Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

Παλινδρόμηση με Μη-Φραγμένες Μεταβλητές υπό Συνθήκες Διαφορικής Ιδιωτικότητας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΙΑΣΩΝ Κ. ΜΗΛΙΩΝΗΣ

Επιβλέπων : Δημήτριος Φωτάκης
Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

Παλινδρόμηση με Μη-Φραγμένες Μεταβλητές υπό Συνθήκες Διαφορικής Ιδιωτικότητας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΙΑΣΩΝ Κ. ΜΗΛΙΩΝΗΣ

Επιβλέπων : Δημήτριος Φωτάκης
Αν. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 5η Ιουλίου 2021.

Δημήτριος Φωτάκης Αν. Καθηγητής ΣΗΜΜΥ Ε.Μ.Π.	Στρατής Ιωαννίδης Αν. Καθηγητής ECE Northeastern University (Boston, Massachusetts, USA)	Αριστείδης Παγουρτζής Καθηγητής ΣΗΜΜΥ Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

Ιάσων Κ. Μηλιώνης

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στην σύγχρονη εποχή η οποία χαρακτηρίζεται από καταιγισμό πληροφοριών, η ιδιωτικότητα είναι από τις κρισιμότερες απαιτήσεις για τα ευαίσθητα δεδομένα κάθε ατόμου, η κατάχρηση των οποίων δεδομένων παραμονεύει σε κάθε γωνιά. Σε αυτό το πλαίσιο, ο σχεδιασμός των αλγορίθμων με την ιδιωτικότητα πρώτη κατά νουν είναι μια νέα τάση, στην οποία ο κατεξοχήν αυστηρός ορισμός της ιδιωτικότητας είναι αυτός της Διαφορικής Ιδιωτικότητας. Εξαιτίας των ισχυρών αποδείξιμων ιδιοτήτων της, η Διαφορική Ιδιωτικότητα έχει προταθεί για να επιλυθεί το εξέχον ζήτημα του κατά πόσο είναι δυνατό να προσφέρουμε ιδιωτικότητα και ταυτόχρονα να εκτελούμε διαδικασίες στατιστικής εκτίμησης. Εντούτοις, υπάρχουν εγγενείς δυσκολίες στους μηχανισμούς με πιθανώς μη-φραγμένη ευαισθησία χειρότερης περίπτωσης, που είχαν δράσει ως εμπόδιο προόδου στο πεδίο της Διαφορικής Ιδιωτικότητας την περασμένη δεκαετία. Πρόσφατες εξελίξεις έχουν επιτρέψει την αποδοτική εισαγωγή της ιδιωτικότητας σε διάφορα θεμελιώδη προβλήματα με μη-φραγμένα δεδομένα, με τις σημαίνουσες εργασίες των Karwa and Vadhan (2018) και Kamath et al. (2019) ως χαρακτηριστικά παραδείγματα. Ωστόσο, τα απαραίτητα για την Μηχανική Μάθηση και την Στατιστική προβλήματα της Ανάλυσης Παλινδρόμησης (Regression Analysis) έχουν παραμείνει ανοικτά στην μη-φραγμένη περίπτωση και οι τρεχόντως επιβεβλημένοι περιορισμοί φραγμένων συμμεταβλητών θεωρούνται συχνά ως υπερβολικά περιοριστικοί (Anonymous, 2019). Σε αυτή την διπλωματική εργασία, καταφέρνουμε να κατασκευάσουμε υπολογιστικά αποδοτικούς, διαφορικά ιδιωτικούς αλγορίθμους για τα κλασσικά περιβάλλοντα της Γραμμικής Παλινδρόμησης, της Προσαρμογής Ελαχίστων Τετραγώνων και της Δυαδικής Παλινδρόμησης, που βασίζονται σε πρόσφατες εργασίες για μη-φραγμένη εκτίμηση μέσης τιμής και συνδιασποράς υπό κανονικές (Gaussian) μεταβλητές. Ως προς το τεχνικό μέρος, επεκτείνουμε τις διαφορικά ιδιωτικές τεχνικές για μη-φραγμένη εκτίμηση μέσης τιμής και συνδιασποράς σε sub-gaussian περιβάλλοντα. Τέλος, στην περίπτωση της Δυαδικής Παλινδρόμησης, που συμπεριλαμβάνει τα ουσιώδη και ενδελεχώς μελετημένα μοντέλα της Λογιστικής Παλινδρόμησης και των γραμμικά διαχωρίσιμων Μηχανών Διανυσμάτων Υποστήριξης, θεμελιώνουμε ότι ο αλγόριθμός μας μαθαίνει έναν αμερόληπτο εκτιμητή των αληθινών παραμέτρων παλινδρόμησης, ως προς έναν πολλαπλασιαστικό συντελεστή.

Λέξεις κλειδιά

Στατιστική Μάθηση, Θεωρία Μάθησης, Διαφορική Ιδιωτικότητα, Παλινδρόμηση, Θεωρία Πιθανοτήτων.

Abstract

In the modern era characterized by a deluge of information, privacy is of utmost significance for every individual's sensitive data, whose misuse lurks around the corner. In this context, the design of algorithms with privacy first in mind is a newly-found concept, for which the prevalent rigorous definition of privacy is that of Differential Privacy. Due to its strong provable properties, Differential Privacy has been set forward to resolve the outstanding issue of how it would be possible to confer privacy and at the same time, perform statistical estimation procedures. Yet, there are inherent difficulties with mechanisms of potentially unbounded worst-case sensitivity, acting as an impediment to progress in the differentially private frontier for the past decade. Recent developments have enabled the efficient introduction of privacy into various fundamental problems with unbounded data, with the seminal works of Karwa and Vadhan (2018) and Kamath et al. (2019) as prominent examples. However, the requisite for Machine Learning and Statistics problems of Regression Analysis have remained open in the unbounded regime, and the currently imposed boundedness constraints on the covariates are often considered as overly restrictive (Anonymous, 2019). In this thesis, we manage to construct computationally efficient, differentially private algorithms for the classical regression settings of Linear Regression, Least Squares Fitting and Binary Regression, that build upon recent previous work on unbounded mean and covariance estimation, under Gaussian marginals. From a technical standpoint, we extend differentially private techniques on mean and covariance estimation to the sub-gaussian regime. Finally, in the case of Binary Regression, which captures the fundamental and widely-studied models of logistic regression and linearly-separable Support Vector Machines, we establish that our algorithm learns an unbiased estimate of the true regression parameters, up to a scaling factor.

Key words

Statistical Learning, Learning Theory, Differential Privacy, Regression, Probability Theory.

Ευχαριστίες

Η πορεία του προπτυχιακού κύκλου σπουδών μου μπορεί να περατώνεται με αυτή την εργασία, όμως η πορεία μου στον μετέπειτα στίβο της ζωής μόλις ξεκινάει. Για το γεγονός αυτό, είμαι ευγνώμων στον μέντορά μου, Καθηγητή κύριο Φωτάκη, η αστείρευτη καθοδήγηση του οποίου είχε ως συνέπεια την τρέχουσα ακαδημαϊκή μου πορεία. Ταυτόχρονα, οφείλω θερμές ευχαριστίες στον συνεργάτη και πάνω από όλα φίλο μου, Άλκη Καλαβάση, υποψήφιο διδάκτορα με τον κύριο Φωτάκη, για την καθοριστική συμβολή του στην διαμόρφωση της ακαδημαϊκής προσωπικότητας μου, τις συμβουλές στο ερευνητικό μου έργο και το γεγονός ότι αποτέλεσε μια ανεξάντλητη πηγή γνώσης για εμένα. Έπειτα, θέλω να ευχαριστήσω θερμά τον Καθηγητή κύριο Ιωαννίδη με τον οποίο συναντήθηκα πρώτη φορά στη Βοστώνη, ο οποίος με ενέπνευσε να συνεργαστούμε με εξαιρετικά πρωτότυπα ερευνητικά αποτελέσματα που αποτελούν και τον κεντρικό πυρήνα για την παρούσα διπλωματική μου εργασία.

