφαρμογή της μηχανικής μάθησης λύνεται με έναν αλγόριθμο βελτιστοποίησης και τα είδη μεθόδων που προκύπτουν.

Στο τέταρτο κεφάλαιο αναλύεται μια σημαντική μέθοδος βελτιστοποίησης στη μηχανική μάθηση μεγάλης κλίμακας, η στοχαστική μέθοδος κλίσης. Γίνεται παρουσίαση του αλγορίθμου, στη μορφή που εμφανίζεται στην εργασία των Bottou, Curtis & Nocedal (2018). Έπειτα, μελετώνται οι προϋποθέσεις ως προς την αντικειμενική συνάρτηση, τη στοχαστική κατεύθυνση και το βήμα, με τη μορφή υποθέσεων και λημμάτων, που θα πρέπει να πληρούνται ώστε η μέθοδος να συγκλίνει στη βέλτιστη λύση. Επακόλουθο αυτών των συνθηκών είναι τα θεωρήματα σύγκλισης της μεθόδου στοχαστικής κλίσης, με δύο περιπτώσεις στην επιλογή βήματος, για κυρτές και μη κυρτές συναρτήσεις. Τέλος, στο πέμπτο κεφάλαιο, δίνονται κάποιες τεχνικές που μπορούν να βελτιώσουν την απόδοση την ταχύτητα της στοχαστικής μεθόδου κλίσης, οι οποίες κάνουν χρήση πληροφοριών δεύτερης τάξης.

Λέξεις - κλειδιά: μηχανική μάθηση, βελτιστοποίηση, αριθμητικές μέθοδοι, στοχαστική μέθοδος κλίσης, μέθοδοι δεύτερης τάξης

Abstract

This thesis is based on the paper recognized by Society for Industrial and Applied Mathematics (SIAM) by L. Bottou, F. E. Curtis and J. Nocedal, ((Optimization Methods in Large-Scale Machine Learning)) (SIAM Review, 2018).

We begin by setting the mathematical foundations needed for better understanding of how the algorithms presented in this paper run. Furthermore, we give the main numerical methods used in unconstrained optimization, namely the gradient, conjugate gradient and steepest descent methods, as well as Newton-Raphson and quasi-Newton methods. We continue with definitions and rules given by statistics and probability theory that are necessary in machine learning. In addition, we give a brief presentation of the tools used in mathematical models of large-scale machine learning problems. As an example of such problem, we present text classification. Based on this, we show how a machine learning application is solved by an optimization algorithm and the types of numerical methods that stem from this process.

In the fourth chapter we analyze stochastic gradient descent method (SG), an important optimization method in the field of large-scale machine learning. We present the algorithm, in the form given in the original paper by Bottou, Curtis & Nocedal (2018). Subsequently, we study the conditions over objective funtions, stochastic directions and stepsizes, in the form of assumptions and lemmas, under which convergence to the optimal solution is guaranteed. As a consequence, we arrive to convergence theorems of the SG method running for convex and non-convex objective functions, with two different cases over stepsize selection. Lastly, the fifth chapter discusses different techniques that use second-order information, which can improve the efficiency and convergence rate of SG method.

Keywords: machine learning, optimization, numerical methods, stochastic gradient descent, second-order methods

Περιεχόμενα

1 Εισαγωγή

Από τα πρώιμα στάδια ανάπτυξης των ηλεκτρονικών υπολογιστών, ένα ερώτημα που δημιουργήθηκε ήταν το εξής: υπάρχει τρόπος ώστε ο υπολογιστής να είναι ικανός να μάθει· Το ερώτημα αυτό, έδωσε το έναυσμα για τη διαδικασία εύρεσης της κατάλληλης τεχνολογίας, αλλά και του λογισμικού, ώστε να δοθεί σε έναν υπολογιστή ή σε ένα σύστημα υπολογιστών η δυνατότητα να μαθαίνει και να παίρνει αποφάσεις. Παρόλο που η έρευνα αναπτύσσεται με ταχύτατους ρυθμούς από τη δεκαετία του 1950 μέχρι και σήμερα, δεν έχουν καταφέρει να δημιουργηθούν τα κατάλληλα εργαλεία ώστε η διαδικασία μάθησης ενός υπολογιστή να είναι ισάξια με την ανθρώπινη. Ωστόσο, η έρευνα αυτή έδωσε την ευκαιρία για την άνθηση ενός νέου πεδίου ενδιαφέροντος, βασισμένο σε τεχνικές δανεισμένες από την στατιστική, τη μηχανική μάθηση. Σύμφωνα με τον Murphy (2012, σ. 1), ως μηχανική μάθηση ορίζεται «ένα σύνολο μεθόδων οι οποίες μπορούν αυτόματα να εντοπίσουν μοτίβα σε δεδομένα και έπειτα να τα χρησιμοποιήσουν ώστε να προβλέψουν μελλοντικά δεδομένα ή να λάβουν διάφορες αποφάσεις σε συνθήκες αβεβαιότητας, όπως το πώς θα συλλέξουμε περισσότερα δεδομένα».

Όπως είναι αναμενόμενο, σε κάθε πρόβλημα στο οποίο επικρατεί αβεβαιότητα εφαρμόζονται στοιχεία της θεωρίας πιθανοτήτων. Τα προβλήματα αυτά προέρχονται από τη σύγχρονη καθημερινότητα, όπως από εφαρμογές της οικονομίας, της πληροφορικής και της ιατρικής, και προκύπτουν από την ανάγκη για ανάλυση μεγάλου όγκου δεδομένων. Η διαδικασία της μηχανικής μάθησης, παρέχει λύσεις σε αυτά τα προβλήματα όταν οι κλασικοί τρόποι αντιμετώπισης δεν είναι αρκετοί ή αποτυγχάνουν. Για την επίλυση των προβλημάτων αυτών, χρησιμοποιούνται στατιστικά μοντέλα που αναπαριστούν με όσο το δυνατόν καλύτερο τρόπο το σύστημα μάθησης με το οποίο μπορούμε να εργαστούμε. Σημαντικό ρόλο, όμως, στην επίλυση των προβλημάτων, παίζουν και οι αριθμητικές μέθοδοι που δίνουν τους κατάλληλους αλγόριθμους για την εξαγωγή συμπερασμάτων.

Συνεπώς, στο κέντρο της μηχανικής μάθησης βρίσκεται η μαθηματική βελτιστοποίηση. Καθώς οι περισσότερες εργασίες μηχανικής μάθησης μοντελοποιούνται, δημιουργούνται μαθηματικά προβλήματα βάσει των δεδομένων που παρέχονται. Στην παρούσα εργασία, η διαδικασία βελτιστοποίησης περιλαμβάνει τη χρήση αριθμητικών αλγορίθμων για την επίλυση προβλημάτων που προκύπτουν από ένα μοντέλο της μηχανικής μάθησης. Οι μέθοδοι βελτιστοποίησης ασχολούνται με την εύρεση της βέλτιστης λύσης προβλημάτων που έχουν ως μεταβλητές παραμέτρους ενός συστήματος, το οποίο λαμβάνει αποφάσεις δεχόμενο νέες πληροφορίες. Δηλαδή, βασισμένο στις πληροφορίες που διαθέτει μέχρι τώρα, επιλέγει τις παραμέτρους που μπορούν να χρησιμοποιηθούν για τη βέλτιστη δυνατή λύση του τρέχοντος προβλήματος.

Με βάση λοιπόν, τους σκοπούς της μηχανικής μάθησης, θα παρουσιάσουμε στη συνέχεια τεχνικές και αλγορίθμους που μπορούν να αντιμετωπίσουν τα ιδιαίτερα μαθηματικά προβλήματα που δημιουργούνται από τη μοντελοποίηση δεδομένων μεγάλης κλίμακας. Ξεκινάμε με τις βασικότερες κλασικές μεθόδους βελτιστοποίησης και τα απαραίτητα μαθηματικά εργαλεία. Το υπόβαθρο αυτό, θα μας βοηθήσει στην καλύτερη κατανόηση της διαδικασίας βελτιστοποίησης πολύπλοκων συναρτήσεων, οι οποίες εμφανίζονται σχεδόν συνέχεια στη μηχανική μάθηση. Στη συνέχεια, γίνεται

μετάβαση στο στοχαστικό περιβάλλον της μηχανικής μάθησης μεγάλης κλίμακας, λαμβάνοντας υπ΄ όψιν τη στατιστική φύση των εργασιών που μας ενδιαφέρουν. Για το σκοπό αυτό, παρουσιάζεται ένα ενδεικτικό παράδειγμα εργασίας, η ταξινόμηση κειμένων. Στο παράδειγμα αυτό δίνεται ο τρόπος με τον οποίο ανακύπτουν προβλήματα κυρτής βελτιστοποίησης μέσα από μια εφαρμογή του πεδίου με το οποίο ασχολούμαστε.

Έπειτα, αναλύεται η μέθοδος βελτιστοποίησης η οποία κυριαρχεί σήμερα στη μηχανική μάθηση με μεγάλο όγκο δεδομένων, η στοχαστική μέθοδος κλίσης. Οι κλασικές μέθοδοι βελτιστοποίησης οι οποίες εξαρτώνται από τον αριθμό των δεδομένων που επεξεργάζεται ένα πρόβλημα, μπορούν να αποδειχθούν αποτελεσματικές για εργασίες μικρής κλίμακας. Ωστόσο, όταν ασχολούμαστε με μεγάλο όγκο παραδειγμάτων και παραμέτρων, η στοχαστική μέθοδος κλίσης εκμεταλλεύεται το γεγονός ότι δεν εξαρτάται από το μέγεθος του προβλήματος. Άρα, καταλήγουμε σε έναν αποδοτικό αλγόριθμο, με χαμηλό υπολογιστικό κόστος, ο οποίος τρέχει ικανοποιητικά για μια μεγάλη γκάμα κυρτών, αλλά και μη κυρτών συναρτήσεων.

Τέλος, δίνουμε μερικά παραδείγματα μεθόδων βελτιστοποίησης που κάνουν χρήση στοιχείων καμπυλότητας της αντικειμενικής συνάρτησης, υπό τη μορφή παραγώγων δεύτερης τάξης. Σκοπός του κεφαλαίου αυτού να ερευνήσουμε τεχνικές με πιθανά ωφέλη στην απόδοση της στοχαστικής μεθόδου κλίσης για πολύπλοκα και ογκώδη προβλήματα. Πρωτού περάσουμε στο κύριο μέρος, πρέπει να αναφέρουμε ότι οι αλγόριθμοι βελτιστοποίησης στη μηχανική μάθηση που παρουσιάζονται στην παρούσα εργασία, καθώς και η ανάλυσή τους, είναι επιλογές από το μεγάλης σημασίας για το αντικείμενο έργο των L. Bottou, F. E. Curtis και J. Nocedal, «*Optimization Methods in Large-Scale Machine Learning*». Η εργασία αυτή δημοσιεύτηκε το 2018 από την Εταιρεία Βιομηχανικών και Εφαρμοσμένων Μαθηματικών (Society for Industrial and Applied Mathematics - SIAM), στο επιστημονικό περιοδικό *SIAM Review*.

2 Κλασικές Αριθμητικές Μέθοδοι

2.1 Μαθηματικό Υπόβαθρο

Στην ενότητα αυτή, ξεκινάμε με μια συνοπτική παρουσίαση των απαραίτητων ορισμών, συνθηκών και θεωρημάτων που θα χρησιμοποιηθούν για την ανάλυση μεθόδων βελτιστοποίησης. Θεωρώντας γνωστές τις έννοιες του διανύσματος, του πίνακα και του διανυσματικού χώρου πάνω στο χώρο $\mathbb{R}^n$, προκύπτει το ερώτημα: πώς μπορούμε να μετρήσουμε το μήκος ενός διανύσματος, αλλά και την απόσταση δύο στοιχείων στο διανυσματικό χώρο· Για τη μέτρηση του μεγέθους ενός διανύσματος, χρησιμοποιείται η συνάρτηση της *νόρμας*. Εάν V είναι ένας πραγματικός διανυσματικός χώρος, μία απεικόνιση $\| \ \| : V \to \mathbb{R}$ καλείται *νόρμα* αν ικανοποιεί τις ιδιότητες:

(i) $\|x\| \geq 0$ για κάθε $x \in V$ και $\|x\| = 0$ αν και μόνο αν $x = 0$.

(ii) $\|\lambda x\| = |\lambda| \cdot \|x\|$ για κάθε $x \in V$ και $\lambda \in \mathbb{R}$.

(iii) $\|x + y\| \leq \|x\| + \|y\|$ για κάθε $x, y \in V$ (τριγωνική ανισότητα).

Παρακάτω δίνονται κάποια παραδείγματα νορμών.

1. Η *ευκλείδεια νόρμα* $\| \ \|_2$ στον $\mathbb{R}^n$ ορίζεται ως εξής: Για $x \in \mathbb{R}^n$ της μορφής $x = (x_1, x_2, \ldots, x_n)$ θέτουμε

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}}.$$

2. Στον $\mathbb{R}^n$ θεωρούμε επίσης τις νόρμες $\| \ \|_1$ και $\| \ \|_\infty$ οι οποίες για $x \in \mathbb{R}^n$ της μορφής $x = (x_1, x_2, \ldots, x_n)$ ορίζονται ως εξής:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\|x\|_\infty = \max\{|x_i| : i = 1, 2, \ldots, n\}.$$

3. Στο χώρο $\mathbb{R}^n$ ορίζονται επίσης οι νόρμες $\| \ \|_p$ για $1 < p < \infty$ ως εξής: Για $x = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^n$ ορίζουμε

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

Κατ΄ αντίστοιχο τρόπο με τη νόρμα διανύσματος, ορίζεται επίσης η *νόρμα πίνακα* στο διανυσματικό χώρο των τετραγωνικών πινάκων $n \times n$, $M_n = \mathbb{R}^{n \times n}$ (Μπακόπουλος & Χρυσοβέργης, 2009, σ. 29). Μια νόρμα στο M_n είναι μία συνάρτηση με πεδίο ορισμού το M_n και τιμές στο $\mathbb{R}$ που ικανοποιεί τις ακόλουθες ιδιότητες:

(i) $\|A\| \geq 0$ για κάθε $A \in M_n$ και $\|A\| = 0$ αν και μόνο αν $A = 0$ (μηδενικός πίνακας).

(ii) $\|\lambda A\| = |\lambda|\|A\|$ για κάθε $A \in M_n$ και $\lambda \in \mathbb{R}$.

(iii) $\|A + B\| \leq \|A\| + \|B\|$ για κάθε $A, B \in M_n$.

(iv) $\|A \cdot B\| \leq \|A\| \, \|B\|$ για κάθε $A, B \in M_n$.

Πρόταση 2.1. *Έστω $\| \ \|_a$ και $\| \ \|_b$, με $1 < a < b < \infty$, δύο οποιεσδήποτε νόρμες στο $\mathbb{R}^n$. Τότε υπάρχουν σταθερές $c > 0$ και $C > 0$ τέτοιες ώστε*

$$c\|x\|_a \leq \|x\|_b \leq C\|x\|_a,$$

δηλαδή όλες οι νόρμες στον χώρο $\mathbb{R}^n$ είναι ισοδύναμες. (Μπακόπουλος & Χρυσοβέργης, 2009, σ. 30)

Απόδειξη. Αρκεί να αποδειχθεί ότι για οποιαδήποτε νόρμα $\| \ \|$ στο $\mathbb{R}^n$, υπάρχουν σταθερές $m > 0$ και $M > 0$ τέτοιες ώστε

$$m\|x\|_\infty \leq \|x\| \leq M\|x\|_\infty.$$

Γνωρίζουμε ότι το σύνολο $S = \{u \in \mathbb{R}^n : \|u\|_\infty = 1\}$ είναι κλειστό και φραγμένο, άρα και συμπαγές, ενώ η συνάρτηση $f : \mathbb{R}^n \to \mathbb{R}$, όπου $f(x) = \|x\|$, είναι συνεχής στο πεδίο ορισμού της ως προς τη νόρμα $\| \ \|$ και τη νόρμα $\| \ \|_\infty$. Επομένως, υπάρχουν m και M τέτοια ώστε

$$m = \min_{u \in S} f(u) = \min_{u \in S} \|u\| > 0 \text{ και } M = \max_{u \in S} f(u) = \max_{u \in S} \|u\| > 0,$$

δηλαδή,

$$m \leq \|u\| \leq M, \text{ για κάθε } u \in S. \tag{2.1}$$

Επίσης, είναι προφανές ότι για $x \in \mathbb{R}^n$, με $x \neq 0$, έχουμε $\frac{x}{\|x\|_\infty} \in S$, άρα από τη σχέση (2.1) προκύπτει

$$m \leq \frac{\|x\|}{\|x\|_\infty} \leq M \tag{2.2}$$

και συνεπώς

$$m\|x\|_\infty \leq \|x\| \leq M\|x\|_\infty. \tag{2.3}$$

□

Η παραπάνω πρόταση αποδεικνύεται παρόμοια και για νόρμες πινάκων. Για την απόδειξή της, εισάγαμε την έννοια του συμπαγούς συνόλου. Ένα σύνολο $S \subset \mathbb{R}^n$ ονομάζεται *συμπαγές* αν κάθε ακολουθία στοιχείων του S περιέχει μια υπακολουθία που συγκλίνει σε ένα στοιχείο του S. Συγκεκριμένα, έστω $\{x_n\}$ ακολουθία στο $\mathbb{R}$ και $x_0 \in \mathbb{R}$. Λέμε ότι το x_0 είναι *όριο* της $\{x_n\}$ και συμβολίζουμε με $\lim_n x_n = x_0$ αν για κάθε $\varepsilon > 0$ υπαρχει $n_0 = x_0(\varepsilon)$ φυσικός αριθμός ώστε για κάθε $n \geq n_0$, $|x_n - x_0| \leq \varepsilon$. Αν η $\{x_n\}$ έχει όριο το x_0 τότε λέμε ότι η $\{x_n\}$ *συγκλίνει* στο x_0 και γράφουμε $x_n \to x_0$. Κατ΄ αντιστοιχία, λέμε ότι μια ακολουθία $\{x_k\}$ στο $\mathbb{R}^n$ συγκλίνει στο $x \in \mathbb{R}^n$ και γράφουμε $\lim_k x_k = x$ ή $x_k \to x$ αν $\lim_k \|x_k - x\| = 0$ για οποιαδήποτε νόρμα στο $\mathbb{R}^n$. Γνωρίζουμε, επίσης, ότι κάθε συγκλίνουσα ακολουθία $\{x_k\}$ στον $\mathbb{R}^n$ είναι φραγμένη, δηλαδή υπάρχει $M > 0$ ώστε για κάθε $n \in \mathbb{N}$, $\|x_k\|_2 \leq M$.

Συνεχίζοντας με ορισμούς συνόλων, ένα σύνολο $V \subset \mathbb{R}$ λέγεται ανοικτό αν για κάθε $x \in V$ υπάρχει $\varepsilon > 0$ ώστε $(x - \varepsilon, x + \varepsilon) \subset V$, ενώ ένα υποσύνολο S του $\mathbb{R}^n$ λέγεται *ανοικτό* αν για κάθε $x \in S$ υπάρχει $\delta > 0$ τέτοιο ώστε $\|y - x\| < \delta$ για κάθε $y \in S$. Το συμπλήρωμα του ανοικτού συνόλου S, $\mathbb{R}^n \setminus S$, λέγεται *κλειστό σύνολο*. *Φραγμένο* ονομάζεται ένα υποσύνολο S του $\mathbb{R}^n$ αν υπάρχει $M > 0$ τέτοιο ώστε $\|x\| \leq M$ για κάθε $x \in S$. Από τον ορισμό που δόθηκε για το συμπαγές σύνολο, αποδεικνύεται ότι το υποσύνολο S του $\mathbb{R}^n$ είναι συμπαγές αν και μόνο αν είναι κλειστό και φραγμένο. Τέλος, παρουσιάζουμε τον ορισμό του κυρτού συνόλου. Ένα υποσύνολο S του $\mathbb{R}^n$ λέγεται *κυρτό* αν

$$(1 - c)x + cy \in S \text{ για κάθε } x, y \in S, c \in [0, 1]. \tag{2.4}$$

Σε αυτό το σημείο, θα δώσουμε κάποιους χρήσιμους ορισμούς συναρτήσεων. Μία συνάρτηση $f : S \subset \mathbb{R}^n \to \mathbb{R}$ είναι *συνεχής σε ένα σημείο* $x_0 \in S$ όταν για κάθε $\varepsilon > 0$ υπάρχει $\delta(x_0, \varepsilon)$ ώστε για κάθε $x \in S$ να ισχύει:

$$\|x - x_0\| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon \tag{2.5}$$

και *συνεχής στο* S αν είναι συνεχής σε κάθε σημείο του S. Μια συνάρτηση $f : S \subset \mathbb{R}^n \to \mathbb{R}$, με το σύνολο S να είναι κυρτό, λέγεται *κυρτή* αν

$$f[(1 - c)x + cy] \leq (1 - c)f(x) + cf(y) \tag{2.6}$$

για κάθε $x, y \in S$, $c \in [0, 1]$ και *αυστηρά κυρτή* αν

$$f[(1 - c)x + cy] < (1 - c)f(x) + cf(y) \tag{2.7}$$

για κάθε $x, y \in S$, $x \neq y$, $c \in (0, 1)$. Τέλος, η συνάρτηση f ονομάζεται *σχεδόν κυρτή* αν ισχύει

$$f(z) \leq \max\{f(x), f(y)\} \text{ για κάθε } x, y, z \text{ με } x \leq z \leq y. \tag{2.8}$$

Θα λέμε ότι το $\bar{x} \in S$ είναι ένα *σημείο ολικού ελαχίστου* για το πρόβλημα ελαχιστοποίησης $\min_{x \in \mathbb{R}^n} f(x)$, όπου $f : \mathbb{R}^n \to \mathbb{R}^n$, εάν

$$f(\bar{x}) \leq f(x), \text{ για κάθε } x \in \mathbb{R}^n, \tag{2.9}$$

ενώ θα ονομάζουμε *σημείο τοπικού ελαχίστου* για το πρόβλημα ελαχιστοποίησης εάν υπάρχει $\epsilon > 0$ τέτοιο ώστε

$$f(\bar{x}) \leq f(x) \text{ για κάθε } x \in B(\bar{x}, \epsilon) \tag{2.10}$$

όπου $B(\bar{x}, \epsilon) \coloneqq \{x \in \mathbb{R}^n : \|x - \bar{x}\| \leq \epsilon\}$ είναι μια περιοχή του $\bar{x}$.

Στη συνέχεια, έστω Ω ένα ανοικτό υποσύνολο του $\mathbb{R}^n$. Αν η συνάρτηση $f : \Omega \to \mathbb{R}$ έχει μερικές παραγώγους πρώτης τάξης $\frac{\partial f(x)}{\partial x_i}$, όπου $i \in \{1, \ldots, n\}$ στο σημείο $x \in \Omega$, θα ονομάζουμε *κλίση* (gradient) της συνάρτησης f στο x το διάνυσμα

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}. \tag{2.11}$$

Επίσης, έστω η συνάρτηση $f : S \subset \mathbb{R}^n \to \mathbb{R}$ και $x \in S, h \in \mathbb{R}^n$. Αν $x + ah \in S$ για κάθε $a \neq 0$ αρκετά μικρό και το όριο

$$\delta f(x, h) = \lim_{a \to 0} \frac{f(x + ah) - f(x)}{a} \tag{2.12}$$

υπάρχει, τότε η δf καλείται παράγωγος της f στο διάνυσμα x *κατά την κατεύθυνση του* h. Για τη βελτιστοποίηση της f, θα πρέπει να αναζητήσουμε την κατεύθυνση κατά την οποία η f πλησιάζει πιο γρήγορα σε ένα ακρότατο. Αυτό μπορεί να επιτευχθεί χρησιμοποιώντας την παράγωγο κατά κατεύθυνση. Αν η συνάρτηση $f : \Omega \to \mathbb{R}^n$ είναι παραγωγίσιμη στο σημείο x του ανοικτού συνόλου Ω, τότε η παράγωγος $\delta f(x, h)$ υπάρχει στο διάνυσμα x για κάθε διάνυσμα $h \in \mathbb{R}^n$ και ισχύει

$$\delta f(x, h) = \nabla f(x)^T h. \tag{2.13}$$

Αν υποθέσουμε ότι $\|h\|_2 = 1$, από τη γνωστή ανισότητα Cauchy - Schwarz προκύπτει

$$|\delta f(x, h)| = |\nabla f(x)^T h| \leq \|\nabla f(x)\|_2 \|h\|_2 = \|\nabla f(x)\|_2. \tag{2.14}$$

Εάν στη συνέχεια διαλέξουμε το μοναδιαίο διάνυσμα $h = u := \frac{\nabla f(x)}{\|\nabla f(x)\|_2}$ και το αντικαταστήσουμε στην εξίσωση (2.3), καταλήγουμε στη σχέση $\delta f(x, u) = \|\nabla f(x)\|_2$. Συνεπώς, για κάθε μοναδιαίο διάνυσμα h έχουμε ότι

$$\delta f(x, h) \leq \delta f(x, u),$$

δηλαδή η κατεύθυνση u της κλίσης $\nabla f(x)$ είναι αυτή κατά την οποία η f έχει το μεγαλύτερο ρυθμό αύξησης $\delta f(x, h)$ κοντά στο x (Μπακόπουλος & Χρυσοβέργης, 2009, σ. 200). Μπορούμε, επομένως, να ελαχιστοποιήσουμε την f κινούμενοι κατά την κατεύθυνση της αρνητικής κλίσης $-\nabla f(x)$ με τον μεγαλύτερο ρυθμό μείωσης.

Εάν έχουμε μια διανυσματική συνάρτηση $f : \mathbb{R}^n \to \mathbb{R}^m$, τότε ορίζουμε τον *Ιακωβιανό* (Jacobian) πίνακα $J \in M^{m \times n}$ των παραγώγων πρώτης τάξης της συνάρτησης ως

$$J = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} & \cdots & \dfrac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}. \tag{2.15}$$

Ο πίνακας

$$\nabla^2 f(a) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{bmatrix}, \tag{2.16}$$

όπου $\frac{\partial^2 f}{\partial x_i \partial x_j}$, με $i, j \in \{1, 2, \ldots, n\}$ είναι μερικές παράγωγοι δεύτερης τάξης της f στο x, ονομάζεται *Εσσιανός* (Hessian) της συνάρτησης f στο x. Εάν οι μερικές παράγωγοι δεύτερης τάξης (και κατά συνέπεια και οι μερικές παράγωγοι πρώτης τάξης) της f είναι συνεχείς, τότε ο Εσσιανός της πίνακας είναι συμμετρικός. Σε αυτό το σημείο αναφέρουμε ότι ένα πίνακας $A \in M_n$ λέγεται *θετικός* όταν $x^T A x \geq 0$, για κάθε $x \in \mathbb{R}^n$ και *θετικά ορισμένος* όταν $x^T A x > 0$, για κάθε $x \in \mathbb{R}^n$.

Έστω A ένας τετραγωνικός πίνακας $n \times n$, με μη μηδενική ορίζουσα, για τον οποίο το γραμμικό σύστημα $Ax = f(x)$ έχει μοναδική λύση και το ομογενές σύστημα $Ax = 0$ έχει μοναδική λύση $x = 0$. Τότε ο πίνακας A καλείται *ομαλός* και η μοναδική λύση του γραμμικού συστήματος $Ax = f(x)$, όπου $f(x) = A^{-1}x$ δίνεται από τη σχέση $x = A^{-1}f(x)$. Επίσης, μία νόρμα πίνακα $\| \ \| : \mathbb{R}^{n\times n} \to \mathbb{R}$ λέγεται *φυσική* όταν $\|A\| = \sup_{0\neq x\in\mathbb{R}^n} \frac{\|Ax\|}{\|x\|}$ και παράγεται από τη νόρμα $\| \ \|$ του $\mathbb{R}^n$. Από τα παραπάνω, για τον ομαλό πίνακα A με νόρμα $\| \ \|$, ο αριθμός $\kappa = \kappa(A) = \|A\| \cdot \|A\|^{-1}$ καλείται *δείκτης κατάστασης* του πίνακα A. Γενικά, ένα γραμμικό σύστημα καλείται *καλής κατάστασης* αν το κ δεν είναι πολύ μεγαλύτερο του 1, διαφορετικά καλείται *κακής κατάστασης*.

Στη συνέχεια, παρουσιάζονται δύο σημαντικά θεωρήματα, το θεώρημα μέσης τιμής (Μπακόπουλος & Χρυσοβέργης, 2009, σ. 200) και ο τύπος Taylor (Nocedal & S. Wright, 2006, σ. 14) για πραγματικές συναρτήσεις.

Θεώρημα 2.2 (Θεώρημα Μέσης Τιμής)**.** *Έστω πραγματική συνάρτηση $f : \Omega \to \mathbb{R}$. Εάν το σημείο $x + ah$ ανήκει στο πεδίο ορισμού Ω της f για κάθε $a \in [0, 1]$ και η f είναι συνεχής σε κάθε σημείο $x + ah$ με $a \in [0, 1]$ και παραγωγίσιμη σε κάθε σημείο $x + ah$ με $a \in (0, 1)$, τότε υπάρχει σταθερά $\mu \in (0, 1)$ τέτοια ώστε*

$$f(x + h) = f(x) + \nabla f(x + \mu h)^T h. \tag{2.17}$$

Θεώρημα 2.3 (Τύπος Taylor)**.** *Έστω $f : \Omega \to \mathbb{R}$ δύο φορές παραγωγίσιμη, με συνεχείς παραγώγους δεύτερης τάξης. Έαν ισχύει ότι $x+ah \in \Omega$ για κάθε $a \in [0, 1]$, τότε υπάρχει σταθερά $\mu \in (0, 1)$ τέτοια ώστε*

$$\nabla f(x + h) = \nabla f(x) + \int_0^1 \nabla^2 f(x + \mu h)h \, d\mu \tag{2.18}$$

και

$$f(x + h) = f(x) + \nabla f(x)^T h + \frac{1}{2}h^T \nabla^2 f(x + \mu h)h. \tag{2.19}$$

Επιπλέον, εάν το $\|h\|$ είναι αρκετά μικρό, έχουμε ότι $x + h \in \Omega$ και

$$f(x + h) = f(x) + \nabla f(x)^T h + \frac{1}{2}h^T \nabla^2 f(x)h + \varepsilon(h)\|h\|^2,$$

όπου $\varepsilon(h) \to 0$, καθώς $\|h\| \to 0$.

Για τον εντοπισμό ενός σημείου τοπικού ελαχίστου, ο άμεσος τρόπος είναι η εξέταση όλων των σημείων μιας περιοχής του υποψήφιου σημείου $\bar{x}$ έτσι ώστε να εξασφαλίσουμε ότι σε κανένα από αυτά η συνάρτηση f δεν έχει μικρότερη τιμή. Ωστόσο, εάν η f είναι δύο φορές συνεχώς παραγωγίσιμη, μπορούμε να αποφασίσουμε εάν το $\bar{x}$ είναι σημείο τοπικού ελαχίστου εξετάζοντας την κλίση της f, $\nabla f(\bar{x})$, σε αυτό το σημείο, καθώς και τον Εσσιανό πίνακα $\nabla^2 f(\bar{x})$. Με βάση το Θεώρημα Μέσης Τιμής και τον τύπο του Taylor όπως παρουσιάστηκαν παραπάνω, θα διατυπώσουμε τις συνθήκες για την εύρεση των σημείων ελαχίστου (Nocedal & S. Wright, 2006, σ. 14-17). Οι αναγκαίες συνθήκες προκύπτουν από την υπόθεση ότι το $\bar{x}$ είναι σημείο τοπικού ελαχίστου και την απόδειξη ιδιοτήτων των $\nabla f(\bar{x})$ και $\nabla^2 f(\bar{x})$.

Θεώρημα 2.4 (Αναγκαίες συνθήκες πρώτης τάξης)**.** *Εάν το $\bar{x}$ είναι σημείο τοπικού ελαχίστου και η f είναι συνεχώς παραγωγίσιμη σε μία ανοιχτή περιοχή του $\bar{x}$, τότε $\nabla f(\bar{x}) = 0$.*

Απόδειξη. Έστω, προς απαγωγή σε άτοπο, ότι $\nabla f(\bar{x}) \neq 0$. Ορίζουμε το διάνυσμα $h = -\nabla f(\bar{x})$ για το οποίο ισχύει $h^T \nabla f(\bar{x}) = -\|\nabla f(\bar{x})\|^2 < 0$. Επειδή η f είναι συνεχής κοντά στο $\bar{x}$, υπάρχει μία σταθερά $T > 0$ τέτοια ώστε

$$h^T \nabla f(\bar{x} + ah) < 0, \text{ για κάθε } a \in [0, T].$$

Για οποιοδήποτε $t \in (0, T]$, από το Θεώρημα Μέσης Τιμής έχουμε

$$f(\bar{x} + th) = f(\bar{x}) + th^T \nabla f(\bar{x} + \mu h), \text{ για κάποιο } \mu \in (0, t).$$

Επομένως, $f(\bar{x} + th) < f(\bar{x})$ για κάθε $t \in (0, T]$. Βρήκαμε, δηλαδή, μία κατεύθυνση η οποία απομακρύνεται από το $\bar{x}$ κατά την οποία η f μειώνεται, συνεπώς το $\bar{x}$ δεν είναι σημείο τοπικού ελαχίστου, άρα έχουμε αντίθεση. □

Ένα σημείο $\bar{x}$ καλείται *στάσιμο* όταν $\nabla f(\bar{x}) = 0$. Σύμφωνα με το παραπάνω θεώρημα, κάθε σημείο τοπικού ελαχίστου πρέπει να είναι στάσιμο σημείο. Τα στάσιμα σημεία για τις αντικειμενικές συναρτήσεις αποτελούν ακρότατα. Ωστόσο, ιδιαίτερη περίπτωση στάσιμων σημείων αποτελούν τα *σημεία σάγματος*, τα οποία δεν είναι ούτε σημεία μεγίστου ούτε ελαχίστου (Goodfellow, Bengio & Courville, 2016).

