ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αλγόριθμοι Langevin και οι εφαρμογές τους στην Υπολογιστική Στατιστική

Ευαγγελία Φραντζεσκάκη

Επιβλέπων

Σαμπάνης Σωτήριος, *Καθηγητής ΣΕΜΦΕ*

Τριμελής Επιτροπή

Βόντα Φιλία	Καρώνη Χρυσηίς	Σαμπάνης Σωτήριος
Καθηγήτρια ΣΕΜΦΕ	*Καθηγήτρια ΣΕΜΦΕ*	*Καθηγητής ΣΕΜΦΕ*

Αθήνα, Ιούλιος 2023

Περιεχόμενα

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω ιδιαίτερα τον επιβλέποντα καθηγητή μου, κύριο Σωτήριο Σαμπάνη για τη καθοδήγησή του και τις πολύτιμες συμβουλές του κατά τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας καθώς και για την προθυμία του να με βοηθήσει γενικώς αλλά και όσον αφορά την συνέχεια των σπουδών μου.

Καθώς εργασία αυτή σηματοδοτεί το τέλος των προπτυχιακών σπουδών μου, θα ήθελα να ευχαριστήσω τους φίλους μου, που ήταν πάντα δίπλα μου και περάσαμε μαζί τα χρόνια αυτά.

Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου για την υπομονή και τη στήριξή τους και για το ότι είναι πάντα οι πιο θερμοί υποστηριχτές μου σε κάθε βήμα και κάθε προσπάθειά μου.

.....

Ευαγγελία Φραντζεσκάκη

Περίληψη

Στην εργασία αυτή μελετάμε τον αλγόριθμο Stochastic Gradient Langevin Dynamics (SGLD), μια μέθοδο για δειγματοληψία και βελτιστοποίηση που απορρέει από την διακριτοποίηση της στοχαστικής διαφορικής εξίσωσης Langevin.

Στο πρώτο κεφάλαιο γίνεται μια σύντομη αναφορά στη θεωρία πιθανοτήτων και στοχαστικών ανελίξεων ώστε να δοθεί το θεωρητικό υπόβαθρο και τα εργαλεία για την κατανόηση των αποδείξεων του κυρίου μέρους.

Στο δεύτερο κεφάλαιο εισάγουμε τον αλγόριθμο προς μελέτη, ξεκινώντας με την περιγραφή του όπως προτάθηκε από τους Welling & Teh. Στην συνέχεια εισαγάγουμε το πρόβλημα δειγματοληψίας από μια κατανομή χρησιμοποιώντας τον αλγόριθμο SGLD, το οποίο θα αναλύσουμε κάτω από συγκεκριμένες υποθέσεις. Σκοπός μας αρχικά είναι να λάβουμε ένα άνω όριο στην απόσταση Wasserstein για την σύγκλιση του αλγορίθμου στην κατανομή από την οποία δειγματοληπτούμε. Στην συνέχεια εξετάζουμε τον αλγόριθμο SGLD ως μέθοδο βελτιστοποίησης και χρησιμοποιώντας την σύγκλιση στην απόσταση Wasserstein που αποδείξαμε, θα λάβουμε ένα μη ασυμπτωτικό φράγμα για το expected excess risk.

Στο τρίτο κεφάλαιο θα ξεκινήσουμε με μια απλή εφαρμογή του αλγορίθμου για προσομοίωση τιμών από μια κατανομή. Στη συνέχεια θα χρησιμοποιήσουμε τον αλγόριθμο ως μέθοδο βελτιστοποίησης για τον υπολογισμό των ελαχίστων σημείων κυρτών συναρτήσεων και τέλος θα κάνουμε μια εφαρμογή σε ένα μοντέλο γραμμικής παλινδρόμησης.

Λέξεις-Κλειδιά: Stochastic Gradient Langevin Dynamics, Στοχαστική Διαφορική Εξίσωση Langevin, Δειγματοληψία, Κυρτή βελτιστοποίηση, Συνθήκη Lipschitz, Απόσταση Wasserstein.

Abstract

In this work, we study the Stochastic Gradient Langevin Dynamics (SGLD) algorithm, which is a method form sampling and optimization that arises from the discretization of the so-called overdamped Langevin stochastic differential equation (SDE).

In the first chapter, a brief analysis is conducted on probability theory and stochastic processes in order to provide the theoretical background and tools for understanding the proofs of the main part.

In the second chapter, we introduce the algorithm under study, starting with its description as proposed by Welling & Teh. We then introduce the problem of sampling from a distribution using the SGLD algorithm, which we will analyze under specific assumptions. Our initial goal is to establish an upper bound on the Wasserstein distance between the target distribution π and its approximations $(Law(\theta_n^\lambda))_{n\in\mathbb{N}}$ generated by the SGLD algorithm. We then examine the SGLD algorithm as an optimization method, and using the Wasserstein-2 convergence result we establish a non asymptotic error bound for the expected excess risk.

In the third chapter, we will begin with some simple applications of the algorithm for sampling from some known distributions. We will then use the algorithm as an optimization method to compute the minimum points of a convex function, and finally, we will apply it to a linear regression model.

Key-Words: Stochastic Gradient Langevin Dynamics, Langevin Stochastic Differential Equation, Sampling , Convex optimization, Lipschitz condition, Wasserstein distance.

1 Μαθηματική Εισαγωγή

Στην πρώτη ενότητα θα επιχειρήσουμε να προσφέρουμε στον αναγνώστη μία γρήγορη και στοχευμένη ανασκόπηση των μαθηματικών εννοιών που απαιτούνται για την θεμελίωση και κατανόηση του κύριου μέρους της εργασίας. Για περαιτέρω ανάγνωση και πηγές, παραπέμπουμε τον αναγνώστη στα [10],[15],[17] και [18].

1.1 σ-άλγεβρες και Χώροι Πιθανότητας

Ορισμός 1.1.1. *Έστω Ω ένα σύνολο. Μια* **σ-άλγεβρα** *$\mathcal{F}$ πάνω στο Ω είναι μια οικογένεια συνόλων του Ω με τις παρακάτω ιδιότητες:*

(i) $\emptyset \in \mathcal{F}$

(ii) *Αν* $A \in \mathcal{F}$, *τότε* $A^c \equiv \Omega \backslash F \in \mathcal{F}$

(iii) *Αν* $A_1, A_2, \cdots \in \mathcal{F}$, *τότε* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Ορισμός 1.1.2. *Έστω X σύνολο και $\mathcal{C} \subset \mathcal{P}(X)$. Το $\mathcal{C}$ δεν είναι απαραίτητα σ-άλγεβρα. Ορίζουμε*

$$\mathcal{J}(\mathcal{C}) \coloneqq \{\mathcal{A} \subset \mathcal{P}(X) : \mathcal{C} \subset \mathcal{A} \text{ και } \mathcal{A} \text{ σ-άλγεβρα}\},$$

δηλαδή το σύνολο των σ-αλγεβρών στο X που κάθε μια τους περιέχει την οικογένεια $\mathcal{C}$. Η σ-άλγεβρα που παράγεται από την $\mathcal{C}$ ορίζεται ως η τομή όλων των σ-αλγεβρών που περιέχουν την $\mathcal{C}$ και συμβολίζεται με $\sigma(\mathcal{C})$, δηλαδή

$$\sigma(\mathcal{C}) = \bigcap_{\mathcal{A} \in \mathcal{J}(\mathcal{C})} \mathcal{A}.$$

Η $\sigma(\mathcal{C})$ είναι η μικρότερη σ-άλγεβρα που περιέχει την οικογένεια $\mathcal{C}$ και θα ονομάζεται η **ελάχιστη σ-άλγεβρα** *που παράγεται από την $\mathcal{C}$.*

Ορισμός 1.1.3. *Η* **άλγεβρα Borel** *την οποία συμβολίζουμε με $\mathcal{B}$ είναι η ελάχιστη σ-άλγεβρα που περιέχει την κλάση C όλων των διαστημάτων της μορφής $(-\infty, x)$, τα οποία μπορεί να θεωρηθούν ως υποσύνολα της πραγματικής ευθείας. Τα στοιχεία της $\mathcal{B}$ ονομάζονται σύνολα Borel.*

Η $\mathcal{B}$ ισούται με την κλάση ισοδυναμίας όλων των διαστημάτων της μορφής (a, b). Περιέχει επίσης όλα τα υποσύνολα που περιέχουν μόνο ένα σημείο (singletons) και μετρήσιμες ενώσεις τέτοιων υποσυνόλων. Έτσι το $\mathcal{B}$ περιέχει πρακτικά όλα τα υποσύνολα του $\mathbb{R}$ τα οποία μας ενδιαφέρουν.

Ορισμός 1.1.4. *Η* ***σ-άλγεβρα Borel*** $\mathcal{B}(\mathbb{R}^{\mathrm{d}})$ *είναι η ελάχιστη σ-άλγεβρα η οποία περιέχει όλα τα παραλληλόγραμμα της μορφής* $(a, b]$ *ή με άλλα λόγια η σ-άλγεβρα που παράγεται από τα παραλληλόγραμμα.*

Κατά αναλογία με τα προηγούμενα η $\mathcal{B}(\mathbb{R}^d)$ περιέχει πρακτικά όλα τα υποσύνολα του $\mathbb{R}^d$ που μας απασχολούν.

Ορισμός 1.1.5. *Έστω ένα σύνολο* Ω *και μια σ-άλγεβρα* $\mathcal{F}$ *που αποτελείται από υποσύνολα του* Ω*. Το ζεύγος* $(\Omega, \mathcal{F})$ *ονομάζεται* ***μετρήσιμος χώρος****.*

Ορισμός 1.1.6. *Ένα* ***μέτρο πιθανότητας*** P *πάνω σε ένα μετρήσιμο χώρο* $(\Omega, \mathcal{F})$ *είναι μια απεικόνιση* $P : \mathcal{F} \to [0, 1]$ *με τις ιδιότητες:*

(i) $P(\emptyset) = 0$, $P(\Omega) = 1$

(ii) Αν $A_1, A_2, \cdots \in \mathcal{F}$ *και τα* $\{A_i\}$ *ανά δύο ξένα, τότε*

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Αν παραλείψουμε την συνθήκη $P(\Omega) = 1$ *τότε λέμε ότι η* P *είναι ένα* ***μέτρο*** *και όχι ένα μέτρο πιθανότητας.*

Ορισμός 1.1.7. *Έστω ο μετρήσιμος χώρος* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$*.* ***Μέτρο Lebesgue*** *στο* $\mathbb{R}$ *θα λέμε το μέτρο* λ *στον χώρο* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *για το οποίο ισχύει*

$$\lambda(I) = length(I)$$

για κάθε διάστημα $I \subset \mathbb{R}$*.*

Ορισμός 1.1.8. *Έστω ένα σύνολο* Ω*, μια σ-άλγεβρα* $\mathcal{F}$ *πάνω σε αυτό και ένα μέτρο πιθανότητας* P*. Η τριάδα* $(\Omega, \mathcal{F}, P)$ *ονομάζεται* ***χώρος πιθανότητας****.*

1.2 Τυχαίες μεταβλητές και Κατανομή τυχαίας μεταβλητής

Ορισμός 1.2.1. *Έστω* $(\Omega, \mathcal{F})$ *και* $(E, \mathcal{E})$ *μετρήσιμοι χώροι. Μια συνάρτηση* $f : \Omega \to E$ *λέγεται* $\mathcal{F}/\mathcal{E}$*-μετρήσιμη (ή απλώς* $\mathcal{F}$*-μετρήσιμη) αν*

$$f^{-1}(A) = \{\omega \in \Omega : f(\omega) \in A\} \in \mathcal{F} \quad \text{για κάθε } A \in \mathcal{E}.$$

Με άλλα λόγια λέμε ότι η συνάρτηση f είναι $\mathcal{F}$-μετρήσιμη αν η αντίστροφη εικόνα ενός υποσυνόλου του E, κάτω από την συνάρτηση αυτή, ανήκει στην σ-άλγεβρα $\mathcal{F}$. Για να απαντήσουμε λοιπόν την ερώτηση αν η συνάρτηση f παίρνει τιμή στο A θα πρέπει να έχουμε στην διάθεσή μας την *πληροφορία* που περιέχεται στην $\mathcal{F}$.

Ορισμός 1.2.2. *Σε έναν χώρο πιθανότητας $(\Omega, \mathcal{F}, P)$ μια $\mathcal{F}$-μετρήσιμη συνάρτηση $X : \Omega \to \mathbb{R}^d$ λέγεται (πραγματική)* ***τυχαία μεταβλητή****.*

Μπορούμε να θεωρήσουμε την τυχαία μεταβλητή X σαν μια μεταβλητή που η τιμή της εξαρτάται από την έκβαση ενός τυχαίου πειράματος. Ο ορισμός μας λέει ότι για κάθε κατάσταση παίρνουμε ένα πραγματικό διάνυσμα $X \in \mathbb{R}^d$ και για να απαντήσουμε την ερώτηση τι τιμή μπορεί να πάρει η τυχαία μεταβλητή X θα πρέπει να έχουμε την πληροφορία σχετικά με τις εκβάσεις του πειράματος που περιέχονται στην σ-άλγεβρα $\mathcal{F}$.

Ορισμός 1.2.3. *Η μικρότερη σ-άλγεβρα για την οποία η τυχαία μεταβλητή X είναι μετρήσιμη, αποκαλείται σ σ-άλγεβρα που παράγεται από την τυχαία μεταβλητή X και συμβολίζεται με $\sigma(X)$.*

Ορισμός 1.2.4. *Έστω $(\Omega, \mathcal{F}, P)$ χώρος πιθανότητας, $(E, \mathcal{E})$ μετρήσιμος χώρος και $X : \Omega \to E$ τυχαία μεταβλητή. Το μέτρο πιθανότητας*

$$P^X(B) \coloneqq P(X^{-1}(B)) = P(X \in B)$$

για κάθε $B \in \mathcal{E}$ λέγεται ***κατανομή της X.***

Ορισμός 1.2.5. *Έστω P ένα μέτρο πιθανότητας στον $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, λ το μέτρο Lebesgue και $f : \mathbb{R} \to [0, \infty]$ μια Borel-μετρήσιμη συνάρτηση. Η f λέγεται* ***πυκνότητα*** *του P αν*

$$P(A) = \int_A f(x) d\lambda(x) \quad \text{για κάθε} \quad A \in \mathcal{B}(\mathbb{R}).$$

Ορισμός 1.2.6. *Έστω $(\Omega, \mathcal{F}, P)$ χώρος πιθανότητας, $X : \Omega \to \mathbb{R}$ τυχαία μεταβλητή και $f : \mathbb{R} \to [0, \infty]$ μια Borel-μετρήσιμη συνάρτηση. Λέμε ότι η f είναι μια* ***πυκνότητα*** *της τυχαίας μεταβλητής X αν είναι πυκνότητα της κατανομής P^X της X.*

1.3 Ανεξαρτησία

Όπως θυμόμαστε από την βασική θεωρία πιθανοτήτων δύο γεγονότα A και B λέγονται ανεξάρτητα αν $P(A \cap B) = P(A)P(B)$. Διαισθητικά λέμε ότι δύο γεγονότα είναι ανεξάρτητα όταν το ένα δεν επηρεάζει το άλλο.

Η έννοια της ανεξαρτησίας μπορεί να οριστεί και για περισσότερα των δύο γεγονότων. Μια άπειρη συλλογή γεγονότων $(A_n)_{n\in I}$ είναι μια ανεξάρτητη συλλογή αν για κάθε πεπερασμένο υποσύνολο J του I ισχύει $P(\bigcap_{n\in J} A_n) = \prod_{n\in J} P(A_n)$. Παρατηρούμε ότι για να έχουμε ανεξαρτησία περισσοτέρων των δύο γεγονότων, φαίνεται από τον ορισμό ότι δεν είναι αρκετό να έχουμε ανεξαρτησία ανά δύο για κάθε ζεύγος.

Ορισμός 1.3.1. ***(Ανεξάρτητες σ-άλγεβρες)*** *Οι υπο-σ-άλγεβρες $\mathcal{F}_i, i \in I$ της $\mathcal{F}$ ονομάζονται ανεξάρτητες αν για κάθε υποσύνολο J του I και κάθε σύνολο $A_i \in \mathcal{F}_i$ έχουμε*

$$P(\bigcap_{n\in J} A_n) = \prod_{n\in J} P(A_n)$$

Ορισμός 1.3.2. ***(Ανεξάρτητες τυχαίες μεταβλητές)*** *Οι τυχαίες μεταβλητές $X_1, X_2, ...$ ονομάζονται ανεξάρτητες αν οι σ-άλγεβρες που παράγονται από αυτές είναι ανεξάρτητες.*

1.4 Μέση τιμή και Διασπορά

Ορισμός 1.4.1. *Έστω χώρος πιθανότητας $(\Omega, \mathcal{F}, P)$ και τυχαία μεταβλητή $X : \Omega \to \mathbb{R}$. Ορίζουμε την* ***μέση τιμή*** *της X ως το ολοκλήρωμα Lebesgue της X και τη συμβολίζουμε με $E[X]$. Δηλαδή,*

$$E[X] \coloneqq \int_\Omega X dP = \int_\Omega X(\omega) P(d\omega)$$

Με τον συμβολισμό $E_P[X]$ θα εννοούμε ότι η μέση τιμή υπολογίζεται ως προς το μέτρο P. Επίσης, αν $E[X] < \infty$, θα λέμε ότι η X είναι ολοκληρώσιμη.

Ορισμένες φορές μας ενδιαφέρει η μέση τιμή μιας τυχαίας μεταβλητής πάνω σε ένα υποσύνολο A του συνόλου γεγονότων. Αυτό θα το συμβολίζουμε ως $E[X; A] \coloneqq \int_A X dP(\omega)$.

Θα δούμε τώρα κάποιες βασικές ιδιότητες της μέσης τιμής.

Ιδιότητες της μέσης τιμής
Έστω $(\Omega, \mathcal{F}, P)$ χώρος πιθανότητας $X, Y : \Omega \to \mathbb{R}$ τυχαίες μεταβλητές με καλά ορισμένη μέση τιμή.

1. Γραμμικότητα : $E[c_1X + c_2Y] = c_1E[X] + c_2E[Y]$ για κάθε $c_1, c_2 \in \mathbb{R}$.

2. Αν $X \leq Y$ σ.β. τότε $E[X] \leq E[Y]$.

3. Αν $X = c \in \mathbb{R}$ σ.β. τότε $E[X] = c$.

4. Αν $X \geq 0$ τότε $E[X] = 0$ αν και μόνο αν $X = 0$ σχεδόν παντού ως προς το μέτρο P.

Ορισμός 1.4.2. *Έστω $(\Omega, \mathcal{F}, P)$ χώρος πιθανότητας και τυχαία μεταβλητή $X : \Omega \to \mathbb{R}$ με $E[X] < \infty$. Ορίζουμε την* **Διασπορά** *της X ως:*

$$Var[X] \coloneqq E[(X - E[X])^2]$$

Η μέση τιμή είναι πραγματικός αριθμός καθώς $E|X| < \infty$. Η διασπορά όμως ενδέχεται να παίρνει την τιμή ∞. Ένας χρήσιμος τύπος για τη διασπορά, που προκύπτει από τον ορισμό της, είναι $Var[X] = E[X^2] - E[X]^2$. Έτσι βλέπουμε ότι αν $E[X^2] < \infty$ τότε $Var[X] < \infty$. Η διασπορά είναι ένα μέτρο της μεταβλητότητας της τυχαίας μεταβλητής γύρω από τη μέση της τιμή. Έτσι, όταν $Var[X] = 0$, αναμένουμε η X να είναι συγκεντρωμένη στη μέση τιμή.