Θα ήθελα ακόμη να εκφράσω την βαθιά εκτίμησή μου προς τον Καθηγητή κύριο Παγουρτζή, ο οποίος με καθοδήγησε από νωρίς στις σπουδές μου και του οποίου το μάθημα της Κρυπτογραφίας, το οποίο επέλεξα να παρακολουθήσω νωρίτερα από το συνηθισμένο, ήταν ένα από τα εναύσματα που με έκανε να ενθουσιαστώ για την Θεωρητική Πληροφορική. Επίσης, τον ευχαριστώ και για την εμπιστοσύνη που μου επέδειξε ως βοηθός στο μάθημα της Κρυπτογραφίας.

Καθώς κλείνει αυτό το σημαντικό κεφάλαιο της ζωής μου, οφείλω αμέτρητη ευγνωμοσύνη στην Σχολή που με στήριξε και συγκεκριμένα στον Καθηγητή κύριο Κοζύρη και την κυρία Κριθινάκη για την υποστήριξη των πολλαπλών και πολυσχιδών συμμετοχών μου σε παγκόσμιους διαγωνισμούς. Χωρίς την καθοδήγηση και την συμβολή της Σχολής, δεν θα καθίσταντο δυνατά τα επιτεύγματα και η αληθινή γνώση που διαθέτω πλέον στη φαρέτρα μου.

Υπό αυτό το πρίσμα, θα ήθελα να αναφερθώ με χρονολογική σειρά σε ορισμένους καθηγητές που γνώρισα στη Σχολή και με έκαναν να αγαπήσω τα πολλαπλά αντικείμενά της: τον Καθηγητή κύριο Παπασπύρου για την περαιτέρω παρακίνηση του ενδιαφέροντος μου για τις γλώσσες προγραμματισμού, τους μεταγλωττιστές και τις πολύπλευρες πτυχές της ανάπτυξης λογισμικού, τον Καθηγητή κύριο Ζάχο για την εύθυμη παρουσία του στα πρώτα έτη και για το εύρωστο υπόβαθρο γνώσης στα μεγαλύτερα έτη, τον Καθηγητή κύριο Σταυρακάκη για το βιβλίο του Μιγαδική Ανάλυση που υπήρξε το εφαλτήριο της ανείπωτης γοητείας που μου άσκησε αυτό το επίπεδο μαθηματικών, τον Καθηγητή κύριο Γλύτση για την φορμαλιστική μορφή της θεωρίας ηλεκτρομαγνητικών πεδίων για την οποία πάντοτε προηγουμένως γνησίως απορούσα, και τον Καθηγητή κύριο Σούντρη για την ενασχόληση που μου προσέφερε απλόχερα με πρακτικά συστήματα.

Πέραν του κεντρικού ρόλου της Σχολής στην παροχή κορυφαίας γνώσης, ο χώρος των αμφιθεάτρων υπήρξε και αφορμή για διαφόρων ειδών κοινωνικές αλληλεπιδράσεις με αρκετούς (τουλάχιστον ενδιαφέροντες) συμφοιτητές. Οι ποικιλότροπες αμφίσημες εμπειρίες που προέκυψαν είχαν ως συνέπεια την καλλιέργειά μου ως ανθρώπου και την κρίσιμη εξέλιξη της προσωπικότητάς μου. Γνωρίζουν ποιοι είναι αυτοί τους οποίους ευχαριστώ.

Τελευταίο και σημαντικότερο, θα ήθελα να κάνω ειδική μνεία στους γονείς μου: πρώτα από όλα για την ύπαρξή μου και έπειτα για την αληθινή στήριξή τους σε ό,τι αποφάσιζα να κάνω και για το γεγονός ότι οι συμβουλές και η καθοδήγησή τους, μεταξύ άλλων, είναι αυτά που με έχουν καταστήσει αυτό που είμαι σήμερα και έχουν διαπλάσει τον τρόπο σκέψης μου.

Ιάσων Κ. Μηλιώνης,

Αθήνα, 5η Ιουλίου 2021

Περιεχόμενα

List of Algorithms

Κεφάλαιο 1

Εκτεταμένη Ελληνική Περίληψη

Σε αυτό το πρώτο κεφάλαιο, θα παρουσιάσουμε συνοπτικά και πλήρως τα περιεχόμενα της διπλωματικής εργασίας. Για περισσότερες λεπτομερείς τεχνικές λεπτομέρειες και αποδείξεις, ο αναγνώστης είναι ευπρόσδεκτος να εξερευνήσει τα περαιτέρω κεφάλαια που αντιστοιχούν με τους τίτλους των επιμέρους ενοτήτων στο τρέχον κεφάλαιο.

1.1 Εισαγωγή

Οι σύγχρονοι αλγόριθμοι αλληλεπιδρούν πολύπλοκα με βάσεις δεδομένων που περιέχουν τα δεδομένα των χρηστών. Σε συγκεκριμένες περιπτώσεις όπως στον τομέα της υγείας και των οικονομικών, ή ακόμη και πιο γενικά, οι χρήστες μπορεί να θεωρούν ότι τα δεδομένα τους είναι ευαίσθητα ή απλά να θέλουν να μην ταυτοποιηθούν μέσω του (δημόσια διαθέσιμου) αποτελέσματος ενός αλγορίθμου. Για παράδειγμα, ένα μοντέλο Μηχανικής Μάθησης που προσπαθεί να προβλέψει έναν συγκεκριμένο τύπο καρκίνου πρέπει να μην είναι δυνατόν να αποκαλύψει την συμμετοχή κάποιου ατόμου στο σύνολο δεδομένων από το οποίο ο αλγόριθμος έχει μάθει το μοντέλο του.

Η διασφάλιση του ιδιωτικού χειρισμού των δεδομένων κάθε ατόμου και η ταυτόχρονη παραγωγή συμπερασμάτων για την συμπεριφορά ενός πληθυσμού είναι δύο στόχοι που είναι σε μεγάλο βαθμό αντικρουόμενοι. Από τη μία πλευρά, η εγγύηση της τέλειας ιδιωτικότητας σημαίνει ότι δεν μπορεί να εξαχθεί καμία χρήσιμη στατιστική πληροφορία, ενώ από την άλλη πλευρά, η εγγύηση ενός τέλεια ακριβούς αποτελέσματος με βάση τα διαθέσιμα δεδομένα επιβάλλει τον περιορισμό (ή ακόμη χειρότερα, την εγκατάλειψη) της ιδιωτικότητας. Για αυτό το λόγο, η αποκάλυψη των συμβιβασμών που θα επέτρεπαν την συνύπραξη της ιδιωτικότητας για τα άτομα που συνεισφέρουν τα δεδομένα τους μαζί με την στατιστική εκτίμηση μιας χρήσιμης ποσότητας με επαρκή ακρίβεια είναι ένα ενθουσιαστικό αντικείμενο προς μελέτη.