Θεώρημα 2.5 (Αναγκαίες συνθήκες δεύτερης τάξης)**.** *Εάν το $\bar{x}$ είναι σημείο τοπικού ελαχίστου της f και κάθε δεύτερη παράγωγος της f υπάρχει και είναι συνεχής σε μια ανοιχτή περιοχή του $\bar{x}$, τότε $\nabla f(\bar{x}) = 0$ και ο Εσσιανός πίνακας $\nabla^2 f(\bar{x})$ είναι θετικός.*

Απόδειξη. Γνωρίζουμε από το προηγούμενο θεώρημα ότι $\nabla f(\bar{x}) = 0$. Έστω ότι ο $\nabla^2 f(\bar{x})$ δεν είναι θετικός. Τότε μπορούμε να βρούμε ένα διάνυσμα h για το οποίο ισχύει ότι $h^T \nabla^2 f(\bar{x}) h < 0$ και επειδή η $\nabla^2 f$ είναι συνεχής κοντά στο $\bar{x}$, υπάρχει σταθερά $T > 0$ τέτοια ώστε $h^T \nabla^2 f(\bar{x} + ah) h < 0$ για κάθε $a \in [0, T]$.

Εφαρμόζοντας τον τύπο του Taylor γύρω από $\bar{x}$, έχουμε ότι για κάθε $t \in (0, T]$ και για κάποιο $\mu \in (0, t)$ είναι

$$f(\bar{x} + th) = f(\bar{x}) + th^T \nabla f(\bar{x}) + \frac{1}{2} h^T \nabla^2 f(\bar{x} + \mu h) h < f(\bar{x})$$

Όπως και στο προηγούμενο θεώρημα, βρήκαμε μία κατεύθυνση από το $\bar{x}$ κατά την οποία η f μειώνεται, επόμενως το $\bar{x}$ δεν είναι σημείο τοπικού ελαχίστου, το οποίο είναι αντίθετο με την αρχική μας υπόθεση. □

Θα διατυπώσουμε τις ικανές συνθήκες για τις παραγώγους της f στο σημείο $\bar{z}$ οι οποίες εξασφαλίζουν ότι το $\bar{x}$ είναι σημείο τοπικού ελαχίστου.

Θεώρημα 2.6 (Ικανές συνθήκες δεύτερης τάξης)**.** *Έστω ότι η f έχει συνεχείς παραγώγους δεύτερης τάξης σε μια ανοιχτή περιοχή του $\bar{x}$ και ότι $\nabla f(\bar{x}) = 0$ και ο Εσσιανός πίνακας της f στο $\bar{x}$ είναι θετικά ορισμένος. Τότε το $\bar{x}$ είναι μοναδικό σημείο τοπικού ελαχίστου της f.*

Απόδειξη. Επειδή ο Εσσιανός πίνακας είναι συνεχής και θετικά ορισμένος στο $\bar{x}$, μπορούμε να επιλέξουμε μία ακτίνα $r > 0$ ώστε ο $\nabla^2 f(x)$ να παραμένει θετικά ορισμένος για κάθε x στην ανοιχτή σφαίρα $\mathcal{D} = \{z : \|z - x\| < r\}$. Παίρνοντας οποιοδήποτε μη μηδενικό διάνυσμα h με $\|h\| < r$, έχουμε ότι $\bar{x} + h \in \mathcal{D}$ και έτσι

$$\begin{aligned} f(\bar{x} + h) &= f(\bar{x}) + h^T \nabla f(\bar{x}) + \frac{1}{2} \nabla^2 f(z) h \\ &= f(\bar{x}) + \frac{1}{2} h^T \nabla^2 f(z) h, \end{aligned}$$

όπου $z = \bar{x} + ah$ για κάποιο $a \in (0, 1)$. Εφόσον $z \in \mathcal{D}$, έχουμε ότι $h^T \nabla^2 f(z) h > 0$ και επομένως, $f(\bar{x} + h) > f(\bar{x})$, που είναι το ζητούμενο. □

Παρατηρούμε ότι οι ικανές συνθήκες δεύτερης τάξης δεν είναι και αναγκαίες: ένα σημείο $\bar{x}$ μπορεί να είναι μοναδικό σημείο τοπικού ελαχίστου και ταυτόχρονα να μην ικανοποιεί τις ικανές συνθήκες. Ένα απλό παράδειγμα δίνεται από τη συνάρτηση $f(x) = x^4$, για την οποία το $\bar{x} = 0$ είναι μοναδικό σημείο τοπικού ελαχίστου όπου ο Εσσιανός παύει να ορίζεται (και συνεπώς δεν είναι θετικά ορισμένος). Όταν η αντικειμενική συνάρτηση (δηλαδή η συνάρτηση την οποία θέλουμε να ελαχιστοποιήσουμε ή μεγιστοποιήσουμε)είναι κυρτή, τα ολικά και τα τοπικά σημεία ελαχίστου είναι εύκολο να εντοπιστούν.

Θεώρημα 2.7. *Εάν η f είναι κυρτή, κάθε σημείο τοπικού ελαχίστου $\bar{x}$ είναι ένα σημείο ολικού ελαχίστου για την f. Εάν, επίσης, η f είναι παραγωγίσιμη, τότε κάθε στάσιμο σημείο $\bar{x}$ είναι σημείο ολικού ελαχίστου της f.*

Απόδειξη. Έστω ότι το $\bar{x}$ είναι σημείο τοπικού αλλά όχι ολικού ελαχίστου. Τότε, υπάρχει ένα σημείο $z \in \Omega$ με $f(z) < f(\bar{x})$. Θεωρούμε το ευθύγραμμο τμήμα που ενώνει το $\bar{x}$ με το z, δηλαδή το

$$x = \lambda z + (1 - \lambda)\bar{x}, \text{ για κάποιο } \lambda \in (0, 1]. \tag{2.20}$$

Από την κυρτότητα της f έχουμε ότι

$$f(x) \leq \lambda f(z) + (1 - \lambda) f(\bar{x}) < f(\bar{x}). \tag{2.21}$$

Κάθε γειτονιά $\mathcal{N}$ του $\bar{x}$ περιέχει ένα τμήμα του ευθύγραμμου τμήματος (2.20), άρα θα υπάρχουν πάντα σημεία $x \in \mathcal{N}$ για τα οποία ικανοποιείται η (2.21). Επομένως, το $\bar{x}$ δεν είναι σημείο τοπικού ελαχίστου, το οποίο είναι αντίθετο με την αρχική υπόθεση.

Έστω, για το δεύτερο μέρος του θεωρήματος, ότι το $\bar{x}$ δεν είναι σημείο ολικού ελαχίστου και το z επιλέγεται όπως παραπάνω. Τότε, από την κυρτότητα της f, έχουμε

$$\begin{aligned} \nabla f(\bar{x})^T (z - \bar{x}) &= \frac{d}{d\lambda} f(\bar{x} + \lambda(z - \bar{x}))|_{\lambda=0} \\ &= \lim_{\lambda \searrow 0} \frac{f(\bar{x} + \lambda(z - \bar{x})) - f(\bar{x})}{\lambda} \\ &\leq \lim_{\lambda \searrow 0} \frac{\lambda f(z) + (1 - \lambda) f(\bar{x})}{\lambda} \\ &= f(z) - f(\bar{x}) < 0. \end{aligned}$$

Συνεπώς, $\nabla f(\bar{x}) \neq 0$ και άρα το $\bar{x}$ δεν είναι στάσιμο σημείο. □

Αυτά τα αποτελέσματα προσφέρουν τις βάσεις για τους αλγορίθμους βελτιστοποίησης χωρίς περιορισμούς. Συνεχίζουμε με τη μελέτη κάποιων σημαντικών κλασικών.

2.2 Μέθοδοι Κλίσης

Έστω η αντικειμενική συνάρτηση $f : \mathbb{R}^n \to \mathbb{R}$. Αναζητούμε σημείο $\bar{x} \in \mathbb{R}^n$ στο οποίο η f ελαχιστοποιείται χωρίς περιορισμούς, δηλαδή $f(\bar{x}) = \min_{x \in \mathbb{R}^n} f(x)$. Υποθέτουμε ότι οι παράγωγοι της f υπάρχουν και είναι συνεχείς, όπως επίσης ότι υπάρχει ένα σημείο $x_0 \in \mathbb{R}^n$ τέτοιο ώστε το κλειστό σύνολο $S = \{x : f(x) \leq f(x_0)\}$ να είναι φραγμένο, άρα και συμπαγές. Στην προηγούμενη ενότητα, είδαμε ότι υπάρχει τουλάχιστον ένα σημείο τοπικού ελαχίστου της f στο S, το οποίο είναι επίσης σημείο τοπικού ελαχίστου της f στο $\mathbb{R}^n$. Επίσης, από τις ικανές και αναγκαίες συνθήκες όπως διατυπώθηκαν στην προηγούμενη ενότητα, για να είναι το $\bar{x}$ σημείο τοπικού ελαχίστου, θα πρέπει $\nabla f(\bar{x}) = 0$.

2.2.1 Μέθοδος της Κλίσης

Οι μέθοδοι της κλίσης είναι μέθοδοι που βασίζονται στην ιδιότητα της καθόδου. Ο ακόλουθος αλγόριθμος περιγράφει τη *μέθοδο της κλίσης* (Μπακόπουλος & Χρυσοβέργης, 2009, σ. 225-226), η οποία είναι μια επαναληπτική μέθοδος καθόδου που υπολογίζει προσεγγιστικά τα σημεία που ικανοποιούν τις αναγκαίες συνθήκες του Θεωρήματος 1.4.

Αλγόριθμος 2.1 Μέθοδος Κλίσης

1: Για $k = 0$, επέλεξε ένα αρχικό διάνυσμα $x_0 \in \mathbb{R}^n$.
2: **for** $k = 1, 2, \ldots$ **do**
3: Υπολόγισε τα $\nabla f(x_k)$ και $\delta_k = -\|\nabla f(x_k)\|_2^2$.
4: **if** $\delta_k = 0$ **then**
5: Επέστρεψε $\bar{x} = x_k$.
6: **else**
7: Υπολόγισε το α_k είτε με ελάχιστο βήμα είτε με βήμα Armijo.
8: Θέσε $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$.
9: **end if**
10: **end for**

Όπως φαίνεται και στο Βήμα 7 του Αλγορίθμου 1.1, υπάρχουν δύο τρόποι υπολογισμού του βήματος α_k.

Βέλτιστο Βήμα. Κατά την εύρεση του *βέλτιστου βήματος*, αναζητούμε σε κάθε επανάληψη το μήκος βήματος α_k για το οποίο ισχύει ότι

$$f(x_k - \alpha_k \nabla f(x_k)) - f(x_k) = \min_{\alpha \geq 0} [f(x_k - \alpha \nabla f(x_k)) - f(x_k)]. \tag{2.22}$$

Για την εύρεση της σταθεράς α που δίνει το ελάχιστο στη σχέση $f(x_k - \alpha \nabla f(x_k)) - f(x_k)$, χρησιμοποιούμε κάποια αριθμητική μέθοδο. Μία μέθοδος για την εύρεση του α που δίνει το ελάχιστο βήμα της μεθόδου κλίσης είναι η *μέθοδος της χρυσής τομής*

(golden section method) (Gill, Murray & M. Wright, 1981/1997). Δίνοντας μια σύντομη περιγραφή της μεθόδου, υποθέτουμε ότι η συνάρτηση f εμφανίζει τοπικό ελάχιστο σε ένα μοναδικό σημείο $\bar{x}$ σε ένα διάστημα της μορφής $[a, b]$. Επιπλέον, υποθέτουμε ότι υπάρχουν δύο σημεία x_1, x_2 για τα οποία ισχύει ότι $a \leq x_1 \leq x_2 \leq b$. Η μέθοδος της χρυσής τομής θέτει την απόσταση των σημείων x_1 και x_2 από τα άκρα του διαστήματος a και b να είναι ίση, δηλαδή

$$x_2 - a = b - x_1 = r(b - a), \tag{2.23}$$

όπου $r = \frac{\sqrt{5}-1}{2} \approx 0.618$ είναι ο γνωστός Λόγος της Χρυσής Τομής. Εάν $f(x_1) < f(x_2)$, τότε το σημείο ελαχίστου βρίσκεται στο $[a, x_2]$, άρα η αναζήτησή μας περιορίζεται σε αυτό το διάστημα. Το νέο σημείο $x_1^{(1)}$ υπολογίζεται σύμφωνα με τη σχέση (2.23), όπου αντικαθιστούμε το b με το x_2:

$$x_2 - x_1^{(1)} = r(x_2 - a) \Rightarrow x_1^{(1)} = x_2 - r(x_2 - a).$$

Κατ' αντίστοιχο τρόπο δουλεύουμε όταν $f(x_2) < f(x_1)$ για την εύρεση του $x_2^{(1)}$. Επαναλαμβάνουμε τη διαδικασία έως ότου η αναζήτηση καταλήξει σε δύο σημεία όπου $f(x_1^{(k)}) = f(x_2^{(k)})$, οπότε υπολογίζουμε το σημείο ελαχίστου ως $\bar{x} = \frac{x_2^{(k)} - x_1^{(k)}}{2}$.

Άλλες μέθοδοι για τον υπολογισμό του ελάχιστου βήματος περιλαμβάνουν τη *μέθοδο Fibonacci*, η οποία παρομοιάζει την προσέγγιση της χρυσής τομής, χρησιμοποιώντας, όμως μεταβαλλόμενο r σε κάθε βήμα, τη *μέθοδο παραβολικής παρεμβολής* καθώς και τη *μέθοδο της διχοτόμησης*. Παρατηρούμε, όμως, ότι η διαδικασία εύρεσης του βέλτιστου βήματος δύσκολα μπορεί να καταλήξει σε έναν ακριβή υπολογισμό και έτσι δεν επιλέγεται συνήθως.

Βήμα Armijo. Για τον υπολογισμό του βήματος της μεθόδου κλίσης, χρησιμοποιούνται επίσης οι απλές και αποτελεσματικές *μέθοδοι τύπου Armijo* (Μπακόπουλος & Χρυσοβέργης, 2009, σ. 228-229). Σε αυτές τις μεθόδους, για κάθε k, επιλέγουμε ένα κατάλληλο $\alpha_k = \alpha > 0$ που ικανοποιεί την ανισότητα

$$\phi_k(\alpha) := f(x_k - \alpha \nabla f(x_k)) - f(x_k) \leq -\alpha b \|\nabla f(x_k)\|_2^2, \tag{2.24}$$

όταν το $b \in (0, 1)$ είναι δεδομένο. Έστω, τώρα, μία σταθερά $c \in (0, 1)$. Η επιλογή του α_k μπορεί να γίνει είτε με την *πρωτότυπη μέθοδο Armijo*, είτε με την *τροποποιημένη μέθοδο Armijo*. Στην πρωτότυπη μέθοδο, ξεκινάμε από ένα αρχικό σταθερό βήμα $\alpha > 0$. Εάν το α ικανοποιεί τη σχέση (2.24), θέτουμε διαδοχικά $\alpha_k \leftarrow \alpha$. Διαφορετικά, θέτουμε διαδοχικά $\alpha \leftarrow \alpha c$ και επιλέγουμε το πρώτο $\alpha_k = \alpha$ που ικανοποιεί την ανισότητα Armijo. Στην τροποποιημένη μέθοδο Armijo, ξεκινάμε πάλι από ένα αρχικό βήμα $\alpha > 0$. Αν το α ικανοποιεί τη σχέση (2.24), θέτουμε διαδοχικά $\alpha \leftarrow \alpha/c$, και διαλέγουμε το τελευταίο $\alpha_k = \alpha$ που ικανοποιεί την επαναληπτική σχέση (2.24). Αλλιώς, δουλεύουμε όπως και στην πρωτότυπη μέθοδο, δηλαδή θέτουμε διαδοχικά $\alpha \leftarrow \alpha c$ και διαλέγουμε το πρώτο $\alpha_k = \alpha$ που ικανοποιεί την ανισότητα.

Είναι απαραίτητο να σημειώσουμε ότι η επιλογή του βήματος Armijo είναι ταχύτερη από τον υπολογισμό του βέλτιστου βήματος για τον λόγο που αναφέρθηκε

παραπάνω. Δηλαδή, η εύρεση του βέλτιστου βήματος αποτελεί μια άπειρη διαδικασία που σπάνια θα μας δώσει μία ακριβή τιμή. Ωστόσο, η εύρεση του βέλτιστου βήματος είναι πιο πιθανό να μας οδηγήσει σε ολικό ελάχιστο για την f.

Η ορθότητα του αλγορίθμου κλίσης έγκειται στο γεγονός ότι εάν ο αλγόριθμος σταματά για κάποιο k, τότε $\nabla f(x_k) = 0$ και στην περίπτωση που αυτό το σύστημα έχει μοναδική λύση $\bar{x}$, τότε ολόκληρη η ακολουθία $\{x_k\}$ του αλγορίθμου συγκλίνει στο $\bar{x}$. Διαφορετικά, η ακολουθία $\{x_k\}$ περιέχει συγκλίνουσες υπακολουθίες, και κάθε όριο $\bar{x}$ τέτοιας ακολουθίας ικανοποιεί τη σχέση $\nabla f(\bar{x})$. Επίσης, εάν η f είναι κυρτή, τότε σύμφωνα με τον ορισμό της κυρτότητας, η μέθοδος της κλίσης υπολογίζει το ολικό ελάχιστο, ενώ αν είναι αυστηρά κυρτή, το σημείο ελαχίστου είναι μοναδικό. Στην πράξη, σταματάμε τις επαναλήψεις όταν $\|\nabla f(x_k)\|^2 \leq \varepsilon$, με ε δεδομένο (Μπακόπουλος & Χρυσοβέργης, 2009, σ. 226, 228).

2.2.2 Μέθοδος των Συζυγών Κλίσεων

Μία συνάρτηση $f : \mathbb{R}^n \to \mathbb{R}$ της μορφής

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c,$$

όπου A είναι ένας τετραγωνικός πίνακας $n \times n$, $b \in \mathbb{R}^n$ ένα διάνυσμα και $c \in \mathbb{R}$ μία σταθερά, λέγεται *τετραγωνική συνάρτηση* (Μπακόπουλος & Χρυσοβέργης, 2009, σ. 209). Η μέθοδος της κλίσης για μία τέτοια τετραγωνική συνάρτηση $f : \mathbb{R}^n \to \mathbb{R}$ έχει επαναληπτικό τύπο

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

όπου η κλίση της αντικειμενικής συνάρτησης και κατεύθυνση καθόδου της μεθόδου είναι η $-\nabla f(x_k) = b - Ax_k$, όταν ο πίνακας A είναι συμμετρικός και θετικά ορισμένος.

Ωστόσο, σε αυτή την περίπτωση μπορούμε να εφαρμόσουμε μία ταχύτερη μέθοδο κλίσης επάνω σε μια κατεύθυνση διαφορετική αυτής του διανύσματος καθόδου $-\nabla f(x_k)$ σε κάθε επανάληψη. Η *μέθοδος των συζυγών κλίσεων* επιλύει το πρόβλημα ελαχιστοποίησης

$$\min \phi(x) \;\text{ όπου }\; \phi(x) \coloneqq \frac{1}{2}x^T Ax - b^T x + c, \tag{2.25}$$

το οποίο ισοδυναμεί με την επίλυση ενός γραμμικού συστήματος $Ax = b$ όπου ο A είναι συμμετρικός και θετικά ορισμένος.

Μία από τις σημαντικότερες ιδιότητες της μεθόδου των συζυγών κλίσεων είναι ότι μπορεί να παράξει ένα σύνολο μη μηδενικών διανυσμάτων $\{y_0, y_1, \ldots, y_l\}$ με την ιδιότητα της *συζυγίας* ως προς τον συμμετρικό πίνακα A, δηλαδή

$$y_i^T A y p_j = 0, \text{ για κάθε } 0 \leq i \leq l, 0 \leq j \leq l, i \neq j, \tag{2.26}$$

για τα οποία ισχύει ότι $x_{k+1} = x_k + y_k$. Τα διανύσματα y_k υπολογίζονται σε κάθε επανάληψη σύμφωνα με τον τύπο

$$y_k = \alpha_k d_k,$$

όπου $\alpha_k > 0$ και τα διανύσματα d_k είναι γραμμικά ανεξάρτητα. Αντικαθιστώντας στον τύπο της ιδιότητας της συζυγίας (2.26) και με κατάλληλη απλοποίηση, καταλήγουμε ότι και για τα διανύσματα κατεύθυνσης d_k ικανοποιούν την ιδιότητα αυτή:

$$d_i^T A d_j = 0, \text{ για κάθε } i \neq j. \tag{2.27}$$

Άρα, η μέθοδος των συζυγών κλίσεων σε κάθε επανάληψη δίνει την τιμή

$$x_{k+1} = x_k + y_k = x_k + \alpha_k d_k. \tag{2.28}$$

Οι τιμές των α_k και d_k δίνονται από τον Αλγόριθμο 1.2 (Μπακόπουλος & Χρυσοβέργης, 2009, σ. 209)

Αλγόριθμος 2.2 Μέθοδος Συζυγών Κλίσεων

1: Για $k = 0$, επέλεξε $x_0 \in \mathbb{R}^n$ και θέσε $d_0 = g_0 = \nabla f(x_0) = Ax_0 - b$.
2: **for** $k = 1, 2, \ldots$ **do**
3: **if** $g_k = 0$ **then**
4: Σταμάτησε.
5: **else**
6: Υπολόγισε το βήμα $\alpha_k = -\frac{g_k^T g_k}{d_k^T A d_k}$
7: Θέσε $x_{k+1} = x_k + \alpha_k d_k$.
8: Θέσε $g_{k+1} = \nabla f(x_{k+1}) = Ax_{k+1} - b = \alpha_k A d_k$.
9: Θέσε $d_{k+1} = g_{k+1} + \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k} d_k$.
10: **end if**
11: **end for**

Επομένως, έχοντας γνωστά τα x_k και d_k, το πρόβλημα ελαχιστοποίησης (2.25) μετατρέπεται σε ένα μονοδιάστατο πρόβλημα βελτιστοποίησης όπου ζητούμενο είναι ένα $\alpha_k \geq 0$ τέτοιο ώστε

$$f(x_k - \alpha_k d_k) = \min_{\alpha \geq 0} f(x_k - \alpha d_k).$$

Από τον υπολογισμό του α_k έπεται το ακόλουθο θεώρημα (Nocedal & S. Wright, 2006, σ. 103-104).

Θεώρημα 2.8. *Για κάθε $x_0 \in \mathbb{R}^n$ η ακολουθία $\{x_k\}$ που παράγεται από τη μέθοδο των συζυγών κλίσεων, συγκλίνει στην ακριβή λύση του γραμμικού συστήματος $Ax = b$, $\bar{x}$, σε n το πολύ βήματα.*

Η απόδειξη του θεωρήματος με επαγωγή βρίσκεται αναλυτικά στο βιβλίο των Nocedal και S. Wright (2006, σ. 109-111). Παρόλο, όμως, που το παραπάνω θεώρημα εξασφαλίζει την ακριβή επίλυση του συστήματος σε σταθερό αριθμό βημάτων, στην πράξη η ιδιότητα της συζυγίας χάνεται γρήγορα λόγω των σφαλμάτων αποκοπής του υπολογιστή. Επομένως, ανάλογα και με τη δομή του πίνακα A, ίσως είναι απαραίτητη η εκτέλεση περαιτέρω βημάτων ώστε να υπολογιστεί η λύση με μεγάλη ακρίβεια.

Όταν έχουμε επαρκείς πληροφορίες για τις κατευθύνσεις, η μέθοδος των συζυγών κλίσεων γενικεύεται για συναρτήσεις όχι απαραίτητα τετραγωνικές. Η κύρια αλλαγή που γίνεται είναι ότι το βήμα α_k υπολογίζεται επαναληπτικά. Συγκεκριμένα, δύο

μέθοδοι συζυγών κλίσεων οι οποίες τρέχουν για γενικές, μη τετραγωνικές συναρτήσεις είναι η *μέθοδος Polak-Ribière* και η *μέθοδος Fletcher-Reeves* (Polak, 1997). Η μέθοδος των συζυγών κλίσεων ανήκει σε μια κατηγορία μεθόδων οι οποίες κατασκευάζουν διανύσματα κατεύθυνσης αναζήτησης χωρίς την αποθήκευση κάποιου πίνακα και χρησιμοποιούνται σε περιπτώσεις που η παραγοντοποίηση δεν είναι εφικτή εξαιτίας του μεγάλου μεγέθους του σχετικού πίνακα (Gill, Murray & M. Wright, 1997).

2.3 Μέθοδοι Γραμμικής Αναζήτησης

Ο ακριβής εντοπισμός ενός σημείου ελαχίστου είναι υπολογιστικά δαπανηρός. Ωστόσο, υπάρχουν συνθήκες οι οποίες εξασφαλίζουν βήματα τέτοια ώστε να μειώνεται σημαντικά η τιμή της συνάρτησης σε κάθε επανάληψη, και τα οποία είναι πεπερασμένου (μη απειροστού) μεγέθους ώστε να επιτυγχάνεται ταχεία σύγκλιση. Η διαδικασία του προσεγγιστικού εντοπισμού του σημείου ελαχίστου, που ικανοποιεί τις συνθήκες αυτές ονομάζεται *γραμμική αναζήτηση* (line search) (Nocedal & S. Wright, 2006, σ. 30-31). Κάθε επανάληψη μιας μεθόδου γραμμικής αναζήτησης, υπολογίζει μία κατεύθυνση αναζήτησης h_k και έπειτα αποφασίζει το πόσο θα κινηθεί σ′ αυτή την κατεύθυνση. Η κάθε επανάληψη δίνει το σημείο

$$x_{k+1} = x_k + a_k h_k, \tag{2.29}$$

όπου a_k είναι το βήμα (όπως είδαμε και στην προηγούμενη παράγραφο). Οι μέθοδοι καθόδου αποτελούν μια περίπτωση των μεθόδων γραμμικής αναζήτησης, όπου ικανοποιείται η σχέση $h_k^T f(x_k) < 0$ που μας εξασφαλίζει ότι η συνάρτηση f ελαττώνεται στην κατεύθυνση h_k. Η κατεύθυνση μπορεί να ισούται με $\nabla f(x_k)$, άρα καταλήγουμε στη μέθοδο της κλίσης. Βλέπουμε, δηλαδή, ότι η μέθοδος κλίσης είναι μία μέθοδος γραμμικής αναζήτησης. Επιπλέον, η κατεύθυνση αναζήτησης έχει συχνά τη μορφή

$$h_k = -B_k^{-1} \nabla f(x_k), \tag{2.30}$$

όπου B_k είναι ένας συμμετρικός και αντιστρέψιμος πίνακας.

2.3.1 Μέθοδος Απότομης Καθόδου

Η πιο απλή επιλογή κατεύθυνσης μιας μεθόδου γραμμικής αναζήτησης, είναι η κατεύθυνση της μέγιστης (ή πιο απότομης) καθόδου $-\nabla f(x_k)$ κατά την οποία, όπως είναι εύκολο να καταλάβουμε διαισθητικά, η f μειώνεται με τον μεγαλύτερο ρυθμό. Εάν η μέθοδος η οποία ελαχιστοποιεί την f κινείται κατά την κατεύθυνση $-\nabla f(x_k)$ σε κάθε βήμα, τότε λέγεται *μέθοδος απότομης καθόδου* (steepest descent method). Η μέθοδος απότομης καθόδου έχει το πλεονέκτημα ότι παρόλο που απαιτεί τον υπολογισμό της κλίσης $-\nabla f(x_k)$, δεν απαιτεί τον υπολογισμό καμίας παραγώγου δευτέρου βαθμού. Η διαφορά της μεθόδου απότομης καθόδου και της μεθόδου κλίσης, έγκειται στο γεγονός ότι το βήμα της μεθόδου απότομης καθόδου πρέπει να επιλέγεται με τέτοιο τρόπο ώστε να επιτυγχάνεται η μεγαλύτερη μείωση της αντικειμενικής συνάρτησης f κατά την κατεύθυνση της αρνητικής κλίσηςη $-\nabla f(x_k)$.

Μπορούμε να συλλέξουμε πληροφορίες για τη συμπεριφορά της μεθόδου απότομης καθόδου από τη βασική περίπτωση της τετραγωνικής αντικειμενικής συνάρτησης

με ακριβείς γραμμικές αναζητήσεις. Η ακριβής γραμμική αναζήτηση δίνει επακριβώς το βήμα a_k το οποίο οδηγεί στην ελαχιστοποίηση της συνάρτησης. Άρα, έστω

$$f(x) = \frac{1}{2}x^T Q x - b^T x + c, \tag{2.31}$$

όπου ο Q είναι ένας συμμετρικός και θετικά ορισμένος πίνακας και c σταθερά. Η κλίση δίνεται από τη σχέση $\nabla f(x) = Qx - b$ και το σημείο ελαχίστου $\bar{x}$ είναι η μοναδική λύση του γραμμικού συστήματος. Το βήμα που ελαχιστοποιεί την $f(x_k - \alpha \nabla f(x_k))$ βρίσκεται παραγωγίζοντας τη συνάρτηση ως προς α

$$f(x_k - \nabla f(x_k)) = \frac{1}{2}(x_k - \alpha \nabla f(x_k))^T Q(x_k - \alpha \nabla f(x_k)) - b^T(x_k - \alpha f(x_k)) \tag{2.32}$$

και θέτοντας την παράγωγο ίση με το μηδέν, άρα παίρνουμε

$$\alpha_k = \frac{\nabla f(x_k)^T \nabla f(x_k)}{\nabla f(x_k)^T Q \nabla f(x_k)}. \tag{2.33}$$

Με τη χρήση αυτού του ακριβούς βήματος, ο επαναληπτικός τύπος της μεθόδου απότομης καθόδου γίνεται

$$x_{k+1} = x_k - \left(\frac{\nabla f(x_k)^T \nabla f(x_k)}{\nabla f(x_k)^T Q \nabla f(x_k)} \right) \nabla f(x_k). \tag{2.34}$$

Ο ρυθμός σύγκλισης της μεθόδου, ο οποίος είναι γραμμικός, υπολογίζεται με τη χρήση της νόρμας $\|x\|_Q^2 = x^T Q x$ και από τη σχέση $Q\bar{x} = b$. Έτσι, βρίσκουμε ότι

$$\frac{1}{2}\|x - \bar{x}\|_Q^2 = f(x) - f(\bar{x}), \tag{2.35}$$

δηλαδή, αυτή η νόρμα δείχνει την απόσταση μεταξύ της τρέχουσας τιμής της αντικειμενικής συνάρτησης από τη βέλτιστη τιμή. Εάν $\lambda_{\min}$ και $\lambda_{\max}$ είναι η ελάχιστη και η μέγιστη ιδιοτιμή του πίνακα Q, αποδεικνύεται ότι

$$\begin{aligned} f(x_{k+1}) - f(\bar{x}) = \frac{1}{2}\|x_{k+1} - \bar{x}\|_Q^2 &\approx \frac{(\lambda_{\max} - \lambda_{\min})^2}{(\lambda_{\max} + \lambda_{\min})^2}(f(x_k) - f(\bar{x})) \\ &= \frac{(\kappa - 1)^2}{(\kappa + 1)^2}(f(x_k) - f(\bar{x})), \end{aligned} \tag{2.36}$$

όπου κ είναι ο δείκτης κατάστασης του Q (Gill, Murray & M. Wright, 1997, σ. 103). Οι σχέσεις (2.35) και (2.36) δείχνουν ότι οι τιμές f_k συγκλίνουν στην ελάχιστη $f(x_k)$ με γραμμικό ρυθμό.

2.3.2 Μέθοδος Newton

Αν η αντικειμενική συνάρτηση f έχει συνεχείς παραγώγους δεύτερης τάξης, μπορούμε να εφαρμόσουμε τη *μέθοδο Newton* (ή *Newton - Raphson*), στην οποία ο πίνακας B_k της σχέσης (2.30) είναι ο Εσσιανός πίνακας $\nabla^2 f(x_k)$ της f. Η κατεύθυνση αναζήτησης της μεθόδου Newton δίνεται από τη σχέση

$$h_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k) \tag{2.37}$$

και για δοθέν αρχικό σημείο $x_0 \in \mathbb{R}^n$, ο επαναληπτικός τύπος είναι

$$x_{k+1} = x_k + h_k = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k). \tag{2.38}$$

Προς αποφυγή υπολογισμού του αντίστροφου πίνακα $\nabla^2 f(x_k)^{-1}$, σε κάθε επανάληψη γίνεται επίλυση του γραμμικού συστήματος

$$\nabla^2 f(x_k) y_k = \nabla f(x_k)$$

ως προς y_k και ύστερα θέτουμε $x_{k+1} = x_k - y_k$.

Θα συζητήσουμε τις ιδιότητες της ταχύτητας τοπικής σύγκλισης της μεθόδου Newton. Γνωρίζουμε ότι για κάθε σημείο x σε μια περιοχή του $\bar{x}$ όπου ο $\nabla^2 f(\bar{x})$ είναι θετικά ορισμένος, ο Εσσιανός πίνακας $\nabla^2 f(x)$ θα είναι επίσης θετικά ορισμένος. Η μέθοδος Newton είναι καλώς ορισμένη σε αυτή την περιοχή και συγκλίνει τετραγωνικά. Λέμε ότι μία συνάρτηση $f : \mathbb{R}^n \to \mathbb{R}$ ικανοποιεί τη συνθήκη Lipschitz ή είναι Lipschitz συνεχής αν υπάρχει σταθερά $L \geq 0$ ώστε για κάθε $x, y \in \mathbb{R}^n$ να ισχύει

$$|f(x) - f(y)| \leq L\|x - y\|.$$

Ισχύει ότι κάθε συνάρτηση f που ικανοποιεί τη συνθήκη Lipschitz είναι συνεχής. Επισημαίνουμε, ακόμα, ότι η ακολουθία $\{x_k\}$ μιας επαναληπτικής μεθόδου συγκλίνει *γραμμικά* στο $\bar{x}$ εάν υπάρχει σταθερά $\mu \in (0, 1)$ τέτοια ώστε

$$\lim_{k\to\infty} \frac{|x_{k+1} - \bar{x}|}{|x_k - \bar{x}|} = \mu.$$

Αντίστοιχα, η ακολουθία $\{x_k\}$ συγκλίνει *υπογραμμικά* (δηλαδή με ταχύτητα μικρότερη από τη γραμμική σύκλιση) εάν

$$\lim_{k\to\infty} \frac{|x_{k+1} - \bar{x}|}{|x_k - \bar{x}|} = 1.$$

Τέλος, λέμε ότι η ακολουθία συκλίνει *τετραγωνικά* εάν για κάποια σταθερά $M > 0$ (όχι απαραίτητα μικρότερη του 1), ισχύει ότι

$$\lim_{k\to\infty} \frac{|x_{k+1} - \bar{x}|}{|x_k - \bar{x}|^2} < M.$$

Θεώρημα 2.9. *(Nocedal & S. Wright, 2006, σ. 44-45) Έστω ότι η f είναι δύο φορές παραγωγίσιμη και ότι ο Εσσιανός πίνακας $\nabla^2 f(x)$ είναι Lipschitz συνεχής σε μια περιοχή ενός σημείου τοπικού ελαχίστου $\bar{x}$ στην οποία οι ικανές συνθήκες του Θεωρήματος 1.6 ικανοποιούνται. Εάν ισχύει ο επαναληπτικός τύπος (2.37), τότε:*

(i) εάν το αρχικό σημείο x_0 είναι αρκετά κοντά στο $\bar{x}$, η ακολουθία $\{x_k\}$ συγκλίνει στο $\bar{x}$,

(ii) η ταχύτητα σύγκλισης της $\{x_k\}$ είναι τετραγωνική και

(iii) η ακολουθία των νορμών των κλίσεων $\{\|\nabla f(x_k)\|\}$ συγκλίνει τετραγωνικά στο μηδέν.