Οι χώροι $\mathcal{L}_p$ με $p \in [1, \infty)$

Ορισμός 1.4.3. *Έστω $(\Omega, \mathcal{F}, P)$ χώρος πιθανότητας, τυχαία μεταβλητή $X : \Omega \to \mathbb{R}$ και $p \in [1, \infty)$. Ορίζουμε*

$$\|X\|_p \coloneqq \left(E[X^p]\right)^{1/p}$$

και

$$\mathcal{L}_p(\Omega, \mathcal{F}, P) \coloneqq \{X | X : \Omega \to \mathbb{R} \quad \text{τυχαία μεταβλητή και} \quad \|X\|_p < \infty\}.$$

Όταν είναι σαφές ποιος είναι ο χώρος Ω, η σ-άλγεβρα $\mathcal{F}$ και το μέτρο P, τότε θα γράφουμε απλώς $\mathcal{L}_p$.

1.5 Δεσμευμένη μέση τιμή

Ορισμός 1.5.1. *Έστω $(\Omega, \mathcal{F}_0, P)$ χώρος πιθανότητας, μια σ-άλγεβρα $\mathcal{F} \subset \mathcal{F}_0$ και μια τυχαία μεταβλητή $X \in \mathcal{F}_0$ για την οποία ισχύει $E|X| < \infty$. Τότε μπορούμε να ορίσουμε την τυχαία μεταβλητή $Y \coloneqq E[X|\mathcal{F}]$ με τις ιδιότητες*

(i) $Y \in \mathcal{F}$ *(δηλαδή είναι $\mathcal{F}$-μετρήσιμη).*

(ii) $\int_A X dP = \int_A Y dP$ *για κάθε $A \in \mathcal{F}$.*

Η τυχαία μεταβλητή Y ονομάζεται ***η δεσμευμένη μέση τιμή της τυχαίας μεταβλητής*** *X ως προς την σ-άλγεβρα $\mathcal{F}$.*

Παρατηρήσεις:

1. Έστω $\mathcal{F} = \{\emptyset, \Omega\}$ η τετριμμένη σ-άλγεβρα. Τότε $E[X|\mathcal{F}] = E[X]$. Αυτό έρχεται σε συμφωνία με την διαισθητική μας αντιμετώπιση του θέματος της δεσμευμένης μέσης τιμής. Εφόσον η τετριμμένη σ-άλγεβρα μας προσφέρει το ελάχιστο της πληροφορίας, τότε με την δεσμευμένη μέση τιμή κάποιας μεταβλητής επάνω στην άλγεβρα αυτή απλά λαμβάνουμε την μέση τιμή της μεταβλητής αυτής.

2. Έστω $\mathcal{F} = \sigma(X)$ η σ-άλγεβρα η οποία παράγεται από την τυχαία μεταβλητή X. Τότε $E[X|\mathcal{F}] = X$. Αυτό προκύπτει από την ορισμό της δεσμευμένης μέσης τιμής αφού $\int\limits_A X dP = \int\limits_A X dP$ για κάθε $A \in \mathcal{F}$. Χρησιμοποιώντας την ερμηνεία της σ-άλγεβρας σαν πληροφορία μπορούμε να ερμηνεύσουμε την $\mathcal{F}$ σαν την μέγιστη πληροφορία που έχουμε στην διάθεσή μας για την τυχαία μεταβλητή X. Άρα χρησιμοποιώντας την μέγιστη αυτή πληροφορία μπορούμε να ανακτήσουμε μέσω της δεσμευμένης μέση τιμής επάνω στην $\mathcal{F}$ την ίδια την τυχαία μεταβλητή.

Ιδιότητες Δεσμευμένης Μέσης Τιμής
Έστω X, Y τυχαίες μεταβλητές με $E|X|, E|Y| < \infty$, $\mathcal{F}$ σ-άλγεβρα και $\mathcal{G}, \mathcal{H}$ υπό-σ-άλγεβρες της $\mathcal{F}$.

(i) Αν $Y = E[X \mid \mathcal{G}]$ τότε $E[X] = E[Y]$.

(ii) *(Γραμμικότητα)* $E[c_1X + c_2Y \mid \mathcal{G}] = c_1E[X \mid \mathcal{G}] + c_2E[Y \mid \mathcal{G}]$ για κάθε $c_1, c_2 \in \mathbb{R}$.

(iii) Αν $X \geq 0$, τότε $E[X \mid \mathcal{G}] \geq 0$ σ.β.

(iv) *(Ανισότητα Jensen για την δεσμευμένη μέση τιμή)* Αν $c : \mathbb{R} \to \mathbb{R}$ κυρτή συνάρτηση και $E[c(X)] < \infty$ τότε

$$E[c(X) \mid \mathcal{G}] \geq c(E[X \mid \mathcal{G}]), \quad \text{σ.β.}$$

Από την ανισότητα Jensen προκύπτει η παρακάτω σημαντική ιδιότητα:

$$\|E[X \mid \mathcal{G}]\|_p \leq \|X\|_p, \quad \text{για κάθε} \quad p \geq 1.$$

(v) *(Ιδιότητα Πύργου)* Αν $\mathcal{H}$ υπό σ-άλγεβρα της $\mathcal{G}$ τότε

$$E[E[X \mid \mathcal{G}] \mid \mathcal{H}] = E[X \mid \mathcal{H}].$$

Επίσης $E[E[X \mid \mathcal{H}] \mid \mathcal{G}] = E[X \mid \mathcal{H}]$. Με άλλα λόγια βλέπουμε ότι η μικρότερη σ-άλγεβρα (δηλαδή η μικρότερη πληροφορία) υπερισχύει.

(vi) Αν $X \in \mathcal{G}$ και $E|X|, E|XY| < \infty$ τότε

$$E[XY \mid \mathcal{G}] = XE[Y \mid \mathcal{G}].$$

Δηλαδή μπορούμε να βγάλουμε έξω από την δεσμευμένη μέση τιμή ότι είναι γνωστό.

(vii) Αν η τυχαία μεταβλητή X είναι ανεξάρτητη από την σ-άλγεβρα $\mathcal{F}$ τότε

$$E[X \mid \mathcal{F}] = E[X].$$

Αυτό σημαίνει ότι καθώς η X είναι ανεξάρτητη της $\mathcal{F}$, η πληροφορία που περιέχεται στην $\mathcal{F}$ δεν χρησιμεύει για την καλύτερη πρόβλεψη της X. Είναι το ίδιο δηλαδή σαν να χρησιμοποιούμε την τετριμμένη σ-άλγεβρα $\{\emptyset, \Omega\}$.

1.6 Στοχαστικές Διαδικασίες και Μαρκοβιανές αλυσίδες

Ορισμός 1.6.1. *Μια* ***στοχαστική διαδικασία*** *είναι μια συλλογή τυχαίων μεταβλητών $\{X_t\}_{t\in T}$ οι οποίες ορίζονται σε έναν χώρο πιθανότητας $(\Omega, \mathcal{F}, P)$ και παίρνουν τιμές στο $\mathbb{R}^d$.*

Μια στοχαστική διαδικασία έχει δύο μεταβλητές, την t και την ω.

- Για κάθε $t \in T$ του συνόλου δεικτών T έχουμε μια τυχαία μεταβλητή

$$\omega \to X_t(\omega), \ \omega \in \Omega.$$

- Θεωρώντας σταθερό το $\omega \in \Omega$ θεωρούμε την συνάρτηση

$$t \to X_t(\omega), \ t \in T,$$

η οποία ονομάζεται τροχιά της X_t.

Το σύνολο δεικτών T συνήθως αντιπροσωπεύει χρονικές στιγμές ή επαναλήψεις σε ένα πείραμα. Αν το T είναι αριθμήσιμο ($T = \mathbb{N}_0 = \{0, 1, 2, ...\}$) λέμε ότι η στοχαστική διαδικασία $\{X_t\}_{t\in T}$ είναι διακριτού χρόνου ενώ αν το T είναι υπεραριθμήσιμο ($T = [0, +\infty)$) λέμε ότι είναι συνεχούς χρόνου.

Ορισμός 1.6.2. ***Φιλτράρισμα (filtration) ή διήθηση*** *στον χώρο πιθανότητας $(\Omega, \mathcal{F}, P)$ ονομάζουμε μια αύξουσα οικογένεια $(\mathcal{F}_t)_{t\geq 0}$ υπο-σ-αλγεβρών της $\mathcal{F}$. Δηλαδή, για κάθε t, η $\mathcal{F}_t$ είναι μια σ-άλγεβρα που περιέχεται στην $\mathcal{F}$ και αν $s \leq t$, τότε $\mathcal{F}_s \subset \mathcal{F}_t$.*

Ορίζουμε επίσης

$$\mathcal{F}_\infty := \sigma\Big(\bigcup_n \mathcal{F}_n\Big) \subset \mathcal{F}.$$

Ο χώρος πιθανότητας $(\Omega, \mathcal{F}, P)$ *μαζί με την διήθηση* $(\mathcal{F}_t)_{t\geq 0}$ *θα καλείται διηθημένος χώρος πιθανότητας* $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$.

Η σ-άλγεβρα $\mathcal{F}_t$ μπορεί να θεωρηθεί σαν την πληροφορία η οποία είναι διαθέσιμη μέχρι την χρονική στιγμή t. Με βάση αυτό μπορούμε να θεωρήσουμε μια διήθηση σαν μια αυξανόμενη δομή πληροφορίας καθώς περνάει ο χρόνος. Μια αρκετά συνηθισμένη έννοια είναι η έννοια της **φυσικής διήθησης**. Αυτή είναι η διήθηση η οποία παράγεται από μια στοχαστική διαδικασία X_t. Όσο περνάει ο χρόνος και παρατηρούμε την εν λόγω στοχαστική διαδικασία, τόσο αυξάνει και η πληροφορία που έχουμε στη διάθεσή μας για την διαδικασία αυτή.

Ορισμός 1.6.3. *Λέμε πως μια στοχαστική διαδικασία* $\{X_t\}_{t\geq 0}$ *είναι* ***προσαρμοσμένη*** *στη διήθηση* $(\mathcal{F}_t)_{t\geq 0}$*, αν και μόνο αν, η τυχαία μεταβλητή* X_t *είναι* $\mathcal{F}_t$*-μετρήσιμη για κάθε* $t \geq 0$.

Αυτό σημαίνει ότι όλη η πληροφορία η οποία αφορά την στοχαστική μεταβλητή X_t μέχρι την χρονική στιγμή t, περιέχεται στην σ-άλγεβρα $\mathcal{F}_t$. Από τον ορισμό της φυσικής διήθησης μπορούμε να δούμε ότι μια στοχαστική διαδικασία είναι προσαρμοσμένη στην φυσική της διήθηση.

Θα δούμε τώρα την κίνηση Brown, μια από τις πιο σημαντικές στοχαστικές διαδικασίες καθώς αποτελεί τον κατεξοχήν τρόπο να μοντελοποιήσουμε την τυχαιότητα.

Ορισμός 1.6.4. *Η* ***κίνηση Brown*** *είναι μια στοχαστική διαδικασία* $(B_t)_{t\geq 0}$ *ορισμένη σε ένα χώρο πιθανότητας χώρος πιθανότητας* $(\Omega, \mathcal{F}, P)$ *η οποία παίρνει τιμές στο* $\mathbb{R}$ *και έχει τις ακόλουθες ιδιότητες:*

(i) Αν $t_0 < t_1 < ... < t_n$ *τότε οι τυχαίες μεταβλητές* $B_{t_0}, B_{t_1} - B_{t_0}, ..., B_{t_n} - B_{t_{n-1}}$ *είναι ανεξάρτητες (ανεξάρτητες προσαυξήσεις).*

(ii) Αν $s, t \geq 0$ *τότε*

$$P(B_{s+t} - B_s \in A) = \int_A \frac{1}{(2\pi t)^{1/2}} exp\big(-\frac{x^2}{2t}\big)$$

όπου A *κάποιο σύνολο Borel.*

Δηλαδή οι μεταβολές της κίνησης Brown είναι κατανεμημένες με την κανονική κατανομή.

(iii) Οι τροχιές της κίνησης Brown είναι συνεχείς με πιθανότητα 1, δηλαδή η $t \to B_t$ είναι συνεχής συνάρτηση.

Οι τρεις αυτές ιδιότητες ορίζουν μια και μοναδική στοχαστική διαδικασία.

Στις περισσότερες μοντελοποιήσεις αυτό που μας ενδιαφέρει είναι η τρέχουσα κατάσταση του συστήματος και όχι το πώς το σύστημα βρέθηκε σε αυτή την κατάσταση. Τα στοχαστικά συστήματα που έχουν αυτή την ιδιότητα χαρακτηρίζονται ως μαρκοβιανά. Θα ορίσουμε λοιπόν τις μαρκοβιανές αλυσίδες και θα δούμε κάποιες βασικές ιδιότητές τους.

Ορισμός 1.6.5. *Έστω $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in T})$ χώρος με διήθηση και $\{X_n\}_{n\in\mathbb{N}_0}$ στοχαστική διαδικασία με τιμές στον μετρήσιμο,αριθμήσιμο χώρο $(E, \mathcal{E})$. Θα λέμε ότι η $X = (X_n, \mathcal{F}_n)_{n\geq 0}$ είναι μια* ***Αλυσίδα Markov*** *αν και μόνο αν:*

1. *Η τυχαία μεταβλητή X_n είναι $\mathcal{F}_n$-μετρήσιμη για κάθε $n \in \mathbb{N}_0$.*
2. *$P(X_{n+1} \in A \mid \mathcal{F}_n) = P(X_{n+1} \in A \mid X_n)$ σ.β. για κάθε $n \geq 0$ και $A \in \mathcal{E}$.*

Με άλλα λόγια μια στοχαστική διαδικασία $\{X_n\}_{n\in\mathbb{N}_0}$ λέγεται *μαρκοβιανή αλυσίδα* αν για κάθε $n \in \mathbb{N}$, η δεσμευμένη κατανομή της X_{n+1} δοθέντων των $(X_0, ..., X_n)$, ταυτίζεται με τη δεσμευμένη κατανομή της X_{n+1} με μόνη δοθείσα την X_n.

Ορισμός 1.6.6. *Έστω $(\mathbb{X}, \mathcal{S})$ μετρήσιμος χώρος και έστω $\{X_n\}_{n\in\mathbb{N}_0}$ μαρκοβιανή αλυσίδα με χώρο καταστάσεων τον $(\mathbb{X}, \mathcal{S})$. Στις περισσότερες μαρκοβιανές αλυσίδες, ο κανόνας $P(X_{n+1} = \cdot | X_n = \cdot)$ που περιγράφει την εξέλιξη της αλυσίδας, δεν εξαρτάται από τη χρονική παράμετρο n. Λέμε τότε ότι η αλυσίδα είναι* ***χρονικά ομοιογενής*** *και ορίζουμε*

$$p : \mathbb{X} \times \mathbb{X} \to [0, 1] \text{ με τύπο } p(X_n, B) = P(X_{n+1} \in B \mid \mathcal{F}_n).$$

Οι $p(X_n, B)$ θα λέγονται ***πιθανότητες μετάβασης*** *της Μαρκοβιανής αλυσίδας. Πιο απλά μπορούμε να γράφουμε $p(x, y) = P(X_{n+1} = y \mid X_n = x)$*

Ορισμός 1.6.7. *Έστω χρονικά ομοιογενής μαρκοβιανή αλυσίδα με πεπερασμένο χώρο καταστάσεων $\mathbb{X} = \{x_1, ..., x_N\}$. Η συλλογή $P = \{p(x, y)\}_{x,y\in\mathbb{X}}$ με $p_{ij} = p(x_i, x_j) = P(X_{n+1} = x_j | X_n = x_i)$ ονομάζεται* ***πίνακας πιθανοτήτων μετάβασης*** *της αλυσίδας.*

Για ολοκληρώσουμε την περιγραφή μιας Μαρκοβιανής αλυσίδας χρειαζόμαστε την αρχική της κατανομή.

Ορισμός 1.6.8. *Έστω $\{X_n\}_{n\in\mathbb{N}_0}$ ομογενής μαρκοβιανή αλυσίδα με χώρο καταστάσεων τον $(\mathbb{X}, \mathcal{S})$. Η κατανομή $\pi_0 : \mathbb{X} \to [0,1]$, $\pi_0 = \pi_0(B), B \in \mathbb{X}$:*

$$\pi_0(B) = P(X_0 \in B) = P(X_0 = x) = \pi_0(x).$$

θα λέγεται ***αρχική κατανομή*** *της μαρκοβιανής αλυσίδας.*

Για να υπολογίζουμε την κατανομή της κατάστασης της αλυσίδας X_n, για κάθε $n \in \mathbb{N}$, μπορούμε να αναγάγουμε τον υπολογισμό της κατανομής της X_{n+1} στον υπολογισμό της κατανομής της X_n, χρησιμοποιώντας τον τύπο της ολικής πιθανότητας.

Για $n \in \mathbb{N}_0$ συμβολίζουμε με π_n τη συνάρτηση μάζας πιθανότητας της X_n, τότε για κάθε $y \in \mathbb{X}$ έχουμε

$$\begin{aligned}
\pi_{n+1}(y) &= P(X_{n+1} = y) \\
&= \sum_{x\in\mathbb{X}} P(X_n = x)P(X_{n+1} = y \mid X_n = x) \\
&= \sum_{x\in\mathbb{X}} \pi_n(x)p(x,y) \\
&= \pi_n(x)P.
\end{aligned}$$

Επαγωγικά προκύπτει ότι $\pi_n = \pi_0 P^n$.

Ορισμός 1.6.9. ***Στάσιμη (ή Αναλλοίωτη) Κατανομή*** *της αλυσίδας $\{X_n\}_{n\in\mathbb{N}_0}$ με πίνακα μετάβασης P λέγεται η κατανομή π για την οποία ισχύει:*

$$\pi = \pi P, \ \ \delta\eta\lambda\alpha\delta\acute{\eta} \ \ \pi(x) = \sum_{y\in\mathbb{X}} \pi(y)p(y,x), \ \ \gamma\iota\alpha \ \kappa\acute{\alpha}\theta\epsilon \ \ x \in \mathbb{X}.$$

2 On Stochastic Gradient Langevin Dynamics

Τα τελευταία χρόνια έχει σημειωθεί ραγδαία αύξηση μεγάλων συνόλων δεδομένων μηχανικής μάθησης, που κυμαίνονται από internet traffic data και δεδομένα δικτύου έως δεδομένα για computer vision και βιοπληροφορική. Το ενδιαφέρον της μηχανικής μάθησης έχει τώρα στραφεί σε αυτά τα μεγάλης κλίμακας δεδομένα, τα οποία δίνουν την ευκαιρία για εκπαίδευση μοντέλων για την επίλυση πολλών εφαρμοσμένων προβλημάτων. Οι πρόσφατες επιτυχίες στη μηχανική μάθηση για δεδομένα μεγάλης κλίμακας έχουν κυρίως βασιστεί σε προσεγγίσεις βασισμένες στην βελτιστοποίηση. Παρόλο που υπάρχουν περίπλοκοι αλγόριθμοι που σχεδιάστηκαν ειδικά για συγκεκριμένους τύπους μοντέλων, ένας από τους πιο επιτυχημένους τύπους αλγορίθμων είναι οι στοχαστικοί αλγόριθμοι βελτιστοποίησης, ή αλγόριθμοι Robbins-Monro. Αυτοί οι αλγόριθμοι επεξεργάζονται μικρά πακέτα (mini-batches) δεδομένων σε κάθε επανάληψη, ανανεώνοντας τις παραμέτρους του μοντέλου παίρνοντας μικρά βήματα κλίσης σε μια συνάρτηση κόστους. Συχνά αυτοί οι αλγόριθμοι εκτελούνται σε ένα περιβάλλον online, όπου τα πακέτα δεδομένων απορρίπτονται μετά την επεξεργασία και γίνεται μόνο ένας περίπατος μέσω των δεδομένων, μειώνοντας δραστικά τις απαιτήσεις μνήμης.