Προς αυτή την κατεύθυνση, έχουν υπάρξει παλαιότερες προσπάθειες για την *ανωνυμοποίηση* συνόλων δεδομένων έτσι ώστε να ανακοινωθούν δημόσια για την εξυπηρέτηση κάποιου κοινού σκοπού, που τελικώς κατέληξαν σε φιάσκα μέσω της ταυτοποίησης συγκεκριμένων χρηστών από αυτά τα "ανωνυμοποιημένα" σύνολα δεδομένων (δείτε, για παράδειγμα, τους Narayanan and Shmatikov (2008) για μια ενδεικτική περίπτωση). Ξεκάθαρα, το να διαθέτουμε μία μη-αυστηρή εγγύηση ότι "όλα τα προσωπικά δεδομένα έχουν σβηστεί πριν από την απελευθέρωση του συνόλου δεδομένων" είναι απολύτως ανεπαρκές: πρέπει να υπάρχει μια μαθηματικώς αυστηρή *εγγύηση ιδιωτικότητας* την οποία ο σχεδιαστής του αλγορίθμου θα μπορεί να διασφαλίσει με απόλυτη ακρίβεια ότι ο αλγόριθμός του ικανοποιεί.

Η *Διαφορική Ιδιωτικότητα* (Dwork et al., 2006) είναι μια έννοια που έχει αναπτυχθεί ως απάντηση ακριβώς στο παραπάνω πλαίσιο: προσφέρει μια αυστηρή εγγύηση ιδιωτικότητας με ιδιαιτέρως ευνοϊκές, ισχυρές ιδιότητες για την διαφύλαξη της ιδιωτικότητας των ατόμων των οποίων τα δεδομένα περιέχονται σε ένα σύνολο δεδομένων που χρησιμοποιείται από κάποιον αλγόριθμο. Βασίζεται στην γενική αρχή ότι, τελικώς, η ιδιωτικότητα για ένα άτομο αφορά την προστασία του από την ταυτοποίηση μέσω των (δημόσια διαθέσιμων) αποτελεσμάτων κάποιου/ων αλγορίθμου/ων. Διαισθητικώς, η Διαφορική Ιδιωτικότητα (Differential Privacy, DP) *εγγυάται*

ότι η συμπερίληψη ή εξαίρεση ενός συγκεκριμένου ατόμου από ένα σύνολο δεδομένων (μέσω τροποποίησης μιας εκ των καταχωρήσεων του συνόλου δεδομένων) δεν θα πρέπει να μπορεί να επηρεάσει "αρκετά" το αποτέλεσμα του αλγορίθμου.

Συγκεκριμένα, ας θεωρήσουμε ένα σύνολο δεδομένων (μια συλλογή από n δείγματα) $X = (X_1, X_2, \ldots, X_n) \in \mathcal{X}^n$. Ορίζουμε δύο σύνολα δεδομένων $X, X' \in \mathcal{X}^n$ ως γειτονικά, αν διαφέρουν σε το πολύ ένα συγκεκριμένο δείγμα (π.χ. το δείγμα $X_i \neq X'_i$ για κάποιο i, ενώ όλα τα άλλα δείγματα είναι τα ίδια). Ένας τυχαιοποιημένος αλγόριθμος M (που επισης καλείται και "μηχανισμός" για ιστορικούς λόγους) θα λέμε ότι είναι *διαφορικά ιδιωτικός*, εάν οι κατανομές των $M(X)$ και $M(X')$ είναι "πολύ παρόμοιες" για κάθε ζεύγος γειτονικών δεδομένων X, X'. Δίνουμε τον μαθηματικώς ακριβή ορισμό του DP στην επόμενη ενότητα, μαζί με το τι εννοούμε υπό τον όρο "ομοιότητα" κατανομών.

Όπως προαναφέρθηκε ανωτέρω, η Διαφορική Ιδιωτικότητα είναι διαδεδομένη λόγω των ισχυρών, ευνοϊκών ιδιοτήτων της που αποδεικνύουν ότι αυτού του τύπου η εγγύηση ιδιωτικότητας *διατηρείται* μέσω μια ευρείας γκάμας πράξεων είτε στο αποτέλεσμα του αλγορίθμου ("μεταεπεξεργασία") ή με την ταυτόχρονη εκτέλεση πολλών (διαφορικά ιδιωτικών) αλγορίθμων σε ένα ιδιωτικό σύνολο δεδομένων ("σύνθεση").

Για να προσφέρουμε έναν διαφορικά ιδιωτικό αλγόριθμο εκτίμησης, η κλασσική προσέγγιση έχει ως εξής: ξεκινούμε με έναν μη-ιδιωτικό αλγόριθμο (ο οποίος υπολογίζει κάποια συνάρτηση $f(X)$ του συνόλου δεδομένων εισόδου X) και προσθέτουμε τυχαίο θόρυβο η ενός προσαρμόσιμου επιπέδου αναλόγως με το επίπεδο ιδιωτικότητας που επιθυμούμε να προσφέρουμε. Η έξοδος του (ιδιωτικού) αλγορίθμου είναι πλέον το $f(X) + \eta$. Ο τυχαίος θόρυβος η μπορεί να ακολουθεί ένα ευρύ σύνολο κατανομών, με μακράν τις πιο κοινές να είναι η Λαπλασιανή και η Κανονική (Γκαουσιανή) κατανομή.

Στην εκτίμηση στατιστικών ποσοτήτων, ενδιαφερόμαστε για δύο στόχους: 1. να εγγυηθούμε ότι ο αλγόριθμος θα είναι *ιδιωτικός* (συγκεκριμένα, διαφορικά ιδιωτικός) και 2. να διασφαλίσουμε ότι ο αλγόριθμός μας θα είναι *ακριβής* ως προς την πληθυσμιακή ποσότητα που επιθυμούμε να εκτιμήσουμε. Από τα παραπάνω προκύπτει ότι επιθυμούμε η εγγύηση ιδιωτικότητας που παρέχουμε να είναι "χειρότερης περίπτωσης", δηλαδή να συμπεριλαμβάνει κάθε πιθανό σημείο εισόδου (ακόμη και αυτά τα δεδομένα εισόδου που ενδέχεται να είναι αρκετά μακριά από τις αναμενόμενες τιμές), εφόσον θέλουμε να προστατεύσουμε τα δεδομένα κάθε ατόμου από την ταυτοποίηση, ανεξαρτήτως από την διακύμανσή τους από τη μέση τιμή του πληθυσμού. Αντιθέτως, ο ιδιωτικός αλγόριθμος εκτίμησης τον οποίο διαθέτουμε μπορεί να προσφέρει μία εγγύηση ακρίβειας με νόημα σε ένα περιβάλλον "μέσης περίπτωσης", δηλαδή όταν το σύνολο δεδομένων που θα λάβει ο αλγόριθμος θα είναι "αρκετά καλό" ως προς κάποια μετρική.