Απόδειξη. Από τον επαναληπτικό τύπο της μεθόδου Newton και τη συνθήκη $\nabla f(\bar{x}) = 0$ έχουμε ότι

$$\begin{aligned} x_k + h_k - \bar{x} &= x_k - \bar{x} - \nabla^2 f(x_k)^{-1} \nabla f(x_k) \\ &= \nabla^2 f(x_k)^{-1} [\nabla^2 f(x_k)(x_k - \bar{x}) - (\nabla f(x_k) - \nabla f(\bar{x}))]. \end{aligned} \tag{2.39}$$

Εφόσον το Θεώρημα Taylor μας λέει ότι

$$\nabla f(x_k) - \nabla f(\bar{x}) = \int_0^1 \nabla^2 f(x_k + \mu(\bar{x} - x_k))(x_k - \bar{x}) d\mu,$$

έχουμε

$$\begin{aligned} &\|\nabla^2 f(x_k)(x_k - \bar{x}) - (\nabla f(x_k) - \nabla f(\bar{x}))\| \\ &\quad = \left\| \int_0^1 [\nabla^2 f(x_k) - \nabla^2 f(x_k + \mu(\bar{x} - x_k))](x_k - \bar{x})\, d\mu \right\| \\ &\quad \le \int_0^1 \left\| \nabla^2 f(x_k) - \nabla^2 f(x_k + \mu(\bar{x} - x_k)) \right\| \|x_k - \bar{b}\|\, d\mu \\ &\quad \le \|x_k - \bar{x}\|^2 \int_0^1 C\mu\, d\mu = \frac{1}{2} L \|x_k - \bar{x}\|^2, \end{aligned} \tag{2.40}$$

όπου L είναι η σταθερά Lipschitz για τον $\nabla^2 f(x)$ για x κοντά στο $\bar{x}$. Εφόσον ο $\nabla^2 f(x)$ είναι αντιστρέψιμος, υπάρχει ακτίνα $r > 0$ τέτοια ώστε $\|\nabla^2 f(x_k)^{-1}\| \le 2\|\nabla^2 f(\bar{x})^{-1}\|$ για κάθε x_k με $\|x_k - \bar{x}\| \le r$. Αντικαθιστώντας στις (2.38) και (2.39), παίρνουμε

$$\|x_k + h_k - \bar{x}\| \le L \|\nabla^2 f(\bar{x})^{-1}\| \|x_k - \bar{x}\|^2 = \tilde{L} \|x_k - \bar{x}\|^2 \tag{2.41}$$

όπου $\tilde{L} = L\|\nabla^2 f(\bar{x})^{-1}\|$. Επιλέγοντας αρχικό σημείο x_0 τέτοιο ώστε $\|x_0 - \bar{x}\| \le \min(r, 1/(2\tilde{L}))$, μπορούμε να χρησιμοποιήσουμε αυτή την ανισότητα ώστε επαγωγικά να καταλήξουμε στο ότι η ακολουθία συγκλίνει στο $\bar{x}$, και ότι η ταχύτητα σύγκλισής της είναι τετραγωνική.

Από τους τύπους της μεθόδου Newton $x_{k+1} = x_k - h_k$ και $y_k = -h_k \Leftrightarrow y_k + h_k = 0 \Leftrightarrow \nabla f(x_k) + \nabla^2 f(x_k) h_k = 0$, έπεται ότι

$$\begin{aligned} \|\nabla f(x_{k+1})\| &= \|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k) h_k\| \\ &= \left\| \int_0^1 \nabla^2 f(x_k + \mu h_k)(x_{k+1} - x_k)\, d\mu - \nabla^2 f(x_k) h_k \right\| \\ &\le \int_0^1 \left\| \nabla^2 f(x_k + \mu h_k) - \nabla^2 f(x_k) \right\| \|h_k\|\, d\mu \\ &\le \frac{1}{2} C \|h_k\|^2 \\ &\le \frac{1}{2} C \|\nabla^2 f(x_k)^{-1}\|^2 \|\nabla f(x_k)\|^2 \\ &\le 2C \|\nabla^2 f(x_k)^{-1}\|^2 \|\nabla f(x_k)\|^2, \end{aligned}$$

αποδεικνύοντας ότι οι νόρμες των κλίσεων συγκλίνουν τετραγωνικά στο μηδέν. □

Παρατηρούμε ότι η μέθοδος Newton ικανοποιεί το γενικό επαναληπτικό τύπο των μεθόδων γραμμικής αναζήτησης με βήμα $a_k = 1$, για το οποίο επιτυγχάνεται τοπικά τετραγωνική ταχύτητα σύγκλισης. Από τον τύπο της κατεύθυνσης της μεθόδου Newton και εφόσον ο Εσσιανός πίνακας $\nabla^2 f(x_k)$ δεν είναι πάντα θετικά ορισμένος, η h_k δεν είναι απαραίτητα κλίση καθόδου και η μέθοδος συγκλίνει σε στάσιμα σημεία, τα οποία μπορεί να είναι είτε σημεία τοπικού μεγίστου είτε σημεία τοπικού ελαχίστου.

2.3.3 Μέθοδοι Quasi-Newton

Έστω ότι s_k είναι το βήμα από το σημείο x_k και έστω ότι η σχέση

$$\nabla f(x_{k+1}) = \nabla f(x_k) + \nabla^2 f(x_k) s_k + \ \ldots$$

είναι το ανάπτυγμα Taylor της κλίσης της συνάρτησης f γύρω από το x_k κατά τη διεύθυνση του s_k (Gill, Murray & M. Wright, 1997). Τότε, η *καμπυλότητα* της συνάρτησης f κατά το s_k, $\nabla^2 f(x_k) s_k$ προσεγγίζεται σύμφωνα με τη σχέση

$$\nabla^2 f(x_k) s_k \approx \nabla f(x_k + s_k) - \nabla f(x_k) \tag{2.42}$$

η οποία χρησιμοποιεί μόνο πληροφορίες πρώτης τάξης (Nocedal & S. Wright, 2006, σ. 24).

Εάν B_k είναι μια προσέγγιση του Εσσιανού πίνακα μιας τετραγωνικής αντικειμενικής συνάρτησης f, τότε ο πίνακας παρέχει στην αρχή της k-οστής επανάληψης τις απαραίτητες πληροφορίες για την καμπυλότητα που αποκτήθηκαν από τις προηγούμενες επαναλήψεις. Μετά από τον υπολογισμό του σημείου x_{k+1}, η νέα προσέγγιση του Εσσιανού πίνακα B_{k+1} ενημερώνεται με βάση την κεντρική ιδέα των μεθόδων quasi-Newton, σύμφωνα με τον τύπο

$$B_{k+1} = B_k + U_k,$$

όπου, U_k είναι ο πίνακας ενημέρωσης. Θέτοντας $s_k = x_{k+1} - x_k = a_k h_k$ και $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$, η κύρια συνθήκη που θα πρέπει να ικανοποιεί ο ενημερωμένος προσεγγιστικός πίνακας B_{k+1} είναι η λεγόμενη *συνθήκη quasi-Newton* (Gill, Murray & M. Wright, 1997, σ. 117-118), η οποία δίνει την προσέγγιση της καμπυλότητας της f κατά το s_k. Επομένως, σύμφωνα με τη σχέση (2.42), θα πρέπει

$$B_{k+1} s_k = y_k. \tag{2.43}$$

Οι *μέθοδοι quasi-Newton* ανήκουν στην κατηγορία των μεθόδων κλίσης, με διάνυσμα κατεύθυνσης (Bertsekas, 1999)

$$h_k = -D_k \nabla f(x_k), \tag{2.44}$$

και επαναληπτικό τύπο

$$x_{k+1} = x_k + a_k h_k, \tag{2.45}$$

όπου ο D_k είναι ένα θετικά ορισμένος πίνακας. Η βασική ιδέα πίσω από τις μεθόδους quasi-Newton είναι ότι ο επαναληπτικός πίνακας D_k ενημερώνεται σε κάθε βήμα μιας μεθόδου κλίσης, λαμβάνοντας υπόψιν τις επιπλέον πληροφορίες για την

καμπυλότητα που προκύπτουν μετά από κάθε επανάληψη. Πολλές από τις μεθόδους quasi-Newton υπολογίζουν τον πίνακα D_k έτσι ώστε βαθμιαία να προσεγγίζει τον αντίστροφο Εσσιανό πίνακα, δηλαδή $D_k = B_k^{-1}$. Άρα η (2.44) γίνεται

$$h_k = -B_k^{-1}\nabla f(x_k). \tag{2.46}$$

Βλέπουμε ότι στις μεθόδους quasi-Newton πρέπει να υπολογίσουμε το νέο επαναληπτικό πίνακα B_{k+1}, αρχικοποιώντας τον B_k είτε με $B_0 = I$ είτε $B_0 = \theta I$, $\theta \in \mathbb{R}/0$.

Στις δύο δημοφιλέστερες κατηγορίες μεθόδων quasi-Newton, ο πίνακας B_{k+1} προκύπτει από τον B_k και τα διανύσματα s_k και y_k μέσω της εξίσωσης που ονομάζεται *συμμετρική ενημέρωση πρώτης τάξης* (symmetric rank-one update) (Gill, Murray & M. Wright, 1997, σ. 118)

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k^T s_k)} \tag{2.47}$$

και του τύπου *BFGS*, από τους εμπνευστές του, Broyden, Fletcher, Goldfarb και Shanno (Nocedal & S. Wright, 2006, σ. 24)

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}. \tag{2.48}$$

Άλλοι τύποι ενημέρωσης του πίνακα B_k, όταν αυτός είναι συμμετρικός (και κατά συνέπεια και ο B_{k+1} είναι συμμετρικός), όπως αναφέρονται στο βιβλίο των Gill, Murray και M. Wright (1997), περιλαμβάνουν τους τύπους *PSB* (Powell-Symmetric-Broyden) και *DFP* (Davidon-Fletcher-Powell - ανάλυση της μεθόδου γίνεται στο βιβλίο του R. Fletcher, 2000).

Στην προηγούμενη ενότητα, είδαμε ότι ένας αποτελεσματικός τρόπος υπολογισμού του βήματος a_k είναι με τον τύπο Armijo (2.24), για κάποια σταθερά $b \in (0, 1)$. Με άλλα λόγια, η μείωση της αντικειμενικής συνάρτησης f εξαρτάται από το βήμα a_k και το τετράγωνο του μέτρου της κλίσης $\nabla f(x_k)$. Μπορούμε να γενικεύσουμε τη συνθήκη αυτή για κατευθύνσεις διαφορετικές της $-\nabla f(x_k)$ σύμφωνα με τον παρακάτω τύπο:

$$\phi(a_k) \coloneqq f(x_k + a_k h_k) \leq f(x_k) + b a_k \nabla f(x_k)^T h_k. \tag{2.49}$$

Ο τύπος Armijo (2.49) είναι γνωστός και ως *επαρκής μείωση*, δηλαδή μας εξασφαλίζει ότι το βήμα που υπολογίζεται, μειώνει ικανοποιητικά τη συνάρτηση f. Όμως, συνθήκη της επαρκούς μείωσης από μόνη της δεν είναι αρκετή ώστε να αποφανθούμε ότι ο αλγόριθμος τρέχει ικανοποιητικά για μικρές τιμές του a. Η επιπλέον συνθήκη την οποία θα πρέπει να ικανοποιεί το βήμα a_k ώστε να αποφευχθούν υπερβολικά μικρά μεγέθη, ονομάζεται *συνθήκη κυρτότητας* (Nocedal & S. Wright, 2006, σ. 33) και δίνεται από τον τύπο

$$\phi'(a_k) \coloneqq \nabla f(x_k - a_k \nabla f(x_k))^T \nabla f(x_k) \leq -c\|\nabla f(x_k)\|_2^2 \tag{2.50}$$

με $c \in (b, 1)$. Κατά την κατεύθυνση h_k, οι ανισότητες (2.49) και (2.50) δίνουν τις *συνθήκες Wolfe* για το βήμα a_k,

$$f(x_k + a_k h_k) \leq f(x_k) + b a_k \nabla f(x_k)^T h_k, \tag{2.51a}$$

$$\nabla f(x_k + a_k h_k)^T h_k \geq c \nabla f(x_k)^T h_k, \tag{2.51b}$$

με $0 < b < c < 1$.

Το παρακάτω αποτέλεσμα δείχνει ότι εάν η κατεύθυνση αναζήτησης μιας μεθόδου quasi-Newton προσεγγίζει ικανοποιητικά την κατεύθυνση Newton, τότε το μοναδιαίο μήκος βήματος θα ικανοποιεί τις συνθήκες Wolfe όσο συκλίνουμε στη λύση. Υποδεικνύει επίσης μία συνθήκη την οποία θα πρέπει να ικανοποιεί η κατεύθυνση αναζήτησης έτσι ώστε επιτευχθεί ένας επαναληπτικός τύπος που συγκλίνει υπεργραμμικά (Nocedal & S. Wright, 2006, σ. 47).

Θεώρημα 2.10. *Έστω $f : \mathbb{R}^n \to \mathbb{R}$ δύο φορές συνεχώς παραγωγίσιμη συνάρτηση. Θεωρούμε τον επαναληπτικό τύπο $x_{k+1} = x_k + h_k$ (δηλαδή, το βήμα a_k είναι σταθερό με μέγεθος 1) και ότι h_k είναι η κατεύθυνση quasi-Newton όπως δίνεται από τη σχέση (2.46). Υποθέτουμε, επίσης, ότι η ακολουθία $\{x_k\}$ συγκλίνει σε ένα σημείο $\bar{x}$ τέτοιο ώστε $\nabla f(\bar{x}) = 0$ και ο Εσσιανός $\nabla^2 f(\bar{x})$ είναι θετικά ορισμένος. Τότε η $\{x_k\}$ συγκλίνει υπεργραμμικά αν και μόνο αν η κατεύθυνση h_k ικανοποιεί τη σχέση*

$$\lim_{k \to \infty} \frac{\|(B_k - \nabla^2 f(\bar{x}))h_k\|}{\|h_k\|} = 0. \tag{2.52}$$

Για την απόδειξη του θεωρήματος, εισάγουμε τους ορισμούς των *ασυμπτωτικών συμβολισμών* O και o. Αν οι f και g είναι πραγματικές συναρτήσεις του $x > 0$ με την g να είναι αυστηρά θετική για μεγάλα x, τότε συμβολίζουμε με $f(x) \in O(g(x))$ ή $f(x) = O(g(x))$ αν και μόνο υπάρχει $x_0 > 0$ και σταθερά $c > 0$ τέτοια ώστε για κάθε $x \geq x_0$ να είναι $f(x) \leq cg(x)$. Δηλαδή, το $O(g(x))$ είναι ένα σύνολο συναρτήσεων με τάξη μεγέθους που δεν υπερβαίνει την τάξη μεγέθους $g(x)$. Επίσης, γράφουμε $f(x) \in o(g(x))$ ή $f(x) = o(g(x))$ αν και μόνο αν για κάθε $c > 0$, υπάρχει $x_0 > 0$ ώστε για κάθε $x \geq x_0$, $f(x) < cg(x)$. Επομένως, το $o(g(x))$ είναι ένα σύνολο συναρτήσεων με τάξη μεγέθους αυστηρά μικρότερη της τάξης μεγέθους $g(x)$.

Απόδειξη. Θα δείξουμε πρώτα ότι η (2.52) είναι ισοδύναμη με τη

$$h_k - h_k^N = o(\|h_k\|), \tag{2.53}$$

όπου $h_k^N = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$ είναι η κατεύθυνση αναζήτησης Newton. Θεωρώντας ότι η (2.52) ισχύει, έχουμε ότι

$$\begin{aligned} h_k - h_k^N &= \nabla^2 f(x_k)^{-1}(\nabla^2 f(x_k) h_k + \nabla f(x_k)) \\ &= \nabla^2 f(x_k)^{-1}(\nabla^2 f(x_k) - B_k) h_k \\ &= O(\|(\nabla^2 f(x_k) - B_k) h_k\|) \\ &= o(\|h_k\|), \end{aligned}$$

όπου χρησιμοποιήσαμε την ιδιότητα ότι η νόρμα $\|\nabla^2 f(x_k)^{-1}\|$ είναι άνω φραγμένη όταν το x_k είναι αρκετά κοντά στο $\bar{x}$, αφού ο οριακός Εσσιανός $\nabla^2 f(\bar{x})$ είναι θετικά ορισμένος. Το αντίστροφο αποδεικνύεται πολλαπλασιάζοντας τα δύο μέρη της (2.53) με $\nabla^2 f(x_k)$ και χρησιμοποιώντας την (2.46).

Συνδυάζοντας τις (2.41) και (2.53), παίρνουμε

$$\|x_k + h_k - \bar{x}\| \leq o(\|x_k - \bar{x}\|),$$

απ' όπου προκύπτει η υπεργραμμική σύγκλιση. □

Στο παραπάνω θεώρημα, η σχέση (2.52) δείχνει ότι η διαφορά μεταξύ της προσέγγισης του Εσσιανού στην τρέχουσα επανάληψη και του πραγματικού Εσσιανού στο σημείο ελαχίστου $\bar{x}$ κατά την κατεύθυνση h_k, τείνει στο μηδέν. Ορισμένες από τις μεθόδους quasi-Newton επιτυγχάνουν μεγαλύτερη ταχύτητα σύγκλισης, λόγω του ότι αποφεύγουν τον υπολογισμό των παραγώγων με την δευτέρας τάξης της f.

3 Βελτιστοποίηση στη Μηχανική Μάθηση

Στο κεφάλαιο αυτό, γίνεται μια συνοπτική παρουσίαση των τεχνικών που χρησιμοποιούνται για την επίλυση προβλημάτων βελτιστοποίησης στη μηχανική μάθηση. Ιδιαίτερα, εξετάζουμε την ταξινόμηση κειμένων, ένα πρόβλημα με μεγάλο όγκο δεδομένων που επιλύεται με τη χρήση σημαντικών εργαλείων και τεχνικών της μηχανικής μάθησης.

Οι εργασίες μηχανικής μάθησης με τις οποίες θα ασχοληθούμε ανήκουν στην κατηγορία της επιβλεπόμενης μάθησης. Η *προγνωστική* (predictive) ή *επιβλεπόμενη* (supervised) μάθηση είναι μια διαδικασία η οποία κατασκευάζει μία συνάρτηση που απεικονίζει εισόδους σε εξόδους από ένα δεδομένο σύνολο $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, το οποίο ονομάζεται *σύνολο εκπαίδευσης* (training set) με μέγεθος n, και στόχος της είναι η γενίκευση της συνάρτησης αυτής για εισόδους με άγνωστη έξοδο. Η επιβλεπόμενη μάθηση χρησιμοποιείται συχνά σε προβλήματα *ταξινόμησης* (classification) για τα οποία θα γίνει περαιτέρω μελέτη στη συνέχεια. Προς το παρόν, αναφέρουμε ότι η διαδικασία της ταξινόμησης απεικονίζει εισόδους, οι οποίες μπορεί να έχουν είτε συνεχείς είτε διακριτές τιμές σε διακριτές εξόδους, δηλαδή σε κατηγορίες. Ωστόσο, μπορούμε μέσω της ταξινόμησης να απεικονίσουμε εισόδους σε συνεχείς εξόδους, οι οποίες εκφράζουν την πιθανότητα ενός δοθέντος παραδείγματος να ανήκει σε κάθε κατηγορία. Στην επόμενη παράγραφο δίνονται οι ορισμοί των διακριτών και των συνεχών τιμών.

3.1 Έννοιες Πιθανοτήτων και Στατιστικής

Όπως αναφέραμε στην εισαγωγή, η μηχανική μάθηση συνδέεται στενά με τη στατιστική και τη θεωρία πιθανοτήτων. Για το λόγο αυτό, δίνεται παρακάτω το κατάλληλο υπόβαθρο αυτών των πεδίων, για την κατανόηση και την ανάπτυξη βασικών εννοιών και τεχνικών που εμφανίζονται στη συνέχεια της εργασίας.

Το σύνολο όλων των δυνατών αποτελεσμάτων από την παρατήρηση ενός φαινομένου με πολλά, τυχαία αποτελέσματα, ονομάζεται *δειγματικός χώρος* και συμβολίζεται με Δ. Καλούμε τα δυνατά αποτελέσματα του φαινομένου, δηλαδή τα στοιχεία του δειγματικού χώρου ως *απλά ενδεχόμενα* ή *απλά γεγόνότα*. Αντίστοιχα, ονομάζουμε *ενδεχόμενο* ή *γεγονός* ένα υποσύνολο του δειγματικού χώρου $A \subset \Delta$ που αποτελείται από απλά γεγονότα με κάποιο κοινό χαρακτηριστικό. Για ένα γεγονός A, «η *πιθανότητα* $\mathbb{P}(A)$, $0 \leq \mathbb{P}(A) \leq 1$, είναι η αριθμητική τιμή που αντιπροσωπεύει το όριο του ποσοστού των φορών που το γεγονός A συμβαίνει όταν μια διαδικασία με αβέβαιο αποτέλεσμα εκτελείται απεριόριστο αριθμό φορών και κάτω από τις ίδιες συνθήκες» (Βόντα, Καραγρηγορίου, 2012, σ. 29). Η πιθανότητα είναι το κριτήριο εκείνο που μέσα σε συνθήκες αβεβαιότητας οδηγεί στη εξαγωγή συμπερασμάτων και στη λήψη αποφάσεων.

Η απεικόνιση $X : \Delta \to \mathbb{R}$, όπου Δ είναι ο δειγματικός χώρος, λέγεται *τυχαία μεταβλητή*. Δηλαδή, μια τυχαία μεταβλητή είναι μια συνάρτηση η οποία αντιστοιχίζει κάθε απλό γεγονός του δειγματικού χώρου σε έναν πραγματικό αριθμό. Θα συμβολίζουμε με κεφαλαία γράμματα, όπως X και Y, τις τυχαίες μεταβλητές, ενώ με μικρά γράμματα (π.χ. $x, y \in \mathbb{R}$) τις πιθανές πραγματικές τιμές που μπορούν να

πάρουν οι τυχαίες μεταβλητές. Αν $\mathbb{P}[X \leq x]$ είναι η πιθανότητα η τυχαία μεταβλητή X να έχει τιμές μικρότερες από την $x \in \mathbb{R}$, τότε η σχέση $\Pi(x) = \mathbb{P}[X \leq x]$ ορίζει την *συνάρτηση κατανομής πιθανότητας*.

Όταν όλες οι πιθανές τιμές μιας τυχαίας μεταβλητής X ανήκουν σε ένα αριθμήσιμο σύνολο $\{x_1, x_2, \ldots, x_n\}$, τότε η X καλείται *διακριτή*. Η ακολουθία $\{p_k\}_{n\in\mathbb{N}}$ με $p_k = \mathbb{P}[X = x_k]$ λέγεται τότε *συνάρτηση μάζας πιθανότητας* της τυχαίας μεταβλητής X και συνδέεται με την συνάρτηση κατανομής πιθανότητας Π με τη σχέση:

$$\Pi(x) = \sum_k p_k, \quad \text{όταν } x_k \leq x.$$

Εάν η X είναι μια διακριτή τυχαία μεταβλητή με συνάρτηση μάζας πιθανότητας $p(x)$, τότε η τιμή $\mathbb{E}[X] = \sum_k x_k p_k$ ονομάζεται *αναμενόμενη ή μέση* τιμή της τυχαίας μεταβλητής X.

Επίσης, η τυχαία μεταβλητή X λέγεται *συνεχής* όταν υπάρχει πραγματική συνάρτηση P με πεδίο ορισμού το $\mathbb{R}$ τέτοια ώστε

$$P(x) \geq 0 \text{ και } \Pi(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} P(y)\, dy,$$

για κάθε $x \in \mathbb{R}$. Η παραπάνω μορφή της συνάρτησης κατανομής πιθανότητας μιας συνεχούς κατανομής δείχνει ότι η P είναι μια συνεχής συνάρτηση στο $\mathbb{R}$. Ακόμα, από τον παραπάνω ορισμό συνεπάγεται ότι η παράγωγος μιας συνεχούς συνάρτησης κατανομής πιθανότητας P, στα σημεία που ορίζεται, είναι ίση με τη *συνάρτηση πυκνότητας πιθανότητας* P, δηλαδή

$$P(x) = \frac{d\Pi(x)}{dx}.$$

Η αναμενόμενη τιμή στην περίπτωση της συνεχούς τυχαίας μεταβλητής X με συνάρτηση πυκνότητας πιθανότητας $P(x)$ είναι

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xP(x)\, dx,$$

και υπάρχει όταν το ολοκλήρωμα αυτό συγκλίνει.

Θα δώσουμε τώρα δύο σημαντικές ιδιότητες της αναμενόμενης τιμής όπως αυτές δίνονται από τα Θεωρήματα 1.1 και 2.1 στο βιβλίο των Κοκολάκη και Σπηλιώτη (2002, σ. 134). Έστω $a, b \in \mathbb{R}$ σταθερές και έστω ότι οι τυχαίες μεταβλητές X και Y έχουν πεπερασμένες αναμενόμενες τιμές. Τότε, η τυχαία μεταβλητή $aX + bY$ έχει μέση τιμή για την οποία ισχύει ότι

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]. \tag{3.1}$$

Επιπλέον, αν $z : \mathbb{R} \to \mathbb{R}$ είναι πραγματική συνάρτηση, τότε

$$\mathbb{E}[z(X)] = \begin{cases} \displaystyle\sum_{i=1}^{\infty} z(x_i)p_i, & \text{αν η } X \text{ είναι διακριτή,} \\ \displaystyle\int_{-\infty}^{\infty} z(x)P(x)\, dx, & \text{αν η } X \text{ είναι συνεχής.} \end{cases}$$

Η *ροπή k-τάξης περί την αρχή*, με $k \in \mathbb{N}$ της τυχαίας μεταβλητής X συμβολίζεται με μ'_k και ορίζεται ως $\mu'_k = \mathbb{E}[X^k]$ και υπάρχει μόνον όταν $\mathbb{E}[|X|^k] < \infty$. Συνεπώς, η ροπή k-τάξης περί την αρχή για διακριτή τυχαία μεταβλητή είναι $\mu'_k = \sum_{i=1}^{\infty} x_i^k p_i$ και για μία συνεχή τυχαία μεταβλητή είναι $\mu'_k = \int_{-\infty}^{\infty} x^k P(x)\, dx$. Από τον ορισμό αυτό, είναι προφανές ότι η ροπή πρώτης τάξης περί την αρχή πρόκειται για την αναμενόμενη τιμή της τυχαίας μεταβλητής X.

Έστω η $n-$διάστατη τυχαία μεταβλητή $X = (X_1, \ldots, X_n)^T$, $X \in \mathbb{R}^n$, όπου κάθε στοιχείο $\{X_1, \ldots, X_n\} \in \mathbb{R}$ είναι επίσης μία τυχαία μεταβλητή. Εάν η X παίρνει τιμές $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, με $\{x_1, \ldots, x_n\} \in \mathbb{R}$, τότε η συνάρτηση

$$\Pi(x) = \mathbb{P}[X_1 \leq x_1, \ldots, X_n \leq x_n]$$

καλείται *από κοινού συνάρτηση κατανομής πιθανότητας* των τυχαίων μεταβλητών $X_1, \ldots, X_n$. Συγκεκριμένα, η από κοινού συνάρτηση κατανομής πιθανότητας $\Pi(x)$ εκφράζει την πιθανότητα να ισχύουν ταυτόχρονα οι σχέσεις $[X_1 \leq x_1], [X_2 \leq x_2], \ldots, [X_n \leq x_n]$. Με βάση τους ορισμούς που δόθηκαν και παραπάνω, ορίζονται αντίστοιχα και οι *από κοινού συνάρτηση μάζας και πυκνότητας πιθανότητας* για την n-διάστατη τυχαία μεταβλητή X με τιμές $x = \{x_1, \ldots, x_n\}$, $x \in \mathbb{R}^n$:

$$p_{1,\ldots,n} = \mathbb{P}[X = x_1, \ldots, X_n = x_n] \geq 0 \quad \text{και}$$

$$P(x_1, \ldots, x_n) \geq 0 \ \text{ και } \ \Pi(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} P(y_1, \ldots, y_n)\, dy_1 \cdots dy_n,$$

δηλαδή,

$$P(x_1, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} \Pi(x_1, \ldots, x_n).$$

Σημειώνουμε ότι, όπως και στον ορισμό της από κοινού συνάρτησης κατανομής πιθανότητας, η από κοινού συνάρτηση μάζας πιθανότητας δίνει την πιθανότητα να ισχύουν ταυτόχρονα οι σχέσεις $[X_1 = x_1], [X_2 = x_2], \ldots, [X_n = x_n]$ για τη διακριτή περίπτωση.

Δύο κατανομές των τυχαίων μεταβλητών X και Y αντίστοιχα, ονομάζονται *δεσμευμένες* όταν υπάρχει αλληλεξάρτηση μεταξύ τους. Συγκεκριμένα, αν οι X, Y είναι συνεχείς και η απο κοινού συνάρτηση πυκνότητας πιθανότητάς τους είναι $P(x, y)$, τότε η δεσμευμένη συνάρτηση πυκνότητας πιθανότητας της X όταν γνωρίζουμε την τιμή της Y είναι $P_{X|Y}(x|y) = \frac{P(x,y)}{P_Y(y)}$. Αντιθέτως, δύο τυχαίες μεταβλητές λέγονται *ανεξάρτητες* όταν οι τιμές που παίρνει η μία δεν επηρεάζουν και δεν παρέχουν καμία πληροφορία για τις τιμές της άλλης. Τότε, από το Θεώρημα 7.1(iiiβ) στο βιβλίο των Κοκολάκη και Σπηλιώτη (2002, σ. 83), η από κοινού συνάρτηση πυκνότητάς πιθανότητας για τις συνεχείς και ανεξάρτητες X και Y είναι $P(x, y) = P_X(x)P_Y(y)$ για κάθε $x, y \in \mathbb{R}$.

Με βάση τον ορισμό των ανεξάρτητων τυχαίων μεταβλητών, μπορούμε να ορίσουμε την *ολική* αναμενόμενη τιμή δύο τυχαίων μεταβλητών. Ολική λέγεται η αναμενόμενη τιμή μιας συνάρτησης ως προς όλες τις τυχαίες μεταβλητές από τις οποίες εξαρτάται. Στην περίπτωση που δύο πραγματικές, συνεχείς τυχαίες μεταβλητές X και Y είναι ανεξάρτητες, είδαμε ότι ισχύει: $P(x, y) = P_X(x)P_Y(y)$. Αν $z : \mathbb{R}^2 \to \mathbb{R}$

είναι μια συνάρτηση των X και Y, τότε για την ολική αναμενόμενη τιμή $\mathbb{E}[z(X,Y)]$ ισχύει ότι

$$\mathbb{E}[z(X,Y)] = \mathbb{E}_Y[\mathbb{E}_X[z(X,Y)]]. \tag{3.2}$$

Πράγματι,

$$\begin{aligned}\mathbb{E}[z(X,Y)] &= \int_Y \int_X z(x,y)P(x,y)\,dx\,dy \\ &= \int_Y \int_X z(x,y)P(x)\,dx\,P(y)\,dy \\ &= \int_Y \mathbb{E}_X[z(x,y)]P(y)\,dy = \mathbb{E}_Y[\mathbb{E}_X[z(X,Y)]].\end{aligned}$$

Όταν θέλουμε να εξετάσουμε τα χαρακτηριστικά ενός συνόλου με πολλά στοιχεία, θα πρέπει να συλλέξουμε πληροφορίες από αυτό. Αντί, όμως, για την έρευνα ολόκληρου του συνόλου, μπορούμε να εξάγουμε ικανοποιητικά συμπεράσματα από την εξέταση ενός μικρότερου υποσυνόλου το οποίο καλούμε *τυχαίο δείγμα*. Το τυχαίο δείγμα με μέγεθος n εκφράζεται από τη συλλογή των τυχαίων μεταβλητών $X_1, X_2, \ldots, X_n$, οι οποίες ακολουθούν μια κοινή κατανομή και η επιλογή της μίας δεν επηρεάζει τη διαδικασία επιλογής των υπολοίπων. Έστω, τώρα, ότι η $n-$διάστατη τυχαία μεταβλητή T ορίζεται ως συνάρτηση της $X \in \mathbb{R}^n$, δηλαδή $T = t(X) = t(X_1, \ldots, X_n)$ και παίρνει τιμές $t(x) = t(x_1, \ldots, x_n)^T \in \mathbb{R}^n$, όπου $x = (x_1, \ldots, x_n)^T$ είναι οι τιμές της X. Θα λέμε στο εξής ότι η τυχαία μεταβλητή - συνάρτηση T και «κάθε συνάρτηση μιας ή περισσότερων τυχαίων μεταβλητών που δεν εξαρτάται από άγνωστες παραμέτρους ονομάζεται *στατιστική συνάρτηση*» (Βόντα και Καραγρηγορίου, 2012, σ. 102) ή *δειγματοσυνάρτηση*.