Ένας τύπος μεθόδων που έχει "μείνει πίσω" από τις πρόσφατες προόδους στη μηχανική μάθηση σε δεδομένα μεγάλης κλίμακας είναι οι Μπεϋζιανές μέθοδοι. Αυτό οφείλεται εν μέρει σε κάποια αρνητικά αποτελέσματα στη Μπευζιανή εκτίμηση παραμέτρων αλλά και στο γεγονός ότι κάθε επανάληψη κλασσικών αλγορίθμων Markov chain Monte Carlo (MCMC) απαιτεί υπολογισμούς σε ολόκληρο το σύνολο δεδομένων. Ωστόσο, οι Μπεϋζιανές μέθοδοι είναι ελκυστικές λόγω της ικανότητάς τους να ποσοτικοποιούν την αβεβαιότητα στις εκτιμήσεις των παραμέτρων και να αποφεύγουν την υπερ-προσαρμογή (overfitting). Πιθανώς σε μεγάλα σύνολα δεδομένων δεν θα υπάρχει σημαντική υπερπροσαρμογή. Ωστόσο, καθώς αποκτάται πρόσβαση σε όλο και μεγαλύτερα σύνολα δεδομένων και περισσότερους υπολογιστικούς πόρους, το ενδιαφέρον στρέφεται στη δημιουργία πιο πολύπλοκων μοντέλων, οπότε θα υπάρχει πάντα η ανάγκη για ποσοτικοποίηση της αβεβαιότητας παραμέτρων.

Σε αυτό το κλίμα, οι Welling & Teh [6], πρότειναν μια μέθοδο για τη Μπεϋζιανή μάθηση από μεγάλα σύνολα δεδομένων. Η μέθοδος που πρότειναν ονομάζεται **Stochastic Gradient Langevin Dynamics (SGLD)** και αποτελεί έναν αλγόριθμο δειγματοληψίας και βελτιστοποίησης που χρησιμοποιείται ευρέως στον τομέα της μηχανικής μάθησης. Βασίζεται στον συνδυασμό δύο αλγορίθμων, τον αλγόριθμο στοχαστικής βελτιστοποίησης (Robbins & Monro, [8]) που βελτιστοποιεί στοχαστικά μια συνάρτηση πιθανοφάνειας, και στην δυναμική Langevin (Neal, [7]) που εισάγει θόρυβο στις παραμέτρους έτσι ώστε να συγκλίνουν στην εκ των προτέρων κατανομή τους. Οι δύο αυτές μέθοδοι έχουν παρόμοια δομή καθώς ανανεώνουν την παράμετρο ενδιαφέροντος θ μέσω gradient steps, όμως στόχος του αλγορίθμου στοχαστικής βελτιστοποίησης είναι η εύρεση της

maximum a posteriori (MAP) τιμή της θ, ενώ η δυναμική Langevin προσομοιώνει τιμές από την εκ των υστέρων κατανομή (posterior distribution) της θ.

Θα δούμε τώρα μια σύντομη περιγραφή των αλγορίθμων στοχαστικής βελτιστοποίησης και δυναμικής Langevin και πώς αυτοί συνδυάζονται για να προκύψει ο SGLD.

Θεωρούμε θ διάνυσμα με παραμέτρους, $p(\theta)$ η εκ των προτέρων κατανομή τους (prior distribution) και $p(x|\theta)$ η πιθανότητα του x δεδομένου ότι το μοντέλο μας έχει παραμετρικοποιηθεί από το θ. Η posterior κατανομή ενός συνόλου $X = \{x_i\}_{i=1}^N$, N δεδομένων είναι :

$$p(\theta|X) \propto \prod_{i=1}^{N} p(x_i|\theta).$$

Ο αλγόριθμος στοχαστικής βελτιστοποίησης των Robbins & Monro λειτουργεί ως εξής:

σε κάθε επανάληψη t, επιλέγεται ένα υποσύνολο $X_t = \{x_{t1}, ..., x_{tn}\}$, n στοιχείων του N και η παράμετρος θ ανανεώνεται με βάση

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{t=1}^{N} \nabla \log p(x_{ti} \mid \theta_t)\right)$$

όπου ϵ_t ακολουθία βημάτων.

Για να εξασφαλιστεί η σύγκλιση του αλγορίθμου, τα βήματα πρέπει να ικανοποιούν τις

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \qquad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty.$$

Ο αλγόριθμος που προκύπτει από την δυναμική Langevin είναι παρόμοιος με τον παραπάνω, προσθέτοντας επιπλέον γκαουσιανό θόρυβο στην παράμετρο, δηλαδή η παράμετρος θ ανανεώνεται ως

$$\Delta\theta_t = \frac{\epsilon}{2}\left(\nabla \log p(\theta_t) + \sum_{t=1}^{N} \nabla \log p(x_i \mid \theta_t)\right) + \eta_t, \quad \eta_t \sim N(0, \epsilon).$$

Ο αλγόριθμος SGLD συνδυάζει τις παραπάνω δύο προσεγγίσεις, ανανεώνοντας το θ όπως ο αλγόριθμος των Robbins & Monro και προσθέτοντας γκαουσιανό θόρυβο ενώ τα βήματα τείνουν στο μηδέν :

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{t=1}^{N} \nabla \log p(x_{ti} \mid \theta_t)\right) + \eta_t$$

$$\eta_t \sim N(0,\epsilon), \quad \sum_{t=1}^{\infty} \epsilon_t = \infty \qquad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty.$$

2.1 Εισαγωγή

Θα μελετήσουμε το πρόβλημα δειγματοληψίας από μια κατανομή π_β που ορίζεται ως

$$\pi_\beta(A) \coloneqq \int_A e^{-\beta U(\theta)} d\theta / \int_{\mathbb{R}^d} e^{-\beta U(\theta)} d\theta, \quad A \in \mathcal{B}(\mathbb{R}^d),$$

όπου $\mathcal{B}(\mathbb{R}^d)$ τα Borel σύνολα του $\mathbb{R}^d$, β μια θετική παράμετρος κλίμακας γνωστή ως παράμετρος θερμοκρασίας (inverse temperature parameter) και $U : \mathbb{R}^d \to \mathbb{R}^+$ αυστηρά κυρτή, συνεχώς παραγωγίσιμη συνάρτηση.

Έστω χώρος πιθανότητας $(\Omega, \mathcal{F}, P)$. Θεωρούμε την overdamped Langevin στοχαστική διαφορική εξίσωση

$$d\theta_t = -h(\theta_t)dt + \sqrt{2\beta^{-1}} dB_t, \quad t > 0, \tag{1}$$

με (τυχαία) αρχική συνθήκη θ_0, όπου $h \coloneqq \nabla U$ και $(B_t)_{t\geq 0}$ μια d-διάστατη κίνηση Brown. Υπό κατάλληλες συνθήκες το μαρκοβιανό semigroup που σχετίζεται με την (1) είναι αντιστρέψιμο ως προς την κατανομή π_β και συγκλίνει γεωμετρικά σε αυτή. Η διακριτοποίηση της (1) με την μέθοδο Euler-Maruyama οδηγεί στο παρακάτω διακριτό σχήμα, γνωστό ως unadjusted Langevin algorithm (ULA)

$$\bar{\theta}_0^\lambda \coloneqq \theta_0, \quad \bar{\theta}_{n+1}^\lambda \coloneqq \bar{\theta}_n^\lambda - \lambda h(\bar{\theta}_n^\lambda) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}, \tag{2}$$

όπου $(\xi_n)_{n\in\mathbb{N}}$ ακολουθία ανεξάρτητων, κανονικών d-διάστατων τυχαίων μεταβλητών, $\lambda > 0$ το βήμα και $\theta_0 \in \mathbb{R}^d$ τυχαία μεταβλητή που δηλώνει την αρχική συνθήκη στις (1) και (2). Με κατάλληλες υποθέσεις για το βήμα λ και το δυναμικό U, η ομογενής Μαρκοβιανή αλυσίδα $(\bar{\theta}_n^\lambda)_{n\in\mathbb{N}}$ συγκλίνει σε μια κατανομή π_λ που διαφέρει ελάχιστα από την π_β για κατάλληλα μικρό λ.

Υιοθετούμε τώρα ένα πλαίσιο στο οποίο ο ακριβής κλίση h είναι άγνωστη, αλλά μπορούμε να παρατηρήσουμε σε κάθε επανάληψη μια αμερόληπτη εκτιμήτρια της. Έστω η $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ μετρήσιμη συνάρτηση και $X \coloneqq (X_n)_{n\in\mathbb{N}}$ μια διαδικασία στον $\mathbb{R}^m$, προσαρμοσμένη σε μια διήθηση $\mathcal{G}_n, n \in \mathbb{N}$ που ικανοποιεί την σχέση

$$h(\theta) = \mathbb{E}[H(\theta, X_n)], \quad \theta \in \mathbb{R}^d, \ n \geq 1. \tag{3}$$

όπου η $(X_n)_{n\geq 1}$ είναι μια αυστηρά στάσιμη διαδικασία. Συμβολίζοντας με μ την κατανομή της $X_n, n \geq 1$ γράφουμε

$$h(\theta) = \int H(\theta, x)\mu(dx), \tag{4}$$

Επίσης υποθέτουμε ότι τα $\theta_0, \mathcal{G}_\infty, (\xi_n)_{n\in\mathbb{N}}$ είναι ανεξάρτητα.

Για κάθε $\lambda > 0$ ορίζουμε αναδρομικά μια τυχαία διαδικασία $(\theta_n^\lambda)_{n\in\mathbb{N}} \in \mathbb{R}^d$ ως

$$\theta_0^\lambda := \theta_0, \quad \theta_{n+1}^\lambda := \theta_n^\lambda - \lambda H(\theta_n^\lambda, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}. \tag{5}$$

Το παραπάνω σχήμα δειγματοληψίας καλείται Stochastic Gradient Langevin Dynamics (SGLD) algorithm.

Προσομοιώνοντας τιμές από την κατανομή π_β και αφήνοντας το β να πάρει επαρκώς μεγάλες τιμές $\beta \to \infty$, η πυκνότητα π_β συγκεντρώνεται ασυμπτωτικά στο ελάχιστο της U (βλέπε [2]), λύνοντας έτσι το πρόβλημα βελτιστοποίησης $\min_{\theta\in\mathbb{R}^d} U(\theta)$. Έτσι η δειγματοληψία μας οδηγεί σε μια μέθοδο βελτιστοποίησης.

Κλείνοντας, παραθέτουμε τον ορισμό της απόστασης Wasserstein καθώς θα μας χρειαστεί στα θεωρήματα που θα εξετάσουμε. *Η απόσταση Wasserstein τάξης $p \geq 1$ μεταξύ δύο μέτρων πιθανότητας μ και ν στον $\mathcal{B}(\mathbb{R}^d)$ ορίζεται ως:*

$$W_p(\mu, \nu) = \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}} \|x - y\|^p d\pi(x, y) \right)^{1/p}.$$

2.2 Υποθέσεις

Υπόθεση 2.1. *Έστω $\mathcal{G}_0 := \{\emptyset, \Omega\}$. Η διαδικασία $(X_n)_{n\in\mathbb{N}}$ είναι conditionally L-mixing ως προς την $(\mathcal{G}_n, \mathcal{G}_n^+)_{n\in\mathbb{N}}$, όπου $(\mathcal{G}_n^+)_{n\in\mathbb{N}}$ φθίνουσα ακολουθία σ-αλγεβρών με $\mathcal{G}_n$ ανεξάρτητη της $\mathcal{G}_n^+$ για κάθε $n \in \mathbb{N}$. Επίσης $\|\theta_0\|_p < \infty$ για κάθε $p \geq 1$.*

$$\textit{Για } (x, \theta) \in \mathbb{R}^m \times \mathbb{R}^d,\ H(x, \theta) = [H^1(x, \theta), ..., H^d(x, \theta)]^T.$$

Υπόθεση 2.2. *Υπάρχουν σταθερές $L_1^i, L_2^i > 0$, $i \in \{1, \ldots, d\}$ έτσι ώστε για κάθε $x, x' \in \mathbb{R}^m$ και $\theta, \theta' \in \mathbb{R}^d$,*

$$|H^i(\theta, x) - H^i(\theta', x')| \leq L_1^i\|\theta - \theta'\| + L_2^i\|x - x'\|.$$

Θέτουμε $L_1 = \sum_{i=1}^d L_1^i$ και $L_2 = \sum_{i=1}^d L_2^i$ και παρατηρούμε ότι $\forall(x, \theta) \in \mathbb{R}^m \times \mathbb{R}^d$

$$\|H(\theta, x) - H(\theta', x')\| \leq L_1\|\theta - \theta'\| + L_2\|x - x'\|.$$

Από την Υπόθεση 2.1 παίρνουμε ότι $\|X_0\| \in L^r$ για κάθε $r \geq 1$. Συνεπώς από τις Υποθέσεις 2.1 και 2.2 προκύπτει ότι η $h(\theta) := \mathbb{E}[H(\theta, X_0)], \theta \in \mathbb{R}^d$, είναι καλά ορισμένη.

Από τις Υποθέσεις 2.1,2.2 προκύπτει η παρακάτω ιδιότητα:

Ιδιότητα 1. *Για κάθε* $\theta \in \mathbb{R}^d, x \in \mathbb{R}^m$

$$\|H(\theta,x)\| \leq L_1\|\theta-\theta^*\| + L_2\|x\| + H^*, \quad H^* = \sum_{i=1}^{d} |H^i(\theta^*,0)|,$$

$$\|H(\theta,x)\| \leq L_1\|\theta\| + L_2\|x\| + H_\star, \quad H_\star = \sum_{i=1}^{d} |H^i(0,0)|,$$

$$\langle \theta-\theta^*, H(\theta,x)\rangle \geq \alpha\|\theta-\theta^*\|^2 + \langle \theta-\theta^*, H(\theta^*,x)\rangle.$$

Απόδειξη.

$$\begin{aligned}\|H(\theta,x)\| &\leq \|H(\theta,x)-H(\theta^*,x)\| + \|H(\theta^*,x)-H(\theta^*,0)\| + \|H(\theta^*,0)\| \\ &\leq L_1\|\theta-\theta^*\| + L_2\|x\| + H^*.\end{aligned}$$

□

Υπόθεση 2.3. *(Μονοτονία της H) Έστω σταθερά* $\alpha > 0$. *Τότε για κάθε* $x \in \mathbb{R}^m$ *και* $\theta, \theta' \in \mathbb{R}^d$,

$$\langle \theta-\theta', H(\theta,x)-H(\theta',x)\rangle \geq \alpha\|\theta-\theta'\|^2.$$

Επίσης από τις Υποθέσεις 2.2 και 2.3 προκύπτουν οι παρακάτω σημαντικές ιδιότητες:

Ιδιότητα 2. *(συνθήκη Lipschitz για την H) Για κάθε* $\theta, \theta' \in \mathbb{R}^d$,

$$\|H(\theta,x)-H(\theta',x)\| \leq L_1\|\theta-\theta'\|.$$

Ιδιότητα 3. *Έστω σταθερά* $\alpha > 0$. *Τότε για κάθε* $\theta, \theta' \in \mathbb{R}^d$,

$$\langle \theta-\theta', h(\theta)-h(\theta')\rangle \geq \alpha\|\theta-\theta'\|^2.$$

Σύμφωνα με το [[12], Theorem 2.1.12] έχουμε ότι υπό αυτές τις υποθέσεις, για κάθε $\theta, \theta' \in \mathbb{R}^d$,

$$\langle \theta-\theta', h(\theta)-h(\theta')\rangle \geq \tilde{a}\|\theta-\theta'\|^2 + \frac{1}{\alpha+L_1}\|h(\theta)-h(\theta')\|^2 \quad \text{όπου} \quad \tilde{a} = \frac{\alpha L_1}{\alpha+L_1}.$$

Όταν οι $(X_n)_{n\in\mathbb{N}}$ είναι ανεξάρτητες και ισόνομες μπορούμε να θεωρήσουμε τις παρακάτω πιο χαλαρές υποθέσεις.

Υπόθεση 2.4. *Έστω L_1, L_2 και ρ θετικές σταθερές. Τότε για κάθε $x, x' \in \mathbb{R}^m$ και $\theta, \theta' \in \mathbb{R}^d$,*

$$\|H(\theta, x) - H(\theta', x)\| \leq L_1(1 + \|x\|)^\rho \|\theta - \theta'\|,$$
$$\|H(\theta, x) - H(\theta, x')\| \leq L_2(1 + \|x\| + \|x'\|)^\rho (1 + \|\theta\|)\|x - x'\|.$$

Από την υπόθεση 2.4 παίρνουμε την παρακάτω ιδιότητα:

Ιδιότητα 4. *(συνθήκη Lipschitz για την h) Για κάθε $\theta, \theta' \in \mathbb{R}^d$,*

$$\|h(\theta) - h(\theta')\| \leq L_1 \mathbb{E}[(1 + \|X_0\|)^\rho] \|\theta - \theta'\|.$$

Απόδειξη.

$$\begin{aligned}\|h(\theta) - h(\theta')\| &\leq \|\mathbb{E}[H(\theta, X_0)] - \mathbb{E}[H(\theta', X_0)]\| \leq \mathbb{E}[\|H(\theta, X_0) - H(\theta, X_0)\|] \\ &\leq \mathbb{E}[L_1(1 + \|X_0\|)^\rho \|\theta - \theta'\|] \leq L_1 \mathbb{E}[(1 + \|X_0\|)^\rho] \|\theta - \theta'\|\end{aligned}$$

□

Υπόθεση 2.5. *Η διαδικασία $(X_n)_{n\in\mathbb{N}}$ αποτελείται από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές και ισχύει $\|X_0\|_{2(\rho+1)} < \infty$, $\|\theta_0\|_2 < \infty$.*

Υπόθεση 2.6. *Έστω $A : \mathbb{R}^m \to \mathbb{R}^{d\times d}$ θετικά ημιορισμένη απεικόνιση δηλαδή*

$$\langle y, A(x)y \rangle \geq 0, \text{ για κάθε } x, y \in \mathbb{R}^d.$$

Τότε για κάθε $x \in \mathbb{R}^m$ και $\theta, \theta' \in \mathbb{R}^d$,

$$\langle \theta - \theta', H(\theta, x) - H(\theta', x) \rangle \geq \langle \theta - \theta', A(x)(\theta - \theta') \rangle$$

και συμβολίζουμε με $\alpha \in \mathbb{R}^+$ την μικρότερη ιδιοτιμή του πίνακα $\mathbb{E}[A(X_0)]$.