Ωστόσο, εξαιτίας της φύσης χειρότερης περίπτωσης της εγγύησης της Διαφορικής Ιδιωτικότητας, είναι γενικώς πολύ απαιτητικό να κατασκευάσουμε αλγορίθμους με μη-φραγμένα δεδομένα εισόδου, δηλαδή όταν τα δείγματα μπορεί να είναι αυθαίρετα μακριά από κάποια "αναμενόμενη τιμή". Για αυτό το λόγο, η αρχική βιβλιογραφία (δείτε, για παράδειγμα, τους Dwork and Roth (2014) για μια ανασκόπηση) στηρίζεται κρίσιμα σε υποθέσεις φραγμένων δειγμάτων εισόδου του συνόλου δεδομένων εισόδου. Ακόμη και για το πρόβλημα της Εκτίμησης Μέσης Τιμής, όταν έχουμε δείγματα που ακολουθούν μια πολυδιάστατη Κανονική κατανομή (για την οποία το αντίστοιχο πρόβλημα θεωρείται ως ένα εκ των θεμελιωδέστερων προβλημάτων της στατιστικής και ταυτόχρονα, ένα πρόβλημα στο οποίο μπορεί να έχουμε *ακραίες τιμές* αυθαίρετα μακριά από την μέση τιμή, παρά το γεγονός ότι η πλειονότητα των δειγμάτων ευρίσκεται γύρω από μία καλώς καθορισμένη μέση τιμή), δεν ήταν παρά προσφάτως που οι Kamath et al. (2019) κατάφεραν να κατασκευάσουν έναν διαφορικά ιδιωτικό αλγόριθμο για την αποδοτική εκτίμηση της μέσης τιμής και του πίνακα συνδιασποράς της πολυδιάστατης Κανονικής κατανομής. Η εργασία τους βασίστηκε σε έρευνα των Karwa and Vadhan (2018), οι οποίοι έδειξαν πώς μπορεί να εκτιμηθεί με διαφορικά ιδιωτικό και αποδοτικό τρόπο η μέση τιμή μιας (μονοδιάστατης) Κανονικής τυχαίας μεταβλητής.

Σε αυτή την εργασία, πραγματοποιούμε ένα περαιτέρω βήμα και εξετάζουμε την ιδιωτική

και αποδοτική στατιστική εκτίμηση σε συνηθισμένα περιβάλλοντα παλινδρόμησης, όπου έχουμε μη-φραγμένα δείγματα στο σύνολο δεδομένων εισόδου. Η παλινδρόμηση είναι ένα σύνολο διαδικασιών που είναι βασική για την στατιστική εκτίμηση, της οποίας αρχικές μορφές είχαν εμφανιστεί ήδη από τους Legendre (1806) και Gauss (δείτε, για παράδειγμα, το ιστορικό σημείωμα του Plackett (1949)). Υπάρχει μια πληθώρα από περιβάλλοντα παλινδρόμησης: εδώ, θα εξετάσουμε αρκετά από αυτά τα περιβάλλοντα (όπως την εύρεση του Εκτιμητή Ελαχίστων Τετραγώνων και την εκτίμηση των συντελεστών παλινδρόμησης ενός υποκρυπτόμενου γραμμικού μοντέλου από δεδομένα με επιγραφές/labels) υπό κανονικές (Gaussian) μεταβλητές. Το σύνολο δεδομένων εισόδου αποτελείται από n σημεία $(\boldsymbol{X}_i, y_i)$ με συμμεταβλητές $\boldsymbol{X}_i$ οι οποίες είναι επισημειωμένες με επιγραφές y_i, όπου είναι καθοριστικό το γεγονός ότι οι συμμεταβλητές επιτρέπεται να είναι *μη-φραγμένες*.

Οι εργασίες επάνω στην μάθηση γραμμικών μοντέλων από μόνες τους είναι άφθονες (δείτε, για παράδειγμα, τους Cai et al. (2020); Wang (2018) για δύο πρόσφατες ανασκοπήσεις). Εντούτοις, παρά το έντονο ενδιαφέρον σε αυτό το ζήτημα (δείτε τις εργασίες των Iyengar et al. (2019); Zhang et al. (2017); Jain and Thakurta (2014), για να ονοματίσουμε μόνον μερικές), *όλη η προϋπάρχουσα μελέτη στην παλινδρόμηση προσφέρει την εγγύηση της Διαφορικής Ιδιωτικότητας υποθέτοντας φραγμένες συμμεταβλητές*. Διαισθητικά, αυτό μπορεί να εξηγηθεί απλώς εξετάζοντας ακόμη και τον απλούστερο εκτιμητή ελαχίστων τετραγώνων που χρησιμοποιείται για την Γραμμική Παλινδρόμηση. Είναι εύκολο να δει κανείς ότι η *ευαισθησία* του εκτιμητή αυτού, δηλαδή η μεταβλητότητά του υπό αλλαγές σε ένα μόνο δείγμα του συνόλου δεδομένων εισόδου, καθορίζεται από τον πίνακα των δειγμάτων. Επειδή η ευαισθησία έχει κεντρικό ρόλο στις εγγυήσεις Διαφορικής Ιδιωτικότητας, η κυρίαρχη τεχνική για τον έλεγχο της ευαισθησίας είναι το να φράσσονται οι ιδιοτιμές του πίνακα των δειγμάτων. Ως εκ τούτου, η υπόθεση των φραγμένων συμμεταβλητών είναι μια πανταχού παρούσα υπόθεση στην DP βιβλιογραφία τόσο για την Γραμμική Παλινδρόμηση όσο και για την μάθηση γενικευμένων γραμμικών μοντέλων.

Αυτή ακριβώς η υπόθεση, όμως, είναι εξόχως περιοριστική και αναγνωρίζεται συχνά ως μια αδυναμία των διαφορικά ιδιωτικών αλγορίθμων παλινδρόμησης από πρακτικής σκοπιάς (Anonymous, 2019). Είναι, επίσης, ένα μείζον μειονέκτημα από θεωρητικής πλευράς, αφού εμποδίζει την μελέτη διαφορικά ιδιωτικών εκτιμητών σε δεδομένα που έχουν ληφθεί από τυχαίες κατανομές με μη-φραγμένο σύνολο ορισμού. Ακόμη και η Κανονική κατανομή, που είναι πιθανώς η συνηθέστερα χρησιμοποιούμενη στην βιβλιογραφία Στατιστικής Μηχανικής Μάθησης (Deng et al., 2021; Daskalakis et al., 2020; Kini and Thrampoulidis, 2020; Diakonikolas et al., 2019b; Nakkiran, 2019; Kreidler et al., 2018) δεν μπορεί να χρησιμοποιηθεί με τους τρέχοντες διαφορικά ιδιωτικούς αλγορίθμους παλινδρόμησης διατηρώντας τις εγγυήσεις Διαφορικής Ιδιωτικότητάς τους.

Η εργασία μας στοχεύει να αντιμετωπίσει άμεσα αυτό το ζήτημα, δίνοντας DP αλγορίθμους για συνηθισμένα περιβάλλοντα παλινδρόμησης υπό την υπόθεση (μη-φραγμένων) Κανονικών συμμεταβλητών.

1.1.1 Η συνεισφορά μας

Η πρώτη μείζων συνεισφορά μας είναι η θετική απάντηση στην ακόλουθη ερώτηση:

Question 1.1.1. *Είναι δυνατή η ιδιωτική Ανάλυση Παλινδρόμησης με μη-φραγμένες συμμεταβλητές;*

Μελετούμε το πρόβλημα αυτό υπό το πρίσμα τριών σεναρίων: της Γραμμικής Παλινδρόμησης, της Προσαρμογής Ελαχίστων Τετραγώνων και της Δυαδικής Παλινδρόμησης. Σε όλα αυτά τα περιβάλλοντα, υποθέτουμε κανονικές (Gaussian) συμμεταβλητές.