Η άγνωστη παράμετρος ενός τυχαίου δείγματος αποτελεί μια χαρακτηριστική ποσότητα ολόκληρου του συνόλου την οποία μας ενδιαφέρει να εξετάσουμε. Αν στην κατανομή του τυχαίου δείγματος $X = (X_1, \ldots, X_n)^T$ οι άγνωστες παράμετροι εκφράζονται από το διάνυσμα $\theta \in \Theta \subset \mathbb{R}^n$, με $\theta = (\theta_1, \ldots, \theta_n)^T$, κάθε δειγματοσυνάρτηση $T = t(X) : \mathbb{R}^n \to \Theta$ που χρησιμοποιείται για την προσέγγιση της παραμέτρου θ ονομάζεται *εκτιμήτρια*. Οι τιμές $t(x) = t(x_1, \ldots, x_n)^T$ της T αποτελούν μια *εκτίμηση* της παραμέτρου θ. Στην περίπτωση που για την εκτιμήτρια $T = t(X)$ της θ ισχύει ότι

$$\mathbb{E}[T;\theta] = \mathbb{E}[t(X);\theta] = \theta \ \text{ για κάθε } \theta \in \Theta,$$

ονομάζουμε την T *αμερόληπτη εκτιμήτρια* της θ. Αντίστοιχα, εάν $a(\theta)$ είναι μια οποιαδήποτε συνάρτηση της παραμέτρου $\theta \in \Theta$, τότε η εκτιμήτρια T είναι αμερόληπτη εάν

$$\mathbb{E}[T;\theta] = a(\theta) \ \text{ για κάθε } \theta \in \Theta. \tag{3.3}$$

3.2 Ταξινόμηση Κειμένου

Αφού εισάγαγαμε τις απαραίτητες έννοιες τις στατιστικής και των πιθανοτήτων που θα μας χρειαστούν στη συνέχεια της μελέτης μας, παρουσιάζομυε ένα από τα γνωρστότερα προβλήματα της μηχανικής μάθησης μεγάλης κλίμακας, την ταξινόμηση κειμένου. Ως πρόβλημα της επιβλεπόμενης μάθησης, η ταξινόμηση κειμένου δίνει μία

στατιστική προσέγγιση της λύσης, δηλαδή της κατάταξης ενός κειμένου σε μία προκαθορισμένη κατηγορία. Η λύση δίνεται μέσω αλγορίθμων που χρησιμοποιούν ζεύγη εισόδων - εξόδων (x_i, y_i) τα οποία ανήκουν στο σύνολο εκπαίδευσης $\mathcal{D}$. Για $i \in 1, \ldots, n$, το διάνυσμα $x_i \in \mathcal{X} \subset \mathbb{R}^n$ αναπαριστά τα *χαρακτηριστικά* του κειμένου, όπως είναι οι διάφορες λέξεις του, ενώ το διακριτό μέγεθος $y_i, i \in \{1, \ldots, n\}$ είναι η *ετικέτα* που δηλώνει εάν το κείμενο ανήκει $(y_i = 1)$ ή δεν ανήκει $(y_i = -1)$ σε μία συγκεκριμένη κατηγορία. Έτσι, μπορεί να κατασκευαστεί ένα πρόγραμμα ταξινόμησης το οποίο χρησιμοποιεί τη γνώση που αποκτά από το σύνολο εκπαίδευσης ώστε να ταξινομήσει σωστά νέες εισόδους με άγνωστη έξοδο. Η διαδικασία αυτή ορίζεται από την *συνάρτηση πρόβλεψης* (prediction function) h, η οποία δίνει τη διακριτή πρόβλεψη $h(x_i)$ για δοθείσα είσοδο x_i. Η συνάρτηση h επιλέγεται από το σύνολο των συναρτήσεων πρόβλεψης $\mathcal{H}$.

Για να μπορέσουμε να εκτιμήσουμε την επίδοση ενός αλγορίθμου της μηχανικής μάθησης, υπάρχουν διάφορα μέτρα των οποίων οι ποσότητες μας δίνουν μια εικόνα για την αποτελεσματικότητά του. Στην επιβλεπόμενη μάθηση, και συγκεκριμένα σε εργασίες όπως η ταξινόμηση, η *ακρίβεια* του μοντέλου παρέχει χρήσιμες πληροφορίες σχετικά με το ποσοστό των παραδειγμάτων, δηλαδή των στοιχείων του συνόλου εκπαίδευσης, για τα οποία ο αλγόριθμος δίνει τη σωστή έξοδο. Αντίστοιχα, με τον *ρυθμό σφάλματος* παίρνουμε το ποσοστό των δεδομένων εισόδου για τα οποία ο αλγόριθμος δίνει λανθασμένη έξοδο. Ο συνηθέστερος τρόπος με τον οποίο προσδιορίζουμε τα μέτρα επίδοσης ενός αλγορίθμου, είναι χρησιμοποιώντας ένα ξεχωριστό σύνολο, το *σύνολο ελέγχου*, αφού έτσι είναι δυνατό να δούμε πώς συμπεριφέρεται ο αλγόριθμος με άγνωστα δεδομένα.

Η επιλογή των μέτρων επίδοσης μπορεί να φαίνεται ως μια απλή διαδικασία, στην πραγματικότητα όμως είναι αρκετά δύσκολο το να βρούμε την ποσότητα εκείνη που απεικονίζει καλύτερα την επιθυμητή συμπεριφορά του αλγορίθμου. Σε ορισμένους αλγόριθμους, η δυσκολία έγκειται στην επιλογή του κατάλληλου μέτρου, οπότε σε αυτή την περίπτωση πρέπει να κοιτάξουμε τις εφαρμογές του αλγορίθμου σε πραγματικά προβλήματα, ενώ σε άλλους η επιλογή της ποσότητας είναι απλή, ωστόσο είναι δύσκολη η μέτρηση των τιμών της. Τότε, θα πρέπει να σχεδιάσουμε ένα εναλλακτικό κριτήριο το οποίο αντιστοιχεί στην συμπεριφορά που θέλουμε ή να βρούμε μία καλή προσέγγιση του μέτρου επίδοσης που μας ενδιαφέρει (Goodfellow, Bengio & Courville, 2016).

Στην ταξινόμηση κειμένου, η επίδοση του προγράμματος εξαρτάται από το πόσο συχνά η πρόβλεψη $h(x_i)$ διαφέρει από την πραγματική ταξινόμηση y_i. Κατ΄ αυτό τον τρόπο, αναζητούμε μια συνάρτηση πρόβλεψης που ελαχιστοποιεί τη συχνότητα εμφάνισης εσφαλμένων ταξινομήσεων, γνωστή ως *εμπειρικό ρίσκο* λανθασμένης ταξινόμησης R_n. Η μαθηματική του διατύπωση, καθώς και καλύτερη επεξήγηση δίνεται στη συνέχεια της εργασίας.

Με βάση, λοιπόν, τα δεδομένα του συνόλου εκπαίδευσης, θα πρέπει να εξασφαλίσουμε καλή γενικευμένη επίδοση του αλγορίθμου. Ένας αποτελεσματικός τρόπος για την επίτευξη καλής επίδοσης είναι μέσω της μεθόδου της *διασταυρωμένης επικύρωσης* (cross-validation), κατά την οποία το σύνολο παραδειγμάτων χωρίζεται σε τρία μέρη: το πρώτο χρησιμοποιείται ως σύνολο εκπαίδευσης, το δεύτερο ως σύνολο ελέγχου και το τρίτο ως *σύνολο επικύρωσης*. Η διαδικασία εύρεσης της συνάρτη-

σης h που ελαχιστοποιεί το εμπειρικό ρίσκο R_n γίνεται από το σύνολο εκπαίδευσης στο σύνολο των υποψήφιων συναρτήσεων πρόβλεψης και σκοπός της είναι η εύρεση του υποσυνόλου με τις επικρατέστερες συναρτήσεις. Έπειτα, για κάθε μία από τις επικρατέστερες συναρτήσεις πρόβλεψης, γίνεται εκτίμηση της επίδοσής της με βάση το σύνολο επικύρωσης και η συναρτήση με την καλύτερη επίδοση είναι η ζητούμενη. Τέλος, με το σύνολο ελέγχου μπορούμε να εκτιμήσουμε την γενικευμένη επίδοση της επιλεγμένης συνάρτησης.

3.3 Προβλήματα Βελτιστοποίησης στη Μηχανική Μάθηση

Θα στρέψουμε τώρα την προσοχή μας στις αριθμητικές μεθόδους οι οποίες επιλύουν προβλήματα βελτιστοποίησης που προκύπτουν από τη μηχανική μάθηση μεγάλης κλίμακας. Για το σκοπό της εργασίας, η διαδικασία βελτιστοποίησης παρουσιάζεται ως μια γενικότερη μορφή του προβλήματος ελαχιστοποίησης σε μία εργασία επιβλεπόμενης ταξινόμησης, όπως είδαμε στην προηγούμενη ενότητα.

3.3.1 Βασική Περιγραφή

Ο κύριος στόχος ενός προβλήματος μηχανικής μάθησης είναι η σωστή μοντελοποίηση του συνόλου των παραδειγμάτων. Η μοντελοποίηση στην επιβλεπόμενη μάθηση είναι η εύρεση μιας συνάρτησης πρόβλεψης $h : \mathcal{X} \to \mathcal{Y}$ η οποία απεικονίζει δεδομένα από το χώρο εισόδων $\mathcal{X}$ στο χώρο εξόδων $\mathcal{Y}$. Η συνάρτηση πρόβλεψης δημιουργείται με βάση τα δεδομένα του συνόλου εκπαίδευσης $\mathcal{D}$ και θα πρέπει, για ένα καινούργιο διάνυσμα εισόδου $x \in \mathcal{X}$ που δεν ανήκει στο σύνολο εκπαίδευσης, να δίνει μια σωστή πρόβλεψη $h(x)$ για την πραγματική τιμή της εξόδου $y \in \mathcal{Y}$.

Η συνάρτηση που θα επιλέξουμε, θα πρέπει να γενικεύει σωστά τις γνώσεις τις οποίες αποκτά κατά τη μάθηση, χωρίς να κάνει στιγνή αποστήθιση των δεδομένων του συνόλου εκπαίδευσης καθώς με αυτό τον τρόπο μπορεί να οδηγηθούμε σε *υπερπροσαρμογή* (overfitting) των δεδομένων: επειδή η συνάρτηση πρόβλεψης προσπαθεί να περιγράψει με τον καλύτερο δυνατό τρόπο τα παραδείγματα, ακόμα και οι μικρότερες αποκλίσεις μπορούν να ληφθούν υπ΄ όψιν για την κατασκευή του μοντέλου. Ως αποτέλεσμα, συμπεριλαμβάνεται υπερβολικά μεγάλος όγκος δεδομένων και καταλήγουμε σε μία συνάρτηση που δεν μας εξασφαλίζει σωστή γενίκευση για τα νέα δεδομένα εισόδου που θα πρέπει να επεξεργαστούμε. Ο λόγος που συμβαίνει αυτό είναι επειδή η συνάρτηση απεικονίζει με μεγάλη ακρίβεια τις πληροφορίες του συνόλου εκπαίδευσης αποκλειστικά.

Από την διαδικασία, λοιπόν, της μάθησης, θα πρέπει να προκύπτει μια συνάρτηση πρόβλεψης h επιλεγμένη από μία κατάλληλη οικογένεια συναρτήσεων $\mathcal{H}$, η οποία δίνει προσεγγιστικές τιμές για τις εξόδους. Ωστόσο, οι τιμές αυτές μπορεί να διαφέρουν από τις πραγματικές εξόδους. Επομένως, μπορούμε να βρούμε μία συνάρτηση, την οποία αποκαλούμε *συνάρτηση απώλειας* (loss function) ή *συνάρτηση σφάλματος* (error function) $\ell(h, y)$ και ορίζουμε σε επόμενη παράγραφο, που θα περιγράφει το πόσο οι τιμές της συνάρτησης πρόβλεψης $h(x)$ παρεκκλίνουν από τις πραγματικές εξόδους y. Δημιουργείται, έτσι, ένα μέτρο το οποίο μπορεί να καθορίσει την αποδοτικότητα του αλγορίθμου μηχανικής μάθησης μέσα από την βελτιστοποίησή του. Συνεπώς,

η ελαχιστοποίηση της συνάρτησης απώλειας μας δίνει τη συνάρτηση πρόβλεψης με τη μικρότερη δυνατή απόκλιση από τις πραγματικές τιμές, δηλαδή τη συνάρτηση που μοντελοποιεί καλύτερα τα δεδομένα του συνόλου εκπαίδευσης.

Ένα μειονέκτημα, όμως, της συνάρτησης απώλειας είναι ότι δίνει ένα απόλυτο μέτρο για τη διαφορά μεταξύ των πραγματικών και των προσεγγιστικών εξόδων χρησιμοποιώντας μόνο το σύνολο εκπαίδευσης, άρα από την ελαχιστοποίησή της μπορεί να οδηγηθούμε σε υπερπροσαρμογή. Έτσι, αντί για τη συνάρτηση απώλειας, μπορούμε να χρησιμοποιήσουμε ως μέτρο για την αποδοτικότητα του αλγορίθμου το *αναμενόμενο ρίσκο* R ως προς τη συνάρτηση h. Το αναμενόμενο ρίσκο αντιστοιχεί στην πιθανότητα εσφαλμένης προσέγγισης των εξόδων για όλες τις δυνατές εισόδους (ακόμα και για αυτές που δεν γνωρίζουμε την πραγματική έξοδο). Περισσότερη ανάλυση για το αναμενόμενο ρίσκο, όπως και για τη διαφορά του από το εμπειρικό ρίσκο που αναφέραμε παραπάνω, γίνεται στη συνέχεια της εργασίας.

Εφόσον το αναμενόμενο ρίσκο R αποτελεί πιθανότητα, βλέπουμε ότι για την ελαχιστοποίησή του είναι χρήσιμη η προηγούμενη γνώση της κατανομής των δεδομένων, δηλαδή της από κοινού συνάρτησης πυκνότητας πιθανότητας $P(x, y)$. Η $P(x, y)$ αντιστοιχεί ταυτόχρονα στην συνάρτηση πυκνότητας πιθανότητας των δεδομένων εισόδου x, $P(x)$, και στη συνάρτηση της δεσμευμένης πιθανότητας $P(x|y)$ που μας δίνεται από το σύνολο εκπαίδευσης. Στην πραγματικότητα όμως, θα πρέπει να είμαστε σε θέση να ελαχιστοποίησουμε το αναμενόμενο ρίσκο χωρίς τη γνώση της P, με μοναδική βάση τα δεδομένα του συνόλου εκπαίδευσης $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$. Επομένως, είναι απαραίτητη η προσεκτική επιλογή της οικογένειας συναρτήσεων πρόβλεψης. Μία προσέγγιση για τη σωστή επιλογή της κατάλληλης οικογένειας συναρτήσεων πρόβλεψης είναι η *ελαχιστοποίηση κατασκευαστικού σφάλματος* (structural risk minimization), η οποία περιγράφεται στο βιβλίο του Vapnik (1998).

Συναρτήσεις Πρόβλεψης και Απώλειας. Αντί για μία αυθαίρετη συνάρτηση πρόβλεψης επιλεγμένη από μια γενική οικογένεια συναρτήσεων, υποθέτουμε ότι η h έχει συγκεκριμένη μορφή. Δηλαδή, εκφράζουμε τη συνάρτηση πρόβλεψης ως προς το ζεύγος εισόδου - εξόδου και το διάνυσμα παραμέτρων $w \in \mathbb{R}^d$, $h(\cdot;\cdot) : \mathbb{R}^{d_x} \times \mathbb{R}^d \to \mathbb{R}^{d_y}$. Τότε, θεωρούμε ότι η h ανήκει στην οικογένεια

$$\mathcal{H} \coloneqq \{h(\cdot; w) : w \in \mathbb{R}^d\}. \tag{3.4}$$

Μέσα από αυτή την οικογένεια, αναζητούμε μια συνάρτηση πρόβλεψης $h \in \mathcal{H} \subset \mathbb{R}^{d_y}$ η οποία θα ελαχιστοποιεί ως προς το διάνυσμα παραμέτρων w τις αποκλίσεις ανάμεσα στις τιμές των εξόδων που προσεγγίζει και των πραγματικών εξόδων. Για τον σκοπό αυτό, οι αποκλίσεις περιγράφονται από μια δεδομένη συνάρτηση απώλειας $\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$ η οποία, δεδομένου ενός ζεύγους εισόδου - εξόδου (x, y), παράγει την απώλεια $\ell(h(x; w), y)$ όπου $h(x; w)$ είναι η προσεγγιστική τιμή της εξόδου ως προς το διάνυσμα εισόδου x και το διάνυσμα παραμέτρου w. Όταν η ℓ εκφράζεται ως συνάρτηση των h και y, μία συνηθισμένη μορφή της είναι η *συνάρτηση τετραγωνικού σφάλματος* (squared error function) $\ell(h(x; w), y) = (y - h(x; w))^2$. Ωστόσο, σε εργασίες ταξινόμησης συγκεκριμένα, επιλέγεται συχνά η συνάρτηση *hinge loss* $\ell(h(x; w), y) = \max(0, 1 - h(x; w)y)$.

Αναμενόμενο Ρίσκο. Ιδανικά, το πρόβλημα βελτιστοποίησης το οποίο καλούμαστε να λύσουμε ελαχιστοποιεί την αναμενόμενη απώλεια για όλα τα παραδείγματα ως προς το διάνυσμα παραμέτρων w. Επομένως, εκφράζουμε το αναμενόμενο ρίσκο στο οποίο αναφερθήκαμε παραπάνω, ως την αναμενόμενη τιμή της συνάρτησης απώλειας ℓ όταν γνωρίζουμε εκ των προτέρων την κατανομή όλων των δεδομένων. Εάν $P : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to [0, 1]$ είναι η συνάρτηση πυκνότητας πιθανότητας που απεικονίζει την πραγματική συσχέτιση μεταξύ εισόδων και εξόδων, η αντικειμενική συνάρτηση που θέλουμε να ελαχιστοποιήσουμε σε όλο το χώρο εισόδων - εξόδων $\mathbb{R}^{d_y} \times \mathbb{R}^{d_y}$ ως προς το διάνυσμα w είναι η

$$R(w) = \int_{\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}} \ell(h(x; w), y)\, dP(x, y) = \mathbb{E}[\ell(h(x; w), y)]. \tag{3.5}$$

Βλέπουμε ότι η $R : \mathbb{R}^d \to \mathbb{R}$ παράγει το αναμενόμενο ρίσκο (δηλαδή την αναμενόμενη απώλεια), δεδομένου ενός διανύσματος παραμέτρων w ως προς τη συνάρτηση πυκνότητας πιθανότητας P.

Εμπειρικό Ρίσκο. Η ακριβής επίλυση του προβλήματος ελαχιστοποίησης της σχέσης (3.5) μπορεί να αποβεί ιδιαίτερα δύσκολη, εξαιτίας του ότι μπορεί να μη γνωρίζουμε την κατανομή των παραδειγμάτων και κατά συνέπεια, ούτε τη συνάρτηση πυκνότητας πιθανότητας P. Άρα, σε μία εργασία επιβλεπόμενης μάθησης, μπορούμε να επιλέξουμε τυχαία παραδείγματα από το σύνολο $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, $n \in \mathbb{N}$ και να τα ορίσουμε ως το σύνολο εκπαίδευσης $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$. Επομένως, έχουμε πρόσβαση στα τυχαία επιλεγμένα παραδείγματα του συνόλου εκπαίδευσης, είτε ταυτόχρονα σε όλα είτε σε μεμονωμένα ζεύγη εισόδου - εξόδου. Με βάση τα παραδείγματα αυτά και τα διανύσματα παραμέτρων τους, ορίζεται η συνάρτηση εμπειρικού ρίσκου $R_n : \mathbb{R}^d \to \mathbb{R}$ ως εξής:

$$R_n(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i; w), y_i), \tag{3.6}$$

η οποία, όπως φαίνεται στον παραπάνω τύπο, μετράει την απώλεια για τα δεδομένα εκπαίδευσης. Η συνάρτηση του εμπειρικού ρίσκου τότε ελαχιστοποιείται, δίνοντας το βέλτιστο διάνυσμα παραμέτρων για το σύνολο εκπαίδευσης. Εφόσον η ακριβής κατανομή των παραδειγμάτων συνήθως δεν είναι γνωστή εκ των προτέρων στη μηχανική μάθηση, μπορούμε να πούμε ότι η ελαχιστοποίηση του εμπειρικού ρίσκου είναι ουσιαστικά το πρόβλημα βελτιστοποίησης που προκύπτει συχνότερα.

Κατά την ελαχιστοποίηση, όμως, του εμπειρικού ρίσκου, θα πρέπει να γενικεύσουμε το αποτέλεσμα για όλα τα παραδείγματα. Με αυτό τον τρόπο προσεγγίζουμε το αναμενόμενο ρίσκο χωρίς γνώση της P και αποφεύγουμε την υπερπροσαρμογή. Επομένως, καταφεύγουμε σε τεχνικές *κανονικοποίησης*, διαδικασία η οποία μειώνει τις αποκλίσεις και τις διαταραχές για δεδομένα εισόδου χωρίς γνωστή έξοδο. Με την κανονικοποίηση, εφαρμόζεται ένας επιπλέον όρος στη συνάρτηση εμπειρικού ρίσκου ο οποίος δίνει μια «ποινή». Όσο πιο περίπλοκη είναι η συνάρτηση του ρίσκου και συνεπώς, όσο πιο επιρρεπές είναι το μοντέλο που ψάχνουμε σε διαταραχές, τόσο μεγαλύτερη είναι η ποινή που εφαρμόζει ο όρος κανονικοποίησης. Ωστόσο, για τους σκοπούς της εργασίας, η συνάρτηση εμπειρικού ρίσκου που θα χρησιμοποιήσουμε θα

έχει τη μη κανονικοποιημένη μορφή (3.6), επισημαίνοντας ότι οι μέθοδοι βελτιστοποίησης που θα μελετήσουμε, εφαρμόζονται άμεσα όταν περιλαμβάνουμε έναν ομαλό όρο κανονικοποίησης.

Σημειώνουμε σε αυτό το σημείο ότι ο χαρακτηρισμός για το εμπειρικό ρίσκο που δόθηκε στην ενότητα της ταξινόμησης κειμένου αφορά το σφάλμα ταξινόμησης για την αντίστοιχη εργασία μηχανικής μάθησης. Από αυτή την παράγραφο και για τη συνέχεια της εργασίας, το αναμενόμενο και το εμπειρικό ρίσκο R και R_n απεικονίζουν γενικότερα μέτρα απώλειας, όπως αυτή ορίζεται από τη συνάρτηση ℓ.

Απλοποιημένος Συμβολισμός. Στις προηγούμενες παραγράφους, οι σχέσεις (3.5) και (3.6) δείχνουν ακριβώς πώς το αναμενόμενο και το εμπειρικό ρίσκο εξαρτώνται από τη συνάρτηση απώλειας. Ωστόσο, για λόγους απλότητας, αλλά και ευκολίας στη γενίκευση των προβλημάτων βελτιστοποίησης, αντικαθιστούμε τις εκφράσεις που δόθηκαν προηγουμένως με νέες, απλούστερες σχέσεις. Αρχικά, θεωρούμε μια νέα τυχαία μεταβλητή ξ ως έναν *σπόρο* (seed), δηλαδή ένα αρχικό σημείο από το οποίο παράγεται μία ακολουθία από τυχαία γνωρίσματα. Μπορούμε, συνεπώς, να συμβολίσουμε με ξ ένα μοναδικό παράδειγμα $(x, y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ και με $\xi_{[i]}$ ένα παράδειγμα (x_i, y_i), $i = 1, \ldots, n$ από το σταθερό σύνολο των τυχαία επιλεγμένων παραδειγμάτων $\{\xi_{[i]}\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. Επομένως, θα εκφράζουμε την απώλεια που προκύπτει από ένα ζεύγος (w, ξ) με μια νέα συνάρτηση $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$. Έτσι, ορίζουμε την συνάρτηση f ως προς ένα τυχαίο, μεμονωμένο παράδειγμα ξ ως

$$f(w; \xi) \coloneqq \ell(h(x; w), y) \tag{3.7}$$

και ως προς το i-οστό παράδειγμα ενός σταθερού συνόλου $\{\xi_{[i]}\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n$ ως

$$f_i(w) \coloneqq f(w; \xi_{[i]}) = \ell(h(x_i; w), y_i), \quad i = 1, \ldots, n. \tag{3.8}$$

Παρατηρούμε, δηλαδή, ότι η f συμβολίζει τη *σύνθεση* της συνάρτησης απώλειας ℓ και της συνάρτησης πρόβλεψης h.

Με βάση τις σχέσεις για τα ρίσκα (3.5), (3.6) και τους ορισμούς (3.7) και (3.8), οι εκφράσεις για το αναμενόμενο και το εμπειρικό ρίσκο αντίστοιχα, διαμορφώνονται κατά τον παρακάτω τρόπο: το αναμενόμενο ρίσκο για δοθέν w είναι η αναμενόμενη τιμή αυτής της σύνθετης συνάρτησης f ως προς την κατανομή του ξ:

$$\text{(Αναμενόμενο Ρίσκο)} \quad R(w) = \mathbb{E}[f(w; \xi)] \tag{3.9}$$

και το εμπειρικό ρίσκο είναι η μέση απώλεια του συνόλου παραδειγμάτων $\{(x_i, y_i)\}_{i=1}^n$:

$$\text{(Εμπειρικό Ρίσκο)} \quad R_n(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w). \tag{3.10}$$

3.4 Είδη Αλγορίθμων Βελτιστοποίησης στη Μηχανική Μάθηση

Οι αλγόριθμοι βελτιστοποίησης στη μηχανική μάθηση αποσκοπούν στην ελαχιστοποίηση του ρίσκου κατά την εκπαίδευση. Παρόλο που η τυπική διαδικασία βελτιστοποίησης αφορά την ελαχιστοποίηση του εμπειρικού ρίσκου (3.10), θα δούμε στη

συνέχεια της εργασίας ότι στην πραγματικότητα μας ενδιαφέρει η ελαχιστοποίηση του αναμενόμενου ρίσκου (3.9). Με βάση αυτή την πληροφορία, θα μελετήσουμε την επίδοση αλγορίθμων που ελαχιστοποιούν το αναμενόμενο ρίσκο R.

Σε πολλά προβλήματα βελτιστοποίησης, δεν μπορούμε να καθορίσουμε ακριβώς το μοντέλο, επειδή αυτό εξαρτάται από ποσότητες οι οποίες είναι άγνωστες κατά τη διάρκεια της δημιουργίας του. Αντί να βασιστούμε σε απλές προβλέψεις για τις αβέβαιες ποσότητες, μπορούμε να ενσωματώσουμε πληροφορίες γι΄ αυτές ώστε να καταλήξουμε σε πιο χρήσιμες λύσεις, όπως για παράδειγμα τα δυνατά σενάρια και τις πιθανότητές τους. Οι αλγόριθμοι *στοχαστικής βελτιστοποίησης* χρησιμοποιούν αυτά τα πιθανολογικά μεγέθη ώστε να βελτιστοποιήσουν την αναμενόμενη συμπεριφορά του μοντέλου.

Για τους σκοπούς αυτής της εργασίας, χωρίζουμε τους αλγόριθμους βελτιστοποίησης στη μηχανικής μάθηση σε δύο κύριες κατηγορίες, στις στοχαστικές μεθόδους και στις μεθόδους batch. Η κύρια στοχαστική μέθοδος βελτιστοποίησης είναι η *στοχαστική μέθοδος κλίσης* (stochastic gradient method - SG), η οποία προτάθηκε αρχικά από τους Robbins και Monro (1951) και ελαχιστοποιεί το εμπειρικό ρίσκο R_n σύμφωνα με τον επαναληπτικό τύπο

$$w_{k+1} \leftarrow w_k - \alpha_k \nabla f_{i_k}(w_k), \tag{3.11}$$

με δοθέν αρχικό διάνυσμα παραμέτρων $w_1 \in \mathbb{R}^d$. Στον επαναληπτικό τύπο (3.11) για κάθε $k \in \mathbb{N} := \{1, 2, \ldots\}$, ο δείκτης i_k που αντιστοιχεί στο παράδειγμα $\xi_{[i_k]} = (x_{i_k}, y_{i_k})$, επιλέγεται τυχαία από το $\{1, \ldots, n\}$ και το α_k είναι ένα θετικό μέγεθος βήματος. Επομένως, κάθε επανάληψη της μεθόδου υπολογίζει την κλίση $\nabla f_{i_k}(w_k)$ για ένα μοναδικό παράδειγμα, δίνοντας στον αλγόριθμο πολύ μικρό κόστος ανά επανάληψη. Παρατηρούμε ότι η εξίσωση του εμπειρικού σφάλματος R_n, η ακολουθία βημάτων $\{\alpha_k\}$ αλλά και το αρχικό διάνυσμα w_1 δεν καθορίζουν αποκλειστικά τον επαναληπτικό τύπο της SG, αφού κάτι τέτοιο θα την καθιστούσε ντετερμινιστική μέθοδο. Ο τύπος εξαρτάται επίσης από την τυχαία ακολουθία δεικτών $\{i_k\}$, έτσι η ακολουθία παραμέτρικών διανυσμάτων $\{w_k\}$ υπολογίζεται από μία στοχαστική διαδικασία.

Αντιθέτως, οι *batch* μέθοδοι χρησιμοποιούν ολόκληρο το σύνολο εκπαίδευσης και σε κάθε επανάληψη επεξεργάζονται όλα τα παραδείγματά του. Μία από τις πιο απλές μεθόδους batch είναι η μέθοδος απότομης καθόδου (steepest descent method), η οποία καλείται και batch μέθοδος κλίσης (batch gradient method). Επειδή η αντικειμενική μας συνάρτηση είναι το εμπειρικό σφάλμα και ελαχιστοποιούμε ως προς w, ο επαναληπτικός τύπος της μεθόδου απότομης κλίσης γίνεται

$$w_{k+1} \leftarrow w_k - \alpha_k \nabla R_n(w_k) = w_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla f_i(w_k). \tag{3.12}$$

Μπορούμε εύκολα να διαπιστώσουμε ότι το κόστος ανά επανάληψη της batch μεθόδου κλίσης για τον υπολογισμό του όρου $-\alpha_k \nabla R_n(w_k)$ είναι σημαντικά μεγαλύτερο από αυτό του όρου $-\alpha_k \nabla f_{i_k}(w_k)$ της μεθόδου SG. Ωστόσο, μπορούμε να περιμένουμε ένα πλεονέκτημα της batch μεθόδου ως προς την ακρίβεια του αποτελέσματος κάθε επανάληψης.

Η επιλογή ανάμεσα στις στοχαστικές μεθόδους και στις μεθόδους batch για κάθε πρόβλημα εξαρτάται από τα χαρακτηριστικά που θέλουμε να πετύχουμε ως προς την ακρίβεια, την χωρική και τη χρονική πολυπλοκότητα του αλγορίθμου και τη σύγκλιση όσον αφορά την ελαχιστοποίηση του εμπειρικού ρίσκου. Θα δούμε όμως και στη συνέχεια της εργασίας, ότι οι στοχαστικές μέθοδοι και συγκεκριμένα η SG προτιμούνται συχνότερα για προβλήματα μηχανικής μάθησης μεγάλης κλίμακας. Υπάρχουν πρακτικοί, θεωρητικοί και διαισθητικοί λόγοι για τους οποίους η μέθοδος SG συγκλίνει γρηγορότερα στην ελαχιστοποίηση του ρίσκου έναντι μιας μεθόδου batch. Δεν θα πρέπει, ωστόσο, να απορρίψουμε τις μεθόδους batch, καθώς κάτω από συγκεκριμένες συνθήκες έχουν διάφορα πλεονεκτήματα για ορισμένα προβλήματα μηχανικής μάθησης.

Με βάση τις προηγούμενες παρατηρήσεις, καταλαβαίνουμε ότι μπορούμε να εξετάσουμε μεθόδους και τεχνικές οι οποίες συνδυάζουν τα δύο είδη με αποτέλεσμα να επωφελούνται από τα πλεονεκτήματα και των δύο. Ένας τρόπος είναι μέσω της *mini-batch* προσέγγισης της μεθόδου SG (3.11). Έτσι, σε κάθε επανάληψη επιλέγουμε τυχαία ένα υποσύνολο $\mathcal{S}_k \subseteq \{1, \dots, n\}$ από τους δείκτες και ο επαναληπτικός τύπος (3.11) γίνεται

$$w_{k+1} \leftarrow w_k - \frac{\alpha_k}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \nabla f_i(w_k). \tag{3.13}$$

Η μέθοδος αυτή είναι ευρέως διαδεδομένη στην πράξη για δύο λόγους. Πρώτον, εμφανίζει πλεονέκτημα στη συνολική ταχύτητα, η οποία είναι μεγαλύτερη από της απλής SG επειδή διευκολύνει τον παράλληλο υπολογισμό των minibatch κλίσεων. Δεύτερον, οι εκτιμήσεις των στοχαστικών κλίσεων παρουσιάζουν μικρότερες αποκλίσεις, επομένως είναι πιο εύκολη επιλογή των βημάτων $\{\alpha_k\}$ (Bottou, Curtis & Nocedal, 2018, σ. 241-242).