2.3 Κύρια αποτελέσματα και αποδείξεις

Στόχος μας αρχικά είναι να καθοριστεί ένα άνω όριο για την απόσταση Wasserstein-2 μεταξύ της κατανομής π και των προσεγγίσεων της $(Law(\theta_n^\lambda))_{n\in\mathbb{N}}$ που δίνει ο αλγόριθμος SGLD (5) για ανεξάρτητα σύνολα δεδομένων. Αυτό επιτυγχάνεται μέσω του θεωρήματος 2.2. Για την απόδειξη του θεωρήματος 2.2 θα χρειαστούμε το παρακάτω λήμμα (Λήμμα 2.1), το οποίο παρέχει ένα άνω όριο για τη διασπορά των προσεγγίσεων του αλγορίθμου (5). Θα χρειαστούμε επίσης το Θεώρημα 2.1 το οποίο μας δίνει τον

ρυθμό σύγκλισης των προσεγγίσεων που παράγει ο αλγόριθμος (2) στην κατανομή π_λ.

Σύμφωνα με το [Theorem 2.1.8,[12]] καθώς η $U(\theta)$ είναι αυστηρά κυρτή, θα έχει ένα μοναδικό σημείο ελαχίστου το οποίο θα συμβολίζουμε με $\theta^* \in \mathbb{R}^d$. Επομένως έχουμε $\nabla U(\theta^*) = h(\theta^*) = 0$ το οποίο θα μας χρειαστεί στην συνέχεια.

Λήμμα 2.1. *Έστω*

$$\lambda_0 \coloneqq \min(\alpha/2L_1^2\mathbb{E}[(1+\|X_0\|)^{2\rho}], 1/\alpha). \tag{6}$$

Για $\lambda \le \lambda_0$ η συνάρτηση $V_1(\theta) \coloneqq \|\theta - \theta^*\|^2$ ικανοποιεί την

$$\mathbb{E}[V_1(\theta_n^\lambda)\big|\theta_{n-1}^\lambda] \le (1-\lambda\alpha)V_1(\theta_{n-1}^\lambda) + \lambda C,$$

όπου

$$C \coloneqq 4L_2^2(1+\|\theta^*\|)^2\mathbb{E}[(1+\|X_0\|)^{2\rho+2}] + 4\{H^*\}^2 + 2d\beta^{-1}.$$

Επομένως, $\sup_{\lambda\le\lambda_0}\sup_{n\in\mathbb{N}}\mathbb{E}[V_1(\theta_n^\lambda)] < \infty$. Επιπλέον, αν $\rho = 0$ στην Υπόθεση2.4 , τότε το παραπάνω αληθεύει για $\lambda \le \min(1/2L_1, 1/\alpha)$.

Απόδειξη.

$$\begin{aligned}
\|\theta_{n+1}^\lambda - \theta^*\|^2 &= \|\theta_n^\lambda - \lambda H(\theta_n^\lambda, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1} - \theta^*\|^2 \\
&= \|\theta_n^\lambda - \theta^*\|^2 + 2\langle\theta_n^\lambda - \theta^*, -\lambda H(\theta_n^\lambda, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}\rangle \\
&\quad + \|-\lambda H(\theta_n^\lambda, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}\|^2 \\
&= \|\theta_n^\lambda - \theta^*\|^2 - 2\langle\theta_n^\lambda - \theta^*, \lambda H(\theta_n^\lambda, X_{n+1})\rangle + 2\langle\theta_n^\lambda - \theta^*, \sqrt{2\lambda\beta^{-1}}\xi_{n+1}\rangle \\
&\quad + \|\lambda H(\theta_n^\lambda, X_{n+1})\|^2 + 2\langle -\lambda H(\theta_n^\lambda, X_{n+1}), \sqrt{2\lambda\beta^{-1}}\xi_{n+1}\rangle + \|\sqrt{2\lambda\beta^{-1}}\xi_{n+1}\|^2 \\
&= \|\theta_n^\lambda - \theta^*\|^2 - 2\lambda\langle\theta_n^\lambda - \theta^*, H(\theta_n^\lambda, X_{n+1}) - H(\theta^*, X_{n+1})\rangle - 2\lambda\langle\theta_n^\lambda - \theta^*, H(\theta^*, X_{n+1})\rangle \\
&\quad + 2\langle\theta_n^\lambda - \theta^*, \sqrt{2\lambda\beta^{-1}}\xi_{n+1}\rangle + \lambda^2\|H(\theta_n^\lambda, X_{n+1})\|^2 \\
&\quad - 2\lambda\langle H(\theta_n^\lambda, X_{n+1}), \sqrt{2\lambda\beta^{-1}}\xi_{n+1}\rangle + 2\lambda\beta^{-1}\|\xi_{n+1}\|^2,
\end{aligned}$$

όπου χρησιμοποιήσαμε ότι

$$\begin{aligned}
\langle\theta_n^\lambda - \theta^*, \lambda H(\theta_n^\lambda, X_{n+1})\rangle &= \lambda\langle\theta_n^\lambda - \theta^*, H(\theta_n^\lambda, X_{n+1}) \pm H(\theta^*, X_{n+1})\rangle \\
&= \lambda\langle\theta_n^\lambda - \theta^*, H(\theta_n^\lambda, X_{n+1}) - H(\theta^*, X_{n+1})\rangle + \lambda\langle\theta_n^\lambda - \theta^*, H(\theta^*, X_{n+1})\rangle.
\end{aligned}$$

Επομένως

$$\begin{aligned}\mathbb{E}[\|\theta_{n+1}^{\lambda}-\theta^{*}\|^{2}\big|\theta_{n}^{\lambda}] &= \mathbb{E}[\|\theta_{n}^{\lambda}-\theta^{*}\|^{2}\big|\theta_{n}^{\lambda}] - 2\lambda\mathbb{E}[\langle\theta_{n}^{\lambda}-\theta^{*}, H(\theta_{n}^{\lambda},X_{n+1})-H(\theta^{*},X_{n+1})\rangle\big|\theta_{n}^{\lambda}]\\ &\quad - 2\lambda\mathbb{E}[\langle\theta_{n}^{\lambda}-\theta^{*}, H(\theta^{*},X_{n+1})\rangle\big|\theta_{n}^{\lambda}] + 2\mathbb{E}[\langle\theta_{n}^{\lambda}-\theta^{*}, \sqrt{2\lambda\beta^{-1}}\xi_{n+1}\rangle\big|\theta_{n}^{\lambda}]\\ &\quad + \lambda^{2}\mathbb{E}[\|H(\theta_{n}^{\lambda},X_{n+1})\|^{2}\big|\theta_{n}^{\lambda}] - 2\lambda\mathbb{E}[\langle H(\theta_{n}^{\lambda},X_{n+1}), \sqrt{2\lambda\beta^{-1}}\xi_{n+1}\rangle\big|\theta_{n}^{\lambda}]\\ &\quad + 2\lambda\beta^{-1}\mathbb{E}[\|\xi_{n+1}\|^{2}\big|\theta_{n}^{\lambda}]\end{aligned} \tag{7}$$

όμως καθώς $\mathbb{E}[\xi_{n+1}^{(i)}|\theta_{n}^{\lambda}] = \mathbb{E}[\xi_{n+1}^{(i)}] = 0$ παίρνουμε ότι

$$\begin{aligned}\mathbb{E}[\langle\theta_{n}^{\lambda}-\theta^{*}, \sqrt{2\lambda\beta^{-1}}\xi_{n+1}\rangle\big|\theta_{n}^{\lambda}] &= \mathbb{E}[\langle\theta_{n}^{\lambda}, \sqrt{2\lambda\beta^{-1}}\xi_{n+1}\rangle\big|\theta_{n}^{\lambda}] - \mathbb{E}[\langle\theta^{*}, \sqrt{2\lambda\beta^{-1}}\xi_{n+1}\rangle\big|\theta_{n}^{\lambda}]\\ &= \mathbb{E}\Big[\sum_{i=1}^{d}\theta_{n}^{\lambda(i)}\sqrt{2\lambda\beta^{-1}}\xi_{n+1}^{(i)}|\theta_{n}^{\lambda}\Big] - \mathbb{E}\Big[\sum_{i=1}^{d}\theta^{*(i)}\sqrt{2\lambda\beta^{-1}}\xi_{n+1}^{(i)}|\theta_{n}^{\lambda}\Big]\\ &= \sum_{i=1}^{d}\sqrt{2\lambda\beta^{-1}}\mathbb{E}[\theta_{n}^{\lambda(i)}|\theta_{n}^{\lambda}]\mathbb{E}[\xi_{n+1}^{(i)}|\theta_{n}^{\lambda}] - \sum_{i=1}^{d}\mathbb{E}[\theta^{*(i)}|\theta_{n}^{\lambda}]\mathbb{E}[\xi_{n+1}^{(i)}|\theta_{n}^{\lambda}]\\ &= 0\end{aligned}$$

Όμοια παίρνουμε και ότι

$$\begin{aligned}\mathbb{E}[\langle H(\theta_{n}^{\lambda},X_{n+1}), \sqrt{2\lambda\beta^{-1}}\xi_{n+1}\rangle\big|\theta_{n}^{\lambda}] &= \mathbb{E}[\sum_{i=1}^{d}H^{(i)}(\theta_{n}^{\lambda},X_{n+1})\sqrt{2\lambda\beta^{-1}}\xi_{n+1}^{(i)}\big|\theta_{n}^{\lambda}]\\ &= \sum_{i=1}^{d}\sqrt{2\lambda\beta^{-1}}\mathbb{E}[H^{(i)}(\theta_{n}^{\lambda},X_{n+1})|\theta_{n}^{\lambda}]\mathbb{E}[\xi_{n+1}^{(i)}|\theta_{n}^{\lambda}]\\ &= 0,\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\|\xi_{n+1}\|^{2}\big|\theta_{n}^{\lambda}] = \mathbb{E}[\|\xi_{n+1}\|^{2}] &= \mathbb{E}[\sum_{i=1}^{d}(\xi_{n+1}^{(i)})^{2}] = \sum_{i=1}^{d}\mathbb{E}[(\xi_{n+1}^{(i)})^{2}]\\ &= \sum_{i=1}^{d}Var[\xi_{n+1}^{(i)}] + \mathbb{E}[\xi_{n+1}^{(i)}]^{2} = \sum_{i=1}^{d}[1+0]^{2}\\ &= d,\end{aligned}$$

οπότε η (7) απλοποιείται και γίνεται

$$\begin{aligned}\mathbb{E}[\|\theta_{n+1}^{\lambda}-\theta^{*}\|^{2}\big|\theta_{n}^{\lambda}] &= \|\theta_{n}^{\lambda}-\theta^{*}\|^{2}-2\lambda\mathbb{E}[\langle\theta_{n}^{\lambda}-\theta^{*},H(\theta_{n}^{\lambda},X_{n+1})-H(\theta^{*},X_{n+1})\rangle\big|\theta_{n}^{\lambda}]\\ &\quad-2\lambda\mathbb{E}[\langle\theta_{n}^{\lambda}-\theta^{*},H(\theta^{*},X_{n+1})\rangle\big|\theta_{n}^{\lambda}]+\lambda^{2}\mathbb{E}[\|H(\theta_{n}^{\lambda},X_{n+1})\|^{2}\big|\theta_{n}^{\lambda}]\\ &\quad+2\lambda d\beta^{-1}\end{aligned} \tag{8}$$

Από Υπόθεση 2.6 και 2.3 παίρνουμε

$$\begin{aligned}-2\lambda\mathbb{E}[\langle\theta_{n}^{\lambda}-\theta^{*},H(\theta_{n}^{\lambda},X_{n+1})-H(\theta^{*},X_{n+1})\rangle\big|\theta_{n}^{\lambda}] &\leq -2\lambda\mathbb{E}[\langle\theta_{n}^{\lambda}-\theta^{*},A(X_{n+1})(\theta_{n}^{\lambda}-\theta^{*})\rangle\big|\theta_{n}^{\lambda}]\\ &\leq -2\lambda\alpha\|\theta_{n}^{\lambda}-\theta^{*}\|^{2},\end{aligned}$$

και χρησιμοποιώντας την σχέση $(x+y)^{2}\leq 2x^{2}+2y^{2}$ παίρνουμε ότι

$$\lambda^{2}\mathbb{E}[\|H(\theta_{n}^{\lambda},X_{n+1})\|^{2}\big|\theta_{n}^{\lambda}]\leq 2\lambda^{2}\mathbb{E}[\|H(\theta_{n}^{\lambda},X_{n+1})-H(\theta^{*},X_{n+1})\|^{2}\big|\theta_{n}^{\lambda}]+2\lambda^{2}\mathbb{E}[\|H(\theta^{*},X_{n+1})\|^{2}\big|\theta_{n}^{\lambda}].$$

Συνολικά η (8) γίνεται

$$\begin{aligned}\mathbb{E}[\|\theta_{n+1}^{\lambda}-\theta^{*}\|^{2}\big|\theta_{n}^{\lambda}] &\leq \|\theta_{n}^{\lambda}-\theta^{*}\|^{2}-2\lambda\mathbb{E}[\langle\theta_{n}^{\lambda}-\theta^{*},A(X_{n+1})(\theta_{n}^{\lambda}-\theta^{*})\rangle\big|\theta_{n}^{\lambda}]-2\lambda\langle\theta_{n}^{\lambda}-\theta^{*},h(\theta^{*})\rangle\\ &\quad+\lambda^{2}\mathbb{E}[\|H(\theta_{n}^{\lambda},X_{n+1})\|^{2}\big|\theta_{n}^{\lambda}]+2\lambda d\beta^{-1}\\ &\leq (1-2\lambda\alpha)\|\theta_{n}^{\lambda}-\theta^{*}\|^{2}+2\lambda^{2}\mathbb{E}[\|H(\theta_{n}^{\lambda},X_{n+1})-H(\theta^{*},X_{n+1})\|^{2}\big|\theta_{n}^{\lambda}]\\ &\quad+2\lambda^{2}\mathbb{E}[\|H(\theta^{*},X_{n+1})\|^{2}\big|\theta_{n}^{\lambda}]+2\lambda d\beta^{-1}\end{aligned} \tag{9}$$

Από την Υπόθεση 2.4 παίρνουμε

$$\|H(\theta_{n}^{\lambda},X_{n+1})-H(\theta^{*},X_{n+1})\|^{2}\leq L_{1}^{2}(1+\|X_{n+1}\|)^{2\rho}\|\theta_{n}^{\lambda}-\theta^{*}\|^{2}\Rightarrow$$

$$\begin{aligned}2\lambda^{2}\mathbb{E}[\|H(\theta_{n}^{\lambda},X_{n+1})-H(\theta^{*},X_{n+1})\|^{2}\big|\theta_{n}^{\lambda}]-\lambda\alpha\|\theta_{n}^{\lambda}-\theta^{*}\|^{2} &\leq 2\lambda^{2}L_{1}^{2}\mathbb{E}[(1+\|X_{n+1}\|)^{2\rho}\big|\theta_{n}^{\lambda}]\|\theta_{n}^{\lambda}-\theta^{*}\|^{2}\\ &\quad-\lambda\alpha\|\theta_{n}^{\lambda}-\theta^{*}\|^{2},\end{aligned}$$

και

$$\begin{aligned}\|H(\theta^{*},X_{n+1})-H(\theta^{*},0)\|^{2} &\leq L_{2}^{2}(1+\|X_{n+1}\|)^{2\rho}(1+\|\theta^{*}\|)^{2}\|X_{n+1}\|^{2}\\ &\leq L_{2}^{2}(1+\|X_{n+1}\|)^{2\rho+2}(1+\|\theta^{*}\|)^{2}\frac{\|X_{n+1}\|^{2}}{(1+\|X_{n+1}\|)^{2}}\\ &\leq L_{2}^{2}(1+\|X_{n+1}\|)^{2\rho+2}(1+\|\theta^{*}\|)^{2},\end{aligned}$$

επίσης

$$\|H(\theta^*, X_{n+1})\|^2 \leq 2\|H(\theta^*, X_{n+1}) - H(\theta^*, 0)\|^2 + 2\|H(\theta^*, 0)\|^2 = 2\|H(\theta^*, X_{n+1}) - H(\theta^*, 0)\|^2 + 2\{\ H^*\}^2.$$

Άρα

$$\begin{aligned} 2\lambda^2 \mathbb{E}[\|H(\theta^*, X_{n+1})\|^2] &\leq 4\lambda^2 \mathbb{E}[\|H(\theta^*, X_{n+1}) - H(\theta^*, 0)\|^2] + 4\lambda^2 \mathbb{E}[\{\ H^*\}^2] \\ &\leq 4\lambda^2 L_2^2 \mathbb{E}[(1 + \|X_{n+1}\|)^{2\rho+2}](1 + \|\theta^*\|)^2 + 4\lambda^2\{\ H^*\}^2 \\ &\leq 4\lambda^2 L_2^2 \mathbb{E}[(1 + \|X_0\|)^{2\rho+2}](1 + \|\theta^*\|)^2 + 4\lambda^2\{\ H^*\}^2, \end{aligned}$$

όπου στην τελευταία ανισότητα χρησιμοποιήσαμε την Υπόθεση 2.5.