Στο περιβάλλον της Προσαρμογής Ελαχίστων Τετραγώνων, δεδομένου ενός συνόλου εισόδου $\{(\boldsymbol{X}_i, y_i)\}$, ο στόχος μας είναι να υπολογίσουμε ιδιωτικά και αποδοτικά έναν εκτιμητή που να είναι κοντά στον Εκτιμητή Ελαχίστων Τετραγώνων (Least Squares Estimate, LSE), δηλαδή τους συντελεστές του βέλτιστα προσαρμοζόμενου γραμμικού μοντέλου στα δεδομένα εισόδου. Σε αυτό το πρόβλημα, υποθέτουμε ότι οι επιγραφές y_i είναι φραγμένες, αλλά δεν κάνουμε καμία

περαιτέρω υπόθεση για το πώς σχετίζονται με τις συμμεταβλητές $\boldsymbol{X}_i \in \mathbb{R}^d$. Το κύριο αποτέλεσμά μας είναι το ακόλουθο:

Informal Theorem 1. *Για ακρίβεια $\alpha > 0$ και εγγυήσεις ιδιωτικότητας $\epsilon, \delta > 0$, υπάρχει ένας αποδοτικός (ϵ, δ)-DP αλγόριθμος ο οποίος με μεγάλη πιθανότητα προσεγγίζει αυθαίρετα α-κοντά τον Εκτιμητή Ελαχίστων Τετραγώνων χρησιμοποιώντας $n = \widetilde{O}\left(d/\alpha^2 + d^{3/2}\log(1/\delta)/(\alpha\epsilon)\right)$ δείγματα.*

Στο δεύτερό μας περιβάλλον (αυτό της απλής Γραμμικής Παλινδρόμησης), αποδεικνύουμε ότι μια τροποποίηση των μεθόδων των Kamath et al. (2019) και Karwa and Vadhan (2018) είναι αρκετή για να μας δώσει έναν ιδιωτικό και ακριβή εκτιμητή των υποκρυπτόμενων συντελεστών του γραμμικού μοντέλου.

Τέλος, στο περιβάλλον της Δυαδικής Παλινδρόμησης, υποθέτουμε περαιτέρω ότι οι επιγραφές είναι δυαδικές (δηλαδή, $y_i = \pm 1$) και ότι οι συμμεταβλητές έχουν μηδενική αναμενόμενη τιμή. Επιπροσθέτως, οι επιγραφές γεννώνται από ένα γενικευμένο γραμμικό μοντέλο της μορφής $\mathbf{Pr}[y_i = +1|\boldsymbol{X}_i] = f(\boldsymbol{\beta}^T\boldsymbol{X}_i)$, όπου $f : \mathbb{R} \to [0, 1]$ είναι η συνάρτηση του μοντέλου και $\boldsymbol{\beta} \in \mathbb{R}^d$ είναι οι αληθινοί συντελεστές του μοντέλου παλινδρόμησης. Αυτό το περιβάλλον συμπεριλαμβάνει κάποιες από τις πιο βασικές διαδικασίες της Μηχανικής Μάθησης, όπως η Λογιστική Παλινδρόμηση και η μάθηση γραμμικά διαχωρίσιμων Μηχανών Διανυσμάτων Υποστήριξης (ΜΔΥ). Το δεύτερό μας κύριο αποτέλεσμα είναι ότι ο ίδιος διαφορικά ιδιωτικός εκτιμητής που χρησιμοποιήσαμε στο περιβάλλον της Προσαρμογής Ελαχίστων Τετραγώνων μπορεί να εφαρμοστεί στην Δυαδική Παλινδρόμηση για να λάβουμε τις ακόλουθες εγγυήσεις:

Informal Theorem 2. *Για ακρίβεια $\alpha > 0$ και εγγυήσεις ιδιωτικότητας $\epsilon, \delta > 0$, υπάρχει ένας αποδοτικός (ϵ, δ)-DP αλγόριθμος, ο οποίος με μεγάλη πιθανότητα προσεγγίζει αυθαίρετα α-κοντά τους πραγματικούς συντελεστές του γραμμικού μοντέλου ως προς έναν πολλαπλασιαστικό παράγοντα χρησιμοποιώντας $n = \widetilde{O}\left(d/\alpha^2 + d^{3/2}\log(1/\delta)/(\alpha\epsilon)\right)$ δείγματα.*

Εξ όσων γνωρίζουμε, αυτά τα αποτελέσματα συνιστούν τους πρώτους ιδιωτικούς και αποδοτικούς αλγορίθμους για Ανάλυση Παλινδρόμησης με μη-φραγμένες συμμεταβλητές. Από τεχνικής απόψεως, η ανάλυσή μας έγκειται στο γεγονός ότι ο LSE απαιτεί τον υπολογισμό ενός αντιστρόφου ενός πίνακα ροπής, καθώς επίσης και την αναμενόμενη τιμή της κεντρικής τυχαίας ποσότητας $y_i\boldsymbol{X}_i$. Το πρώτο μπορεί να εκτιμηθεί με ιδιωτικό τρόπο χρησιμοποιώντας τις πρόσφατες τεχνικές που αναπτύχθηκαν από τους Kamath et al. (2019) και Karwa and Vadhan (2018). Εμείς αποδεικνύουμε επίσης ότι το δεύτερο μπορεί να εκτιμηθεί με έναν παρόμοιο τρόπο, επεκτείνοντας τις εργασίες των Kamath et al. (2019) και Karwa and Vadhan (2018) σε sub-gaussian τυχαία διανύσματα. Δείχνουμε, ακόμη, ότι η εξάρτηση της δειγματικής πολυπλοκότητας στην διάσταση των συμμεταβλητών d μπορεί να βελτιωθεί στα περιβάλλοντα που μελετούμε (από $\widetilde{O}(d^2)$ σε $\widetilde{O}(d^{3/2})$) εν συγκρίσει με την εκτίμηση συνδιασποράς.

1.2 Υπόβαθρο στην Διαφορική Ιδιωτικότητα

Η έννοια της Διαφορικής Ιδιωτικότητας διατυπώθηκε μαθηματικά το 2006 (Dwork et al., 2006). Ήταν η αυστηρή μορφή της ιδιωτικότητας δεδομένων που άντεξε στη διάρκεια του χρόνου και έχει γίνει πλέον η κυρίαρχη έννοια ιδιωτικότητας για στατιστικές εφαρμογές. Η Διαφορική Ιδιωτικότητα επιλύει την πρόκληση που ετέθη στην προηγούμενη ενότητα: οποιοσδήποτε αντίπαλος, όση υπολογιστική ισχύ και να διαθέτει, δε θα μπορεί να ξεχωρίσει την συμμετοχή ενός συγκεκριμένου ατόμου σε ένα (ιδιωτικό) σύνολο δεδομένων μέσω των αποτελεσμάτων ενός διαφορικά ιδιωτικού αλγορίθμου που εκτελέστηκε σε αυτό το σύνολο δεδομένων. Εξαιτίας των ιδιοτήτων που θα εξερευνήσουμε συντόμως, η Διαφορική Ιδιωτικότητα χρησιμοποιείται ευρέως, με το πιο πρόσφατο διακεκριμένο παράδειγμα να είναι η χρήση της για τα αποτελέσματα της δεκαετούς απογραφής του 2020 των Ηνωμένων Πολιτειών της Αμερικής (US Census Bureau, 2020).