4 Στοχαστικές Μέθοδοι Κλίσης

Στην ενότητα αυτή, θα μελετήσουμε τις ιδιότητες σύγκλισης και την υπολογιστική πολυπλοκότητα χειρότερης περίπτωσης μίας από τις μεθόδους SG. Ξεκινάμε αναλύοντας τη γνωστή μέθοδο SG για μία ισχυρά κυρτή συνάρτηση όπου μπορούμε να έχουμε ολική σύγκλιση της μεθόδου και όχι απλά τοπική. Στη συνέχεια, αναλύουμε τη μέθοδο για γενικότερες, μη κυρτές συναρτήσεις. Προκειμένου να αποδείξουμε τη γενικότητα των αποτελεσμάτων του παρόντος κεφαλαίου, θεωρούμε ότι η αντικειμενική συνάρτηση απεικονίζει είτε το αναμενόμενο ρίσκο (3.9) είτε το εμπειρικό ρίσκο (3.10). Άρα, για την αντικειμενική συνάρτηση $F : \mathbb{R}^d \to \mathbb{R}$ έχουμε ότι

$$F(w) = \begin{cases} R(w) = \mathbb{E}[f(w;\xi)] \\ \qquad \text{ή} \\ R_n(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w). \end{cases} \tag{4.1}$$

Οι παρακάτω αποδείξεις ισχύουν ισάξια και για τις δύο περιπτώσεις, με μόνη διαφορά στις εκτιμήσεις των κλίσεων ανάλογα με το πρόβλημα που μας ενδιαφέρει. Αναφέρουμε εδώ ότι η επιλογή παραδειγμάτων από ένα πεπερασμένο σύνολο εκπαίδευσης και η αντικατάστασή τους σε κάθε επανάληψη, αντιστοιχεί σε μία διακριτή κατανομή των παραδειγμάτων. Σε αυτή την περίπτωση, ο αλγόριθμος SG ελαχιστοποιεί το εμπειρικό ρίσκο $F = R_n$, Διαφορετικά, η επιλογή των παραδειγμάτων από την κατανομή P γίνεται σε κάθε επανάληψη και ο αλγόριθμος ελαχιστοποιεί το αναμενόμενο ρίσκο $F = R$.

Η γενικευμένη μέθοδος στοχαστικών κλίσεων τρέχει σύμφωνα με τον παρακάτω αλγόριθμο (Bottou, Curtis & Nocedal, 2018, σ. 243).

Αλγόριθμος 4.1 Στοχαστική Μέθοδος Κλίσης (SG)

1: Επέλεξε ένα αρχικό διάνυσμα w_1.
2: **for** $k = 1, 2, \ldots$ **do**
3: Δημιούργησε μια νέα τυχαία μεταβλητή ξ_k.
4: Υπολόγισε το $g(w_k, \xi_k)$.
5: Επέλεξε ένα βήμα $\alpha_k > 0$.
6: Θέσε $w_{k+1} \leftarrow w_k - \alpha_k g(w_k, \xi_k)$.
7: **end for**

Ο αλγόριθμος αυτός υποθέτει ότι προϋπάρχουν οι μηχανισμοί που θα επιλέξουν την τυχαία μεταβλητή ξ_k, το στοχαστικό διάνυσμα $g(w_k, \xi_k)$ και το βαθμωτό βήμα $\alpha_k > 0$. Σημειώνουμε εδώ ότι με ξ_k συμβολίζεται η τυχαία μεταβλητή ξ που επιλέγεται κατά την $k-$οστή επανάληψη του αλγορίθμου. Έτσι, η μέθοδος παράγει την ακολουθία ανεξάρτητων τυχαίων μεταβλητών $\{\xi_k\}$. Ως προς τη γενικότητά του, στον Αλγόριθμο 4.1 μπορούμε να θεωρήσουμε είτε την ξ_k ως ένα μοναδικό παράδειγμα εισόδου - εξόδου (x_k, y_k) όπως στην απλή μέθοδο SG (3.11) είτε την $\xi_{[k,i]}$ ως στοιχείο ενός συνόλου παραδειγμάτων $\{(x_{k,i}, y_{k,i})\}_{i\in\mathcal{S}_k}$ όπως αυτό που χρησιμοποιεί ο επαναληπτικός τύπος της minibatch SG (3.13).

Το διάνυσμα κατεύθυνσης $g(w_k, \xi_k)$ μπορεί να είναι μία στοχαστική κλίση, δηλαδή μία αμερόληπτη εκτιμήτρια της κλίσης $\nabla F(w_k)$. Σε αυτή την περίπτωση, από τον

ορισμό της αμερόληπτης εκτιμήτριας θα πρέπει $\mathbb{E}[g(w_k;\xi_k)] = \nabla F(w_k)$. Δηλαδή, όταν $F(w_k) = R(w_k) = \mathbb{E}[f(w_k;\xi_k)]$, τότε

$$\nabla F(w_k) = \nabla R(w_k) = \nabla \mathbb{E}[f(w_k;\xi_k)] = \mathbb{E}[\nabla f(w_k;\xi_k)],$$

αφού η κλίση της αντικειμενικής συνάρτησης υπολογίζεται ως προς το παραμετρικό διάνυσμα w_k και μόνο. Κατά τον ίδιο τρόπο, εάν

$$F(w_k) = R_n(w_k) = \frac{1}{n_k}\sum_{i=1}^{n_k} f_i(w_k) = \frac{1}{n_k}\sum_{i=1}^{n_k} f(w_k;\xi_{k,i}),$$

η στοχαστική κλίση θα είναι $g(w_k;\xi_k) = \frac{1}{n_k}\sum_{i=1}^{n_k}\nabla f(w_k;\xi_{k,i})$, όπου n_k με $k \in \mathbb{N}$ είναι το μέγεθος του minibatch. Επομένως, έχουμε τρεις επιλογές για το διάνυσμα g. Η πρώτη αφορά την απλή μέθοδο SG, η οποία χρησιμοποιεί μόνο ένα παράδειγμα για τον υπολογισμό της στοχαστικής κλίσης. Έπειτα, η δεύτερη αφορά τη minibatch μέθοδο SG, η οποία χρησιμοποιεί n_k το πλήθος παραδείγματα στον υπολογισμό. Τέλος, η τρίτη περίπτωση αφορά τη minibatch μέθοδο SG με τη χρήση ενός πίνακα μετασχηματισμών. Αυτή η περίπτωση στοχαστικής κατεύθυνσης αφορά τις στοχαστικές μεθόδους Newton ή quasi-Newton, οι οποίες παρουσιάζονται στο πέμπτο κεφάλαιο. Οι παραπάνω επιλογές για το g συνοψίζονται στον παρακάτω τύπο:

$$g(w_k,\xi_k) = \begin{cases} \nabla f(w_k;\xi_k), \\ \dfrac{1}{n_k}\displaystyle\sum_{i=1}^{n_k}\nabla f(w_k;\xi_{k,i}), \\ H_k\dfrac{1}{n_k}\displaystyle\sum_{i=1}^{n_k}\nabla f(w_k;\xi_{k,i}), \end{cases} \tag{4.2}$$

όπου H_k είναι ένας θετικά ορισμένος πίνακας μετασχηματισμών. Καταλαβαίνουμε, λοιπόν, ότι ο Αλγόριθμος 4.1 μπορεί να τρέξει με τρεις διαφορετικούς τρόπους, ανάλογα με την περίπτωση της στοχαστικής κλίσης που επιλέγουμε κάθε φορά.

Επίσης, ο Αλγόριθμος 4.1 επιτρέπει διάφορες επιλογές για την ακολουθία βημάτων $\{\alpha_k\}$. Ωστόσο, στην παρούσα εργασία θα επικεντρώσουμε την προσοχή μας σε δύο επιλογές, μία με σταθερό μέγεθος βημάτων και μία με βήματα μειούμενου μεγέθους. Τέλος, σημειώνουμε ότι ο Αλγόριθμος 4.1 καλύπτει επίσης και περιπτώσεις στις οποίες το σημείο της τρέχουσας επανάληψης w_k επηρεάζει την επιλογή των παραδειγμάτων. Η υπόθεση ότι τα στοιχεία της τυχαίας μεταβλητής $\{\xi_k\}$ είναι ανεξάρτητα διευκολύνει την ανάλυση της στοχαστικής μεθόδου κλίσης. Ωστόσο, παρόμοια αποτελέσματα ισχύουν και όταν οι τυχαίες μεταβλητές ξ_k δεν είναι ανεξάρτητες και οι αναμενόμενες τιμές αφορούν τη δεσμευμένη κατανομή της ξ_k δεδομένων των $\{\xi_1, \ldots, \xi_{k-1}\}$.

Προς αποφυγή σύγχυσης, από εδώ και στο εξής θα αναφερόμαστε στον Αλγόριθμο 4.1 ως τη μέθοδο SG, στη σχέση (3.11) ως *απλή μέθοδο SG* και στη σχέση (3.13) ως *minibatch SG*. Επομένως, έχοντας μία πρώτη εικόνα για τη μέθοδο SG, μπορούμε να παρουσιάσουμε τις ιδιότητές της ως προς τη σύγκλιση. Οι υποθέσεις και τα λήμματα που δίνονται στην επόμενη ενότητα εφαρμόζονται ικανοποιητικά για οποιοδήποτε είδος αντικειμενικής συνάρτησης (εμπειρικό ή αναμενόμενο ρίσκο) και στοχαστικής κλίσης (τύπος (4.2» επιλέξουμε για τη μέθοδο SG, χωρίς βλάβη της γενικότητας.

4.1 Θεμελιώδεις Υποθέσεις και Λήμματα

Όπως προαναφέραμε, στην ενότητα αυτή δίνονται κάποια σημαντικά λήμματα με τις απαραίτητες υποθέσεις τους ώστε να μπορέσουμε να εξάγουμε αποτελέσματα για τη σύγκλιση της μεθόδου SG. Μία προϋπόθεση που θα μας βοηθήσει στην ανάλυση των υποθέσεων και των λημμάτων είναι ότι η αντικειμενική συνάρτηση F είναι λεία. Μια συνάρτηση $f : \mathbb{R}^d \subset \mathbb{R}^n \to \mathbb{R}$ λέγεται *λεία* όταν όλες οι μερικές της παράγωγοι όλων των τάξεων υπάρχουν και είναι συνεχείς. Επομένως είναι προφανές ότι εάν η συνάρτηση f είναι λεία, τότε θα είναι και *συνεχώς παραγωγίσιμη*, δηλαδή, οι μερικές της παράγωγοι πρώτης τάξης $\frac{\partial}{\partial x_d}$ υπάρχουν και είναι συνεχείς. Με βάση αυτή την προϋπόθεση, προχωράμε στην πρώτη υπόθεση (Bottou, Curtis & Nocedal, σ. 244, 2018).

Υπόθεση 4.1 (Lipschitz κλίσεις)**.** *Η αντικειμενική συνάρτηση $F : \mathbb{R}^d \to \mathbb{R}$ είναι συνεχώς παραγωγίσιμη και η κλίση της $\nabla F : \mathbb{R}^d \to \mathbb{R}^d$ είναι Lipschitz συνεχής με σταθερά Lipschitz $L > 0$, δηλαδή,*

$$\|\nabla F(w) - \nabla F(v)\|_2 \leq L \|w - v\|_2 \quad \textit{για κάθε } \{w, v\} \subset \mathbb{R}^d. \tag{4.3}$$

Η συνθήκη Lipschitz διαισθητικά μας δείχνει πόσο απότομη μπορεί να γίνει μία συνάρτηση. Άρα, η Υπόθεση 4.1 εξασφαλίζει ότι η κλίση δεν αλλάζει με άπειρα γρήγορο ρυθμό ως προς το παραμετρικό διάνυσμα. Έτσι, δημιουργείται ένα μέτρο ως προς το πόσο μπορούμε να κινηθούμε κατά μία κατεύθυνση έτσι ώστε η F να ελαττώνεται. Ως προς το αντίστροφο της υπόθεσης, επειδή η συνθήκη της Lipschitz συνέχειας είναι πιο ισχυρή από αυτή της συνέχειας, τότε κάθε παραγωγίσιμη συνάρτηση με Lipschitz συνεχή κλίση είναι συνεχώς παραγωγίσιμη. Συνέπεια της παραπάνω υπόθεσης είναι η παρακάτω σχέση

$$F(w) \leq F(v) + \nabla F(v)^T (w - v) + \frac{1}{2} L \|w - v\|_2^2 \quad \text{για κάθε } \{w, v\} \subset \mathbb{R}^d. \tag{4.4}$$

Απόδειξη. Για την απόδειξη της σχέσης (4.4), έχουμε ότι

$$\begin{aligned} F(w) &= F(v) + (F(w) - F(v)) \\ &= F(v) + (F(v + (w - v)) - F(v)) \\ &= F(v) + (F(v + (w - v)) - F(v)) \int_0^1 dt \\ &= F(v) + \int_0^1 (F(v + (w - v)) - F(v))\, dt, \end{aligned}$$

το οποίο από το Θεώρημα μέσης τιμής (2.17) γίνεται

$$\begin{aligned} F(w) &= F(v) + \int_0^1 \nabla F(v + t(w - v))^T (w - v)\, dt \\ &= F(v) + \int_0^1 \nabla F(v)^T (w - v)\, dt \\ &\quad + \int_0^1 [\nabla F(v + t(w - v))^T (w - v) - \nabla F(v)^T (w - v)]\, dt \end{aligned}$$

$$
\begin{aligned}
&= F(v) + \nabla F(v)^T(w-v) + \int_0^1 [\nabla F(v+t(w-v)) - \nabla F(v)]^T(w-v)\,dt \\
&\le F(v) + \nabla F(v)^T(w-v) + \left\| \int_0^1 [\nabla F(v+t(w-v)) - \nabla F(v)]^T(w-v)\,dt \right\|_2 \\
&\le F(v) + \nabla F(v)^T(w-v) + \int_0^1 \left\| [\nabla F(v+t(w-v)) - \nabla F(v)]^T(w-v) \right\|_2 dt \\
&= F(v) + \nabla F(v)^T(w-v) + \int_0^1 \|\nabla F(v+t(w-v)) - \nabla F(v)\| \, \|w-v\|_2 \, dt
\end{aligned}
$$

και από τη σχέση (4.3) της Υπόθεσης 4.1 έχουμε

$$
\begin{aligned}
F(w) &\le F(v) + \nabla F(v)^T(w-v) + \int_0^1 Lt\|w-v\|_2\|w-v\|_2 dt \\
&= F(v) + \nabla F(v)^T(w-v) + L\,\|w-v\|_2^2 \int_0^1 t\,dt \\
&= F(v) + \nabla F(v)^T(w-v) + \frac{1}{2}L\,\|w-v\|_2^2
\end{aligned}
$$

που είναι η ζητούμενη σχέση (4.4). □

Από την Υπόθεση 4.1 προκύπτει το παρακάτω λήμμα (Bottou, Curtis & Nocedal, σ. 244-245, 2018), όπου με $\mathbb{E}_{\xi_k}[\ ]$ συμβολίζουμε την αναμενόμενη τιμή ως προς την κατανομή της τυχαίας μεταβλητής ξ_k δοθέντος w_k. Επομένως, η ποσότητα $\mathbb{E}_{\xi_k}[F(w_{k+1})]$ έχει νόημα εφόσον το w_{k+1} εξαρτάται από την ξ_k όπως φαίνεται και στο βήμα 6 του Αλγορίθμου 4.1.

Λήμμα 4.2. *Δεδομένης της Υπόθεσης 4.1, τα διανύσματα $\{w_k\}$ της SG (Αλγόριθμος 4.1) ικανοποιούν την ακόλουθη ανισότητα για κάθε $k \in \mathbb{N}$:*

$$
\begin{aligned}
\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \le &-\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k,\xi_k)] \\
&+ \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k,\xi_k)\|_2^2].
\end{aligned} \tag{4.5}
$$

Απόδειξη. Το Βήμα 6 του Αλγορίθμου 4.1 δίνει τον επαναληπτικό τύπο της SG:

$$
w_{k+1} = w_k - \alpha_k g(w_k,\xi_k). \tag{4.6}
$$

Από τη σχέση (4.4), για $w = w_{k+1}$ και $v = w_k$, τα παραμετρικά διανύσματα w_{k+1} που προκύπτουν από κάθε επανάληψη της μεθόδου SG ικανοποιούν την παρακάτω σχέση:

$$
\begin{aligned}
F(w_{k+1}) &\le F(w_k) + \nabla F(w_k)^T(w_{k+1}-w_k) + \frac{1}{2}L\,\|w_{k+1}-w_k\|_2^2 \quad \text{ή} \\
F(w_{k+1}) - F(w_k) &\le \nabla F(w_k)^T(w_{k+1}-w_k) + \frac{1}{2}L\,\|w_{k+1}-w_k\|_2^2 .
\end{aligned}
$$

Ο επαναληπτικός τύπος (4.6) δίνει $w_{k+1} - w_k = -\alpha_k g(w_k,\xi_k)$. Άρα,

$$
F(w_{k+1}) - F(w_k) \le \nabla F(w_k)^T(w_{k+1}-w_k) + \frac{1}{2}L\,\|w_{k+1}-w_k\|_2^2
$$

$$= \nabla F(w_k)(-\alpha_k g(w_k, \xi_k)) + \frac{1}{2} L \| - \alpha_k g(w_k, \xi_k)\|^2$$
$$= -\alpha_k \nabla F(w_k)^T g(w_k, \xi_k) + \frac{1}{2}\alpha_k^2 L \left\| g(w_k, \xi_k)\right\|_2^2 .$$

Παίρνοντας τις αναμενόμενες τιμές για κάθε μέλος της παραπάνω ανισότητας ως προς την κατανομή του ξ_k έχουμε

$$\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)] \leq \mathbb{E}_{\xi_k}[-\alpha_k \nabla F(w_k)^T g(w_k, \xi_k) + \frac{1}{2}\alpha_k^2 L \left\| g(w_k, \xi_k)\right\|_2^2]. \quad (4.7)$$

Παρατηρούμε, όμως, ότι το w_k δεν εξαρτάται από το ξ_k, επομένως, από την ιδιότητα (3.1) της αναμενόμενης τιμής έχουμε ότι

$$\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)] = \mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k)$$

και

$$\mathbb{E}_{\xi_k}[-\alpha_k \nabla F(w_k)^T g(w_k, \xi_k) + \frac{1}{2}\alpha_k^2 L \left\| g(w_k, \xi_k)\right\|_2^2]$$
$$= -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2].$$

Από τις δύο προηγούμενες σχέσεις, η (4.7) γίνεται

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2].$$

□

Το λήμμα αυτό δείχνει ότι ανεξάρτητα με το πώς η μέθοδος SG έφτασε στο w_k, η μείωση που περιμένουμε στην αντικειμενική συνάρτηση μέχρι το k-οστό βήμα είναι άνω φραγμένη από μία ποσότητα που εξαρτάται από την αναμενόμενη παράγωγο της F στο w_k κατά την κατεύθυνση του $-g(w_k, \xi_k)$ και από τη ροπή δεύτερης τάξης του διανύσματος $g(w_k, \xi_k)$. Για παράδειγμα, εάν το $g(w_k, \xi_k)$ είναι μια αμερόληπτη εκτιμήτρια της παραγώγου $\nabla F(w_k)$, τότε από τον ορισμό της αμερόληπτης εκτιμήτριας (3.3) έχουμε ότι $\mathbb{E}_{\xi_k}[g(w_k, \xi_k)] = \nabla F(w_k)$. Άρα, η σχέση (4.5) του Λήμματος 4.2 δίνει:

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) &\leq -\alpha_k \nabla F(w_k)^T \nabla F(w_k) + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \\ &= -\alpha_k \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2].\end{aligned} \quad (4.8)$$

Θα δούμε στη συνέχεια ότι η μέθοδος SG συγκλίνει εφόσον οι στοχαστικές κατευθύνσεις $\{g(w_k, \xi_k)\}$ και τα βήματα α_k επιλέγονται κατά τρόπο τέτοιο ώστε το δεξιό μέλος της ανισότητας (4.5) να είναι φραγμένο από μία ντετερμινιστική ποσότητα που δίνει ικανοποιητική ασυμπτωτική μείωση στην F. Η συγκεκριμένη προϋπόθεση μπορεί να εξασφαλιστεί παίρνοντας επιπλέον συνθήκες για τις πρώτες και δεύτερες ροπές των στοχαστικών διευθύνσεων. Εάν η διακύμανση του διανύσματος $g(w_k, \xi_k)$ είναι

$$\mathbb{V}_{\xi_k}[g(w_k, \xi_k)] \coloneqq \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] - \left\|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\right\|_2^2 , \quad (4.9)$$

η παρακάτω υπόθεση περιορίζει την ποσότητα (4.9) ώστε η αρνητική επίδραση του τελευταίου όρου της ανισότητας (4.8) (μη ντετερμινιστικός) να είναι μικρότερη.

Υπόθεση 4.3 (όρια ροπής πρώτης και δεύτερης τάξης)**.** *Η αντικειμενική συνάρτηση και η SG όπως περιγράφεται από τον Αλγόριθμο 4.1 ικανοποιούν τις παρακάτω συνθήκες:*

(α) *Η ακολουθία των παραμετρικών διανυσμάτων* $\{w_k\}$ *περιορίζεται σε ένα ανοιχτό σύνολο επί του οποίου η* F *είναι κάτω φραγμένη από ένα βαθμωτό μέγεθος* $F_{\inf}$*.*

(β) *Υπάρχουν σταθερές* $\mu_G \geq \mu > 0$ *τέτοιες ώστε για κάθε* $k \in \mathbb{N}$*,*

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \geq \mu \left\|\nabla F(w_k)\right\|_2^2 \quad \textit{και} \tag{4.10a}$$

$$\left\|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\right\|_2 \leq \mu_G \left\|\nabla F(w_k)\right\|_2 . \tag{4.10b}$$

(γ) *Υπάρχουν σταθερές* $M \geq 0$ *και* $M_V \geq 0$ *τέτοιες ώστε για κάθε* $k \in \mathbb{N}$*,*

$$\mathbb{V}_{\xi_k}[g(w_k, \xi_k)] \leq M + M_V \left\|\nabla F(w_k)\right\|_2^2 . \tag{4.11}$$

Σύμφωνα με την πρώτη συνθήκη, η αντικειμενική συνάρτηση πρέπει να είναι φραγμένη στο σύνολο στο οποίο τρέχει ο αλγόριθμος. Η δεύτερη συνθήκη υποδεικνύει ότι κατά την αναμενόμενη τιμή του, το διάνυσμα $-g(w_k, \xi_k)$ μας δίνει μια διεύθυνση στη οποία η F, ξεκινώντας από το w_k, μειώνεται ικανοποιητικά και η νόρμα της $\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]$ είναι συγκρίσιμη με την νόρμα της κλίσης της F. Η τρίτη συνθήκη δηλώνει ότι η διακύμανση της στοχαστικής κατεύθυνσης $g(w_k, \xi_k)$ είναι περιορισμένη, αλλά σχετικά λίγο. Για παράδειγμα, εάν η F είναι μια κυρτή, τετραγωνική συνάρτηση, τότε η διακύμανση μπορεί να είναι μη μηδενική για κάθε στάσιμο σημείο της F και μπορεί να αυξάνεται τετραγωνικά προς κάθε κατεύθυνση.

Εφαρμόζοντας τον ορισμό της διακύμανσης κατά την κατανομή της τυχαίας μεταβλητής ξ_k για την αμερόληπτη εκτιμήτρια $g(w_k, \xi_k)$ (4.9) στη συνθήκη (4.11) παίρνουμε

$$\begin{aligned} &\mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] - \|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2^2 \leq M + M_V \|\nabla F(w_k)\|_2^2 \qquad \text{ή} \\ &\mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \leq M + M_V \|\nabla F(w_k)\|_2^2 + \|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2^2 . \end{aligned} \tag{4.12}$$

Όμως, από τη συνθήκη (4.10b) της Υπόθεσης 3 έχουμε ότι

$$\begin{aligned} &\left\|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\right\|_2 \leq \mu_G \left\|\nabla F(w_k)\right\|_2 \Rightarrow \\ &\left\|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\right\|_2^2 \leq \mu_G^2 \left\|\nabla F(w_k)\right\|_2^2 . \end{aligned}$$

Άρα, η σχέση (4.12) γίνεται

$$\begin{aligned} \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] &\leq M + M_V \|\nabla F(w_k)\|_2^2 + \|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2^2 \\ &\leq M + M_V \|\nabla F(w_k)\|_2^2 + \mu_G^2 \left\|\nabla F(w_k)\right\|_2^2 \\ &= M + \|\nabla F(w_k)\|_2^2 (M_V + \mu_G^2). \end{aligned}$$

Συνεπώς, καταλήγουμε στο ότι η ροπή δεύτερης τάξης της εκτιμήτριας $g(w_k, \xi_k)$ ικανοποιεί την εξής ανισότητα

$$\mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \leq M + M_G \left\|\nabla F(w_k)\right\|_2^2 \ \text{ με } \ M_G \coloneqq M_V + \mu_G^2 \geq \mu^2 > 0, \tag{4.13}$$

επειδή από την Υπόθεση 4.3 έχουμε $\mu_G \geq \mu > 0$, δηλαδή $\mu_G^2 \geq \mu^2 \Rightarrow M_V + \mu_G^2 \geq \mu^2 > 0$. Επομένως, από το Λήμμα 4.2 και τις επιπλέον υποθέσεις που κάναμε, προκύπτει το επόμενο λήμμα.

Λήμμα 4.4. *Δεδομένων των Υποθέσεων 4.1 και 4.3, τα διανύσματα* $\{w_k\}$ *που προκύπτουν από τον αλγόριθμο της SG ικανοποιούν τις ακόλουθες ανισότητες για κάθε* $k \in \mathbb{N}$*:*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq \mu\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \tag{4.14a}$$

$$\leq -(\mu - \frac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 LM. \tag{4.14b}$$

Απόδειξη. Η πρώτη σχέση της δεύτερης συνθήκη της Υπόθεσης 4.3, (4.10a) δίνει

$$\begin{aligned} \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] &\geq \mu \|\nabla F(w_k)\|_2^2 \Rightarrow \\ -\nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] &\leq -\mu \|\nabla F(w_k)\|_2^2 \Rightarrow \\ -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] &\leq -\mu\alpha_k \|\nabla F(w_k)\|_2^2 \end{aligned}$$

Συνεπώς, η σχέση (4.8) που προκύπτει από το Λήμμα 4.2 γίνεται

$$\begin{aligned} \mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) &\leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \\ &\leq -\mu\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2], \end{aligned}$$

δηλαδή καταλήξαμε στην (4.14a). Για την απόδειξη της σχέσης (4.14b), εφαρμόζουμε στην (4.14a) την ανισότητα (4.13) που προκύπτει από την Υπόθεση 4.3, δηλαδή

$$\begin{aligned} \mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) &\leq -\mu\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \\ &\leq -\mu\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L(M + M_G \|\nabla F(w_k)\|_2^2) \\ &= -\mu\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L M_G \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 LM \\ &= -(\mu - \frac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 LM. \end{aligned}$$

□

Το λήμμα αυτό υποδεικνύει ότι ανεξάρτητα με το πώς καταλήγουμε στο διάνυσμα w_k, η διαδικασία της βελτιστοποίησης συνεχίζει με το διάνυσμα w_{k+1} να εξαρτάται μόνο από το w_k, την τυχαία μεταβλητή ξ_k και το βήμα α_k και από κανένα από τα προηγούμενα διανύσματα. Το δεύτερο μέλος των ανισοτήτων του Λήμματος 4.4 είναι μία ντετερμινιστική ποσότητα η οποία φράσσει τη διαφορά $\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k)$. Επίσης, για αρκετά μικρό α_k ο πρώτος όρος της σχέσης (4.14b) είναι αρνητικός, επομένως η αντικειμενική συνάρτηση ελαττώνεται σε μέγεθος ανάλογο της νόρμας $\|\nabla F(w_k)\|_2^2$. Όμως, θα πρέπει γενικά να προσέχουμε κατά την εκτέλεση της μεθόδου SG καθώς ο δεύτερος όρος της (4.14b) μπορεί να είναι αρκετά μεγάλος και έτσι η αντικειμενική συνάρτηση τελικά να αυξάνεται.

Η υπόθεση ότι η αντικειμενική συνάρτηση είναι λεία, έχει το πλεονέκτημα ότι εξασφαλίζει τις ιδιότητες σύγκλισης για κυρτές αλλά και για μη κυρτές συναρτήσεις, οι

οποίες εμφανίζονται αρκετά συχνά στη μηχανική μάθηση. Επίσης, όταν η αντικειμενική συνάρτηση F είναι λεία, τα αποτελέσματα των παραπάνω λημμάτων ισχύουν για όποια σχέση επιλέξουμε για το στοχαστικό διάνυσμα g (4.2). Το βασικό μειονέκτημα, όμως, αυτής της υπόθεσης είναι ότι θα πρέπει να χειριστούμε την ελαχιστοποίηση των μη λείων μοντέλων ξεχωριστά.

Μία λύση είναι να χρησιμοποιήσουμε συναρτήσεις οι οποίες είναι κυρτές, έτσι ώστε τα απαραίτητα λήμματα για την ανάλυση της SG να ισχύουν για λείες και μη λείες συναρτήσεις ταυτόχρονα. Η συνθήκη της κυρτότητας μας εξασφαλίζει παρόμοιες εγγυήσεις (λήμματα) με τις παραπάνω. Κατ΄ αυτό τον τρόπο, μπορούμε να αναλύσουμε τις ιδιότητες σύγκλισης μιας μεθόδου στοχαστικών κλίσεων, ειδικά όταν ισχύει η υπόθεση των Lipschitz συνεχών κλίσεων. Μία διεξοδική ανάλυση για μη παραγωγίσιμες, μη λείες αλλά κυρτές συνάρτήσεις γίνεται στην εργασία των Nemirovski et al. (2009). Στη μελέτη αυτή, γίνεται εκτίμηση της απόστασης της τρέχουσας τιμής της αντικειμενικής συνάρτησης από τη βέλτιστη λύση. Ωστόσο, η τεχνική αυτή μπορεί να αποδειχθεί δύσχρηστη σε μη κυρτές συναρτήσεις ή όταν το βήμα της μεθόδου επιλέγεται σύμφωνα με τις μεθόδους Newton και quasi-Newton. Επομένως, είναι πολλές φορές απαραίτητη η εκ των προτέρων γνώση του σημείου ελαχίστου $\bar{w}$.

4.2 Μέθοδος SG για Ισχυρά Κυρτές Συναρτήσεις

Σε αυτή την εργασία, παρουσιάζονται κυρίως ιδιότητες σύγκλισης για ισχυρά κυρτές αντικειμενικές συναρτήσεις, καθώς αποτελούν την πιο απλή περίπτωση για τη μέθοδο SG. Μία αντικειμενική συνάρτηση $F : \mathbb{R}^d \to \mathbb{R}$ λέγεται *ισχυρά κυρτή* εάν υπάρχει μία σταθερά $c > 0$ για την οποία ισχύει ότι:

$$F(v) \geq F(w) + \nabla F(w)^T (v - w) + \frac{1}{2} c \left\| v - w \right\|_2^2 \quad \textit{για κάθε } (v, w) \in \mathbb{R}^d \times \mathbb{R}^d. \tag{4.15}$$

Η επιλογή αυτή γίνεται διότι τα αποτελέσματα που προκύπτουν από την ανάλυση τέτοιων συναρτήσεων γενικεύονται εύκολα για διάφορα προβλήματα της μηχανικής μάθησης. Για παράδειγμα, ένα μοντέλο μηχανικής μάθησης όπου γίνεται ελαχιστοποίηση μιας (απλά) κυρτής συνάρτησης, συχνά κανονικοποιείται από μία ισχυρά κυρτή συνάρτηση έτσι ώστε να διευκολυνθεί η εξαγωγή των αποτελεσμάτων. Επίσης, σε πολλές περιπτώσεις το μοντέλο περιγράφεται από μία συνάρτηση η οποία δεν είναι ισχυρά κυρτή σε όλο το πεδίο ορισμού, αλλά μόνο σε περιοχές των σημείων τοπικών ελαχίστων. Συνεπώς τα αποτελέσματα περιγράφουν καλύτερα τη συμπεριφορά του αλγορίθμου σε αυτές τις περιοχές. Τέλος, οι ιδιότητες που περιγράφουμε εδώ μπορούν εύκολα να γενικευτούν και για την ελαχιστοποίηση μη ισχυρά κυρτών αντικειμενικών συναρτήσεων.