Σύμφωνα με τα παραπάνω η (9) γίνεται

$$\begin{aligned} \mathbb{E}[\|\theta_{n+1}^\lambda - \theta^*\|^2 \big| \theta_n^\lambda] \leq (1 - \lambda\alpha)\|\theta_n^\lambda - \theta^*\|^2 + (2\lambda^2 L_1^2 \mathbb{E}[(1 + \|X_0\|)^{2\rho}] - \lambda\alpha)\|\theta_n^\lambda - \theta^*\|^2 \\ + 4\lambda^2 L_2^2 \mathbb{E}[(1 + \|X_0\|)^{2\rho+2}](1 + \|\theta^*\|)^2 + 4\lambda^2\{\ H^*\}^2 + 2\lambda d\beta^{-1}, \end{aligned}$$

οπότε για $\lambda \leq \lambda_0 \coloneqq \min(\alpha/2L_1^2\mathbb{E}[(1 + \|X_0\|)^{2\rho}], 1/\alpha)$

$$\begin{aligned} \mathbb{E}[\|\theta_{n+1}^\lambda - \theta^*\|^2 \big| \theta_n^\lambda] &\leq (1 - \lambda\alpha)\|\theta_n^\lambda - \theta^*\|^2 + 4\lambda^2 L_2^2 \mathbb{E}[(1 + \|X_0\|)^{2\rho+2}](1 + \|\theta^*\|)^2 \\ &\quad + 4\lambda^2\{\ H^*\}^2 + 2\lambda d\beta^{-1} \\ \Rightarrow \mathbb{E}[\|\theta_{n+1}^\lambda - \theta^*\|^2 \big| \theta_n^\lambda] &\leq (1 - \lambda\alpha)\|\theta_n^\lambda - \theta^*\|^2 + \lambda C, \end{aligned}$$

όπου $C = 4\lambda_0 L_2^2 \mathbb{E}[(1 + \|X_0\|)^{2\rho+2}](1 + \|\theta^*\|)^2 + 4\lambda_0\{\ H^*\}^2 + 2d\beta^{-1}$.
Δηλαδή για $V_1(\theta_n^\lambda) = \|\theta_n^\lambda - \theta^*\|^2$ δείξαμε ότι

$$\mathbb{E}[V_1(\theta_n^\lambda) \big| \theta_{n-1}^\lambda] \leq (1 - \lambda\alpha)V_1(\theta_{n-1}^\lambda) + \lambda C.$$

Συνεχίζοντας, $\forall n \geq 0$

$$\begin{aligned}
\mathbb{E}[\|\theta_{n+1}^{\lambda} - \theta^*\|^2] &\leq (1-\lambda\alpha)\mathbb{E}[\|\theta_n^{\lambda} - \theta^*\|^2] + \lambda C \\
&\leq (1-\lambda\alpha)\big\{(1-\lambda\alpha)\mathbb{E}[\|\theta_{n-1}^{\lambda} - \theta^*\|^2] + \lambda C\big\} + \lambda C \\
&\leq (1-\lambda\alpha)^2\mathbb{E}[\|\theta_{n-1}^{\lambda} - \theta^*\|^2] + \lambda(1-\lambda\alpha)C + \lambda C \\
&\leq (1-\lambda\alpha)^n\mathbb{E}[\|\theta_0^{\lambda} - \theta^*\|^2] + \sum_{i=0}^{n-1}(1-\lambda\alpha)^i\lambda C \\
&\leq (1-\lambda\alpha)^n\mathbb{E}[\|\theta_0^{\lambda} - \theta^*\|^2] + \lambda C\frac{1-(1-\lambda\alpha)^n}{\lambda\alpha} \\
&\leq (1-\lambda\alpha)^n\mathbb{E}[\|\theta_0^{\lambda} - \theta^*\|^2] + \frac{C}{\alpha} \Rightarrow
\end{aligned}$$

$$\mathbb{E}[\|\theta_{n+1}^{\lambda} - \theta^*\|^2] \leq (1-\lambda\alpha)^n\mathbb{E}[\|\theta_0^{\lambda} - \theta^*\|^2] + \frac{C}{\alpha} < \infty$$

δηλαδή $\sup_{\lambda\leq\lambda_0}\sup_{n\in\mathbb{N}}\mathbb{E}[V_1(\theta_n^{\lambda})] < \infty$.

Επίσης παρατηρούμε ότι για $\rho = 0$ στην Υπόθεση 2.4 η H είναι co-coercive δηλαδή $\forall x \in \mathbb{R}^m$ και $\forall \theta, \theta' \in \mathbb{R}^d$,

$$\langle \theta - \theta', H(\theta, x) - H(\theta', x)\rangle \geq \frac{1}{L_1}\|H(\theta, x) - H(\theta', x)\|^2. \tag{10}$$

Οπότε επιστρέφουμε στην (8)

$$\begin{aligned}
\mathbb{E}[\|\theta_{n+1}^{\lambda} - \theta^*\|^2\big|\theta_n^{\lambda}] &= \|\theta_n^{\lambda} - \theta^*\|^2 - 2\lambda\mathbb{E}[\langle\theta_n^{\lambda} - \theta^*, H(\theta_n^{\lambda}, X_{n+1}) - H(\theta^*, X_{n+1})\rangle\big|\theta_n^{\lambda}] \\
&\quad - 2\lambda\mathbb{E}[\langle\theta_n^{\lambda} - \theta^*, H(\theta^*, X_{n+1})\rangle\big|\theta_n^{\lambda}] + \lambda^2\mathbb{E}[\|H(\theta_n^{\lambda}, X_{n+1})\|^2\big|\theta_n^{\lambda}] + 2\lambda d\beta^{-1} \\
&\leq \|\theta_n^{\lambda} - \theta^*\|^2 - 2\lambda\mathbb{E}[\langle\theta_n^{\lambda} - \theta^*, H(\theta_n^{\lambda}, X_{n+1}) - H(\theta^*, X_{n+1})\rangle\big|\theta_n^{\lambda}] \\
&\quad - 2\lambda\langle\theta_n^{\lambda} - \theta^*, h(\theta^*)\rangle + \lambda^2\mathbb{E}[\|H(\theta_n^{\lambda}, X_{n+1})\|^2\big|\theta_n^{\lambda}] + 2\lambda d\beta^{-1} \\
&\leq \|\theta_n^{\lambda} - \theta^*\|^2 - \lambda\alpha\|\theta_n^{\lambda} - \theta^*\|^2 - \frac{\lambda}{L_1}\mathbb{E}[\|H(\theta_n^{\lambda}, X_{n+1}) - H(\theta^*, X_{n+1})\|^2\big|\theta_n^{\lambda}] \\
&\quad + 2\lambda^2\mathbb{E}[\|H(\theta_n^{\lambda}, X_{n+1}) - H(\theta^*, X_{n+1})\|^2\big|\theta_n^{\lambda}] + 2\lambda^2\mathbb{E}[\|H(\theta^*, X_{n+1})\|^2\big|\theta_n^{\lambda}] \\
&\quad + 2\lambda d\beta^{-1} \Rightarrow \\
\mathbb{E}[\|\theta_{n+1}^{\lambda} - \theta^*\|^2\big|\theta_n^{\lambda}] &\leq (1-\lambda\alpha)\|\theta_n^{\lambda} - \theta^*\|^2 + (2\lambda^2 - \frac{\lambda}{L_1})\mathbb{E}[\|H(\theta_n^{\lambda}, X_{n+1}) - H(\theta^*, X_{n+1})\|^2] \\
&\quad + 2\lambda^2\mathbb{E}[\|H(\theta^*, X_{n+1})\|^2\big] + 2\lambda d\beta^{-1},
\end{aligned}$$

οπότε πάλι για $\lambda \leq \lambda_0' \coloneqq \min(\frac{1}{2L_1}, 1/a)$,

$$\begin{aligned}\mathbb{E}[\|\theta_{n+1}^{\lambda} - \theta^*\|^2 \big| \theta_n^{\lambda}] &\leq (1-\lambda\alpha)\|\theta_n^{\lambda} - \theta^*\|^2 + 4\lambda^2 L_2^2 \mathbb{E}[(1+\|X_0\|)^2](1+\|\theta^*\|)^2 \\ &\quad + 4\lambda^2\{\ H^*\}^2 + 2\lambda d\beta^{-1} \Rightarrow \\ \mathbb{E}[\|\theta_{n+1}^{\lambda} - \theta^*\|^2 \big| \theta_n^{\lambda}] &\leq (1-\lambda\alpha)\|\theta_n^{\lambda} - \theta^*\|^2 + \lambda C.\end{aligned}$$

όπου $C = 4\lambda_0 L_2^2 \mathbb{E}[(1+\|X_0\|)^2](1+\|\theta^*\|)^2 + 4\lambda_0\{\ H^*\}^2 + 2d\beta^{-1}$, δηλαδή το αποτέλεσμα ισχύει για $\lambda \leq \lambda_0'$ όταν το $\rho = 0$. □

<u>*απόδειξη της*</u> (10) *(co-coercivity of stochastic gradient)* :

Έχουμε $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ μετρήσιμη συνάρτηση με $U(\theta) \coloneqq \mathbb{E}[f(\theta, x)]$ και $H(\theta, x) = \nabla_\theta f(\theta, x)$.

Για $\rho = 0$ στην Υπόθεση 2.4 : $\|H(\theta,x) - H(\theta',x)\| \leq L_1(1+\|x\|)^\rho \|\theta - \theta'\| = L_1\|\theta - \theta'\|$, δηλαδή η $H(\theta, x) = \nabla_\theta f(\theta, x)$ είναι L_1-Lipschitz.

Επίσης από το [Lemma 2,[11]] η Υπόθεση 2.3 δίνει ότι η $f(\theta, x)$ είναι αυστηρά κυρτή ως προς θ για κάθε $x \in \mathbb{R}^m$.

Από θεμελιώδες θεώρημα ολοκληρωτικού λογισμού παίρνουμε για κάθε $\theta, \theta' \in \mathbb{R}^d$ και για κάθε $x \in \mathbb{R}^m$,

$$f(\theta,x) - f(\theta',x) = \int_0^1 \frac{d}{dt} f(\theta' + t(\theta - \theta'), x)dt = \int_0^1 \langle \nabla_\theta f(\theta' + t(\theta - \theta'), x), \theta - \theta' \rangle dt \Rightarrow$$

$$|f(\theta,x)-f(\theta',x)-\langle\nabla_\theta f(\theta',x),\theta-\theta'\rangle| = |\int_0^1 \langle\nabla_\theta f(\theta'+t(\theta-\theta'),x)-\nabla_\theta f(\theta',x),\theta-\theta'\rangle dt|$$

$$\leq \int_0^1 \left|\langle\nabla_\theta f(\theta'+t(\theta-\theta'),x)-\nabla_\theta f(\theta',x),\theta-\theta'\rangle\right| dt$$

$$\overset{\text{C-S}}{\leq} \int_0^1 \|\nabla_\theta f(\theta'+t(\theta-\theta'),x)-\nabla_\theta f(\theta',x)\|\cdot\|\theta-\theta'\| dt$$

$$\overset{\text{Lip}}{\leq} \int_0^1 L_1\|(\theta'+t(\theta-\theta'))-\theta'\|\cdot\|\theta-\theta'\| dt = \int_0^1 L_1 t\|\theta-\theta'\|^2 dt$$

$$\leq \frac{L_1}{2}\|\theta-\theta'\|^2 \tag{11}$$

Έστω $\theta^* \in \mathbb{R}^d$. Ορίζουμε $g(\theta) = f(\theta,x) - \langle\nabla_\theta f(\theta^*,x),\theta\rangle$. Η g είναι κυρτή και $\nabla g(\theta^*) = \nabla_\theta f(\theta^*,x) - \nabla_\theta f(\theta^*,x) = 0$ συνεπώς το θ^* είναι σημείο ολικού ελαχίστου της g.

Επιστρέφουμε στην (11) αφού ισχύει για κάθε $\theta,\theta' \in \mathbb{R}^d$, την εφαρμόζουμε για $\theta - \frac{1}{L_1}\nabla g(\theta) \in \mathbb{R}^d$ και $\theta \in \mathbb{R}^d$,

$$f(\theta-\frac{1}{L_1}\nabla g(\theta),x)-f(\theta,x)-\langle\nabla_\theta f(\theta,x),-\frac{1}{L_1}\nabla g(\theta)\rangle \leq \frac{L_1}{2}\|-\frac{1}{L}\nabla g(\theta)\|^2 \Rightarrow$$

$$g(\theta-\frac{1}{L_1}\nabla g(\theta))-g(\theta)-\langle\nabla_\theta f(\theta^*,x)-\nabla_\theta f(\theta,x),-\frac{1}{L_1}\nabla g(\theta) \leq \frac{L_1}{2}\frac{1}{L_1^2}\|\nabla g(\theta)\rangle\|^2 \Rightarrow$$
$$g(\theta-\frac{1}{L_1}\nabla g(\theta))-g(\theta)-\langle-\nabla g(\theta),-\frac{1}{L_1}\nabla g(\theta)\rangle \leq \frac{1}{2L_1}\|\nabla g(\theta)\|^2 \Rightarrow$$
$$g(\theta-\frac{1}{L_1}\nabla g(\theta))-g(\theta)+\frac{1}{L_1}\|\nabla g(\theta)\|^2 \leq \frac{1}{2L_1}\|\nabla g(\theta)\|^2 \Rightarrow$$
$$g(\theta-\frac{1}{L_1}\nabla g(\theta)) \leq g(\theta)-\frac{1}{2L_1}\|\nabla g(\theta)\|^2.$$

Το θ^* είναι σημείο ολικού ελαχίστου της g επομένως

$$g(\theta^*) \leq g(\theta - \frac{1}{L_1}\nabla g(\theta)) \leq g(\theta) - \frac{1}{2L_1}\|\nabla g(\theta)\|^2 \Rightarrow g(\theta^*) \leq g(\theta) - \frac{1}{2L_1}\|\nabla g(\theta)\|^2 \Rightarrow$$

$$f(\theta^*, x) - \langle \nabla_\theta f(\theta^*, x), \theta^* \rangle \leq f(\theta, x) - \langle \nabla_\theta f(\theta^*, x), \theta \rangle - \frac{1}{2L_1}\|\nabla_\theta f(\theta, x) - \nabla_\theta f(\theta^*, x)\|^2 \Rightarrow$$

$$f(\theta^*, x) - f(\theta, x) - \langle \nabla_\theta f(\theta^*, x), \theta - \theta^* \rangle \leq -\frac{1}{2L_1}\|\nabla_\theta f(\theta, x) - \nabla_\theta f(\theta^*, x)\|^2,$$

Αφού ισχύει για κάθε $\theta \in \mathbb{R}^d$ και τυχαίο $\theta^* \in \mathbb{R}^d$ παίρνουμε επίσης

$$f(\theta, x) - f(\theta^*, x) - \langle \nabla_\theta f(\theta, x), \theta - \theta^* \rangle \leq -\frac{1}{2L_1}\|\nabla_\theta f(\theta, x) - \nabla_\theta f(\theta^*, x)\|^2.$$

Προσθέτουμε τις δύο τελευταίες σχέσεις οπότε

$$\langle \nabla_\theta f(\theta, x) - \nabla_\theta f(\theta^*, x), \theta - \theta^* \rangle \geq \frac{1}{L_1}\|\nabla_\theta f(\theta, x) - \nabla_\theta f(\theta^*, x)\|^2 \Rightarrow$$

$$\langle \theta - \theta', H(\theta, x) - H(\theta', x) \rangle \geq \frac{1}{L_1}\|H(\theta, x) - H(\theta', x)\|^2.$$

□

Λήμμα 2.2. *Έστω $\mathcal{F}, \mathcal{G}, \mathcal{H}$ σ-άλγεβρες με $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ και X, Y πραγματικές τυχαίες μεταβλητές στον $L^p, p \geq 1$ όπου η Y είναι $\mathcal{H} \vee \mathcal{G}$-μετρήσιμη. Τότε*

$$\mathbb{E}^{1/p}[\|X - \mathbb{E}[X|\mathcal{H} \vee \mathcal{G}]\|^p|\mathcal{G}] \leq 2\mathbb{E}^{1/p}[\|X - Y\|^p|\mathcal{G}].$$

Θεώρημα 2.1. *Έστω ότι οι Υποθέσεις 2.1,2.2,2.3 ισχύουν και έστω $\lambda \leq \bar{\lambda}$ όπου $\bar{\lambda} := \frac{2}{\alpha + L_1}$. Τότε η μαρκοβιανή αλυσίδα $(\bar{\theta}_n^\lambda)_{n\in\mathbb{N}}$ δέχεται ένα αναλλοίωτο μέτρο π_λ τέτοιο ώστε για κάθε $n \in \mathbb{N}$,*

$$W_2(Law(\bar{\theta}_n^\lambda), \pi_\lambda) \leq \hat{c}e^{-\alpha\lambda n}, \quad n \in \mathbb{N},$$

όπου $\hat{c} := \sqrt{2}(\|\theta_0 - \theta\|^2 + d/\tilde{\alpha})^{1/2}$.

Επιπλέον

$$W_2(\pi_\beta, \pi_\lambda) \leq c\sqrt{\lambda},$$

όπου $c := (L_1^2\tilde{\alpha}^{-1}(2\lambda + \tilde{\alpha}^{-1})(d + \lambda^2 L_1^2 d/12 + L_1^2\lambda d/2a))^{1/2}$.

Θεώρημα 2.2. *Έστω ότι οι Υποθέσεις 2.4,2.5,2.6 ισχύουν και έστω* $\bar{\lambda} := \dfrac{2}{\alpha + L_1}$. *Τότε υπάρχουν σταθερές* $c_1, c_2 > 0$ *έτσι ώστε για κάθε* $0 < \epsilon \leq 1/2$,

$$W_2(Law(\theta_n^\lambda), \pi_\beta) \leq \epsilon,$$

όταν το $\lambda \leq \min\left(\alpha/2L_1^2\mathbb{E}[(1+\|X_0\|)^{2\rho}], 1/\alpha\right)$ *ικανοποιεί τις*

$$\lambda \leq c_1\epsilon^2 \quad \textit{και} \quad n \geq \frac{c_2}{\epsilon^2}\ln(1/\epsilon),$$

όπου οι σταθερές c_1, c_2 *εξαρτώνται μόνο από το* $d, \beta, \alpha, \mathbb{E}[\|X_0\|^{2\rho+2}], L_1$ *και* L_2. *Επιπλέον, αν* $\rho = 0$ *στην Υπόθεση2.4 , τότε το παραπάνω αληθεύει για* $\lambda \leq \min(\dfrac{1}{2L_1}, 1/\alpha)$.

Απόδειξη.

Αρχικά παρατηρούμε ότι

$$\|\theta_n^\lambda\|^2 \leq 2\|\theta_n^\lambda - \theta^*\|^2 + 2\|\theta^*\|^2 \Rightarrow \mathbb{E}[\|\theta_n^\lambda\|^2] \leq 2\mathbb{E}[\|\theta_n^\lambda - \theta^*\|^2] + 2\mathbb{E}[\|\theta^*\|^2] \Rightarrow$$
$$\mathbb{E}[\|\theta_n^\lambda\|^2] \leq 2(1-\lambda\alpha)^n\mathbb{E}[\|\theta_0 - \theta^*\|^2] + \frac{2C}{\alpha} + 2\|\theta^*\|^2,$$

άρα παίρνουμε $\sup_{\lambda\leq\lambda_0}\sup_{n\in\mathbb{N}}\mathbb{E}[\|\theta_n^\lambda\|^2] < c_0$,

όπου $c_0 := 2\mathbb{E}[\|\theta_0 - \theta^*\|^2] + \dfrac{2C}{\alpha} + 2\|\theta^*\|^2$ και το C δίνεται στο Λήμμα 2.1.