Η Διαφορική Ιδιωτικότητα είναι μια ιδιότητα ενός τυχαιοποιημένου αλγορίθμου/μηχανισμού, που συμβολίζεται με M. Διαισθητικά, η Διαφορική Ιδιωτικότητα εγγυάται ότι όποιο (ένα) άτομο και αν αλλάζει σε μια μεγάλη συλλογή από δεδομένα ατόμων, η έξοδος του αλγορίθμου θα παραμείνει "περίπου η ίδια". Πιο συγκεκριμένα, η πιθανοτική κατανομή των πιθανών εξόδων του αλγορίθμου θα παραμείνει περίπου η ίδια.

Μετακινούμαστε τώρα προς τον αυστηρό ορισμό της Διαφορικής Ιδιωτικότητας (DP). Λέμε ότι ένας τυχαιοποιημένος αλγόριθμος/μηχανισμός M ο οποίος παίρνει ως είσοδο ένα σύνολο δεδομένων X είναι (ϵ, δ) διαφορικά ιδιωτικός (DP), εάν για όλα τα σύνολα δεδομένων X, X' που είναι γειτονικά (δηλαδή μόνο ένα δείγμα είναι διαφορετικό, κάτι που μπορεί να συμβολιστεί ως $\|X - X'\| = 1$ για κάποια κατάλληλα ορισμένη νόρμα) και όλα τα πιθανά υποσύνολα του συνόλου τιμών της εξόδου του αλγορίθμου $Y \subseteq \mathrm{Im}(M)$, ισχύει ότι η πιθανότητα η έξοδος του αλγορίθμου (όταν εκτελεστεί στο σύνολο δεδομένων X) να ανήκει στο σύνολο Y είναι κοντά στην πιθανότητα η έξοδος του αλγορίθμου (όταν εκτελεστεί στο *γειτονικό* σύνολο δεδομένων X') να ανήκει στο ίδιο σύνολο Y. Συγκεκριμένα, έχουμε τον ακόλουθο αυστηρό ορισμό.

Definition 1.2.1 (Διαφορική Ιδιωτικότητα (Dwork et al., 2006))**.** Ένας τυχαιοποιημένος αλγόριθμος $M : \mathcal{X}^n \to \mathcal{Y}$ ικανοποιεί την (ϵ, δ) Διαφορική Ιδιωτικότητα (ισοδύναμα, λέγεται ότι είναι (ϵ, δ)-DP) εάν για κάθε ζεύγος γειτονικών συνόλων δεδομένων $X, X' \in \mathcal{X}^n$ τα οποία διαφέρουν σε το πολύ ένα στοιχείο/δείγμα, ισχύει ότι

$$\mathbf{Pr}[M(X) \in Y] \leq \exp(\epsilon)\, \mathbf{Pr}[M(X') \in Y] + \delta\,, \forall Y \subseteq \mathcal{Y}\,.$$

Έπειτα, θα εξετάσουμε εδώ ορισμένες ιδιότητες των διαφορικά ιδιωτικών αλγορίθμων που κάνουν την παρεχόμενη εγγύηση ιδιωτικότητας *εύρωστη σε διάφορες μεταβολές* στον ίδιο τον αλγόριθμο, σε ένα σύνολο ατόμων στους οποίους θέλουμε να προσφέρουμε ιδιωτικότητα και όταν συνδυάζουμε αποτελέσματα μιας σειράς αλγορίθμων που εκτελούνται σε (ιδιωτικά) σύνολα δεδομένων εισόδου.

Αντοχή στην μετα-επεξεργασία. Η πρώτη ιδιότητα είναι αυτή της "μετα-επεξεργασίας," η οποία διαισθητικά λέει ότι είναι αδύνατον να αδυνατιστεί η εγγύηση ιδιωτικότητας που παρέχεται από τον ορισμό της Διαφορικής Ιδιωτικότητας χωρίς επιπρόσθετη γνώση για το ιδιωτικό σύνολο δεδομένων που χρησιμοποιήθηκε για να υπολογισθεί η διαφορικά ιδιωτική στατιστική ποσότητα.

Lemma 1.2.1 (Μετα-επεξεργασία)**.** *Δεδομένου ενός (ϵ, δ)-διαφορικά ιδιωτικού τυχαιοποιημένου αλγορίθμου $M : \mathcal{X}^n \to \mathcal{Y}$, έστω $f : \mathcal{Y} \to \mathcal{R}$ μία αυθαίρετη τυχαιοποιημένη απεικόνιση. Τότε, ο μηχανισμός $f \circ M : \mathcal{X}^n \to \mathcal{R}$ είναι (ϵ, δ)-διαφορικά ιδιωτικός.*

Οποιοσδήποτε αλγόριθμος μετα-επεξεργασίας και να εφαρμοστεί στο αποτέλεσμα ενός (ϵ, δ)-DP αλγορίθμου δεν θα ως συνέπεια απολύτως καμία εξασθένιση της εγγύησης ιδιωτικότητας. Ουσιαστικά, αυτό που λέει η ιδιότητα αυτή είναι πως θα μπορούσαμε να απελευθερώσουμε (δημόσια) ένα αποτέλεσμα ενός διαφορικά ιδιωτικού αλγορίθμου, επιτρέποντας σε οποιονδήποτε να κάνει οποιονδήποτε επιπρόσθετο υπολογισμό χωρίς να χρειαστεί να ανησυχούμε μήπως καταφέρει να παραβιάσει την (ϵ, δ) εγγύηση ιδιωτικότητας.

Ιδιωτικότητα ομάδας. Η δεύτερη επιθυμητή ιδιότητα αφορά το τι συμβαίνει όταν επιθυμούμε να εγγυηθούμε ιδιωτικότητα για ομάδες από k άτομα (και όχι μόνο ένα, όπως λέει ο ορισμός της Διαφορικής Ιδιωτικότητας). Σε αυτή την περίπτωση, μπορούμε να εγγυηθούμε κάτι παρόμοιο με την εγγύηση ιδιωτικότητας του DP, κάτι που είναι ιδιαιτέρως χρήσιμο, για παράδειγμα, για μέλη του ίδιου νοικοκυριού (οικογένειας). Στο ακόλουθο Λήμμα, αναφερόμαστε μόνο στην περίπτωση του "γνήσιου" $(\epsilon, 0)$-DP, αφού οι εγγυήσεις για την περίπτωση του γενικού (ϵ, δ)-DP είναι πιο σύνθετες.

Lemma 1.2.2 (Ιδιωτικότητα ομάδας)**.** *Δεδομένου ενός* $(\epsilon, 0)$*-διαφορικά ιδιωτικού αλγορίθμου* $M : \mathcal{X}^n \to \mathcal{Y}$*, για κάθε ζεύγος συνόλων δεδομένων* $X, X' \in \mathcal{X}^n$ *που διαφέρουν σε το πολύ* k *στοιχεία (δηλαδή* k *από τα αρχικά* n *στοιχεία έχουν αλλάξει), ισχύει ότι*

$$\mathbf{Pr}[M(X) \in Y] \leq \exp(k\epsilon)\, \mathbf{Pr}[M(X') \in Y]\,, \forall Y \subseteq \mathcal{Y}\,.$$

Σύνθεση. Συνεχίζουμε τώρα στα θεωρήματα σύνθεσης της Διαφορικής Ιδιωτικότητας, τα οποία λένε ότι οι διαφορικά ιδιωτικοί αλγόριθμοι μπορούν να συνδυαστούν και αναλύουν επακριβώς πώς οι εγγυήσεις ιδιωτικότητας εξασθενίζουν από αυτή την "σύνθεση."