Υπόθεση 4.5 (μοναδικό σημείο ελαχίστου)**.** *Εάν η αντικειμενική συνάρτηση F είναι ισχυρά κυρτή, δηλαδή υπακούει στη σχέση (4.15), τότε εμφανίζει μοναδικό σημείο ελαχίστου $\bar{w} \in \mathbb{R}^d$, με ελάχιστη τιμή την $\bar{F} \coloneqq F(\bar{w})$.*

Παρατηρούμε ότι από τις σχέσεις (4.15) και (4.4), οι σταθερές c και L θα πρέπει να ικανοποιούν την ανισότητα $c \leq L$. Επίσης, από την ισχυρή κυρτότητα προκύπτει μία χρήσιμη πληροφορία για την απόσταση μεταξύ της τιμής της F σε ενα σημείο

w και της ελάχιστης τιμής της. Θα δείξουμε ότι, δεδομένης της Υπόθεσης 4.5, μπορούμε να φράξουμε τη διαφορά τους με το τετράγωνο της Ευκλείδιας νόρμας της κλίσης της αντικειμενικής συνάρτησης στο σημείο w, δηλαδή:

$$2c(F(w) - \bar{F}) \leq \|\nabla F(w)\|_2^2 \text{ για κάθε } w \in \mathbb{R}^d. \tag{4.16}$$

Απόδειξη. (Bottou, Curtis & Nocedal, 2018, σ. 303) Δοθέντος ενός διανύσματος $w \in \mathbb{R}^d$, το τετραγωνικό μοντέλο που ορίζεται από τη συνάρτηση $q : \mathbb{R}^d \to \mathbb{R}$ με

$$q(v) \coloneqq F(w) + \nabla F(w)^T(v - w) + \frac{1}{2}c\|v - w\|_2^2,$$

υπακούει στη σχέση (4.15) για την ισότητα, άρα και στην Υπόθεση 4.5. Επομένως, η q εμφανίζει μοναδικό ελάχιστο, το οποίο υπολογίζεται με τον παρακάτω τρόπο:

$$\frac{dq}{dv} = 0 \Rightarrow \nabla F(w)^T + \frac{1}{2}2c(v - w) = 0,$$

απ΄ όπου προκύπτει ότι το μοναδικό σημείο ελαχίστου είναι το $\bar{v} \coloneqq w - \frac{1}{c}\nabla F(w)$ με

$$\begin{aligned} q(\bar{v}) &= F(w) + \nabla F(w)^T(\bar{v} - w) + \frac{1}{2}c\|\bar{v} - w\|_2^2 \\ &= F(w) + \nabla F(w)^T\left(w - \frac{1}{c}\nabla F(w) - w\right) + \frac{1}{2}c\left\|w - \frac{1}{c}\nabla F(w) - w\right\|_2^2 \\ &= F(w) - \frac{1}{c}\|\nabla F(w)\|_2^2 + \frac{1}{2c}\|\nabla F(w)\|_2^2, \end{aligned}$$

δηλαδή $q(\bar{v}) = F(w) - \frac{1}{2c}\|\nabla F(w)\|_2^2$. Άρα, η ανισότητα της ισχυρής κυρτότητα (4.15) με $v = \bar{w}$ και οποιοδήποτε $w \in \mathbb{R}^d$ δίνουν

$$q(\bar{w}) = F(\bar{w}) = \bar{F} \geq F(w) - \frac{1}{2c}\|\nabla F(w)\|_2^2,$$

δηλαδή,

$$\begin{aligned} \bar{F} - F(w) \geq -\frac{1}{2c}\|\nabla F(w)\|_2^2 \Rightarrow \\ 2c(F(w) - \bar{F}) \leq \|\nabla F(w)\|_2^2, \end{aligned}$$

που είναι η ζητούμενη ανισότητα. □

Θα πρέπει να προστεθεί σε αυτό το σημείο ότι τα αποτελέσματα που παρουσιάζονται σε αυτό το κεφάλαιο αφορούν την αναμενόμενη συμπεριφορά της SG ως προς τη σύγκλιση. Επομένως θα συναντήσουμε συχνά τις αναμενόμενες τιμές της αντικειμενικής συνάρτησης που μελετάμε. Στο παρακάτω θεώρημα (Bottou, Curtis & Nocedal, 2018, σ. 247-248) βλέπουμε ότι όταν το βήμα δεν είναι πολύ μεγάλο, τότε αναμένουμε ότι η ακολουθία $\{F(w_k)\}$ θα συγκλίνει κοντά στο σημείο ελαχίστου. Με $\mathbb{E}[\cdot]$ θα συμβολίζουμε από εδώ και στο εξής την ολική αναμενόμενη τιμή ως προς όλες τις τυχαίες μεταβλητές ξ_k. Δηλαδή, εάν θεωρήσουμε ότι το διάνυσμα w_k εξαρτάται

από τις ανεξάρτητες τυχαίες μεταβλητές $\{\xi_1, \xi_2, \ldots, \xi_{k-1}\}$, τότε η ολική αναμενόμενη τιμή της $F(w_k)$ για κάθε $k \in \mathbb{N}$ είναι

$$\mathbb{E}[F(w_k)] = \mathbb{E}_{\xi_1}[\mathbb{E}_{\xi_2}[\,\ldots[\mathbb{E}_{\xi_{k-1}}[F(w_k)]\;\ldots]].$$

Το αποτέλεσμα αυτό προκύπτει ως γενίκευση της σχέσης (3.2) για $k-1$ ανεξάρτητες τυχαίες μεταβλητές.

Στην ενότητα αυτή, λοιπόν, εξετάζουμε δύο θεωρήματα για τη σύγκλιση της μεθόδου SG, το Θεώρημα 4.6 και το Θεώρημα 4.7. Τα θεωρήματα αυτά αντιστοιχούν στις περιπτώσεις που η SG τρέχει με σταθερό και με μη σταθερό μέγεθος βήματος. Ξεκινάμε με την πρώτη, δηλαδή όταν χρησιμοποιούμε σταθερό βήμα για την εκτέλεση τη μεθόδου. Τα αποτελέσματα του θεωρήματος ισχύουν μόνο για μια περιοχή του σημείου ελαχίστου. Παρατηρούμε ότι αυτό φαίνεται και στη σχέση (4.14b) όπου, παρόλο που όσο η κλίση $\nabla F(w_k)$ τείνει στο μηδέν πλησιάζουμε στη λύση, ο δεύτερος όρος του δεξιού μέλους της ανισότητας παραμένει σταθερός. Επομένως, από κάποιο σημείο και έπειτα δεν περιμένουμε περαιτέρω ελάττωση της τιμής της αντικειμενικής συνάρτησης.

Θεώρημα 4.6 (ισχυρά κυρτή αντικειμενική συνάρτηση, σταθερά βήματα)**.** *Από τις Υποθέσεις 4.1, 4.3 και 4.5 (με $F_{\inf} = \bar{F}$), έστω ότι η μέθοδος SG του Αλγορίθμου 4.1 τρέχει με σταθερό βήμα, $\alpha_k = \bar{\alpha}$ για κάθε $k \in \mathbb{N}$, ικανοποιώντας την*

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}. \tag{4.17}$$

Τότε η αναμενόμενη τιμή της διαφοράς της τρέχουσας και της ελάχιστης τιμής ικανοποιεί την παρακάτω ανισότητα για κάθε $k \in \mathbb{N}$:

$$\begin{aligned}\mathbb{E}[F(w_k) - \bar{F}] &\leq \frac{\bar{\alpha}LM}{2c\mu} + (1 - \bar{\alpha}c\mu)^{k-1}\left(F(w_1) - \bar{F} - \frac{\bar{\alpha}LM}{2c\mu}\right) \\ &\xrightarrow{k\to\infty} \frac{\bar{\alpha}LM}{2c\mu}.\end{aligned} \tag{4.18}$$

Απόδειξη. Θα αποδείξουμε το θεώρημα επαγωγικά. Εφόσον οι τιμές $F(w_1)$ και $\bar{F}$ δεν εξαρτώνται από τις τυχαίες μεταβλητές ξ_k για κανένα $k \in \mathbb{N}$, έχουμε ότι $\mathbb{E}[F(w_1)] = F(w_1), \mathbb{E}[\bar{F}] = \bar{F}$ και $\mathbb{E}[F(w_1) - \bar{F}] = F(w_1) - \bar{F}$. Άρα, εύκολα βλέπουμε ότι η σχέση (4.18) ισχύει για $k = 1$ στην ισότητά της. Για την επαγωγική υπόθεση, θεωρούμε ότι η ανισότητα ισχύει για $k \in \mathbb{N}$. Αποδεικνούμε, τώρα, τη σχέση για $k + 1 \in \mathbb{N}$.

Εφόσον το w_k δεν εξαρτάται από το ξ_k, το αριστερό μέλος της σχέσης (4.14a) του Λήμματος 4.4 γράφεται $\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) = \mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)]$. Επομένως η σχέση (4.14a) γίνεται

$$\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)] \leq -(\mu - \frac{1}{2}\bar{\alpha}LM_G)\bar{\alpha}\,\|\nabla F(w_k)\|_2^2 + \frac{1}{2}\bar{\alpha}^2 LM.$$

Στην παραπάνω σχέση, για σταθερό βήμα $\alpha_k = \bar{\alpha}$, η ανισότητα (4.17) δίνει

$$\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)] \leq -(\mu - \frac{1}{2}\bar{\alpha}LM_G)\bar{\alpha}\,\|\nabla F(w_k)\|_2^2 + \frac{1}{2}\bar{\alpha}^2 LM$$

$$\leq -\frac{1}{2}\bar{\alpha}\mu \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\bar{\alpha}^2 LM.$$

Όμως, για την ανισότητα (4.16) ισχύει ότι

$$2c(F(w) - \bar{F}) \leq \left\|\nabla F(w)\right\|_2^2 \Rightarrow -2c(F(w) - \bar{F}) \geq -\left\|\nabla F(w)\right\|_2^2.$$

Συνεπώς, για κάθε $k \in \mathbb{N}$ έχουμε ότι

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)] &\leq -\frac{1}{2}\bar{\alpha}\mu \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\bar{\alpha}^2 LM \\ &\leq -\bar{\alpha}c\mu(F(w_k) - \bar{F}) + \frac{1}{2}\bar{\alpha}^2 LM.\end{aligned}$$

Αφαιρούμε την ελάχιστη τιμή $\bar{F}$ από τα δύο μέλη της παραπάνω σχέσης:

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)] - \bar{F} &= \mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) - \bar{F} \\ &\leq -\bar{\alpha}c\mu(F(w_k) - \bar{F}) - \bar{F} + \frac{1}{2}\bar{\alpha}^2 LM,\end{aligned}$$

δηλαδή

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(w_{k+1})] - \bar{F} &\leq F(w_k) - \bar{\alpha}c\mu F(w_k) - \bar{F} - \bar{\alpha}c\mu\bar{F} + \frac{1}{2}\bar{\alpha}^2 LM \\ &\leq (1 - \bar{\alpha}c\mu)(F(w_k) - \bar{F}) + \frac{1}{2}\bar{\alpha}^2 LM.\end{aligned}$$

Παίρνουμε, τώρα, την ολική αναμενόμενη τιμή στα δύο μέλη της προηγούμενης ανισότητας:

$$\mathbb{E}[F(w_{k+1}) - \bar{F}] \leq (1 - \bar{\alpha}c\mu)\mathbb{E}[(F(w_k) - \bar{F})] + \frac{1}{2}\bar{\alpha}^2 LM.$$

Αφαιρώντας τη σταθερά $\bar{\alpha}LM/(2c\mu)$ και από τα δύο μέλη, παίρνουμε

$$\begin{aligned}\mathbb{E}[F(w_{k+1}) - \bar{F}] - \frac{\bar{\alpha}LM}{2c\mu} &\leq (1 - \bar{\alpha}c\mu)\mathbb{E}[F(w_k) - \bar{F}] + \frac{\bar{\alpha}^2 LM}{2} - \frac{\bar{\alpha}LM}{2c\mu} \\ &= (1 - \bar{\alpha}c\mu)\left(\mathbb{E}[F(w_k) - \bar{F}] - \frac{\bar{\alpha}LM}{2c\mu}\right). \qquad (4.19)\end{aligned}$$

Εφαρμόζοντας επαναληπτικά την παραπάνω ανισότητα παίρνουμε

$$\begin{aligned}\mathbb{E}[F(w_{k+1}) - \bar{F}] - \frac{\bar{\alpha}LM}{2c\mu} &\leq (1 - \bar{\alpha}c\mu)\left(\mathbb{E}[F(w_k) - \bar{F}] - \frac{\bar{\alpha}LM}{2c\mu}\right) \\ &\leq (1 - \bar{\alpha}c\mu)\left((1 - \bar{\alpha}c\mu)\left(\mathbb{E}[F(w_{k-1}) - \bar{F}] - \frac{\bar{\alpha}LM}{2c\mu}\right)\right) \\ &= (1 - \bar{\alpha}c\mu)^2\left(\mathbb{E}[F(w_{k-1}) - \bar{F}] - \frac{\bar{\alpha}LM}{2c\mu}\right) \\ &\vdots\end{aligned}$$

$$\leq (1-\bar{\alpha}c\mu)^k \left(\mathbb{E}[F(w_1)-\bar{F}] - \frac{\bar{\alpha}LM}{2c\mu}\right).$$

Εφόσον οι τιμές $F(w_1)$ και $\bar{F}$ δεν εξαρτώνται από τις τυχαίες μεταβλητές ξ_k για κανένα $k \in \mathbb{N}$, έχουμε ότι $\mathbb{E}[F(w_1)] = F(w_1), \mathbb{E}[\bar{F}] = \bar{F}$ και $\mathbb{E}[F(w_1) - \bar{F}] = F(w_1) - \bar{F}$. Άρα, καταλήγουμε στη ζητούμενη ανισότητα

$$\mathbb{E}[F(w_k)-\bar{F}] \leq \frac{\bar{\alpha}LM}{2c\mu} + (1-\bar{\alpha}c\mu)^{k-1}\left(F(w_1)-\bar{F}-\frac{\bar{\alpha}LM}{2c\mu}\right).$$

Παρατηρούμε ότι η (4.13) δίνει $M_G \geq \mu^2$ ή $1/M_G \leq 1/\mu^2$ και από την (4.17),

$$0 < \bar{\alpha}c\mu \leq \frac{c\mu^2}{LM_G} \leq \frac{c\mu^2}{L\mu^2} = \frac{c}{L}. \tag{4.20}$$

Ο ορισμός της ισχυρής κυρτότητας (4.15) δίνει

$$F(v) \geq F(w) + \nabla F(w)^T(v-w) + \frac{1}{2}c\|v-w\|_2^2$$

ενώ από τη σχέση (4.4) έχουμε ότι

$$F(v) \leq F(w) + \nabla F(w)^T(v-w) + \frac{1}{2}L\|v-w\|_2^2,$$

επομένως για τις σταθερές c και L ισχύει ότι $c \leq L$. Συνεπώς, η ανισότητα (4.20) γίνεται $0 < \bar{\alpha}c\mu \leq \frac{c}{L} \leq 1$, δηλαδή $0 \leq 1-\bar{\alpha}c\mu < 1$, άρα, για $k \to \infty$ έχουμε ότι $(1-\bar{\alpha}c\mu)^{k-1} \to 0$ και τότε η ανισότητα (4.18) γίνεται

$$\begin{aligned}\mathbb{E}[F(w_k)-\bar{F}] &\leq \frac{\bar{\alpha}LM}{2c\mu} + (1-\bar{\alpha}c\mu)^{k-1}\left(F(w_1)-\bar{F}-\frac{\bar{\alpha}LM}{2c\mu}\right)\\ &\xrightarrow{k\to\infty} \frac{\bar{\alpha}LM}{2c\mu}.\end{aligned}$$

□

Εάν η $g(w_k, \xi_k)$ είναι αμερόληπτη εκτίμηση της $\nabla F(w_k)$, τότε από την ισότητα της σχέσης (4.10α) είναι $\mu = 1$, και αν η $g(w_k, \xi_k)$ δεν έχει μεγάλη διακύμανση, μπορούμε να θεωρήσουμε ότι $M_G = 1$, κάτι το οποίο προκύπτει από την (4.13). Σε αυτή την περίπτωση, το βήμα που θεωρήσαμε στην (4.17) περιορίζεται στο $\bar{\alpha} \in (0, 1/L]$, βήμα το οποίο εμφανίζεται συχνά στην ανάλυση των μεθόδων κλίσης και απότομης καθόδου.

Το Θεώρημα 4.6 δείχνει τη σχέση μεταξύ του βήματος και του φράγματος της διακύμανσης των στοχαστικών κατευθύνσεων g. Όταν δεν υπάρχει καθόλου θόρυβος στον υπολογισμό των κλίσεων ή όταν έχουμε $M = 0$ στις σχέσεις (4.11) και (4.13), δηλαδή όταν το φράγμα είναι ανάλογο του τετραγώνου $\|\nabla F(w_k)\|_2^2$, τότε η μέθοδος SG συγκλίνει γραμμικά στη βέλτιστη λύση. Όταν εμφανίζεται θόρυβος στις κλίσεις, τότε η ιδιότητα της γραμμικής σύγκλισης παύει να ισχύει. Μπορούμε ακόμα να χρησιμοποιήσουμε σταθερό βήμα στον υπολογισμό επιτυγχάνοντας γραμμική σύγκλιση

κοντά στη βέλτιστη λύση, όμως ο θόρυβος θα σταματήσει την σύγκλιση έπειτα από κάποιο σημείο.

Μία λύση σε αυτό πρόβλημα είναι να τρέξουμε την SG με κάποιο σταθερό βήμα και όταν δούμε ότι η πρόοδος σταματήσει, επιλέγουμε ένα μικρότερο βήμα και επαναλαμβάνουμε τη διαδικασία. Για παράδειγμα, μπορούμε να μειώσουμε το βήμα στο μισό έτσι ώστε η αναμενόμενη διαφορά μεταξύ της τρέχουσας τιμής της αντικειμενικής συνάρτησης και της βέλτιστης λύσης να τείνει στο μηδέν. Όμως, το αποτέλεσμα αυτό δεν συμβαίνει με τον υποδιπλασιασμό του βήματος σε κάθε επανάληψη. Αποδεικνύεται ότι ο κατάλληλος ρυθμός μείωσης του βήματος για την SG με μειούμενα βήματα είναι *υπογραμμικός*, άρα και η ταχύτητα σύγκλισης της μεθόδου θα είναι υπογραμμική. Βασική προϋπόθεση, όμως, για τη σωστή επιλογή ακολουθίας βημάτων είναι να ικανοποιεί τις συνθήκες

$$\sum_{k=1}^{\infty} \alpha_k = \infty \ \text{ και } \ \sum_{k=1}^{\infty} \alpha_k^2 < \infty, \tag{4.21}$$

δηλαδή η φθίνουσα ακολουθία των μεγεθών των βημάτων $\{a_k\}$ να είναι τύπου $1/k$, όπως φαίνεται στην εργασία των Robbins και Monro (1951, σ. 403 - 404). Οι συγκεκριμένες συνθήκες μας επιτρέπουν πιο ελεύθερη επιλογή ακολουθίας βημάτων και το αποτέλεσμά τους δίνει το παρακάτω θεώρημα σύγκλισης (Bottou, Curtis & Nocedal, 2018, σ. 249-250).

Θεώρημα 4.7 (ισχυρά κυρτή αντικειμενική συνάρτηση, μειούμενα βήματα)**.** *Δεδομένων των Υποθέσεων 4.1, 4.3 και 4.5 (με $F_{\inf} = \bar{F}$), έστω ότι η μέθοδος SG τρέχει με ακολουθία βημάτων τέτοια ώστε, για κάθε $k \in \mathbb{N}$,*

$$\alpha_k = \frac{\beta}{\gamma + k} \ \text{ για κάποια } \ \beta > \frac{1}{c\mu} \ \text{ και } \ \gamma > 0 \ \text{ τέτοια ώστε } \ \alpha_1 \leq \frac{\mu}{LM_G}. \tag{4.22}$$

Τότε, για κάθε $k \in \mathbb{N}$, η αναμενόμενη διαφορά της τρέχουσας τιμής και της ελάχιστης τιμής ικανοποιεί την

$$\mathbb{E}[F(w_k) - \bar{F}] \leq \frac{\nu}{\gamma + k}, \tag{4.23}$$

όπου

$$\nu := \max\left\{ \frac{\beta^2 LM}{2(\beta c\mu - 1)}, (\gamma + 1)(F(w_1) - \bar{F}) \right\}. \tag{4.24}$$

Απόδειξη. Επειδή η ακολουθία βημάτων $\{\alpha_k\}$ είναι φθίνουσα, η ανισότητα (4.22) δίνει για κάθε $k \in \mathbb{N}$ ότι $\alpha_k LM_G \leq \alpha_1 LM_G \leq \mu$. Άρα, η σχέση (4.14b) του Λήμματος 4.4 γίνεται

$$\begin{aligned} \mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) &\leq -(\mu - \frac{1}{2}\alpha_k LM_G)\alpha_k \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha_k^2 LM \\ &\leq -\frac{1}{2}\alpha_k \mu \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha_k^2 LM \end{aligned}$$

και από τη σχέση (4.16), έχουμε ότι για κάθε $k \in \mathbb{N}$,

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq \ -\alpha_k c\mu (F(w_k) - \bar{F}) + \frac{1}{2}\alpha_k^2 LM.$$

Αφαιρώντας την $\bar{F}$ και από τα δύο μέλη έχουμε

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) - \bar{F} \leq & -\alpha_k c\mu(F(w_k) - \bar{F}) + \frac{1}{2}\alpha_k^2 LM - \bar{F} \quad \text{ή} \\ \mathbb{E}_{\xi_k}[F(w_{k+1})] - \bar{F} \leq & \; F(w_k) - \alpha_k c\mu F(w_k) - \alpha_k c\mu\bar{F} - \bar{F} + \frac{1}{2}\alpha_k^2 LM \\ & = (1 - \alpha_k c\mu)(F(w_k) - \bar{F}) + \frac{1}{2}\alpha_k^2 LM\end{aligned}$$

και παίρνοντας τις ολικές αναμενόμενες τιμές, καταλήγουμε στη σχέση

$$\mathbb{E}[F(w_{k+1}) - \bar{F}] \leq (1 - \alpha_k c\mu)\mathbb{E}[F(w_k) - \bar{F}] + \frac{1}{2}\alpha_k^2 LM. \tag{4.25}$$

Αποδεικνύουμε τώρα την (4.23) με επαγωγή. Πρώτα, ο ορισμός του ν (4.24) όπου $\nu = (\gamma + 1)(F(w_1) - \bar{F})$ εξασφαλίζει ότι για $k = 1$ είναι

$$\mathbb{E}[F(w_1) - \bar{F}] = F(w_1) - \bar{F} = \frac{(\gamma + 1)(F(w_1) - \bar{F})}{\gamma + 1} = \frac{\nu}{\gamma + 1}.$$

Άρα η συνθήκη ισχύει (στην ισότητά της) για $k = 1$. Έπειτα, υποθέτοντας ότι η (4.23) ισχύει για κάποιο $k > 1$, έπεται από την (4.22) $\left(\alpha_k = \frac{\beta}{\gamma + k}\right)$ και την (4.25) ότι

$$\begin{aligned}\mathbb{E}[F(w_{k+1}) - \bar{F}] &\leq (1 - \alpha_k c\mu)\mathbb{E}[F(w_k) - \bar{F}] + \frac{1}{2}\alpha_k^2 LM \\ &= \left(1 - \frac{\beta c\mu}{\hat{k}}\right)\mathbb{E}[F(w_k) - \bar{F}] + \frac{\beta^2 LM}{2\hat{k}^2} \quad \text{με } \hat{k} \coloneqq \gamma + k\end{aligned}$$

και από την επαγωγική υπόθεση (4.23) παίρνουμε

$$\begin{aligned}\mathbb{E}[F(w_{k+1}) - \bar{F}] &\leq \left(1 - \frac{\beta c\mu}{\hat{k}}\right)\frac{\nu}{\hat{k}} + \frac{\beta^2 LM}{2\hat{k}^2} \\ &= \left(\frac{\hat{k} - \beta c\mu}{\hat{k}^2}\right)\nu + \frac{\beta^2 LM}{2\hat{k}^2} \\ &= \left(\frac{\hat{k} - 1 + 1 - \beta c\mu}{\hat{k}^2}\right)\nu + \frac{\beta^2 LM}{2\hat{k}^2} \\ &= \left(\frac{\hat{k} - 1}{\hat{k}^2}\right)\nu - \left(\frac{\beta c\mu - 1}{\hat{k}^2}\right)\nu + \frac{\beta^2 LM}{2\hat{k}^2}.\end{aligned}$$

Από τον ορισμό του ν έπεται ότι, αν $\nu = \frac{\beta^2 LM}{2(\beta c\mu - 1)}$, τότε ο τρίτος όρος του δεξιού μέλους της παραπάνω ανισότητας γίνεται

$$\left(\frac{\beta c\mu - 1}{\hat{k}^2}\right)\nu = \left(\frac{\beta c\mu - 1}{\hat{k}^2}\right)\frac{\beta^2 LM}{2(\beta c\mu - 1)} = \frac{\beta^2 LM}{2\hat{k}^2}$$

δηλαδή $-\left(\frac{\beta c\mu - 1}{\hat{k}^2}\right)\nu + \frac{\beta^2 LM}{2\hat{k}^2} = 0$. Στην περίπτωση που $\nu = (\gamma + 1)(F(w_1) - \bar{F})$, τότε

$$\nu > \frac{\beta^2 LM}{2(\beta c\mu - 1)} \Rightarrow \left(\frac{\beta c\mu - 1}{\hat{k}^2}\right)\nu > \frac{\beta^2 LM}{2\hat{k}^2}$$

Επομένως, σε κάθε περίπτωση είναι $-\left(\frac{\beta c\mu-1}{\hat{k}^2}\right)\nu+\frac{\beta^2 LM}{2\hat{k}^2}\leq 0$. Άρα, έχουμε

$$\left(\frac{\hat{k}-1}{\hat{k}^2}\right)\nu-\left(\frac{\beta c\mu-1}{\hat{k}^2}\right)\nu+\frac{\beta^2 LM}{2\hat{k}^2}\leq\left(\frac{\hat{k}-1}{\hat{k}^2}\right)\nu\leq\frac{\nu}{\hat{k}+1}$$

όπου η τελευταία ανισότητα προκύπτει επειδή $\hat{k}^2\geq\hat{k}^2-1=(\hat{k}+1)(\hat{k}-1)$. Άρα, αποδείξαμε τη σχέση για $k+1$:

$$\mathbb{E}[F(w_{k+1})-\bar{F}]\leq\frac{\nu}{\hat{k}+1}=\frac{\nu}{\gamma+k+1}$$

και από τη μαθηματική επαγωγή καταλήξαμε στη ζητούμενη ανισότητα. □

Τα συμπεράσματα που μπορούμε να βγάλουμε από τα Θεωρήματα 4.6 και 4.7 προκύπτουν από την ισχυρή κυρτότητα της αντικειμενικής συνάρτησης και την επιλογή του αρχικού σημείου. Παρατηρούμε από τις σχέσεις (4.19) και (4.25) ότι η παράμετρος $c>0$ της ισχυρής κυρτότητας πρέπει να είναι θετική έτσι ώστε η διαφορά μεταξύ της τρέχουσας τιμής της F και της ελάχιστης τιμής να μειώνεται. Ωστόσο, η παράμετρος c επηρεάζει την επιλογή του βήματος διαφορετικά για τα δύο θεωρήματα. Στην περίπτωση του σταθερού βήματος, οι πιθανές τιμές του $\bar{\alpha}$ φράσσονται από τη σχέση (4.17), η οποία δεν εξαρτάται από το c. Αντίθετα, για την περίπτωση που έχουμε μια φθίνουσα ακολουθία βημάτων, το αρχικό βήμα α_1 φράσσεται από την ίδια ποσότητα, ενώ η παράμετρος βήματος β πρέπει να είναι μεγαλύτερη από την ποσότητα $1/(c\mu)$, όπως φαίνεται και στη σχέση (4.22).

Επιπλέον, σημαντική είναι η επιλογή του αρχικού διανύσματος w_1 για τη διαφορά $\mathbb{E}[F(w_1)-\bar{F}]$. Στην περίπτωση που το μέγεθος βήματος είναι σταθερό, η αρχική διαφορά μειώνεται εκθετικά, όπως φαίνεται στο Θεώρημα 4.6. Αντιθέτως, όταν έχουμε μειούμενα βήματα, η διαφορά αυτή εμφανίζεται μόνο στη δεύτερη περίπτωση της ποσότητας ν. Όμως, με κατάλληλη αρχικοποίηση μπορούμε να εξασθενήσουμε τον ρόλο που παίζει ο όρος αυτός. Έστω, για παράδειγμα, ότι υπολογίζουμε προσεγγιστικά το αρχικό σημείο w_1 τρέχοντας τη μέθοδο SG με σταθερό βήμα $\bar{\alpha}$ μέχρι το σημείο αυτό. Τότε, το Θεώρημα 4.6 μας εξασφαλίζει ότι για το σημείο αυτό θα ισχύει $\mathbb{E}[F(w_1)-\bar{F}]\leq\bar{\alpha}LM/(2c\mu)$. Θέτοντας, λοιπόν, $\alpha_1=\bar{\alpha}$ και τρέχοντας την SG με μειούμενο βήμα, το Θεώρημα 4.7 μας δίνει $\gamma+1=\beta\alpha_1^{-1}$, άρα

$$\begin{aligned}(\gamma+1)\mathbb{E}[F(w_1-\bar{F})]&\leq(\gamma+1)\frac{\alpha_1 LM}{2c\mu}=\beta\alpha_1^{-1}\frac{\alpha_1 LM}{2c\mu}\\&=\frac{\beta LM}{2c\mu}=\frac{\beta^2 LM}{2\beta c\mu}<\frac{\beta^2 LM}{2(\beta c\mu-1)}.\end{aligned}$$

Βλέπουμε, δηλαδή, ότι μεγαλύτερη σημασία έχει η πρώτη περίπτωση στον ορισμό του ν στο φράγμα της αναμενόμενης απόστασης της τιμής $F(w_1)$ και της ελάχιστης $\bar{F}$.

Με βάση τα παραπάνω θεωρήματα, μπορούμε να κάνουμε μια σύγκριση μεταξύ των απλών και των minibatch στοχαστικών μεθόδων κλίσης, (3.11) και (3.13), όταν αυτές ελαχιστοποιούν το εμπειρικό ρίσκο, δηλαδή όταν $F=R_n$. Στην πιο βασική μορφή της, η απλή μέθοδος SG επιλέγει ένα τυχαίο δειγμα με δείκτη i_k σε κάθε επανάληψη και υπολογίζει την κατεύθυνση

$$g(w_k,\xi_k)=\nabla f_{i_k}(w_k). \tag{4.26}$$

Αντιθέτως, η minibatch SG δεν επιλέγει ένα μοναδικό δείγμα, αλλά ένα υποσύνολο δεικτών $\mathcal{S}_k$ των παραδειγμάτων και υπολογίζει την

$$g(w_k, \xi_k) = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \nabla f_i(w_k). \tag{4.27}$$

Χάριν απλότητας, έστω ότι σε κάθε επανάληψη επιλέγουμε τον ίδιο αριθμό παραδειγμάτων ώστε τα minibatches να έχουν σταθερό μέγεθος $|\mathcal{S}_k| = n_{\mathrm{mb}}$. Υποθέτουμε τότε ότι το μέγεθος του minibatch είναι $n_{\mathrm{mb}} \ll n$. Ο υπολογισμός της στοχαστικής κατεύθυνσης $g(w_k, \xi_k)$ σε κάθε επανάληψη από την (4.27) είναι προφανώς n_{mb} φορές πιο δύσκολος απ΄ ότι στην απλή μέθοδο SG (4.26). Όμως, όπως φαίνεται και στο βιβλίο του Freund (Miller & Miller, 2014, σ. 235, Θεώρημα 1), η διακύμανση της κατεύθυνσης στη minibatch SG ελαττώνεται σύμφωνα με το συντελεστή $1/n_{\mathrm{mb}}$. Επομένως, οι σταθερές M και M_V που εμφανίζονται στην Υπόθεση 4.3 θα ελαττώνονται και αυτές κατά τον ίδιο συντελεστή, άρα θα γίνουν M/n_{mb} και M_V/n_{mb} για τη μέθοδο minibatch SG. Είναι φυσικό να αναρωτηθεί κανείς εάν αυτή η μείωση στη διακύμανση αποζημειώνει για το υψηλότερο υπολογιστικό κόστος ανά επανάληψη.

Για να απαντήσουμε σε αυτό το ερώτημα, θεωρούμε την περίπτωση ενός αρκετά μικρού σταθερού βήματος $\bar{\alpha} > 0$. Για τη minibatch SG, το Θεώρημα 4.6 οδηγεί στη σχέση

$$\mathbb{E}[F(w_k) - \bar{F}] \leq \frac{\bar{\alpha} L M}{2c\mu n_{\mathrm{mb}}} + [1 - \bar{\alpha} c \mu]^{k-1} \left(F(w_1) - \bar{F} - \frac{\bar{\alpha} L M}{2c\mu n_{\mathrm{mb}}} \right) .$$

Η απλή μέθοδος SG με βήμα $\bar{\alpha}/n_{\mathrm{mb}}$ δίνει παρόμοια ασυμπτωτική διαφορά:

$$\mathbb{E}[F(w_k) - \bar{F}] \leq \frac{\bar{\alpha} L M}{2c\mu n_{\mathrm{mb}}} + [1 - \frac{\bar{\alpha} c \mu}{n_{\mathrm{mb}}}]^{k-1} \left(F(w_1) - \bar{F} - \frac{\bar{\alpha} L M}{2c\mu n_{\mathrm{mb}}} \right) .$$

Η σταθερά $1 - \bar{\alpha} c \mu$ δείχνει ότι πρέπει να τρέξουμε n_{mb} φορές περισσότερες επαναλήψεις του αλγορίθμου της απλής SG ώστε να καταλήξουμε στην ίδια αναμενόμενη διαφορά $F(w_k) - \bar{F}$. Παρ΄ όλα αυτά, από τη στιγμή που ο υπολογισμός σε μια επανάληψη της απλής SG είναι n_{mb} φορές γρηγορότερος, αντιστοιχεί αθροιστικά στο ίδιο συνολικό υπολογιστικό κόστος με τη μέθοδο minibatch SG. Μία παρόμοια ανάλυση με εφαρμογή του αποτελέσματος του Θεωρήματος 4.7 μπορεί να γίνει με την χρήση μειούμενων βημάτων.

Από την παραπάνω ανάλυση, καταλήγουμε ότι με την κατάλληλη επιλογή αρχικού βήματος, οι δύο μέθοδοι αποδίδουν ισάξια. Ωστόσο, προϋπόθεση των δύο θεωρημάτων για το αρχικό βήμα είναι η σχέση $\alpha_1 \leq \mu/(LM_G)$ και από την ανισότητα (4.13) έχουμε ότι $M_G \leq \mu^2$. Άρα

$$\alpha_1 \leq \frac{\mu}{LM_G} = \frac{\mu^2}{M_G} \frac{1}{L\mu} \leq \frac{1}{L\mu}.$$

Δηλαδή, δεν μπορούμε απλά να επιλέξουμε αρχικό βήμα n_{mb} φορές μεγαλύτερο για τη minibatch SG σε κάθε περίπτωση για να αντισταθμίσουμε το μεγαλύτερο κόστος ανά επανάληψη. Μπορούμε, όμως, να επωφεληθούμε από τα πλεονεκτήματα της χρήσης των minibatches στην πράξη, καθώς αυτά διευκολύνουν τον παράλληλο υπολογισμό. Για παράδειγμα, η χρήση μεγάλων minibatches είναι συχνά ο μόνος τρόπος πλήρους αξιοποίησης ενός επεξεργαστή GPU. Επιπλέον, τα minibatches με μεγέθη που αλλάζουν δυναμικά μπορούν να χρησιμοποιηθούν αντί των μειούμενων βημάτων.

4.3 Μέθοδος SG για Μη Κυρτές Συναρτήσεις

Όπως αναφέρθηκε και παραπάνω, πολλά μοντέλα της μηχανικής μάθησης καταλήγουν σε προβλήματα βελτιστοποίησης με όχι απαραίτητα κυρτές αντικειμενικές συναρτήσεις. Η ανάλυση σύγκλισης για την SG όταν αυτή τρέχει για μια μη κυρτή αντικειμενική συνάρτηση, μπορεί να αποδειχθεί δύσκολη, λόγω των ιδιομορφιών που παρουσιάζουν τέτοιες συναρτήσεις (πολλά τοπικά ελάχιστα και άλλα στάσιμα σημεία). Όπως και στην προηγούμενη ενότητα, παρουσιάζονται δύο θεωρήματα, ένα για την περίπτωση του σταθερού βήματος και ένα για μειούμενα βήματα. Η σχέση (4.2), οι Υποθέσεις 4.1 και 4.3, καθώς και τα αντίστοιχα λήμματα συνεχίζουν να ισχύουν για το στοχαστικό διάνυσμα $g(w_k, \xi_k)$, χωρίς όμως να λαμβάνουμε υπ΄ όψιν την υπόθεση της ισχυρής κυρτότητας.