Χρησιμοποιώντας τις (2),(5) υπολογίζουμε

$$\begin{aligned}
\|\theta_{n+1}^\lambda - \bar{\theta}_{n+1}^\lambda\|^2 &= \|\theta_n^\lambda - H(\theta_n^\lambda, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1} - (\bar{\theta}_n^\lambda - h(\bar{\theta}_n^\lambda) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1})\|^2 \\
&= \|\theta_n^\lambda - \bar{\theta}_n^\lambda - \lambda(H(\theta_n^\lambda, X_{n+1}) - h(\bar{\theta}_n^\lambda))\|^2 \\
&\leq \|\theta_n^\lambda - \bar{\theta}_n^\lambda\|^2 - 2\lambda\langle\theta_n^\lambda - \bar{\theta}_n^\lambda, H(\theta_n^\lambda, X_{n+1}) - h(\bar{\theta}_n^\lambda)\rangle + \lambda^2\|H(\theta_n^\lambda, X_{n+1}) - h(\bar{\theta}_n^\lambda)\|^2 \\
&\leq \|\theta_n^\lambda - \bar{\theta}_n^\lambda\|^2 - 2\lambda\langle\theta_n^\lambda - \bar{\theta}_n^\lambda, H(\theta_n^\lambda, X_{n+1}) - h(\theta_n^\lambda)\rangle - 2\lambda\langle\theta_n^\lambda - \bar{\theta}_n^\lambda, h(\theta_n^\lambda) - h(\bar{\theta}_n^\lambda)\rangle \\
&\quad + 2\lambda^2\|H(\theta_n^\lambda, X_{n+1}) - h(\theta_n^\lambda)\|^2 + 2\lambda^2\|h(\theta_n^\lambda) - h(\bar{\theta}_n^\lambda)\|^2
\end{aligned} \tag{12}$$

όπου για την τελευταία ανισότητα χρησιμοποιήσαμε ότι

$$\langle\theta_n^\lambda - \bar{\theta}_n^\lambda, H(\theta_n^\lambda, X_{n+1}) - h(\bar{\theta}_n^\lambda)\rangle = \langle\theta_n^\lambda - \bar{\theta}_n^\lambda, H(\theta_n^\lambda, X_{n+1}) - h(\theta_n^\lambda)\rangle + \langle\theta_n^\lambda - \bar{\theta}_n^\lambda, h(\theta_n^\lambda) - h(\bar{\theta}_n^\lambda)\rangle,$$

και $\|H(\theta_n^\lambda, X_{n+1}) - h(\bar{\theta}_n^\lambda)\|^2 \leq 2\|H(\theta_n^\lambda, X_{n+1}) - h(\theta_n^\lambda)\|^2 + 2\|h(\theta_n^\lambda) - h(\bar{\theta}_n^\lambda)\|^2$.

Στην (12) μέσω των *Ιδιοτήτων 2* και *3* έχουμε

$$\begin{aligned}
\mathbb{E}[\|\theta_{n+1}^\lambda - \bar{\theta}_{n+1}^\lambda\|^2|\theta_n^\lambda, \bar{\theta}_n^\lambda] &\leq \mathbb{E}[\|\theta_n^\lambda - \bar{\theta}_n^\lambda\|^2|\theta_n^\lambda, \bar{\theta}_n^\lambda] - 2\lambda\mathbb{E}[\langle\theta_n^\lambda - \bar{\theta}_n^\lambda, H(\theta_n^\lambda, X_{n+1}) - h(\theta_n^\lambda)\rangle|\theta_n^\lambda, \bar{\theta}_n^\lambda] \\
&\quad - 2\lambda\mathbb{E}[\langle\theta_n^\lambda - \bar{\theta}_n^\lambda, h(\theta_n^\lambda) - h(\bar{\theta}_n^\lambda)\rangle|\theta_n^\lambda, \bar{\theta}_n^\lambda] \\
&\quad + 2\lambda^2\mathbb{E}[\|H(\theta_n^\lambda, X_{n+1}) - h(\theta_n^\lambda)\|^2|\theta_n^\lambda, \bar{\theta}_n^\lambda] \\
&\quad + 2\lambda^2\mathbb{E}[\|h(\theta_n^\lambda) - h(\bar{\theta}_n^\lambda)\|^2|\theta_n^\lambda, \bar{\theta}_n^\lambda] \\
&\leq \|\theta_n^\lambda - \bar{\theta}_n^\lambda\|^2 - 2\lambda\tilde{\alpha}\|\theta_n^\lambda - \bar{\theta}_n^\lambda\|^2 - \frac{2\lambda}{\alpha + L_1}\|h(\theta_n^\lambda) - h(\bar{\theta}_n^\lambda)\|^2 \\
&\quad + 2\lambda^2\mathbb{E}[\|H(\theta_n^\lambda, X_{n+1}) - \mathbb{E}[H(\theta_n^\lambda, X_{n+1})]\|^2|\theta_n^\lambda, \bar{\theta}_n^\lambda] \\
&\quad + 2\lambda^2\|h(\theta_n^\lambda) - h(\bar{\theta}_n^\lambda)\|^2
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \mathbb{E}[\|\theta_{n+1}^\lambda - \bar{\theta}_{n+1}^\lambda\|^2|\theta_n^\lambda, \bar{\theta}_n^\lambda] &\leq (1 - \lambda\tilde{\alpha})\|\theta_n^\lambda - \bar{\theta}_n^\lambda\|^2 - \lambda\tilde{\alpha}\|\theta_n^\lambda - \bar{\theta}_n^\lambda\|^2 \\
&\quad + (2\lambda^2 - \frac{2\lambda}{\alpha + L_1})\|h(\theta_n^\lambda) - h(\bar{\theta}_n^\lambda)\|^2 \\
&\quad + 2\lambda^2\mathbb{E}[\|H(\theta_n^\lambda, X_{n+1}) - \mathbb{E}[H(\theta_n^\lambda, X_{n+1})]\|^2|\theta_n^\lambda, \bar{\theta}_n^\lambda]
\end{aligned}$$

οπότε για $2\lambda^2 - \frac{2\lambda}{\alpha + L_1} \leq 0 \Rightarrow \lambda \leq \frac{1}{\alpha + L_1}$,

$$\mathbb{E}[\|\theta_{n+1}^\lambda - \bar{\theta}_{n+1}^\lambda\|^2|\theta_n^\lambda, \bar{\theta}_n^\lambda] \leq (1 - \lambda\tilde{\alpha})\|\theta_n^\lambda - \bar{\theta}_n^\lambda\|^2 + 2\lambda^2\mathbb{E}[\|H(\theta_n^\lambda, X_{n+1}) - \mathbb{E}[H(\theta_n^\lambda, X_{n+1})]\|^2|\theta_n^\lambda, \bar{\theta}_n^\lambda]. \tag{13}$$

Μέσω του Λήμματος 2.2 των Υποθέσεων 2.4 και 2.5 φράσσουμε τον τελευταίο όρο της (13) ως

$$\begin{aligned}
&2\lambda^2\mathbb{E}[\|H(\theta_n^\lambda, X_{n+1}) - \mathbb{E}[H(\theta_n^\lambda, X_{n+1})]\|^2|\theta_n^\lambda, \bar{\theta}_n^\lambda] \\
&\leq 2\lambda^2(2^2)\mathbb{E}[\|H(\theta_n^\lambda, X_{n+1}) - H(\theta_n^\lambda, \mathbb{E}[X_{n+1}|\theta_n^\lambda, \bar{\theta}_n^\lambda])\|^2|\theta_n^\lambda, \bar{\theta}_n^\lambda] \\
&\leq 8\lambda^2 L_2^2\mathbb{E}[(1 + \|X_{n+1}\| + \|\mathbb{E}[X_{n+1}]\|)^{2\rho}(1 + \|\theta_n^\lambda\|)^2\|X_{n+1} - \mathbb{E}[X_{n+1}]\|^2] \\
&\leq 8\lambda^2 L_2^2\mathbb{E}[(1 + \|X_0\| + \|\mathbb{E}[X_0]\|)^{2\rho}\|X_0 - \mathbb{E}[X_0]\|](1 + \|\theta_n^\lambda\|)^2.
\end{aligned}$$

Θέτουμε

$$Var_{\mathcal{W}}(X_0) \coloneqq \mathbb{E}[(1 + \|X_0\| + \|\mathbb{E}[X_0]\|)^{2\rho}\|X_0 - \mathbb{E}[X_0]\|^2], \tag{14}$$

οπότε η (13) γίνεται

$$\begin{aligned}
&\mathbb{E}[\|\theta_{n+1}^\lambda - \bar{\theta}_{n+1}^\lambda\|^2|\theta_n^\lambda, \bar{\theta}_n^\lambda] \leq (1 - \lambda\tilde{\alpha})\|\theta_n^\lambda - \bar{\theta}_n^\lambda\|^2 + 8\lambda^2 L_2^2 Var_{\mathcal{W}}(X_0)(1 + \|\theta_n^\lambda\|)^2 \\
&\quad \Rightarrow \mathbb{E}[\|\theta_{n+1}^\lambda - \bar{\theta}_{n+1}^\lambda\|^2] \leq (1 - \lambda\tilde{\alpha})\|\theta_n^\lambda - \bar{\theta}_n^\lambda\|^2 + 8\lambda^2 L_2^2 Var_{\mathcal{W}}(X_0)\mathbb{E}[(1 + \|\theta_n^\lambda\|)^2]
\end{aligned}$$

και για $1-\lambda\tilde{\alpha}\leq 0$ παίρνουμε

$$\begin{aligned}\mathbb{E}[\|\theta_{n+1}^{\lambda}-\bar{\theta}_{n+1}^{\lambda}\|^2] &\leq 8\frac{\lambda}{\tilde{\alpha}}L_2^2 Var_{\mathcal{W}}(X_0)\mathbb{E}[(1+\|\theta_n^{\lambda}\|)^2]\\ \mathbb{E}[\|\theta_{n+1}^{\lambda}-\bar{\theta}_{n+1}^{\lambda}\|^2] &\leq 8\frac{\lambda}{\tilde{\alpha}}L_2^2 Var_{\mathcal{W}}(X_0)(1+\sup_{n\geq 0}\mathbb{E}[\|\theta_n^{\lambda}\|^2])\end{aligned} \tag{15}$$

Επίσης στην αρχή δείξαμε ότι
$\sup_{\lambda\leq\lambda_0}\sup_{n\in\mathbb{N}}\mathbb{E}[\|\theta_n^{\lambda}\|^2] < c_0 \coloneqq 2\mathbb{E}[\|\theta_0-\theta^*\|^2]+\dfrac{2C}{\alpha}+2\|\theta^*\|^2$,
άρα για $\bar{c}\coloneqq[\dfrac{8}{\tilde{\alpha}}L_2^2(1+c_0)Var_{\mathcal{W}}(X_0)]^{1/2}$ η (15) μας δίνει $\mathbb{E}[\|\theta_{n+1}^{\lambda}-\bar{\theta}_{n+1}^{\lambda}\|^2]\leq\lambda\bar{c}^2$

$$\Rightarrow W_2(Law(\theta_n^{\lambda}),Law(\bar{\theta}_n^{\lambda}))\leq[\lambda\bar{c}^2]^{1/2}=\lambda^{1/2}\bar{c}.$$

Με βάση αυτό το αποτέλεσμα και το Θεώρημα 2.1

$$\begin{aligned}W_2(Law(\theta_n^{\lambda}),\pi_{\beta}) &\leq W_2(Law(\theta_n^{\lambda}),Law(\bar{\theta}_n^{\lambda}))+W_2(Law(\bar{\theta}_n^{\lambda}),\pi_{\lambda})+W_2(\pi_{\lambda},\pi_{\beta})\\ &\leq \lambda^{1/2}\bar{c}+\hat{c}e^{-\lambda\alpha n}+c\sqrt{\lambda}\\ &\leq \bar{C}[\lambda^{1/2}+e^{-\lambda\alpha n}],\end{aligned}$$

όπου $\bar{C}=\max\{c,\hat{c},\bar{c}\}$.
Για κάθε $0\leq\epsilon\leq 1/2$

$$\bar{C}\lambda^{1/2}<\epsilon/2\Rightarrow\lambda<\frac{\epsilon^2}{4\bar{C}^2}\Rightarrow\lambda<c_1\epsilon^2,\quad c_1\coloneqq\frac{1}{4\bar{C}^2}$$

και $\bar{C}e^{-\lambda\alpha n}\leq\epsilon/2\Rightarrow$

$$\begin{aligned}n &\geq \frac{1}{\lambda\alpha}\ln\big(\frac{2\bar{C}}{\epsilon}\big)>\frac{1}{\alpha c_1\epsilon^2}\ln\big(\frac{2\bar{C}}{\epsilon}\big)>\frac{1}{\alpha c_1\epsilon^2}(\ln(2\bar{C})-\ln\epsilon)\\ &>\frac{1}{\alpha c_1\epsilon^2}\left(\frac{\ln(2\bar{C})}{-\ln\epsilon}+1\right)\ln\left(\frac{1}{\epsilon}\right)\Rightarrow\\ n &>\frac{1}{\alpha c_1\epsilon^2}\big(\ln(2\bar{C})+1\big)\ln\left(\frac{1}{\epsilon}\right)=\frac{c_2}{\epsilon^2}\ln\left(\frac{1}{\epsilon}\right),\quad c_2\coloneqq\frac{1}{\alpha c_1}\big(\ln(2\bar{C})+1\big).\end{aligned}$$

□

2.4 Κύρια αποτελέσματα και αποδείξεις στην βελτιστοποίηση

Έστω τώρα $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ μια μετρήσιμη συνάρτηση για την οποία ισχύει $\mathbb{E}[|f(\theta, X)|] < \infty$ για κάθε $\theta \in \mathbb{R}^d$ με $U(\theta) \coloneqq \mathbb{E}[f(\theta, X)]$ και $h \coloneqq \nabla U$.

Θεωρούμε το πρόβλημα ελαχιστοποίησης της κυρτής συνάρτησης U,

$$\min_{\theta \in \mathbb{R}^d} U(\theta) \coloneqq \mathbb{E}[f(\theta, X)].$$

Σκοπός μας είναι να υπολογίσουμε ένα ασυμπτωτικό άνω όριο για το αναμενόμενο υπερβάλλον ρίσκο (expected excess risk) $\mathbb{E}[U(\hat{\theta})] - \inf_{\theta \in \mathbb{R}^d} U(\theta)$.

Για να το πετύχουμε αυτό θα χρησιμοποιήσουμε το αποτέλεσμα της προηγούμενης παραγράφου για την απόσταση Wasserstein μεταξύ της κατανομής π_β και των προσεγγίσεων της (Θεώρημα 2.2).

Λήμμα 2.3. *Έστω μ, ν δύο μέτρα πιθανότητας στον $\mathbb{R}^d$ με πεπερασμένες δεύτερες ροπές και έστω συνάρτηση $g : \mathbb{R}^d \to \mathbb{R}$, $g \in C^1$ για την οποία ισχύει*

$$\|\nabla g(\theta)\| \leq c_1 \|\theta\| + c_2, \quad \text{για κάθε } \theta \in \mathbb{R}^d.$$

Τότε

$$\Big| \int_{\mathbb{R}^d} g d\mu - \int_{\mathbb{R}^d} g d\nu \Big| \leq (c_1 \sigma + c_2) W_2(\mu, \nu),$$

όπου $\sigma^2 \coloneqq \int_{\mathbb{R}^d} \|\theta\|^2 \mu(d\theta) \wedge \int_{\mathbb{R}^d} \|\theta\|^2 \nu(d\theta)$.

Πόρισμα 2.1. *Υπάρχουν σταθερές $C^\sharp, \bar{C}_1 > 0$ έτσι ώστε για κάθε $\beta > 0, 0 < \lambda \leq \lambda_0$ και $n \in \mathbb{N}$,*

$$\mathbb{E}[U(\theta_n^\lambda)] - \inf_{\theta \in \mathbb{R}^d} U(\theta) \leq \bar{C}_1 [e^{-\lambda \alpha n} + \lambda^{1/2}] + C^\sharp.$$

Απόδειξη. Έχουμε $\theta_\infty \sim \pi_\beta$ και

$$\mathbb{E}[U(\theta_n^\lambda)] - \inf_{\theta \in \mathbb{R}^d} U(\theta) = \Big(\mathbb{E}[U(\theta_n^\lambda)] - \mathbb{E}[U(\theta_\infty)] \Big) + \Big(\mathbb{E}[U(\theta_\infty)] - \inf_{\theta \in \mathbb{R}^d} U(\theta) \Big).$$

Θα φράξουμε ξεχωριστά τα $\Big(\mathbb{E}[U(\theta_n^\lambda)] - \mathbb{E}[U(\theta_\infty)] \Big)$ και $\Big(\mathbb{E}[U(\theta_\infty)] - \inf_{\theta \in \mathbb{R}^d} U(\theta) \Big)$.

Θέλουμε να εφαρμόσουμε το Λήμμα 2.3 για την U.

$$\begin{aligned} \|\nabla U(\theta)\| = \|h(\theta)\| = \|\mathbb{E}[H(\theta, X_0)]\| &\leq \mathbb{E}[\|H(\theta, X_0)\|] \\ &\leq \mathbb{E}[L_1(1 + \|X_0\|)^\rho \|\theta\| + L_2(1 + \|X_0\|)^\rho \|x\| + H_\star] \\ &\leq L_1 \mathbb{E}[(1 + \|X_0\|)^\rho] \|\theta\| + L_2 \mathbb{E}[(1 + \|X_0\|)^\rho] + H_\star \end{aligned}$$

Άρα

$$\begin{aligned}
&\mathbb{E}[U(\theta_n^\lambda)] - \mathbb{E}[U(\theta_\infty)] \\
&\leq (L_1\mathbb{E}[(1+\|X_0\|)^\rho]\mathbb{E}[\|\theta_n^\lambda\|^2] + L_2\mathbb{E}[(1+\|X_0\|)^\rho] + H_\star)W_2(Law(\theta_n^\lambda), \pi_\beta) \\
&\leq (L_1\mathbb{E}[(1+\|X_0\|)^\rho]c_0 + L_2\mathbb{E}[(1+\|X_0\|)^\rho] + H_\star)W_2(Law(\theta_n^\lambda), \pi_\beta) \\
&\leq (L_1\mathbb{E}[(1+\|X_0\|)^\rho]c_0 + L_2\mathbb{E}[(1+\|X_0\|)^\rho] + H_\star) \cdot (\bar{C}[e^{-\lambda\alpha n} + \lambda^{1/2}]) \\
&\leq \bar{C}_1[e^{-\lambda\alpha n} + \lambda^{1/2}]
\end{aligned}$$

όπου $\bar{C}_1 = \bar{C}(L_1\mathbb{E}[(1+\|X_0\|)^\rho]c_0 + L_2\mathbb{E}[(1+\|X_0\|)^\rho] + H_\star)$

Θα φράξουμε τώρα το $\left(\mathbb{E}[U(\theta_\infty)] - \inf_{\theta\in\mathbb{R}^d} U(\theta)\right)$.
Έστω

$$p_\beta(\theta) = e^{-\beta U(\theta)} / \int_{\mathbb{R}^d} e^{-\beta U(\theta)} d\theta = \frac{1}{Z} e^{-\beta U(\theta)}$$

η πυκνότητα του μέτρου π_β όπου Z η σταθερά κανονικοποίησης.
Ορίζουμε την διαφορική εντροπία της πυκνότητας p_β ως

$$h(p_\beta) = -\int_{\mathbb{R}^d} p_\beta(\theta)\log p_\beta(\theta) d\theta = -\int_{\mathbb{R}^d} \frac{e^{-\beta U(\theta)}}{Z} \log \frac{e^{-\beta U(\theta)}}{Z} d\theta$$

$$\begin{aligned}
\mathbb{E}[U(\theta)] &= \int_{\mathbb{R}^d} U(\theta)\pi_\beta(d\theta) = \int_{\mathbb{R}^d} U(\theta) d\pi_\beta(\theta) = \int_{\mathbb{R}^d} U(\theta)\frac{e^{-\beta U(\theta)}}{Z} d\theta \\
&= \frac{-1}{\beta}\int_{\mathbb{R}^d} \log e^{-\beta U(\theta)} \frac{e^{-\beta U(\theta)}}{Z} d\theta \\
&= \frac{-1}{\beta}\int_{\mathbb{R}^d} \log\left(\frac{e^{-\beta U(\theta)}}{Z}\right)\frac{e^{-\beta U(\theta)}}{Z} + \frac{e^{-\beta U(\theta)}}{Z}\log(Z) d\theta \\
&= \frac{-1}{\beta}\int_{\mathbb{R}^d} \log\left(\frac{e^{-\beta U(\theta)}}{Z}\right)\frac{e^{-\beta U(\theta)}}{Z} d\theta - \frac{1}{\beta}\int_{\mathbb{R}^d} \log(Z)\frac{e^{-\beta U(\theta)}}{Z} d\theta \\
&= \frac{-1}{\beta}\int_{\mathbb{R}^d} \log\left(\frac{e^{-\beta U(\theta)}}{Z}\right)\frac{e^{-\beta U(\theta)}}{Z} d\theta - \frac{1}{\beta}\log(Z)\int_{\mathbb{R}^d} \frac{e^{-\beta U(\theta)}}{Z} d\theta \\
&= \frac{1}{\beta}\big(h(p_\beta) - \log Z\big)
\end{aligned}$$

Για να φράξουμε την $h(p_\beta)$ θα χρειαστούμε ένα άνω φράγμα για την 2η ροπή της π_β.