Γιατί είναι επιθυμητή η σύνθεση; Η μεγαλύτερη απειλή στην ιδιωτικότητα προέρχεται από τον συνδυασμό διαφορετικών αποτελεσμάτων, καθένα εκ των οποίων μπορεί (με μια πρώτη ματιά) να μοιάζουν ιδιωτικά από μόνα τους. Ας επαναλάβουμε το διαφωτιστικό παράδειγμα των Steinke and Ullman (2020): έστω ένας εργοδότης ο οποίος απασχολεί 1000 άτομα και τους έχει ασφαλίσει σε ένα προσαρμοσμένο ασφαλιστικό πλάνο ενός παρόχου ασφαλειών. Ο εργοδότης μπορεί να επιθυμεί περιληπτικά δεδομένα για το πόσοι από τους συνολικούς εργαζομένους του έχουν κάποια συγκεκριμένη ιατρική ανάγκη. Ας υποθέσουμε ότι αυτό το ερώτημα απαντάται και η απάντηση είναι 42 από τους τρέχοντες εργαζομένους. Αυτός ο αριθμός από μόνος του δεν συνιστά παραβίαση της ιδιωτικότητας, εφόσον δεν είναι δυνατόν για τον εργοδότη να γνωρίζει εάν κάποιος συγκεκριμένος εργαζόμενος έχει ή όχι την ερωτηθείσα ιατρική ανάγκη. Ο εργοδότης διαθέτει μόνο συνοψισμένα δεδομένα, τα οποία μπορεί με μια πρώτη ματιά να φαίνονται ιδιωτικά. Ωστόσο, τι συμβαίνει εάν κάποιος άλλος εργαζόμενος προσληφθεί στην εταιρεία του εργοδότη; Τότε, ο εργοδότης θα μπορούσε να ρωτήσει ξανά την ίδια (περιληπτική) ερώτηση και έστω ότι το αποτέλεσμα αυτή τη φορά ήταν 43. Άμεσα, ο εργοδότης γνωρίζει ότι ο συγκεκριμένος εργαζόμενος τον οποίο μόλις προσέλαβε έχει την συγκεκριμένη ιατρική ανάγκη. Αυτό συνιστά μία ακραία παραβίαση της ιδιωτικότητας. Παρατηρείστε ότι ο εργοδότης μπόρεσε να λάβει αυτή την κρίσιμη πληροφορία μέσω του *συνδυασμού* αποτελεσμάτων που εκ πρώτης όψεως φαίνονταν ιδιωτικά. Αυτός είναι και ο λόγος που η σύνθεση, ένα παράδειγμα της οποίας είναι η χρήση του ίδιου αλγορίθμου πολλαπλές φορές, είναι κρίσιμη για την απόδειξη μιας έννοιας διατήρησης της ιδιωτικότητας. Ο συνδυασμός πληροφοριών από διάφορες πηγές που έχουν απελευθερωθεί χωρίς μαθηματικά αυστηρές εγγυήσεις μπορεί να προβεί καταστροφικός για ορισμένα άτομα, όπως οι Narayanan and Shmatikov (2008) έδειξαν στην πράξη, χρησιμοποιώντας δημόσια δεδομένα από το IMDb για να σπάσουν την ανωνυμία ενός συνόλου δεδομένων του Netflix που είχε απελευθερωθεί για έναν διαγωνισμό.

Lemma 1.2.3 (Απλή σύνθεση $(\epsilon, 0)$-DP μηχανισμών)**.** *Έστω* $M_1 : \mathcal{X}^n \to \mathcal{R}_1$ *να είναι ένας* $(\epsilon_1, 0)$*-DP αλγόριθμος και* $M_2 : \mathcal{X}^n \to \mathcal{R}_2$ *ένας* $(\epsilon_2, 0)$*-DP αλγόριθμος, των οποίων οι εσωτερικές τυχαίες ρίψεις είναι ανεξάρτητες μεταξύ τους. Ορίζουμε τον αλγόριθμο* $M : \mathcal{X}^n \to \mathcal{R}_1 \times \mathcal{R}_2$ *ως την απεικόνιση* $M(X) = (M_1(X), M_2(X))$*. Τότε, ο αλγόριθμος* M *είναι* $(\epsilon_1 + \epsilon_2, 0)$*-DP.*

Εντούτοις, η εγγύηση $\left(\sum_{i=1}^{k} \epsilon_i, \sum_{i=1}^{k} \delta_i\right)$-DP για k μηχανισμούς από τους Dwork and Lei (2009) είναι μάλλον περιοριστική, εφόσον οι αλγόριθμοι πρέπει οπωσδήποτε να είναι ανεξάρτητοι ο ένας από τον άλλο, δηλαδή δεν γίνεται ο ένας να λαμβάνει το αποτέλεσμα του άλλου και κάποια επιπρόσθετη πληροφορία από το (ιδιωτικό) σύνολο δεδομένων και να εξάγει ένα καινούριο αποτέλεσμα. Ευτυχώς, υπάρχουν ευκρινέστερα θεωρήματα σύνθεσης και συγκεκριμένα, το κάτωθι Προχωρημένο Θεώρημα Σύνθεσης το οποίο αφορά το ισχυρότερο περιβάλλον που αναζητούμε: αυτό της προσαρμοστικής σύνθεσης. Το όνομα "προσαρμοστική' προέρχεται από το γεγονός ότι κάθε ερώτημα (στην πραγματικότητα, αλγόριθμος) υποβάλλεται ένα-ένα και μπορεί να παραμετροποιείται / εξαρτάται από τα αποτελέσματα προηγούμενων ερωτημάτων / αλγορίθμων.

Συγκεκριμένα, το κάτωθι θεώρημα χαρακτηρίζει τις ιδιότητες ιδιωτικότητας μιας ακολουθίας αλγορίθμων $M_1(X), M_2(X), \ldots, M_N(X)$, όπου ο i-οστός αλγόριθμος ενδέχεται να εξαρτάται από τα αποτελέσματα των αλγορίθμων $M_1(X), M_2(X), \ldots, M_{i-1}(X)$, για κάθε $i \in [N]$.

Theorem 1.2.4 (Προχωρημένο Θεώρημα Σύνθεσης (Dwork and Roth, 2014))**.** *Εάν M είναι μια προσαρμοστική σύνθεση των διαφορικά ιδιωτικών αλγορίθμων $M_1, \ldots, M_N$, όπου ο αλγόριθμος M_i είναι (ϵ, δ_i)-DP για κάθε $i \in [N]$, τότε ισχύει ότι ο M είναι $(\epsilon N, \sum_{i=1}^{N} \delta_i)$-DP και επιπροσθέτως, για κάθε $\delta > 0$, ο M είναι $(\epsilon\sqrt{6N\log(1/\delta)}, \delta + \sum_{i=1}^{N} \delta_i)$-DP.*

Ένας δημοφιλής τρόπος να κατασκευάζουμε διαφορικά ιδιωτικούς αλγορίθμους είναι να προσθέτουμε κάποιου είδους θόρυβο σε ένα "τέλειο" αποτέλεσμα το οποίο θα περίμενε κανείς από τον αντίστοιχο μη-ιδιωτικό αλγόριθμο. Αυτή η γενική μέθοδος συμπεριλαμβάνει έναν ευρή αριθμό τεχνικών και θα εξετάσουμε εδώ τους πιο κλασσικούς τρόπους για αριθμητικές συναρτήσεις.