Θεώρημα 4.8 (μη κυρτή αντικειμενική συνάρτηση, σταθερό βήμα)**.** *Δεδομένων των Υποθέσεων 4.1 και 4.3, έστω ότι η μέθοδος SG τρέχει με σταθερό βήμα $\alpha_k = \bar{\alpha}$ για κάθε $k \in \mathbb{N}$, ικανοποιώντας την ανισότητα*

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}. \tag{4.28}$$

Τότε η κλίση της συνάρτησης F σε κάθε επανάληψη της μεθόδου SG ικανοποιεί τις ακόλουθες ανισότητες για κάθε $K \in \mathbb{N}$:

$$\mathbb{E}\left[\sum_{k=1}^{K} \|\nabla F(w_k)\|_2^2\right] \leq \frac{K\bar{\alpha}LM}{\mu} + \frac{2(F(w_1) - F_{\text{inf}})}{\mu\bar{\alpha}} \tag{4.29a}$$

$$\text{και επομένως } \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(w_k)\|_2^2\right] \leq \frac{\bar{\alpha}LM}{\mu} + \frac{2(F(w_1) - F_{\text{inf}})}{K\mu\bar{\alpha}} \tag{4.29b}$$

$$\xrightarrow{K\to\infty} \frac{\bar{\alpha}LM}{\mu}.$$

Απόδειξη. Από τη σχέση (4.14b) και το φράγμα στο μέγεθος του βήματος (4.28) έχουμε

$$\begin{aligned}
\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)] &\leq -(\mu - \frac{1}{2}\bar{\alpha}LM_G)\bar{\alpha}\|\nabla F(w_k)\|_2^2 + \frac{1}{2}\bar{\alpha}^2 LM \\
&\leq -(\mu - \frac{1}{2}\frac{\mu}{LM_G}LM_G)\|\nabla F(w_k)\|_2^2 + \frac{1}{2}\bar{\alpha}^2 LM \\
&= -\frac{1}{2}\mu\bar{\alpha}\|\nabla F(w_k)\|_2^2 + \frac{1}{2}\bar{\alpha}^2 LM.
\end{aligned}$$

Η ολική αναμενόμενη τιμή και στα δύο μέλη και η ιδιότητα $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ δίνουν

$$\mathbb{E}[F(w_{k+1})] - \mathbb{E}[F(w_k)] \leq -\frac{1}{2}\mu\bar{\alpha}\mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}\bar{\alpha}^2 LM.$$

Το άθροισμα των δύο μελών της ανισότητας για $k \in \{1, \ \ldots, K\}$ είναι

$$\sum_{k=1}^{K}(\mathbb{E}[F(w_{k+1})] - \mathbb{E}[F(w_k)]) \leq \sum_{k=1}^{K}\left(-\frac{1}{2}\mu\bar{\alpha}\mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}\bar{\alpha}^2 LM)\right).$$

Όμως, έχουμε ότι

$$\begin{aligned}\sum_{k=1}^{K}\left(\mathbb{E}[F(w_{k+1})]-\mathbb{E}[F(w_k)]\right) &= \mathbb{E}[F(w_2)]-\mathbb{E}[F(w_1)]+\mathbb{E}[F(w_3)]-\mathbb{E}[F(w_2)]+\dots\\ &\quad+\mathbb{E}[F(w_K)]-\mathbb{E}[F(w_{K-1})]+\mathbb{E}[F(w_{K+1})]-\mathbb{E}[F(w_K)]\\ &= \mathbb{E}[F(w_{K+1})]-\mathbb{E}[F(w_1)].\end{aligned}$$

Η πρώτη συνθήκη της Υπόθεσης 4.3 μας λέει ότι στο πεδίο ορισμού της η F είναι κάτω φραγμένη από την ποσότητα $F_{\inf}$, άρα ισχύει η σχέση $F_{\inf}\leq F(w_{K+1})\Rightarrow F_{\inf}\leq\mathbb{E}[F(w_{K+1})]\Rightarrow F_{\inf}-F(w_1)\leq\mathbb{E}[F(w_{K+1})]-F(w)_1)$. Επομένως, έχουμε

$$\begin{aligned}F_{\inf}-F(w_1)\leq\mathbb{E}[F(w_{K+1})]-F(w_1) &\leq \sum_{k=1}^{K}\left(-\frac{1}{2}\mu\bar{\alpha}\mathbb{E}[\|\nabla F(w_k)\|_2^2]+\frac{1}{2}\bar{\alpha}^2LM\right)\\ &= -\frac{1}{2}\mu\bar{\alpha}\sum_{k=1}^{K}\mathbb{E}[\|\nabla F(w_k)\|_2^2]+\frac{1}{2}K\bar{\alpha}^2LM\\ &= -\frac{1}{2}\mu\bar{\alpha}\mathbb{E}\left[\sum_{k=1}^{K}\|\nabla F(w_k)\|_2^2\right]+\frac{1}{2}K\bar{\alpha}^2LM\Rightarrow\\ F(w_1)-F_{\inf}\geq\ &\frac{1}{2}\mu\bar{\alpha}\mathbb{E}\left[\sum_{k=1}^{K}\|\nabla F(w_k)\|_2^2\right]+\frac{1}{2}K\bar{\alpha}^2LM\Rightarrow\\ \mathbb{E}\left[\sum_{k=1}^{K}\|\nabla F(w_k)\|_2^2\right]\leq\ &\frac{K\alpha LM}{\mu}+\frac{2(F(w_1)-F_{\inf})}{\mu\alpha},\end{aligned}$$

δηλαδή καταλήξαμε στην (4.29a) και διαιρώντας με K προκύπτει η (4.29b). □

Όταν $M=0$, τότε δεν υπάρχει θόρυβος ή ο θόρυβος μειώνεται αναλογικά με το $\|\nabla F(w_k)\|_2^2$. Τότε η σχέση (4.29a) εξασφαλίζει ότι το άθροισμα των τετραγώνων των κλίσεων παραμένει πεπερασμένο με $\{\|\nabla F(w_k)\|_2\}\to 0$, το οποίο είναι χαρακτηριστικό αποτέλεσμα της batch μεθόδου κλίσης για μη κυρτές συναρτήσεις. Ωστόσο, όταν υπάρχει θόρυβος ($M>0$), σε αντίθεση την περίπτωση της κυρτής συνάρτησης, δεν μπορούμε να φράξουμε την αναμενόμενη διαφορά ανάμεσα στην τρέχουσα τιμή $F(w_k)$ και στην ελάχιστη τιμή. Συνεπώς, το Θεώρημα 4.8 και η σχέση (4.29b) εξασφαλίζει ότι θα υπάρχει ένα φράγμα για τη μέση νόρμα των κλίσεων των διανυσμάτων $\{w_k\}_{k=1}^K$. Το γεγονός, όμως, ότι το φράγμα αυτό μικραίνει ποσοτικά όσο αυξάνεται το K, μπορεί να αυξήσει τον χρόνο που χρειάζεται η μέθοδος SG για να συγκλίνει. Επίσης, όπως και στην κυρτή περίπτωση, η πρόοδος της μεθόδου ως προς τη σύγκλιση σταματάει ύστερα από κάποιον αριθμό επαναλήψεων.

Συνεπώς, προχωράμε και πάλι στην περίπτωση που η μέθοδος SG τρέχει με μία φθίνουσα ακολουθία βημάτων. Υπενθυμίζουμε ότι η ακολουθία αυτή θα πρέπει να ικανοποιεί τις συνθήκες (4.21). Οι ιδιότητες σύγκλισης σε αυτή την περίπτωση φαίνονται στο παρακάτω θεώρημα (Bottou, Curtis & Nocedal, 2018, σ. 254).

Θεώρημα 4.9 (μη κυρτή αντικειμενική συνάρτηση, μειούμενα βήματα)**.** *Δεδομένων των Υποθέσεων 4.1 και 4.3, έστω ότι η μέθοδος SG τρέχει με μία ακολουθία βημάτων*

που ικανοποιεί τις συνθήκες (4.21). Αν $A_K \coloneqq \sum_{k=1}^{K} \alpha_k$, *τότε*

$$\lim_{K\to\infty} \mathbb{E}\left[\sum_{k=1}^{K} \alpha_k \|\nabla F(w_k)\|_2^2\right] < \infty \tag{4.30a}$$

$$\textit{και επομένως } \mathbb{E}\left[\frac{1}{A_K}\sum_{k=1}^{K} \alpha_k \|\nabla F(w_k)\|_2^2\right] \xrightarrow{K\to\infty} 0. \tag{4.30b}$$

Απόδειξη. Η δεύτερη συνθήκη της (4.21), $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, εξασφαλίζει ότι $\alpha_k \to 0$, δηλαδή, χωρίς βλάβη της γενικότητας, μπορούμε να υποθέσουμε ότι με τη βοηθεια της συνθήκης (4.22) ότι $\alpha_k \leq \alpha_1 \leq \frac{\mu}{LM_G}$, δηλαδή, $\alpha_k L M_G \leq \mu$ για κάθε $k \in \mathbb{N}$. Τότε, η ολική αναμενόμενη τιμή της σχέσης (4.14b) του Λήμματος 4.4 είναι

$$\begin{aligned}
\mathbb{E}_{\xi_k}[F(w_{k+1})] - \mathbb{E}_{\xi_k}[F(w_k)] &\leq -\left(\mu - \frac{1}{2}\alpha_k L M_G\right)\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 LM \Rightarrow \\
\mathbb{E}[F(w_{k+1})] - \mathbb{E}[F(w_k)] &\leq -\left(\mu - \frac{1}{2}\alpha_k L M_G\right)\alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}\alpha_k^2 LM \\
&\leq -\frac{1}{2}\mu\alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}\alpha_k^2 LM.
\end{aligned}$$

Παίρνουμε το άθροισμα για $k \in \{1, \ldots, K\}$ και για τα δύο μέλη:

$$\begin{aligned}
\sum_{k=1}^{K} (\mathbb{E}[F(w_{k+1})] - \mathbb{E}[F(w_k)]) &\leq \sum_{k=1}^{K}\left(-\frac{1}{2}\mu\alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}\alpha_k^2 LM\right) \\
&= -\frac{1}{2}\mu \sum_{k=1}^{K}\left(\alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2]\right) + \frac{1}{2}LM\sum_{k=1}^{K}\alpha_k^2.
\end{aligned}$$

Όπως δείξαμε και στο Θεώρημα 4.8, έχουμε ότι

$$F_{\inf} - F(w_1) = F_{\inf} - \mathbb{E}[F(w_1)] \leq \mathbb{E}[F(w_{K+1})] - \mathbb{E}[F(w_1)]$$

Επομένως καταλήγουμε στο ότι

$$\begin{aligned}
F_{\inf} - \mathbb{E}[F(w_1)] &\leq \mathbb{E}[F(w_{K+1})] - \mathbb{E}[F(w_1)] \\
&\leq -\frac{1}{2}\mu \sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}LM\sum_{k=1}^{K}\alpha_k^2,
\end{aligned}$$

δηλαδή,

$$\sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] \leq 2(\mathbb{E}[F(w_1)] - F_{\inf}) + \frac{1}{2}LM\sum_{k=1}^{K}\alpha_k^2$$

Διαιρούμε με $\mu/2$, οπότε παίρνουμε

$$\sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] \leq \frac{2(\mathbb{E}[F(w_1)] - F_{\inf})}{\mu} + \frac{LM}{\mu}\sum_{k=1}^{K}\alpha_k^2. \tag{4.31}$$

Η δεύτερη συνθήκη της (4.21) συνεπάγεται ότι το άθροισμα $\sum_{k=1}^{K} \alpha_k^2$ συγκλίνει σε ένα πεπερασμένο όριο όταν το K αυξάνεται, αποδεικνύοντας έτσι την σχέση (4.30a), εφόσον ισχύει ότι $\sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] = \mathbb{E}\left[\sum_{k=1}^{K} \alpha_k \|\nabla F(w_k)\|_2^2\right]$. Η υπόθεση του θεωρήματος μας δίνει $A_K \coloneqq \sum_{k=1}^{K} \alpha_k$, άρα, διαιρώντας με A_K την ανισότητα (4.31) έχουμε

$$\frac{1}{A_k} \sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] \leq \frac{1}{A_k} \left(\frac{2(\mathbb{E}[F(w_1)] - F_{\text{inf}})}{\mu} + \frac{LM}{\mu} \sum_{k=1}^{K} \alpha_k^2 \right).$$

Από την πρώτη, όμως, συνθήκη της σχέσης (4.21), είναι $\sum_{k=1}^{K} \alpha_k = \infty$, άρα $A_K \to \infty$ και $\frac{1}{A_K} \to 0$ όσο $K \to \infty$, άρα καταλήγουμε στη σχέση (4.30a). □

Το παραπάνω θεώρημα χρησιμοποιεί τα διαφορετικά μεγέθη των βημάτων $\{\alpha_k\}$ ως συντελεστές των τετραγωνισμένων κλίσεων της F σε κάθε επανάληψη. Η σχέση (4.30b) του Θεωρήματος 4.10 δείχνει ότι ακόμα και όταν υπάρχει θόρυβος στα δεδομένα ($M > 0$) η μέση τιμή των τετραγώνων των κλίσεων τείνει στο μηδέν, σε αντίθεση με τη σχέση (4.29b). Παρόλο που η (4.30b) ορίζει μια ιδιότητα σύγκλισης του μέσου των τετραγωνικών κλίσεων με συντελεστές την ακολουθία $\{\alpha_k\}$, μπορούμε να πούμε ότι και οι αναμενόμενες νόρμες των κλίσεων θα πρέπει ασυμπτωτικά να τείνουν στο μηδέν.

Με βάση την ανάλυση των ιδιοτήτων σύγκλισης των παραπάνω ενοτήτων, μπορούμε να δούμε ότι η μέθοδος SG είναι πιο αργή σε σχέση με πολλές από τις μεθόδους batch, καθώς επιτυγχάνει υπογραμμικό ρυθμό σύγκλισης. Επίσης, οι πληροφορίες που χρησιμοποιεί για την αντικειμενική συνάρτηση αφορούν εκτιμήσεις των $F(w_k)$ και $\nabla F(w_k)$ οι οποίες είναι επιρρεπείς στον θόρυβο. Ωστόσο, προτιμάται για την επίλυση προβλημάτων μεγάλης κλίμακας, καθώς η ελαχιστοποίηση του εμπειρικού ρίσκου μέσω της SG δεν εξαρτάται από το μέγεθος n του συνόλου εκπαίδευσης. Οι ασυμπτωτικές ιδιότητες σύγκλισης όπως παρουσιάστηκαν στα προηγούμενα θεωρήματα, μπορούν να εφαρμοστούν στην πράξη όταν ο αλγόριθμος προσαρμόζεται στις τοπικές ιδιότητες κυρτότητας της αντικειμενικής συνάρτησης. Οι σταθεροί συντελεστές που εμφανίζονται στα θεωρήματα (για παράδειγμα οι L, M, c και οι λόγοι M/c και L/c) δίνουν ικανοποιητικές συνθήκες σύγκλισης. Θα δούμε, όμως, στο επόμενο κεφάλαιο ότι τα αποτελέσματα αυτά μπορούν να βελτιωθούν όταν στα διανύσματα των στοχαστικών διευθύνσεων g εφαρμόζεται κάποιος πίνακας μετασχηματισμών.

5 Μέθοδοι Δεύτερης Τάξης

Εκτός από τη μέθοδο SG που μελετήθηκε στο προηγούμενο κεφάλαιο, υπάρχουν μέθοδοι βελτιστοποίησης στη μηχανική μάθηση οι οποίες αντιμετωπίζουν καλύτερα αντικειμενικές συναρτήσεις μη γραμμικές και με κακή κατάσταση. Η έννοια της κατάστασης μιας συνάρτησης, περιγράφει τον ρυθμό με τον οποίο αλλάζει η συνάρτηση όταν εμφανίζονται μικρές διαταραχές στα δεδομένα εισόδου της. Οι συναρτήσεις οι οποίες αλλάζουν γρήγορα με μικρές μεταβολές των δεδομένων μπορούν να δημιουργήσουν πολλά προβλήματα σε επαναληπτικές διαδικασίες όπου μικρά σφάλματα στρογγυλοποιήσης στην είσοδο επιφέρουν μεγάλες αλλαγές στην έξοδο.

Οι μέθοδοι που θα αναλύσουμε στο κεφάλαιο αυτό αντιμετωπίζουν τέτοιου είδους αντικειμενικές συναρτήσεις με τη χρήση πληροφοριών δεύτερης τάξης. Θα δείξουμε ότι αυτές οι μέθοδοι, κάτω από συγκεκριμένες συνθήκες, βελτιώνουν τους ρυθμούς σύγκλισης των στοχαστικών και των batch μεθόδων. Τα πλεονεκτήματα που προσφέρουν οι μέθοδοι δεύτερης τάξης έναντι των πρωτοβάθμιων είναι ότι δεν επηρεάζονται από μετασχηματισμούς. Για παράδειγμα, η batch μέθοδος κλίσης που ελαχιστοποιεί μια συνεχώς παραγωγίσιμη συνάρτηση $F : \mathbb{R}^d \to \mathbb{R}$ έχει επαναληπτικό τύπο που δίνεται από τη σχέση (3.12), δηλαδή

$$w_{k+1} \leftarrow w_k - \alpha_k \nabla F(w_k). \tag{5.1}$$

Θεωρούμε τώρα, ένα παραμετρικό διάνυσμα $\tilde{w} = B^{-1}w$, όπου ο B είναι ένας συμμετρικός και θετικά ορισμένος πίνακας γραμμικών μετασχηματισμών. Ένας εναλλακτικός τύπος για την batch μέθοδο κλίσης δημιουργείται θεωρώντας το πρόβλημα $\min_{\tilde{w}} F(B\tilde{w})$. Τότε έχουμε ότι

$$\tilde{w}_{k+1} \leftarrow \tilde{w}_k - \alpha_k B \nabla F(B\tilde{w}_k).$$

Με βάση τον ορισμό του $\tilde{w}$, θέτουμε $\{w_k\} \coloneqq \{B\tilde{w}_k\}$. Επομένως, ο νέος επαναληπτικός τύπος είναι για το διάνυσμα w είναι

$$\begin{aligned} B^{-1}w_{k+1} &\leftarrow B^{-1}w_k - \alpha_k B \nabla F(w_k) \\ w_{k+1} &\leftarrow w_k - \alpha_k B^2 \nabla F(w_k). \end{aligned} \tag{5.2}$$

Από τη διαφορά στους τύπους (5.1) και (5.2), περιμένουμε ότι ο αλγόριθμος θα έχει διαφορετική συμπεριφορά μετά από μια αλλαγή στις μεταβλητές. Για παράδειγμα, εάν η F είναι μια ισχυρά κυρτή, τετραγωνική συνάρτηση με μοναδικό σημείο ελαχίστου $\bar{w}$, ο επαναληπτικός τύπος (5.1) χρειάζεται πολλές επαναλήψεις για την προσέγγιση του ελαχίστου. Όταν, όμως, ξεκινάμε από ένα αρχικό σημείο w_1, ο επαναληπτικός τύπος (5.2) με $B = (\nabla^2 F(w_1))^{-1/2}$, αρχικό βήμα $\alpha_1 = 1$ και $k = 1$, δίνει

$$w_2 = w_1 - (\nabla^2 F(w_1))^{-1} \nabla F(w_1),$$

το οποίο συμπίπτει με μία επανάληψη της μεθόδου Newton (2.38) και, λόγω της μοναδικότητας του σημείου ελαχίστου, είναι $w_2 = \bar{w}$.

Παρατηρούμε, επίσης, ότι κάθε επανάληψη της μορφής (5.1) ή (5.2) επιλέγει το επόμενο παραμετρικό διάνυσμα υπολογίζοντας πρώτα το σημείο ελαχίστου ενός δευτεροβάθμιου αναπτύγματος Taylor $q_k : \mathbb{R}^d \to \mathbb{R}$ της F στο w_k:

$$q_k(w) = F(w_k + (w - w_k))$$
$$= F(w_k) + \nabla F(w_k)^T(w - w_k) + \frac{1}{2}(w - w_k)^T B^{-2}(w - w_k). \tag{5.3}$$

Για τη batch μέθοδο κλίσης είναι $B^{-2} = I$, ενώ για τη μέθοδο Newton είναι $B^{-2} = \nabla^2 F(w_k)$ (υποθέτοντας ότι ο Εσσιανός $\nabla^2 F(w_k)$ είναι θετικά ορισμένος). Άρα, βλέπουμε ότι μία batch μέθοδος κλίσης τρέχει με ακρίβεια μόνο σε μοντέλα πρώτης τάξης, ενώ μία μέθοδος Newton εφαρμόζει τοπικά μετασχηματισμούς με σκοπό την ελαχιστοποίηση μιας δευτεροβάθμιας προσέγγισης της F σε κάθε επανάληψη.

Επομένως, μπορούμε να δούμε ότι οι ντετερμινιστικές μέθοδοι επωφελούνται από τη χρήση πληροφοριών δεύτερης τάξης, ενώ οι στοχαστικές μέθοδοι παραμένουν υπογραμμικά συγκλίνουσες. Θα εξετάσουμε στη συνέχεια τρόπους με τους οποίους διαδοχικοί μετασχηματισμοί βασισμένοι σε παραγώγους δευτέρας τάξης μπορούν να φανούν χρήσιμοι σε batch αλλά και στοχαστικούς αλγορίθμους. Οι τεχνικές που θα χρησιμοποιηθούν σε αυτό το κεφάλαιο έχουν σκοπό τον συνδυασμό των δευτεροβάθμιων πληροφοριών και της στοχαστικότητας των μεθόδων.

5.1 Ανακριβείς Μέθοδοι Newton Χωρίς Εσσιανό Πίνακα

Είδαμε παραπάνω ότι οι μέθοδοι Newton παραμένουν αμετάβλητες μετά από μετασχηματισμούς και γνωρίζουμε από το δεύτερο κεφάλαιο ότι συγκλίνουν τετραγωνικά κοντά στο σημείο ελαχίστου. Οι ιδιότητες αυτές τις καθιστούν καλές μεθόδους βελτιστοποίησης, όχι όμως σε προβλήματα μεγάλων διαστάσεων. Επομένως, θα εξετάσουμε παρακάτω διάφορους τρόπους ώστε να προσαρμόσουμε τους αλγορίθμους σε μεγάλη κλίμακα ακόμα και για μη κυρτές συναρτήσεις.

Ο επαναληπτικός τύπος της μεθόδου Newton (2.38) για την αντικειμενική συνάρτηση F ως προς το διάνυσμα w γράφεται ισοδύναμα

$$w_{k+1} \leftarrow w_k + \alpha_k p_k, \tag{5.4a}$$
$$\text{όπου} \quad \nabla^2 F(w_k) p_k = -\nabla F(w_k). \tag{5.4b}$$

Ο ακριβής υπολογισμός της σχέσης (5.4b) είναι υπολογιστικά απαιτητικός. Επομένως, μπορούμε να λύσουμε με ανακριβή τρόπο (προσεγγιστικά) το σύστημα Newton με τη χρήση μιας επαναληπτικής μεθόδου, όπως η μέθοδος συζυγών κλίσεων και έτσι να επιτύχουμε μείωση του υπολογιστικού κόστους και γρηγορότερη σύγκλιση στη λύση.

Το γεγονός ότι η μέθοδος συζυγών κλίσεων δεν χρειάζεται τον υπολογισμό του Εσσιανού πίνακα για την εκτέλεσή της, την καθιστά μία μέθοδο *χωρίς Εσσιανό* (Hessian-free). Αντί του Εσσιανού και μόνο, η μέθοδος χρησιμοποιεί το γινόμενο μεταξύ του Εσσιανού και των διανυσμάτων, το οποίο μπορεί να υπολογιστεί ξεχωριστά. Γενικότερα, για μια λεία αντικειμενική συνάρτηση F, μπορούμε να υπολογίσουμε το $\nabla^2 F(w)d$, όπου $d \in \mathbb{R}^d$ διάνυσμα, με κόστος υποπολλαπλάσιο αυτού του ∇F και χωρίς τον υπολογισμό του Εσσιανού που χρειάζεται $\mathcal{O}(d^2)$ αποθηκευτικό χώρο.

5.1.1 Μέθοδοι Newton Μικρότερου Δείγματος Χωρίς Εσσιανό

Θα αναλύσουμε τώρα έναν τύπο μεθόδων που προκύπτουν από το γεγονός ότι οι ανακριβείς μέθοδοι Newton συγκλίνουν ικανοποιητικά στη λύση χωρίς την ακριβή εκτίμηση του Εσσιανού πίνακα. Στις εφαρμογές της μηχανικής μάθησης μεγάλης κλίμακας αυτό σημαίνει ότι η μέθοδος αποδίδει καλύτερα ακόμα και όταν υπάρχει μεγαλύτερη διακύμανση στην εκτίμηση του Εσσιανού απ΄ ό, τι στην εκτίμηση της κλίσης. Βασισμένοι στην παραπάνω παρατήρηση, θα εξετάσουμε μια τεχνική η οποία επιλέγει ένα μικρότερο δείγμα για την εκτίμηση του Εσσιανού από αυτό της στοχαστικής κλίσης. Από τις σχέσεις (4.2), μια εκτίμηση της στοχαστικής κλίσης για ένα μικρότερο σύνολο παραδειγμάτων μεγέθους $n_k = |\mathcal{S}_k|$ είναι:

$$\nabla f_{\mathcal{S}_k}(w_k; \xi_k) = g(w_k; \xi_k) = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \nabla f(w_k; \xi_{k,i}). \tag{5.5}$$

Σύμφωνα με αυτόν το συμβολισμό για το στοχαστικό διάνυσμα κλίσης, η εκτίμηση του στοχαστικού Εσσιανού πίνακα για ένα διαφορετικό δείγμα μεγέθους $|\mathcal{S}_k^H|$ στο w_k είναι:

$$\nabla^2 f_{\mathcal{S}_k^H}(w_k; \xi_k^H) = \frac{1}{|\mathcal{S}_k^H|} \sum_{i \in \mathcal{S}_k^H} \nabla^2 f(w_k; \xi_{k,i}). \tag{5.6}$$

Κάτω από συγκεκριμένες προϋποθέσεις, το $\mathcal{S}_k^H$ δεν εξαρτάται από το $\mathcal{S}_k$. Το μέγεθος $|\mathcal{S}_k^H|$ του μικρότερου δείγματος θα πρέπει να επιλεχθεί αρκετά μικρό ώστε να μειώνει το κόστος μιας επανάληψης της μεθόδου συζυγών κλίσεων και αρκετά μεγάλο ώστε οι πληροφορίες καμπυλότητας από το γινόμενο Εσσιανού - διανύσματος να είναι χρήσιμες. Παρακάτω δίνεται ένας αλγόριθμος (Bottou, Curtis & Nocedal, 2018, σ. 273) για την ανακριβή μέθοδο Newton για μικρότερο δείγμα χωρίς υπολογισμό του Εσσιανού πίνακα.

Αλγόριθμος 5.1 Ανακριβής Newton Μικρότερου Δείγματος Χωρίς Εσσιανό

1: Επέλεξε μια αρχική τιμή w_1.
2: Επέλεξε τις σταθερές $\rho \in (0,1), \gamma \in (0,1),\ \eta \in (0,1),\ \text{και}\ \max_{cg} \in \mathbb{N}$.
3: **for** $k = 1, 2, \ldots$ **do**
4: Θέσε τις τυχαίες μεταβλητές ξ_k και ξ_k^H που αντιστοιχούν στα $\mathcal{S}_k$ και $\mathcal{S}_k^H$.
5: Υπολόγισε το p_k με συζυγείς κλίσεις για την επίλυση του

$$\nabla^2 f_{\mathcal{S}_k^H}(w_k; \xi_k^H) p_k = -\nabla f_{\mathcal{S}_k}(w_k; \xi_k) \tag{5.7}$$

έως ότου εκτελεστούν $\max_{cg}$ επαναλήψεις ή μια δοκιμαστική λύση δώσει

$$\|r_k\|_2 \coloneqq \|\nabla^2 f_{\mathcal{S}_k^H}(w_k; \xi_k^H) p_k + \nabla f_{\mathcal{S}_k}(w_k; \xi_k)\|_2 \leq \rho \|\nabla f_{\mathcal{S}_k}(w_k; \xi_k)\|_2.$$

6: Θέσε $w_{k+1} \leftarrow w_k + \alpha_k p_k$, όπου $\alpha_k \in \{\gamma^0, \gamma^1, \gamma^2, \ldots\}$ το μέγιστο στοιχείο με

$$f_{\mathcal{S}_k}(w_{k+1}; \xi_k) \leq f_{\mathcal{S}_k}(w_k; \xi_k) + \eta \alpha_k \nabla f_{\mathcal{S}_k}(w_k; \xi_k)^T p_k. \tag{5.8}$$

7: **end for**

Ο παραπάνω αλγόριθμος αποφεύγει τον υπολογισμό του πλήρους Εσσιανού πίνακα, χρησιμοποιώντας μόνο την εκτίμηση που δόθηκε από τη σχέση (5.6) και τη μέθοδο συζυγών κλίσεων χωρίς Εσσιανό για τον υπολογισμό της κλίσης Newton. Επίσης, το Βήμα 6 δείχνει ότι το μέγεθος του βήματος σε κάθε επανάληψη υπολογίζεται με τη μέθοδο Armijo.

Έστω g_{cost} το κόστος υπολογισμού της εκτίμησης της κλίσης $\nabla f_{\mathcal{S}_k}(w;\xi_k)$ και $c \times g_{cost}$ το κόστος υπολογισμού του γινομένου Εσσιανού και διανύσματος για τον Αλγόριθμο 5.1, όπου $c > 0$ είναι ένας παράγοντας κόστους. Εάν σε κάθε κύρια επανάληψη (της ανακριβούς Newton) του αλγορίθμου εκτελείται ο μέγιστος αριθμός επαναλήψεων της μεθόδου συζυγών κλίσεων $\max_{cg}$, τότε το υπολογιστικό κόστος κάθε βήματος του Αλγορίθμου 5.1 είναι

$$\max_{cg} \times c \times g_{cost} + g_{cost}.$$

Δουλεύοντας με έναν batch αλγόριθμο για την ανακριβή μέθοδο Newton με συζυγείς κλίσεις που υπολογίζει το εμπειρικό ρίσκο R_n, θα πρέπει για τα σύνολα των παραδειγμάτων να ισχύει ότι $|\mathcal{S}_k^H| = |\mathcal{S}_k| = n$. Βλέπουμε, δηλαδή, ότι και πάλι οδηγούμαστε σε μεγάλο κόστος ανά επανάληψη του αλγορίθμου. Αντιθέτως, σε ένα στοχαστικό αλγόριθμο που χρησιμοποιεί μικρότερο δείγμα για τον Εσσιανό, μπορούμε να θέσουμε έναν παράγοντα c αρκετά μικρό έτσι ώστε $\max_{cg} \times c \approx 1$, το οποίο δίνει υπολογιστικό κόστος ανά επανάληψη ανάλογο αυτού της SG.

Η μέθοδος που περιγράφεται από τον Αλγόριθμο 5.1 τρέχει ικανοποιητικά καλά μόνο όταν το μέγεθος του δείγματος για τον υπολογισμό των κλίσεων $|\mathcal{S}_k|$ είναι μεγάλο. Σε αυτή την περίπτωση, περιέχονται αρκετά παραδείγματα ώστε να κάνουμε επιλογή υποσυνόλων τους για την σωστή προσέγγιση του Εσσιανού πίνακα. Διαφορετικά, αν ο αλγόριθμος έτρεχε όπως η SG με μικρό μέγεθος $|\mathcal{S}_k|$ και οι κλίσεις εμφάνιζαν μεγάλη διακύμανση, τότε θα έπρεπε να επιλέξουμε $|\mathcal{S}_k^H| > |\mathcal{S}_k|$ για τη σωστή εκτίμηση του Εσσιανού. Η επιλογή αυτή θα είχε επίπτωση στην αποτελεσματικότητα του αλγορίθμου.

Η σύγκλιση του αλγορίθμου αποδεικνύεται εύκολα όταν η συνάρτηση του εμπειρικού ρίσκου $F = R_n$ είναι ισχυρά κυρτή, θέτοντας $\mathcal{S}_k \leftarrow \{1, \ldots, n\}$ για κάθε $k \in \mathbb{N}$. Τότε, βλέπουμε ότι η μέθοδος επωφελείται από τη χρήση συζυγών κλίσεων, σε αντίθεση με τις πλήρεις κλίσεις, εφόσον ελαχιστοποιούμε ένα τετραγωνικό μοντέλο της αντικειμενικής συνάρτησης της μορφής (5.3). Επομένως, από το Θεώρημα 2.8 για τη μέθοδο συζυγών κλίσεων, θα έχουμε σύγκλιση στη σχέση (5.7) σε d το πολύ επαναλήψεις. Επιπλέον, όταν γίνεται εκτίμηση του Εσσιανού πίνακα με μικρότερο δείγμα, δηλαδή όταν $\mathcal{S}_k^H \subset \mathcal{S}_k$, δεν έχει αποδειχθεί εάν επιτυγχάνεται ρυθμός σύγκλισης καλύτερος του γραμμικού. Όμως, η μείωση του αριθμού των επαναλήψεων που χρειάζονται για μία καλή προσέγγιση της λύσης σε αυτή την περίπτωση είναι μικρότερη σε σύγκριση με τις μεθόδους χωρίς υπολογισμό του Εσσιανού πίνακα.