Από την Ιδιότητα 3 παίρνουμε για την h :

$$\text{Για κάθε } \theta \in \mathbb{R}^d,\ \langle h(\theta), \theta\rangle \geq m\|\theta\|^2 - b \text{ , για κάποια } m > 0, b \geq 0.$$

Για να προσδιορίσουμε ακριβώς τα m, b εργαζόμαστε ως εξής:

$$\begin{aligned}\langle \theta - \theta', h(\theta) - h(\theta')\rangle \geq \alpha\|\theta - \theta'\|^2 \overset{\theta'=0}{\Rightarrow} \langle \theta, h(\theta)\rangle &\geq \alpha\|\theta\|^2 + \langle \theta, h(0)\rangle \\ \langle \theta, h(\theta)\rangle &\geq \alpha\|\theta\|^2 - |\langle \theta, h(0)\rangle| \\ \langle \theta, h(\theta)\rangle &\geq \alpha\|\theta\|^2 - \varepsilon^2/2\|\theta\|^2 - 1/2\varepsilon^2\|h(0)\|^2 \\ \langle \theta, h(\theta)\rangle &\geq (\alpha - \varepsilon^2/2)\|\theta\|^2 - 1/2\varepsilon^2\|h(0)\|^2\end{aligned}$$

όπου χρησιμοποιήσαμε την σχέση $|\langle x, y\rangle| \leq \frac{1}{2a}\|x\|^2 + \frac{a}{2}\|y\|^2$

Συνεπώς $m \coloneqq \alpha - \varepsilon^2/2$ και $b \coloneqq 1/\varepsilon^2\|h(0)\|^2$

Έτσι από την (3.19) στο [4] παίρνουμε για τη 2η ροπή της π_β

$$\int_{\mathbb{R}^d} \|\theta\|^2 \pi_\beta(d\theta) \leq \frac{b + d/\beta}{m}$$

Επίσης η διαφορική εντροπία μιας πυκνότητας με πεπερασμένη δεύτερη ροπή φράσσεται από την γκαουσιανή διαφορική εντροπία [[9],Theorem 8.6.5] επομένως

$$h(p_\beta) \leq \frac{1}{2}\log\left((2\pi e)^d(\sigma^2)^d\right) \leq \frac{d}{2}\log\left(\frac{2\pi e(b + d/\beta)}{md}\right) \tag{16}$$

Επιπλέον για μια συνάρτηση με Lipschitz gradient ισχύει

$$U(\theta) - U(\theta') - \langle \nabla U(\theta), \theta - \theta'\rangle \leq \frac{L}{2}\|\theta - \theta'\|^2$$

οπότε για το σημείο ελαχίστου $\theta^* \in \mathbb{R}^d$ της U ισχύει $U(\theta) - U(\theta^*) \leq \frac{L}{2}\|\theta - \theta^*\|^2$ και

$$\begin{aligned}\log Z = \log\int_{\mathbb{R}^d} e^{-\beta U(\theta)}d\theta &= \log\int_{\mathbb{R}^d} e^{-\beta U(\theta^*)}e^{\beta U(\theta^*) - \beta U(\theta)}d\theta \\ &= -\beta U(\theta^*) \log\int_{\mathbb{R}^d} e^{\beta(U(\theta^*) - U(\theta))}d\theta \\ &\geq -\beta U(\theta^*) + \log\int_{\mathbb{R}^d} e^{\frac{-\beta L_1}{2}\|\theta - \theta^*\|^2}d\theta\end{aligned}$$

Θα χρησιμοποιήσουμε τη μέθοδο του Laplace για την προσέγγιση του ολοκληρώματος

$$\int_{\mathbb{R}^d} e^{\frac{-\beta L_1}{2}\|\theta-\theta^*\|^2} d\theta = \frac{1}{\sqrt{2d}} \left(\frac{2\pi}{\beta L_1/2} \right)^{d/2}$$

οπότε

$$\log \int_{\mathbb{R}^d} e^{\frac{-\beta L_1}{2}\|\theta-\theta^*\|^2} d\theta = \frac{d}{2} \log \left(\frac{2\pi}{\beta L_1/2} \right) - \log \sqrt{2d} \geq \frac{d}{2} \log \left(\frac{2\pi}{\beta L_1} \right)$$

και τελικά

$$\log Z \geq -\beta U(\theta^*) + \frac{d}{2} \log \left(\frac{2\pi}{\beta L_1} \right) \tag{17}$$

$$\int_{\mathbb{R}^d} U(\theta) \pi_\beta(d\theta) = \frac{1}{\beta} \big(h(p_\beta) - \log Z \big) \overset{(16)(17)}{\leq} \frac{d}{2\beta} \log \left(\frac{2\pi e(b + d/\beta)}{md} \right) + \beta U(\theta^*) - \frac{d}{2\beta} \log \left(\frac{2\pi}{\beta L_1} \right)$$

$$\int_{\mathbb{R}^d} U(\theta) \pi_\beta(d\theta) - \beta U(\theta^*) \leq \frac{d}{2\beta} \log \left(e\beta L_1 \frac{(b + d/\beta)}{md} \right) \Rightarrow$$

$$\mathbb{E}[U(\theta_\infty)] - \inf_{\theta \in \mathbb{R}^d} U(\theta) \leq \frac{d}{2\beta} \log \left(\frac{eL_1}{m} \left(\frac{b\beta}{d} + 1 \right) \right)$$

Θέτουμε $C^\sharp := \frac{d}{2\beta} \log \left(\frac{eL_1}{m} \left(\frac{b\beta}{d} + 1 \right) \right)$ και συνολικά έχουμε το ζητούμενο

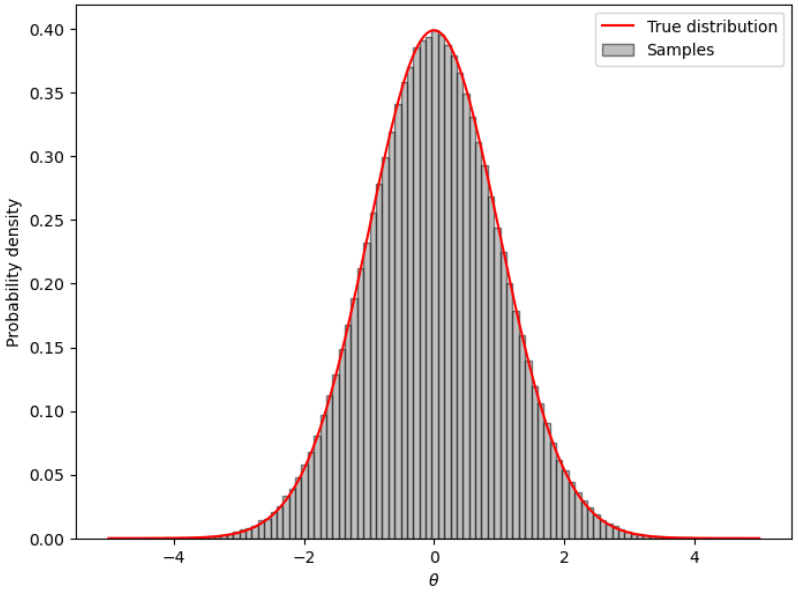
$$\mathbb{E}[U(\theta_n^\lambda)] - \inf_{\theta \in \mathbb{R}^d} U(\theta) \leq \bar{C}_1 [e^{-\lambda \alpha n} + \lambda^{1/2}] + C^\sharp.$$

□

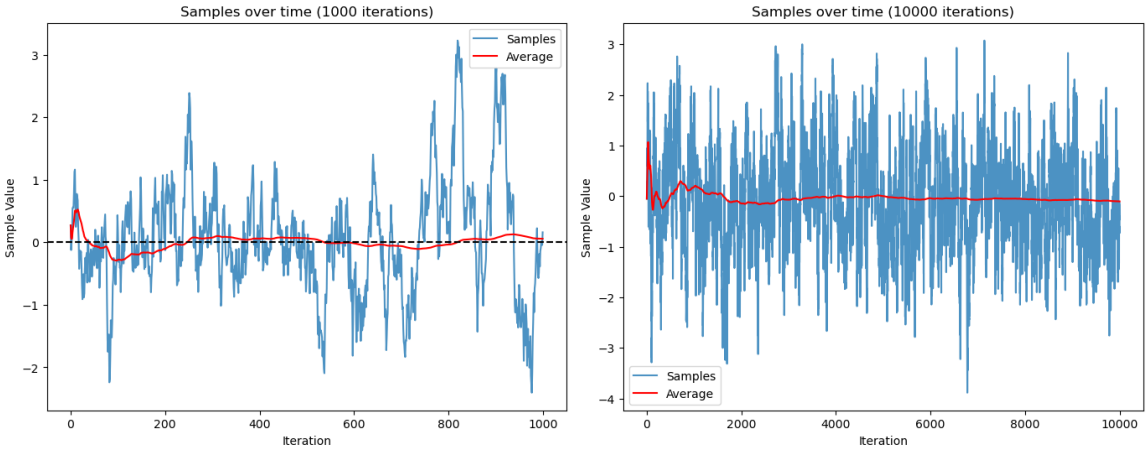
3 Εφαρμογές

3.1 Προσομοίωση από κανονική κατανομή

Θα ξεκινήσουμε με μια απλή εφαρμογή του αλγορίθμου για δειγματοληψία από την τυποποιημένη κανονική κατανομή.



Σχήμα 1: 100000 επαναλήψεις του αλγορίθμου, step size=0.05

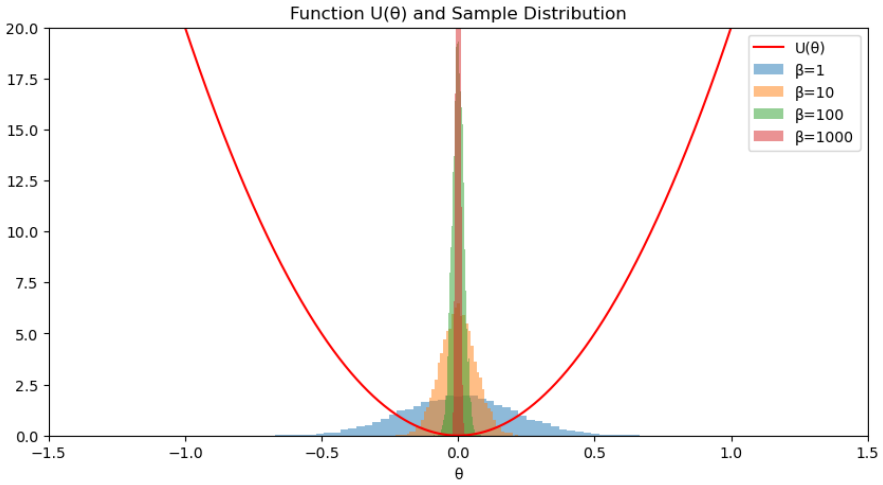


Σχήμα 2: Γράφημα των προσομοιωμένων τιμών όπου η κόκκινη γραμμή αναπαριστά την μέση τιμή των δειγμάτων, step size=0.05.

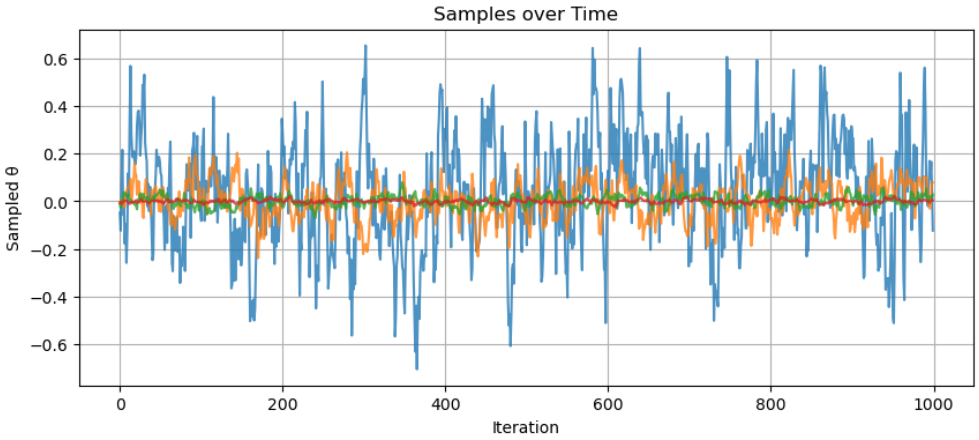
3.2 Ελαχιστοποίηση κυρτής συνάρτησης

Σε αυτή την εφαρμογή θα χρησιμοποιήσουμε τον αλγόριθμο για να υπολογίσουμε το ελάχιστο της συνάρτησης $U(\theta) = \theta^2$. Η U είναι αυστηρά κυρτή με Lipschitz gradient.

Τρέχουμε τον αλγόριθμο 4 φορές εισάγοντας κάθε φορά μεγαλύτερη τιμή για το β.



Σχήμα 3: 1000 επαναλήψεις του αλγορίθμου, step size=0.05



Σχήμα 4: Γράφημα των δειγμάτων για 4 προσομοιώσεις όπου αυξάνουμε το β.

Παρατηρούμε ότι για μεγαλύτερες τιμές του β η πυκνότητα των δειγμάτων συγκεντρώνεται στο ελάχιστο της U.

3.3 Γραμμική Παλινδρόμηση

Η μέθοδος των ελαχίστων τετραγώνων για την εκτίμηση των παραμέτρων θ στο γενικό γραμμικό μοντέλο βασίζεται στην ελαχιστοποίηση της παράστασης

$$\min_{\theta} \mathbb{E}[|z_n - \langle y_n, \theta \rangle|^2]$$

όπου $\theta \in \mathbb{R}^d$ και $x_n = (y_n, z_n), -\infty < n < \infty$ η από κοινού διαδικασία των $y_n \in \mathbb{R}^d$ και $z_n \in \mathbb{R}$.

Έχουμε $f(\theta, x_n) = |z_n - \langle y_n, \theta \rangle|^2$ και $H(\theta, x_n) = \nabla f(\theta, x_n) = -2y_n z_n + 2y_n \langle y_n, \theta \rangle$

Υπολογίζουμε

$$\begin{aligned} \|H(\theta, x_n) - H(\theta', x_n)\| &= \|2y_n \langle y_n, \theta - \theta' \rangle\| \leq 2\|y_n\|^2 \|\theta - \theta'\| \\ &\leq 2(1 + \|x_n\|)^2 \|\theta - \theta'\| \end{aligned}$$

Παρατηρούμε ότι ισχύει η Υπόθεση 2.4 με $L_1 = 2$ και $\rho = 2$.

Εφαρμογή 1: Θα εφαρμόσουμε τον αλγόριθμο σε ένα παράδειγμα γραμμικής παλινδρόμησης. Σκοπός μας αρχικά είναι να ελέγξουμε τη σωστή λειτουργία του αλγορίθμου. Θα παράξουμε κάποια δεδομένα όπου θα θέσουμε εμείς τις τιμές των παραμέτρων θ και στην συνέχεια θα εκτιμήσουμε τα θ με τον αλγόριθμο ώστε να επαληθεύσουμε ότι δίνει σωστά αποτελέσματα.

Θα προσαρμόσουμε το γραμμικό μοντέλο

$$Y = X\theta + \epsilon$$

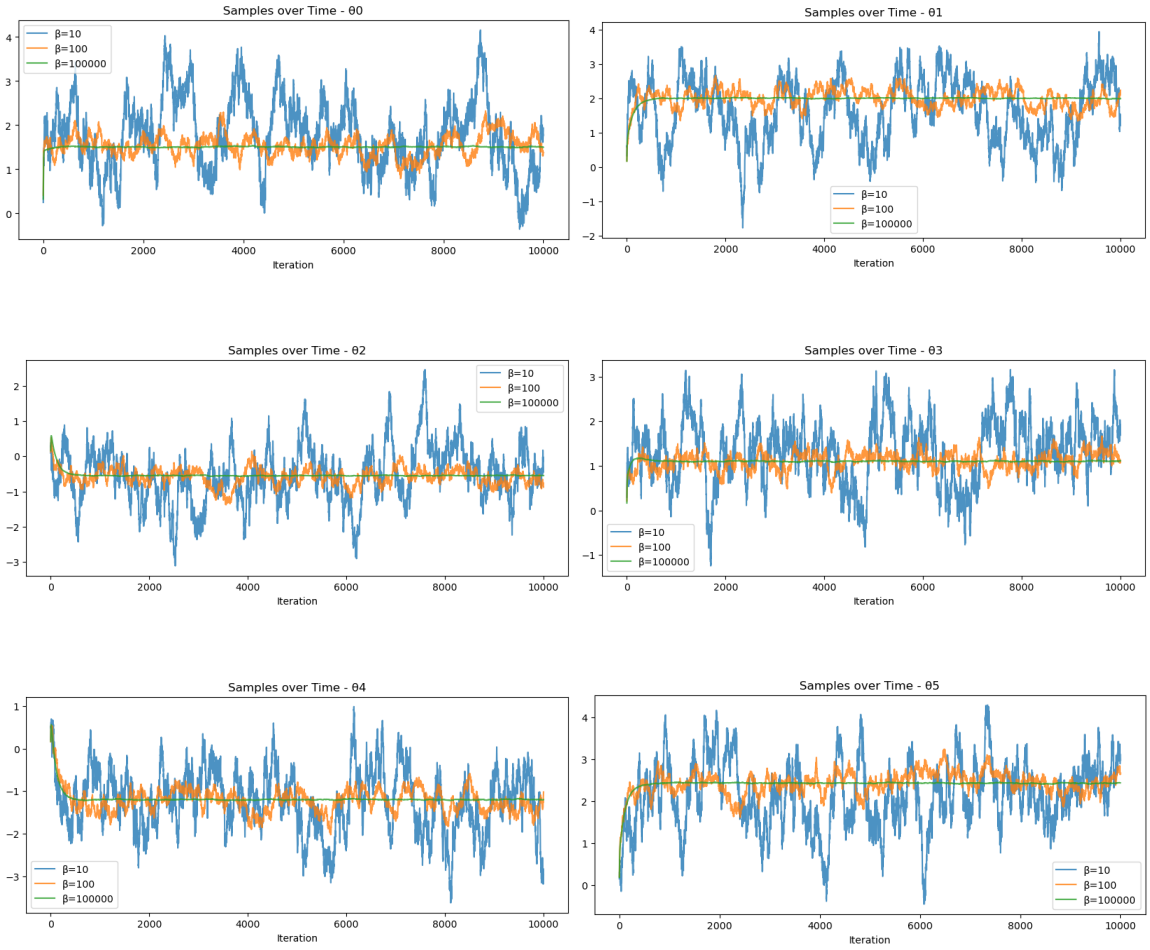
$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} & x_{15} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & x_{n4} & x_{15} \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \end{bmatrix}$$

Θεωρούμε

$$\theta_0 = 1.5, \ \theta_1 = 2, \ \theta_2 = -0.5, \ \theta_3 = 1, \ \theta_4 = -1.2, \ \theta_5 = 2.5$$

και ένα πίνακα X με τυχαίες τιμές και με βάση αυτά παράγουμε $n = 100$ τιμές για την μεταβλητή απόκρισης Y.

Τρέχουμε τον αλγόριθμο για 10.000 επαναλήψεις με step size=0.05 και για 3 τιμές του β.



Σχήμα 7: Κάθε διάγραμμα αποτυπώνει την τροχιά της παραμέτρου θ_i για τις διαφορετικές τιμές της παραμέτρου θερμοκρασίας β

Πράγματι παρατηρούμε ότι αυξάνοντας το β ο αλγόριθμος συγκλίνει και οι εκτιμήσεις που δίνει για $\beta = 10000$ είναι

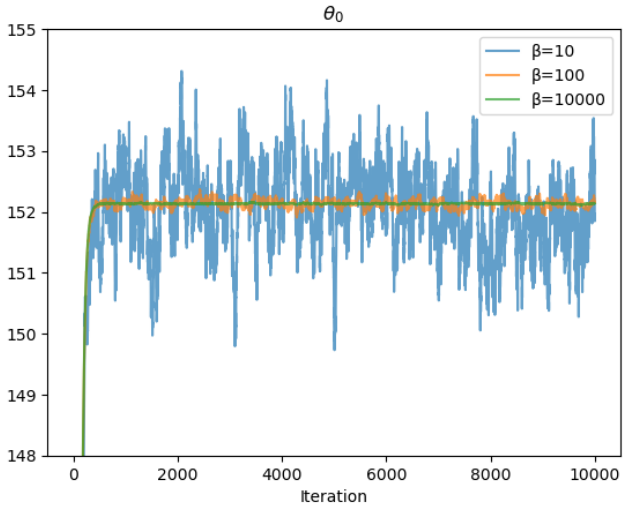
$$\hat{\theta} = \begin{bmatrix} 1.50 \\ 1.98 \\ -0.48 \\ 1.03 \\ -1.18 \\ 2.47 \end{bmatrix}$$ τα οποία είναι πολύ κοντά στις τιμές που είχαμε ορίσει.

Εφαρμογή 2: Θα εφαρμόσουμε τώρα τον αλγόριθμο σε ένα παράδειγμα από πραγματικά δεδομένα. Έχουμε δεδομένα για $n = 442$ ασθενείς με διαβήτη [1]. Θα προσαρμόσουμε το γραμμικό μοντέλο

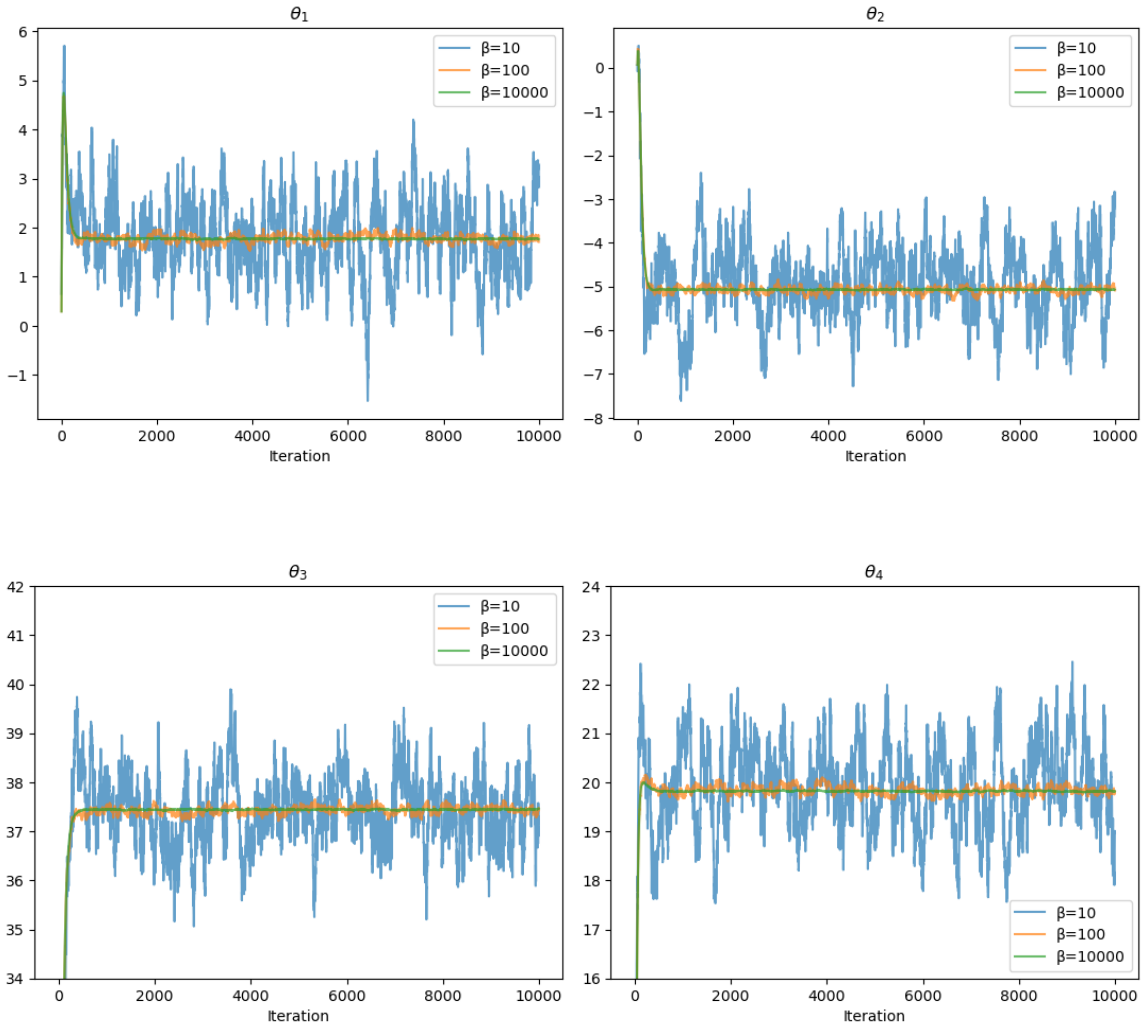
$$Y = X\theta + \epsilon$$

με 4 επεξηγηματικές μεταβλητές (ηλικία, φύλο, δείκτης μάζας σώματος bmi, πίεση) και μια μεταβλητή απόκρισης Y (δείκτης της εξέλιξης της νόσου).

Τρέχουμε τον αλγόριθμο για 10000 επαναλήψεις με βήμα 0.01 και 3 τιμές της παραμέτρου θερμοκρασίας $\beta = 10$, $\beta = 100$ και $\beta = 10000$. Στα παρακάτω σχήματα βλέπουμε τις τροχιές των εκτιμημένων παραμέτρων. Πάλι παρατηρούμε σύγκλιση στο ελάχιστο της συνάρτησης κόστους καθώς το β αυξάνεται.



[1] Δεδομένα:`https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html#sklearn.datasets.load_diabetes`



Σχήμα 9

Ο κλειστός τύπος για την εκτίμηση των παραμέτρων με την μέθοδο των ελαχίστων τετραγώνων δίνει

$$\hat{\theta}_{OLS} = (X^T X)^{-1} X^T y = \begin{bmatrix} 152.13 \\ 1.77 \\ -5.07 \\ 37.44 \\ 19.82 \end{bmatrix}$$ ενώ από την μέση τιμή των δειγμάτων $\hat{\theta}_{SGLD} = \begin{bmatrix} 152.18 \\ 1.73 \\ -4.98 \\ 37.43 \\ 19.79 \end{bmatrix}$

για $\beta = 10000$.

Βιβλιογραφία

[1] Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, 2004.

[2] Chii-Ruey Hwang. *Laplace's method revisited: weak convergence of probability measures.* The Annals of Probability, 8(6):1177–1182, 1980.

[3] Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations.* Springer-Verlag New York, 2014.

[4] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. *Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis*. In Conference on Learning Theory, pages 1674–1703, 2017.

[5] M. Barkhagen, N. Chau, E. Moulines, M. Rasonyi, S. Sabanis, and Y. Zhang. *On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case*, arXiv:1812.02709, 2018.

[6] M. Welling and Y. W. Teh. *Bayesian learning via stochastic gradient Langevin dynamics.* In: Proceedings of the 28th International Conference on Machine Learning, 681–688, 2011

[7] Neal, R. M. *MCMC using Hamiltonian dynamics.* In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L.(eds.), Handbook of Markov Chain Monte Carlo. Chapman & Hall / CRC Press, 2010.

[8] Robbins, H. and Monro, S. *A stochastic approximation method.* Annals of Mathematical Statistics, 22(3):400–407, 1951.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, New York, 2nd edition, 2006.

[10] Williams, David. *Probability With Martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, UK, 1991.

[11] Xingyu Zhou. *On the Fenchel Duality between Strong Convexity and Lipschitz Continuous Gradient*, arXiv:1803.06573v1, 2018.

[12] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Applied Optimization. Springer, 2004.

[13] Y. W. Teh, A. H. Thiery, and S. J. Vollmer. *Consistency and fluctuations for stochastic gradient Langevin dynamics.*, arXiv:1409.0578 , 2014.

[14] Y. Zhang, O.D. Akyildiz, T. Damoulas, S. Sabanis. *Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization*, arXiv:1910.02008, 2022.

[15] Γιαννακόπουλος Αθανάσιος. *Στοχαστική Ανάλυση και Εφαρμογές στη Χρηματοοικονομική*, Τόμος Ι: Εισαγωγή Στην Στοχαστική Ανάλυση [ηλεκτρ. βιβλ.], 2003.

[16] Λουλάκης, Μ. *Στοχαστικές Διαδικασίες* [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις, 2015.

[17] Χελιώτης, Δ. *Εισαγωγή στον στοχαστικό λογισμό*. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[18] Χελιώτης, Δ. *Ένα δεύτερο μάθημα στις πιθανότητες*. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

A Παράρτημα

Κώδικας 3.1

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

def U(theta):
    return theta**2/2

def grad_U(theta):
    return theta

def sgld(epsilon, T, theta0):

    theta = theta0
    samples = np.zeros(T)

    for t in range(T):

        noise = np.random.normal(0,1)

        theta = theta - epsilon * grad_U(theta) + np.sqrt(2*epsilon) * noise

    samples[t] = theta

    return samples
```

```
theta_range = np.linspace(-5, 5, num=1000)

# Compute the unnormalized log-probability density for each value of theta
log_U = np.array([-U(theta) for theta in theta_range])

# Compute the normalized probability density
prob = np.exp(log_U - log_U.max())
prob /= np.trapz(prob, theta_range)

# Generate samples from the posterior distribution using SGLD
epsilon=0.01
T=100000
samples = sgld(epsilon, T, 0)

# Plot the  probability density and the samples
fig, ax = plt.subplots(figsize=(8, 6))
ax.plot(theta_range, prob, label='True distribution',color='red')
ax.hist(samples, bins=100, density=True, alpha=0.5, label='Samples',ec='black',↩
    color='grey')
ax.set_xlabel(r'$\theta$')
ax.set_ylabel('Probability density')
ax.legend()
plt.show()
```

Κώδικας 3.2

```
import numpy as np
import matplotlib.pyplot as plt

def U(theta):
    return 20*theta ** 2

def grad_U(theta):
    return 20*theta

def sgld_minimize(step_size,  num_iterations, beta):
    theta = 0
    samples = np.zeros(num_iterations)

    for i in range(num_iterations):
        noise = np.random.normal(0, 1)
        theta = theta -  step_size * grad_U(theta) + np.sqrt(2*step_size/beta)*↩
            noise

    samples[i]=theta

    return samples

step_size = 0.01
num_iterations = 10000

#create samples for beta=1 10 100 1000
samples1 = sgld_minimize(step_size, num_iterations , 1)
samples2 = sgld_minimize(step_size, num_iterations , 10)
samples3 = sgld_minimize(step_size, num_iterations , 100)
samples4 = sgld_minimize(step_size, num_iterations , 1000)
```

```

theta_values = np.linspace(-5, 5, num_iterations)
U_values = U(theta_values)
# Plot the function U(x) = x^2
plt.plot(theta_values, U_values, color='red')

# Plot the histogram of the samples
plt.hist(samples1, bins=50, density=True, alpha=0.5, label='beta=1' )
plt.hist(samples2, bins=50, density=True, alpha=0.5, label='beta=10')
plt.hist(samples3, bins=50, density=True, alpha=0.5, label='beta=100')
plt.hist(samples4, bins=50, density=True, alpha=0.5, label='beta=1000')

plt.xlabel('theta')
plt.legend()
plt.title('Function U and Sample Distribution')

plt.xlim(-1.5, 1.5)
plt.ylim(0, 20)
plt.show()

# Plot of samples over time
plt.figure(figsize=(10, 4))
plt.plot(range(num_iterations), samples1, alpha=0.8, label='beta=1' )
```

```
    plt.plot(range(num_iterations), samples2, alpha=0.8, label='beta=10' )
    plt.plot(range(num_iterations), samples3, alpha=0.8, label='beta=100' )
    plt.plot(range(num_iterations), samples4, alpha=0.8, label='beta=1000' )
    plt.xlabel('Iteration')
    plt.ylabel('Sampled theta')
    plt.title('Samples over Time')
    plt.grid(True)
    plt.show()
```

Κώδικας 3.3

Εφαρμογή 1

```
    import numpy as np
    import matplotlib.pyplot as plt

    #Generate data
    np.random.seed(0)  # For reproducibility

    # True parameter values
    theta_true = np.array([1.5, 2.0, -0.5, 1, -1.2, 2.5])

    # Generate X values
    X = np.random.rand(100, 5)

    # Add columnn of ones
    X = np.concatenate([np.ones((X.shape[0], 1)), X], axis=1)

    # Generate Y values
    epsilon = np.random.randn(100)*0.2  # Gaussian noise
    Y = np.dot(X, theta_true) + epsilon
```

```
    def sgld(X, Y, learning_rate, num_iterations, beta):
        num_samples, num_features = X.shape
        theta = np.zeros(num_features)  # Initialize theta with zeros
        theta_trace = np.zeros((num_iterations, num_features))  # Store theta ↩
            values

        for iteration in range(num_iterations):

            gradient = 2 / num_samples * np.dot(X.T, np.dot(X, theta) - Y)

            noise = np.random.randn(num_features)
            theta = theta - learning_rate * gradient + np.sqrt(2 * learning_rate / ↩
                beta) * noise

            # Store theta value
            theta_trace[iteration] = theta

    return theta_trace

    # Run the SGLD algorithm with different beta values
    learning_rate = 0.05
    num_iterations = 10000
```

```
    betas = [10, 100, 100000]

    theta_traces = []
    for beta in betas:
        theta_trace = sgld(X, Y, learning_rate, num_iterations, beta)
        theta_traces.append(theta_trace)

    # Plot of samples over time
    for i in [0, 1, 2, 3, 4, 5]:
        plt.figure(figsize=(10, 4))
        plt.plot(range(num_iterations), theta_traces[0][:, i], alpha=0.8, label='↩
            beta=10' )
        plt.plot(range(num_iterations), theta_traces[1][:, i], alpha=0.8, label='↩
            beta=100' )
        plt.plot(range(num_iterations), theta_traces[2][:, i], alpha=0.8, label='↩
            beta=100000' )

        plt.xlabel('Iteration')
        plt.title('Samples over Time')
        plt.legend()
        plt.show()
```

Εφαρμογή 2

```
    import numpy as np
    import matplotlib.pyplot as plt
    import numpy as np
    from sklearn.datasets import load_diabetes
    from sklearn.preprocessing import StandardScaler
    from sklearn.linear_model import LinearRegression

    # Load the diabetes dataset
    diabetes = load_diabetes()
    X, y = diabetes.data[:, :4], diabetes.target  # Use only the first 4 ↩
        explanatory variables

    # Standardize the features
    scaler = StandardScaler()
    X = scaler.fit_transform(X)

    # Perform linear regression using the builtin function for OLS
    reg = LinearRegression()
    reg.fit(X, y)

    # Print the coefficients from the built-in function for OLS
    coefficients_builtin = reg.coef_
    intercept_builtin = reg.intercept_
    print("Coefficients from OLS Regression:")
    for i, coeff in enumerate(coefficients_builtin):
        print(f"Theta{i+1}: {coeff}")
        print(f"Intercept: {intercept_builtin}")

    def sgld(X, Y, learning_rate, num_iterations, beta):
        num_samples, num_features = X.shape
        theta = np.zeros(num_features)  # Initialize theta with zeros
        theta_trace = np.zeros((num_iterations, num_features))  # Store theta ↩
            values
```

```
        for iteration in range(num_iterations):

            gradient = 2 / num_samples * np.dot(X.T, np.dot(X, theta) - Y)

            noise = np.random.randn(num_features)
            theta = theta - learning_rate * gradient + np.sqrt(2 * learning_rate / ↩
                beta) * noise

            theta_trace[iteration] = theta

        return theta_trace
    # Standardize the features
    scaler = StandardScaler()
    X = scaler.fit_transform(X)

    # Add column of 1s for the intercept term
    X = np.hstack((np.ones((X.shape[0], 1)),X))

    learning_rate = 0.01
    num_iterations = 100000

    # Run SGLD for linear regression for different betas
    theta_chain1 = sgld(X, y, learning_rate, num_iterations,1)
    # Convert the chain to a NumPy array
    theta_chain1 = np.array(theta_chain1)

    theta_chain2 = sgld(X, y, learning_rate, num_iterations,100)
    theta_chain2 = np.array(theta_chain2)

    theta_chain3 = sgld(X, y, learning_rate, num_iterations,10000)

    theta_chain3 = np.array(theta_chain3)

    # Print the coefficients from SGLD
    final_coefficients_sgld = theta_chain3[-1, :]
    print("Coefficients from SGLD:")
    for i, coeff in enumerate(final_coefficients_sgld):
        print(f"Theta{i}: {coeff}")

    # trace plots for each thea
    for i in [0,1,2,3,4]:
        plt.plot(theta_chain1[:10000, i], alpha=0.7, label="beta=10")
        plt.plot(theta_chain2[:10000, i], alpha=0.7, label="beta=100")
        plt.plot(theta_chain3[:10000, i], alpha=0.7, label="beta=10000")

        plt.legend()
        plt.xlabel('Iteration')
        plt.show()
```