5.2 Στοχαστικές Μέθοδοι Quasi-Newton

Όπως είδαμε και στο δεύτερο κεφάλαιο, οι μέθοδοι quasi-Newton έχουν δώσει σημαντικά ωφέλη στον τομέα της μη γραμμικής βελτιστοποίησης. Είναι σημαντικό να αποφασίσουμε κάτω από ποιες συνθήκες μπορούν οι μέθοδοι αυτές να χρησιμοποιηθούν σε στοχαστικό πλαίσιο. Σε αυτή την παράγραφο θα γίνει μια ανάλυση μιας

από τις πιο διαδεδομένες μεθόδους quasi-Newton, της BFGS. Όπως και στη μέθοδο Newton (5.4a) και (5.4b), ο επαναληπτικός τύπος BFGS για την ελαχιστοποίηση μιας δύο φορές συνεχώς παραγωγίσιμης συνάρτησης $F : \mathbb{R}^d \to \mathbb{R}$ είναι

$$w_{k+1} \leftarrow w_k - \alpha_k H_k \nabla F(w_k), \tag{5.9}$$

όπου ο H_k είναι ένας συμμετρικός και θετικά ορισμένος πίνακας που προσεγγίζει τον πίνακα $(\nabla^2 F(w_k))^{-1}$. Η διαφορά του παραπάνω επαναληπτικού τύπου που τον καθιστά quasi-Newton είναι ότι η ακολουθία $\{H_k\}$ ενημερώνεται κατά την εκτέλεση του αλγορίθμου δυναμικά και δεν είναι απλά ένας υπολογισμός παραγώγου δεύτερης τάξης σε κάθε επανάληψη. Συγκεκριμένα, ο νέος αντίστροφος Εσσιανός δίνεται από τη διαφορά στα παραμετρικά διανύσματα που προκύπτουν από την επαναληπτική διαδικασία και τη διαφορά στις κλίσεις σε αυτά:

$$s_k \coloneqq w_{k+1} - w_k \ \text{ και } \ y_k \coloneqq \nabla F(w_{k+1}) - \nabla F(w_k).$$

Ο τύπος ενημέρωσης του αντιστρόφου του Εσσιανού πίνακα για τη μέθοδο BFGS δίνεται στο βιβλίο των Nocedal και Wright (2006, Κεφάλαιο 6.1, σ. 139-140):

$$H_{k+1} \leftarrow \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right)^T H_k \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right) + \frac{s_k s_k^T}{s_k^T y_k}. \tag{5.10}$$

Είναι σημαντικό να δούμε ότι ο τύπος (5.10) ικανοποιεί τη συνθήκη quasi-Newton (2.43) με $B_{k+1} = H_{k+1}^{-1}$, δηλαδή $H_{k+1}^{-1} s_k = y_k$. Πράγματι,

$$\begin{aligned}
H_{k+1} y_k &= \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right)^T H_k y_k \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right) + \frac{s_k s_k^T}{s_k^T y_k} y_k \\
&= \left(H_k y_k - \frac{y_k^T s_k}{s_k y_k^T} H_k y_k\right) \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right) + \frac{s_k s_k^T}{s_k^T y_k} y_k \\
&= (H_k y_k - H_k y_k) \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right) + s_k = s_k \Rightarrow \\
H_{k+1}^{-1} s_k &= y_k.
\end{aligned}$$

Στην εργασία των Dennis και Moré (1974) (Bottou, Curtis & Nocedal, 2018) αποδεικνύεται ότι η BFGS που δίνεται από τους τύπους (5.9) και (5.10) έχει τοπικά υπεργραμμικό ρυθμό σύγκλισης. Επίσης, η ταχύτητα αυτή επιτυγχάνεται μόνο από πληροφορίες πρώτης τάξης, χωρίς την ανάγκη επίλυσης κάποιου γραμμικού συστήματος. Μία τέτοια μέθοδος συνήθως χρησιμοποιείται σε προβλήματα μικρού μεγέθους, λόγω των πυκνών πινάκων που παράγονται από τη σχέση (5.10). Ωστόσο, έχουν αναπτυχθεί τεχνικές οι οποίες μειώνουν το κόστος ανά επανάληψη της μεθόδου για μεγαλύτερα προβλήματα και εξασφαλίζουν γραμμική σύγκλιση.

5.2.1 Μετατροπή σε Στοχαστικές Μεθόδους

Στη μηχανική μάθηση οι μέθοδοι quasi-Newton επεκτείνονται από το ντετερμινιστικό περιβάλλον στο οποίο κινούνται, σε στοχαστικό. Άρα, ο επαναληπτικός τύπος γίνεται

$$w_{k+1} \leftarrow w_k - \alpha_k H_k g(w_k, \xi_k). \tag{5.11}$$

Εφόσον μας ενδιαφέρουν προβλήματα μεγάλης κλίμακας, υποθέτουμε ότι η (5.11) εφαρμόζει μία τεχνική L-BFGS, η οποία αποφεύγει την αναλυτική κατασκευή του H_k. Η μέθοδος *limited memory BFGS* ή *L-BFGS* είναι μία μέθοδος quasi-Newton με μειωμένο αποθηκευτικό χώρο που χρησιμοποιείται σε μεγάλα προβλήματα ή σε προβλήματα που ο Εσσιανός πίνακας υπολογίζεται δύσκολα.

Η μέθοδος αυτή, αντί να αποθηκεύει πλήρεις, $n \times n$ εκτιμήσεις των Εσσιανών πινάκων, κρατάει μόνο μερικά αντιπροσωπευτικά διανύσματα μήκους n. Συγκεκριμένα, η μέθοδος L-BFGS, χρησιμοποιει πληροφορίες καμπυλότητας (δευτεροβάθμιες) μόνο από τις πιο πρόσφατες επαναλήψεις για την προσέγγιση του Εσσιανού. Παλαιότερες πληροφορίες για την καμπυλότητα διαγράφονται, καθώς δεν συμβάλλουν ουσιαστικά στον υπολογισμό του τρέχοντος πίνακα, ελευθερώνοντας έτσι αποθηκευτικό χώρο. Ο Αλγόριθμος 5.2 περιγράφει τη ντετερμινιστική μέθοδο L-BFGS για μια δύο φορές συνεχώς παραγωγίσιμη συνάρτηση $f : \mathbb{R}^d \to \mathbb{R}$ του $w \in \mathbb{R}^d$ (Nocedal & Wright, 2006, σ. 179).

Αλγόριθμος 5.2 L-BFGS

Επέλεξε αρχικό σημείο w_0 και ακέραιο $m > 0$.
for $k = 0, 1, \ldots$ **do**
    Επέλεξε έναν H_k^0.
    Υπολόγισε το διάνυσμα $p_k \leftarrow -H_k \nabla f(w_k)$.
    Επέλεξε το βήμα α_k με βάση τις συνθήκες Wolfe (2.51a) και (2.51β).
    Υπολόγισε την τιμή $w_{k+1} \leftarrow w_k + \alpha_k p_k$.
    if $k > m$ **then**
        Διάγραψε το ζεύγος $\{s_{k-m}, y_{k-m}\}$ από τη μνήμη.
    end if
    Υπολόγισε και αποθήκευσε τα

$$s_k \leftarrow w_{k+1} - w_k \quad \text{και} \quad y_k \leftarrow \nabla f(w_{k+1}) - \nabla f(w_k).$$

end for

Το γινόμενο $H_k g(w_k, \xi_k)$ υπολογίζεται σύμφωνα με τον αλγόριθμο *L-BFGS two-loop recursion* (Αλγόριθμος 7.4, Nocedal & Wright, 2006, σ. 178). Επίσης, ένας τρόπος με τον οποίο μπορούμε να επιλέξουμε το H_k^0 είναι θέτοντας $H_k^0 = \gamma_k I$, όπου

$$\gamma_k = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}.$$

Η τιμή γ είναι ο συντελεστής μετασχηματισμού που προσπαθεί να εκτιμήσει το μέγεθος του πραγματικού Εσσιανού πίνακα κατά την πιο πρόσφατη κατεύθυνση αναζήτησης.

Συγκριτικά με τη μέθοδο SG, όπως είδαμε στο τέταρτο κεφάλαιο, το υπολογιστικό κόστος κάθε επανάληψης της SG είναι πολύ μικρό, αφού απαιτεί μόνο τον υπολογισμό του $g(w_k, \xi_k)$. Αντίθετα, ο επαναληπτικός τύπος (5.11) υπολογίζει το γινόμενο $H_k g(w_k, \xi_k)$, το οποίο, σύμφωνα με την ανάλυση που γίνεται στο βιβλίο των Nocedal και Wright (2006, Κεφάλαιο 7), χρειάζεται $4md$ πράξεις. Για παράδειγμα, για ακριβώς ένα δείγμα εκτελούνται d πράξεις για την κλίση $g(w_k, \xi_k)$ και αν η παράμετρος μνήμης τεθεί $m = 5$, τότε βρίσκουμε ότι η μέθοδος L-BFGS έχει 20 φορές περισσότερες πράξεις από την SG.

Παρατηρούμε, ωστόσο, ότι ο παραπάνω αριθμός πράξεων στον υπολογισμό της κλίσης, αφορά ένα μόνο παράδειγμα. Επομένως, το πρόβλημα του επιπλέον υπολογιστικού κόστους ανά επανάληψη στις στοχαστικές μεθόδους quasi-Newton βελτιώνεται συνήθως υπολογίζοντας τις κλίσεις με minibatches. Δηλαδή, μπορεί να γίνει χρήση μικρότερων δειγμάτων για τον υπολογισμό της κλίσης, όπως γίνεται στην προηγούμενη ενότητα για τις μεθόδους Newton χωρίς Εσσιανό. Με αυτό τον τρόπο επωφελούμαστε από τα πλεονεκτήματα που προσφέρει μια στοχαστική μέθοδος quasi-Newton στη διαχείριση συναρτήσεων επιρρεπών στο θόρυβο, με κόστος ανά επανάληψη οριακά μεγαλύτερο της SG.

Η πιο απλή μετατροπή, λοιπόν, της L-BFGS σε στοχαστική μέθοδο είναι με την αντικατάσταση των ντετερμινιστικών κλίσεων με στοχαστικές κλίσεις. Τότε, τα διανύσματα μετατόπισης γίνονται

$$s_k \coloneqq w_{k+1} - w_k \;\text{ και }\; y_k \coloneqq \nabla f_{\mathcal{S}_k}(w_{k+1}, \xi_k) - \nabla f_{\mathcal{S}_k}(w_k, \xi_k), \tag{5.12}$$

όπου η $\nabla f_{\mathcal{S}_k}$ είναι ορισμένη από τον τύπο (5.5). Σύμφωνα, όμως, με την εργασία των Mokhtari και Ribeiro (2014), κατά την εκτέλεση της μεθόδου L-BFGS με τη χρήση του τύπου (5.12) είναι αναγκαίος ο υπολογισμός των δύο στοχαστικών κλίσεων $g(w_{k+1}, \xi_k)$ και $g(w_{k+1}, \xi_{k+1})$ σε κάθε βήμα, ώστε να εξασφαλιστεί η σύγκλισή της. Η $g(w_{k+1}, \xi_k)$ χρησιμοποιείται στον υπολογισμό της μετατόπισης y_k και στην ενημέρωση του Εσσιανού και η $g(w_{k+1}, \xi_{k+1})$ για την εκτέλεση του επόμενου βήματος του αλγορίθμου. Επομένως, έχουμε περισσότερο υπολογιστικό κόστος ανά επανάληψη. Επίσης, υπάρχει περίπτωση να μην γίνεται ενημέρωση στην προσέγγιση του αντιστρόφου του Εσσιανού σε κάθε επανάληψη του αλγορίθμου.

Μία λύση για την καλύτερη προσέγγιση της συμπεριφοράς του Εσσιανού είναι δίνοντας έναν εναλλακτικό ορισμό της διαφοράς y_k. Συγκεκριμένα, εφόσον από τη σχέση (2.42) έχουμε $\nabla f_{\mathcal{S}_k}(w_{k+1}) - \nabla f_{\mathcal{S}_k}(w_k) \approx \nabla^2 F(w_k)(w_{k+1} - w_k)$, μπορούμε να γράψουμε το y_k ως:

$$y_k = \nabla^2 f_{\mathcal{S}_k^H}(w_k; \xi_k^H) s_k, \tag{5.13}$$

όπου $\nabla^2 f_{\mathcal{S}_k^H}(w_k; \xi_k^H)$ είναι ο στοχατικός Εσσιανός μικρότερου δείγματος όπως δίνεται από τη σχέση (5.6) και το μέγεθος $|\mathcal{S}_k^H|$ είναι αρκετά μεγάλο ώστε να δίνει καλύτερες πληροφορίες για την καμπυλότητα $\nabla^2 F(w_k)(w_{k+1} - w_k)$. Όπως και στην περίπτωση της μεθόδου Newton χωρίς Εσσιανό, το γινόμενο (5.13) μπορεί να χρησιμοποιηθεί χωρίς τον ακριβή υπολογισμό του $\nabla^2 f_{\mathcal{S}_k^H}(w_k; \xi_k^H)$.

Ανεξάρτητα από τον ορισμό του y_k, όταν το μέγεθος $|\mathcal{S}_k^H|$ είναι αρκετά μεγαλύτερο από το $|\mathcal{S}_k|$, το κόστος κάθε επανάληψης μιας στοχαστικής μεθόδου quasi-Newton είναι μεγάλο, εξαιτίας του μεγάλου κόστους του υπολογισμού του y_k. Μπορούμε να αντιμετωπίσουμε αυτό το πρόβλημα με τον υπολογισμό του y_k μόνο μετά από κάποιο αριθμό επαναλήψεων. Με βάση αυτή την ιδέα, διαχωρίζεται η quasi-Newton ενημέρωση του αντιστρόφου του Εσσιανού από τον επαναληπτικό τύπο της μεθόδου quasi-Newton. Κατ΄ αυτό τον τρόπο, δημιουργείται ένας γενικότερος αλγόριθμος στοχαστικών μεθόδων quasi-Newton (Bottou, Curtis & Nocedal, 2018, σ. 278), ο οποίος δίνεται παρακάτω. Συγκεκριμένα, εκτελείται ένας αριθμός επαναλήψεων του τύπου (5.1) για σταθερό πίνακα H_k. Έπειτα, υπολογίζεται το διάνυσμα s_k σύμφωνα με τον τύπο (5.12) και το y_k σύμφωνα με κάποιον από τους τύπους (5.12) ή (5.13).

Το νέο ζεύγος αντικαθιστά ένα από αυτά που βρίσκονται στη μνήμη και έτσι ολοκληρώνεται μια επανάληψη L-BFGS. Στον παρακάτω αλγόριθμο, το $\mathcal{P} = \{s_j, y_j\}$, $j \in \{1, \ldots, m\}$ συμβολίζει μία σύλλογή από $m \in \mathbb{N}$ ζεύγη μετατοπίσεων.

Αλγόριθμος 5.3 Στοχαστικές Μέθοδοι Quasi-Newton

1: Επέλεξε αρχικό διάνυσμα w_1, σταθερά $m \in \mathbb{N}$ και θέσε $\mathcal{P} \leftarrow \emptyset$.
2: Επέλεξε μία ακολουθία βημάτων $\{\alpha_k\}$ με $\alpha_k > 0$ για κάθε $k \in \mathbb{N}$.
3: **for** $k = 1, 2, \ldots$ **do**
4: Θέσε τις τυχαίες μεταβλητές ξ_k και ξ_k^H που αντιστοιχούν στα $\mathcal{S}_k$ και $\mathcal{S}_k^H$.
5: Υπολόγισε το $p_k = H_k g(w_k, \xi_k)$.
6: Θέσε $s_k \leftarrow -\alpha_k p_k$.
7: Θέσε $w_{k+1} \leftarrow w_k + s_k$.
8: **if** `update pairs = true` **then**
9: Υπολόγισε τα s_k και y_k (βάσει του δείγματος $\mathcal{S}_k^H$).
10: Πρόσθεσε το νέο ζεύγος μετατοπίσεων (s_k, y_k) στο $\mathcal{P}$.
11: Εάν $|\mathcal{P}| > m$, τότε αφαίρεσε το παλαιότερο ζεύγος από το $\mathcal{P}$.
12: **end if**
13: **end for**

Στον παραπάνω αλγόριθμο, το γινόμενο $H_k g(w_k, \xi_k)$ υπολογίζεται από τον αλγόριθμο two-loop recursion που έχουμε ήδη αναφέρει και συνθήκη `update pairs` θέτεται `true` όταν πρέπει να ενημέρωση στα διανύσματα s_k και y_k. Για έναν αλγόριθμο quasi-Newton, όπως είναι αυτός της BFGS, ο επαναληπτικός τύπος είναι καλώς ορισμένος μόνο όταν για κάθε ζεύγος (s_j, y_j) ισχύει ότι $s_j^T y_j > 0$. Παρόλο που σε κάποιο ντετερμινιστικό αλγόριθμο κάτι τέτοιο εξασφαλίζεται εύκολα, στο στοχαστικό περιβάλλον δεν έχουμε ακριβή υπολογισμό της κλίσης, αλλά μόνο μια εκτίμησή της, άρα το ζεύγος διανυσμάτων (s_j, y_j) μπορεί να εμφανίζει θόρυβο. Συνεπώς, καταφεύγουμε σε διαφορετικές τεχνικές για την αντιμετώπιση αυτού του προβλήματος. Ένας τρόπος ο οποίος εξετάζεται στην επόμενη ενότητα είναι με την αντικατάσταση του Εσσιανού πίνακα από έναν πίνακα Gauss-Newton.

5.3 Μέθοδοι Gauss-Newton

Η κλασική μέθοδος Gauss-Newton χρησιμοποιείται για τη βελτιστοποίηση προβλημάτων με ελάχιστα τετράγωνα. Συγκεκριμένα, η μέθοδος ελαχιστοποιεί αντικειμενικές συναρτήσεις οι οποίες αποτελούν άθροισμα τετραγώνων. Ανάλυση της μεθόδου και των ιδιοτήτων σύγκλισής της γίνεται γίνεται στο βιβλίο των Nocedal και Wright (2006, Κεφάλαιο 10). Το πλεονέκτημα της μεθόδου έγκειται στο γεγονός ότι μπορούμε να δημιουργήσουμε μια προσέγγιση του Εσσιανού με ικανοποιητικές ιδιότητες, βασισμένοι μόνο σε πληροφορίες πρώτης τάξης. Στη μηχανική μάθηση, λοιπόν, η μέθοδος μπορεί να ελαχιστοποιήσει συναρτήσεις απώλειας με ελάχιστα τετράγωνα, τύπος συναρτήσεων απώλειας αρκετά συνηθισμένος.

Συνεχίζουμε με την παρουσίαση της *κλασικής μεθόδου Gauss-Newton*, δεδομένου ενός ζεύγους εισόδου - εξόδου (x, y). Θεωρούμε ότι η απώλεια που προκύπτει από το διάνυσμα παραμέτρων w είναι ανάλογη του τετραγώνου της νόρμας της διαφοράς μεταξύ της πρόβλεψης $h(x; w) \in \mathbb{R}^d$ και της πραγματικής εξόδου $y \in \mathbb{R}^d$. Άρα

γράφουμε τη συνάρτηση απώλειας με το ζεύγος εισόδου - εξόδου τυχαία επιλεγμένο από την κατανομή της τυχαίας μεταβλητής ξ ως

$$f(w;\xi) = \ell(h(x_\xi; w), y_\xi) = \frac{1}{2}\|h(x_\xi; w) - y_\xi\|_2^2.$$

Ο δείκτης ξ δηλώνει ότι τα διανύσματα εισόδου και εξόδου (x, y) ανήκουν στην κατανομή της τυχαίας μεταβλητής ξ. Αναπτύσσοντας τη συνάρτηση αυτή σύμφωνα με το θεώρημα Taylor γύρω από το w_k, προκύπτει μία τετραγωνική συνάρτηση της μορφής (5.3). Όπως είδαμε και παραπάνω, η ελαχιστοποίησή της γίνεται σύμφωνα με τη μέθοδο Newton με την αντίστοιχη προσέγγιση του Εσσιανού πίνακα στο w_k. Εναλλακτικά, με τη μέθοδο Gauss-Newton ξεκινάμε με ένα ανάπτυγμα Taylor πρώτου βαθμού της συνάρτηση πρόβλεψης μέσα στην τετραγωνική συνάρτηση απώλειας. Αν $J_h(\cdot;\xi)$ είναι ο Ιακωβιανός πίνακας της $h(x_\xi;\cdot)$ ως προς w, τότε από το ανάπτυγμα Taylor πρώτου βαθμού της $h \in \mathbb{R}^d$ γύρω από το w_k είναι

$$h(x_\xi, w) \approx h(x_\xi; w) + J_h(w_k;\xi)(w - w_k),$$

το οποίο οδηγεί στη συνάρτηση

$$\begin{aligned} f(w;\xi) &\approx \frac{1}{2}\|h(x_\xi; w_k) + J_h(w_k;\xi)(w - w_k) - y_\xi\|_2^2 \\ &= \frac{1}{2}\|h(x_\xi; w_k) - y_\xi\|_2^2 + (h(x_\xi; w_k) - y_\xi)^T J_h(w_k;\xi)(w - w_k) \\ &+ \frac{1}{2}(w - w_k)^T J_h(w_k;\xi)^T J_h(w_k;\xi)(w - w_k). \end{aligned}$$

Παρατηρούμε ότι στην παραπάνω σχέση δεν υπάρχουν παράγωγοι δεύτερης τάξης της συνάρτησης h ως προς το παραμετρικό διάνυσμα w. Ο δευτεροβάθμιος όρος που παραμένουν είναι το γινόμενο $J_h(w_k;\xi)^T J_h(w_k;\xi)$. Επομένως, μπορούμε να αντικαταστήσουμε τον στοχαστικό Εσσιανό πίνακα μικρότερου δείγματος (5.6) με τον πίνακα Gauss - Newton (Bottou, Curtis & Nocedal, 2018, σ. 279):

$$G_{\mathcal{S}_k^H}(w_k; \xi_k^H) = \frac{1}{|\mathcal{S}_k^H|} \sum_{i \in \mathcal{S}_k^H} J_h(w_k; \xi_{k,i})^T J_h(w_k; \xi_{k,i}). \tag{5.14}$$

Ο πίνακας Gauss-Newton διαφέρει από τον πραγματικό Εσσιανό πίνακα στο ότι περιέχει όρους με τη διαφορά ανάμεσα στην πρόβλεψη $h(x_\xi; w_k)$ και στην πραγματική έξοδο y_ξ. Επομένως, στην περίπτωση που η συνάρτηση απώλειας δίνει απολύτως ακριβή αποτελέσματα ($h(x_\xi; w_k) = y_\xi$) και δεν υπάρχει απώλεια, οι δύο πίνακες είναι ίσοι. Ένα πλεονέκτημα της τεχνικής αυτής είναι ότι μας εξασφαλίζει ότι ο πίνακας Gauss-Newton θα είναι θετικός. Ωστόσο, ένα σύνηθες εμπόδιο που εμφανίζεται είναι ότι οι πίνακες Gauss-Newton είναι συνήθως μη αντιστρέψιμοι. Το χαρακτηριστικό αυτό διορθώνεται με την πρόσθεση ενός θετικού πολλαπλάσιου του μοναδιαίου πίνακα ως όρο κανονικοποίησης. Έτσι, μπορούμε να αντικαταστήσουμε τον κανονικοποιημένο πίνακα Gauss-Newton στις μεθόδους Newton και quasi-Newton που παρουσιάστηκαν πιο πάνω στο κεφάλαιο. Με αυτό τον τρόπο, κατοχυρώνουμε ότι ο πίνακας μετασχηματισμού θα είναι θετικός.

Η μέθοδος Gauss-Newton μπορεί επίσης να γενικευτεί και για μη τετραγωνικές συναρτήσεις απώλειας. Για παράδειγμα, έστω μία τυχαία, κυρτή συνάρτηση απώλειας $\ell(h, y)$ και μία συνάρτηση πρόβλεψης $h(x; w)$. Τότε, ο γενικευμένος πίνακας Gauss-Newton είναι

$$G_{\mathcal{S}_k^H}(w_k; \xi_k^H) = \frac{1}{|\mathcal{S}_k^H|} \sum_{i \in \mathcal{S}_k^H} J_h(w_k; \xi_{k,i})^T H_\ell(w_k; \xi_{k,i}) J_h(w_k; \xi_{k,i}), \tag{5.15}$$

όπου το $H_\ell(w_k; \xi) = \frac{\partial^2 \ell}{\partial h^2}(h(x_\xi; w_k), y_\xi)$ εκφράζει την καμπυλότητα (δευτεροβάθμιες πληροφορίες) της συνάρτησης απώλειας ℓ. Επομένως, το πίνακας (5.15) πρόκειται για μια γενίκευση του πίνακα (5.14), όπου έχουμε ότι $H_\ell = I$.

Το υπολογιστικό κόστος της μεθόδου Gauss-Newton εξαρτάται από τις διαστάσεις της συνάρτησης πρόβλεψης. Όταν η h είναι μονοδιάστατη, ο Ιακωβιανός της πίνακας J_h είναι πίνακας - γραμμή με στοιχεία που είναι ήδη γνωστά. Ο υπολογισμός τους προκύπτει μέσω του υπολογισμού της στοχαστικής κλίσης $\nabla f(w; \xi)$. Για μεγαλύτερες, όμως, διαστάσεις της h αυτό δεν ισχύει, καθώς ο υπολογισμός του στοχαστικού διανύσματος $\nabla f(w; \xi)$ δεν απαιτεί τον υπολογισμό όλων των γραμμών του Ιακωβιανού πίνακα.

6 Συμπεράσματα

Είδαμε, λοιπόν, στην παρούσα εργασία, τα βασικά στοιχεία που χρειάζονται για την κατανόηση και την αξιοποίηση των δυνατοτήτων της μαθηματικής βελτιστοποίησης στον τομέα της μηχανικής μάθησης. Συγκεκριμένα, έγινε μια αναλυτική παρουσίαση των μαθηματικών θεμελίων, από το πεδίο της αριθμητικής ανάλυσης και των κλασικών αλγορίθμων βελτιστοποίησης, αλλά και των βασικών ορισμών στατιστικής. Επίσης, με τη βοήθεια της ταξινόμησης κειμένων, δείξαμε πώς ένα πρόβλημα της μηχανικής μάθησης μετατρέπεται σε πρόβλημα βελτιστοποίησης. Η αριθμητική του επίλυση, μας δίνει τη βέλτιστη δυνατή προσέγγιση της σωστής ταξινόμησης ενός κειμένου στην κατάλληλη κατηγορία. Κατ΄ επέκταση, για μια γενικότερη διαδικασία μηχανικής μάθησης, τα γνωστά δεδομένα της μοντελοποιούνται σύμφωνα με μια στατιστική κατανομή και παραμετροποιούνται. Τότε, ένας κατάλληλος αλγόριθμος να μπορεί να εξάγει συμπεράσματα για άγνωστα στοιχεία.

Το κύριο μέρος της εργασίας αφορά τον αλγόριθμο SG, ο οποίος σύμφωνα με τα σημερινά δεδομένα προτιμάται για την επίλυση προβλημάτων μεγάλης κλίμακας. Οι συνθήκες που θέσαμε τόσο για την αντικειμενική συνάρτηση, όσο και για τη στοχαστική κατεύθυνση, καθιστούν τη μέθοδο SG εξαιρετικά χρήσιμη για την ελαχιστοποίηση κυρτών και μη κυρτών συναρτήσεων. Επίσης, με τη βοήθεια θεωρητικών εργαλείων, αποδείξαμε ότι για μεγάλο αριθμό δεδομένων, η μέθοδος αυτή συγκλίνει αρκετά γρήγορα στην προσεγγιστική λύση και με μικρό υπολογιστικό κόστος, σε σύγκριση με άλλες μεθόδους. Παρουσιάσαμε, ωστόσο, και τεχνικές οι οποίες μπορούν, έστω και θεωρητικά, να αντιμετωπίσουν διάφορα μειονεκτήματα που εμφανίζει η μέθοδος SG όταν πρέπει να τρέξει για συναρτήσεις με κακή κατάσταση και δεδομένα με θόρυβο. Οι μέθοδοι δεύτερης τάξης κάνουν χρήση μιας προσέγγισης του Εσσιανού πίνακα ή του γινομένου της προσέγγισης με κάποιο συζυγές διάνυσμα. Έτσι, δίνουν λύση στα εμπόδια που καλούμαστε να ξεπεράσουμε όταν η απλή μέθοδος SG δεν επαρκεί.

Παρόλο που μέχρι σήμερα η μέθοδος SG είναι ευρέως διαδεδομένη στη μηχανική μάθηση μεγάλης κλίμακας, η συνεχής έρευνα επάνω στο αντικείμενο δείχνει ότι υπάρχουν περιθώρια και για εναλλακτικές μεθόδους. Μέθοδοι που μειώνουν τον θόρυβο των δεδομένων, βελτιστοποιούν κανονικοποιημένα μοντέλα και χρησιμοποιούν διαφορετικές μεθόδους εύρεσης της κατεύθυνσης αναζήτησης, μελετώνται ολοένα και περισσότερο. Τα θεωρητικά τους πλεονεκτήματα συνεχίζουν να παίρνουν πραγματικές διαστάσεις με την εφαρμογή τους σε πρακτικά προβλήματα της μηχανικής μάθησης. Τέλος, παρά τα πλεονεκτήματα της μεθόδου SG, πλεόν κατασκευάζονται εναλλακτικοί αλγόριθμοι βελτιστοποίησης, οι οποίοι εκμεταλλεύονται σε μεγαλύτερο βαθμό νέες δυνατότητες των υπολογιστικών συστημάτων, όπως ο παράλληλος και κατανεμημένος υπολογισμός. Με βάση αυτά τα σχόλια, βλέπουμε ότι η μαθηματική βελτιστοποίηση στη μηχανική μάθηση μεγάλης κλίμακας αποτελεί ένα ταχύτατα αναπτυσσόμενο πεδίο έρευνας, με πολλές δυνατότητες για μεγαλύτερη εξέλιξη.

Βιβλιογραφία

[1] Lee J. Bain, Max Engelhardt. *Introduction to Probability and Mathematical Statistics*. Second Edition, Duxbury Press, 1992, California.

[2] S. Becker, Y. Le Cun. *Improving the convergence of back-propagation learning with second-order methods*. Proceedings of the 1988 Connectionist Models Summer School, pp. 29-37, 1989.

[3] Dimitri P. Bertsekas. *Nonlinear Programming*. Second Edition, Athena Scientific, 1999, Massachusetts.

[4] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Calendon Press, 1995, Oxford.

[5] Leon Bottou, Frank E. Curtis, Jorge Nocedal. *Optimization Methods in Large-Scale Machine Learning*. SIAM Review, Vol. 60, No. 2, pp. 223-311, 2018.

[6] K. L. Chung. *On a Stochastic Approximation Method*. Annuals of Mathematical Statistics, Vol. 25, No. 3, pp. 463-483, 1954.

[7] Philippe Dennery, André Krzywicki. *Mathematics for Physicists*. Dover Publications, Inc., 1995, New York.

[8] J. E. Dennis, Jr, Jorge J. Moré. *A Characterization of Superlinear Convergence and Its Application to Quasi-Newton Methods*. Mathematics of Computation, Vol. 28, No. 126, pp. 549-560, 1974.

[9] Pilip E. Gill, Walter Murray, Margaret H. Wright. *Practical Optimization*. Academic Press, 1997, London.

[10] Ian Goodfellow, Yoshua Bengio, Aaron Courville. *Deep Learning*. The MIT Press, 2016, Massachusetts.

[11] I. Guyon, V. Vapnik, B. Boser, L. Bottou, S. A. Solla. *Structural Risk Minimization for Character Recognition*. Advances in Neural Information Processing Systems 4, pp. 471-479, 1991.

[12] Simon Haykin. *Neural Networks and Learning Machines*. Third Edition, Rev. ed. of: *Neural Networks*. 2nd ed. (1999), Pearson Prentice Hall, 2009, New Jersey.

[13] Rie Johnson, Tong Zhang. *Accelerating stochastic gradient descent using predictive variance reduction*. Advances in Neural Information Processing Systems 26, pp. 315-323, 2013.

[14] Irwin Miller, Marylees Miller. *John E. Freund's Mathematical Statistics with Applications*. Eighth Edition, Pearson Education, 2014, Essex, England.

[15] Aryan Mokhtari, Alejandro Ribeiro. *RES: Regularized stochastic BFGS algorithm*. IEEE Transactions on Signal Processing 62, pp. 6089–6104, 2014.

[16] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012, Massachusetts.

[17] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, Alexander Shapiro. *Robust Stochastic Approximation Approach to Stochastic Programming*. SIAM J. Optim., Vol. 19, No. 4, pp. 1574–1609, 2009.

[18] Jorge Nocedal, Stephen Wright. *Numerical Optimization*. Second ed., Springer, 2006, New York.

[19] Elijah Polak. *Optimization: Algorithms and Consistent Approximations*. Springer-Verlag, 1997, New York.

[20] B. T. Polyak. *Some methods of speeding up the convergence of iteration methods*. USSR Computational Mathematics and Mathematical Physics 4, pp. 1-17, 1964.

[21] Herbert Robbins, Sutton Monro. *A Stochastic Approximation Method*. Annals of Mathematical Statistics, Vol. 22, No. 3, pp. 400-407, 1951.

[22] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970, New Jersey.

[23] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998, New York.

[24] Ίλια Βόντα, Αλέξανδρος Καραγρηγορίου. *Εφαρμοσμένη Στατιστική και Στοιχεία Πιθανοτήτων*. Αναθεωρημένη Έκδοση, Ιδιωτική Έκδοση, 2002, Αθήνα.

[25] Ν. Καδιανάκης, Σ. Καρανάσιος, Α. Φελλούρης. *Ανάλυση ΙΙ: Συναρτήσεις Πολλών Μεταβλητών*. Έκδοση Ένατη, Ιδιωτική Έκδοση, 2013, Αθήνα.

[26] Γ. Κοκολάκης, Ι. Σπηλιώτης. *Εισαγωγή στις Πιθανότητες*. Εκδόσεις Συμεών, 2002, Αθήνα.

[27] Γ. Κοκολάκης, Δ. Φουσκάκης. *Στατιστική: Θεωρία και Εφαρμογές*. Εκδόσεις Συμεών, 2009, Αθήνα.

[28] Α. Μπακόπουλος, Ι. Χρυσοβέργης. *Εισαγωγή στην Αριθμητική Ανάλυση*. Εκδόσεις Συμεών, 2009, Αθήνα.