Προτού προχωρήσουμε στον πρώτο τέτοιο μηχανισμό (τον Laplacian μηχανισμό), θα εισαγάγουμε την έννοια της ℓ_1 ευαισθησίας μιας διανυσματικής συνάρτησης, η οποία διαισθητικά υποδηλώνει την μέγιστη αλλαγή (χειρότερης περίπτωσης) που μπορεί να προκαλέσουν τα δεδομένα ενός ατόμου στην συνάρτηση της οποίας την τιμή πάνω σε ένα σύνολο δεδομένων θέλουμε να εκτιμήσουμε. Επομένως, είναι λογικό ότι θα χρειαστεί να προσθέσουμε θόρυβο ανάλογο προς αυτή την ευαισθησία, για να αποκρύψουμε την συμμετοχή καθενός συγκεκριμένου ατόμου στο σύνολο δεδομένων.

Definition 1.2.2 (ℓ_1-ευαισθησία μιας συνάρτησης)**.** Η ℓ_1 ευαισθησία μιας συνάρτησης $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^d$ είναι η μέγιστη διαταραχή $\Delta_1 \boldsymbol{f}$ που μπορεί να προκαλέσει στην συνάρτηση μια αλλαγή συνόλου δεδομένων σε κάποιο γειτονικό (αλλαγή η οποία συμβολίζεται ως $\|x - y\| = 1$) στην ℓ_1-νόρμα:

$$\Delta_1 \boldsymbol{f} = \max_{\|x-y\|=1} \|\boldsymbol{f}(x) - \boldsymbol{f}(y)\|_1 \ .$$

Definition 1.2.3 (Κατανομή Laplace)**.** Λέμε ότι η τυχαία μεταβλητή Z είναι κατενεμημένη σύμφωνα με την κατανομή Laplace $\mathcal{L}(\mu, b)$ κεντραρισμένη στο μ με παράμετρο κλίμακας b εάν η συνάρτηση κατανομής πιθανότητάς της είναι:

$$g(z) = \frac{1}{2b} \exp\left(-\frac{|z-\mu|}{b}\right), \ \forall z \in \mathbb{R}\,.$$

Lemma 1.2.5 (Laplacian Μηχανισμός)**.** *Για μια συνάρτηση $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^d$ και για κάθε $\epsilon > 0$, η αναφορά της τιμής $\boldsymbol{f}(x) + \boldsymbol{\eta}$ είναι ένας $(\epsilon, 0)$-DP αλγόριθμος, όπου $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_d)$ και οι μεταβλητές $\{\eta_i\}_{1 \le i \le d}$ είναι ανεξάρτητες και ισόνομα κατανεμημένες τυχαίες μεταβλητές που ακολουθούν κατανομή Laplace ώστε $\eta_i \sim \mathcal{L}(0, \Delta_1 \boldsymbol{f}/\epsilon)$.*

Αποδεικνύεται επίσης (Dwork and Roth, 2014) ότι η ακρίβεια του Laplacian Μηχανισμού έχει ως εξής: εάν ονοματίσουμε το αποτέλεσμα του μηχανισμού ως $\boldsymbol{y}$ και υπενθυμίζοντας ότι το "τέλειο" αποτέλεσμα σε ένα σύνολο δεδομένων x είναι η τιμή $\boldsymbol{f}(x)$, τότε για κάθε $\gamma \in (0, 1)$, σύμφωνα με τις ουρές της κατανομής Laplace και ένα φράγμα ένωσης, προκύπτει ότι

$$\mathbf{Pr}\left[\|\boldsymbol{y} - \boldsymbol{f}(x)\|_\infty \le \ln\left(\frac{d}{\gamma}\right) \cdot \frac{\Delta_1 \boldsymbol{f}}{\epsilon}\right] \ge 1 - \gamma\,,$$

όπου $\|\boldsymbol{a}\|_\infty = \max_{i \in [d]} |a_i|$ είναι η max-νόρμα ενός διανύσματος $\boldsymbol{a} = (a_1, a_2, \ldots, a_d) \in \mathbb{R}^d$.

Στη συνέχεια, αναφέρουμε τον Gaussian μηχανισμό, ο οποίος αλλάζει αναλόγως με την ℓ_2 ευαισθησία μιας συνάρτησης, αντί για την ℓ_1 ευαισθησία, και ο οποίος είναι ένας εκ των σημαντικότερων μηχανισμών για την παροχή (ϵ, δ) Διαφορικής Ιδιωτικότητας με $\delta > 0$. Παρατηρείστε ότι είναι μια σημαίνουσα διαφορά του Laplacian με τον Gaussian μηχανισμό ότι ο δεύτερος δεν μπορεί να χρησιμοποιηθεί για την παροχή "γνήσιας" ($\delta = 0$) Διαφορικής Ιδιωτικότητας. Αυτό το ζήτημα είναι εγγενές στον Gaussian μηχανισμό και τον τρόπο με τον οποίο προσθέτει θόρυβο αναλογικά προς την ℓ_2 ευαισθησία μιας συνάρτησης (Dwork and Roth, 2014).

Definition 1.2.4 (ℓ_2-ευαισθησία μιας συνάρτησης)**.** Η ℓ_2 ευαισθησία μιας συνάρτησης $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^d$ είναι η μέγιστη διαταραχή $\Delta_2 \boldsymbol{f}$ που μπορεί να προκαλέσει στην συνάρτηση μια αλλαγή συνόλου δεδομένων σε κάποιο γειτονικό (αλλαγή η οποία συμβολίζεται ως $\|x - y\| = 1$) στην ℓ_2-νόρμα:

$$\Delta_2 \boldsymbol{f} = \max_{\|x-y\|=1} \|\boldsymbol{f}(x) - \boldsymbol{f}(y)\|_2 \ .$$

Lemma 1.2.6 (Gaussian Μηχανισμός)**.** *Για μια συνάρτηση $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^d$ και κάθε $\epsilon \in (0, 1), \delta > 0$, η αναφορά της τιμής $\boldsymbol{f}(x) + \boldsymbol{\eta}$ είναι ένας (ϵ, δ)-DP αλγόριθμος, όπου το τυχαίο διάνυσμα $\boldsymbol{\eta} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbb{I}_d\right)$ αποτελείται από d ανεξάρτητες κανονικές (Gaussian) τυχαίες μεταβλητές με τυπική απόκλιση $\sigma = \sqrt{2 \ln(1.25/\delta)} \Delta_2 \boldsymbol{f} / \epsilon$.*

1.3 Sub-gaussian τυχαίες μεταβλητές

Καθ' όλη τη διάρκεια του ερευνητικού μας έργου, χρησιμοποιούμε εκτενώς μία γενίκευση των Κανονικών (Gaussian) μεταβλητών, τις sub-gaussian τυχαίες μεταβλητές. Αυτή η κλάση κατανομών συμπεριλαμβάνει πολλές κρίσιμες κατανομές, όπως η Κανονική, η Bernoulli και *όλες* οι φραγμένες κατανομές. Η χρησιμότητά τους έγκειται κυρίως σε ένα αποτέλεσμα συγκέντρωσης που προσομοιάζει στο Hoeffding το οποίο εφαρμόζεται –σχεδόν εξ ορισμού– σε τυχαίες μεταβλητές που ακολουθούν κάποια sub-gaussian κατανομή.

Ξεκινούμε με τους ορισμούς των αντίστοιχων sub-gaussian κατανομών και της "sub-gaussian norm" για μοναδιάστατες και πολυδιάστατες τυχαίες μεταβλητές και συνεχίζουμε παραθέτοντας ορισμένες εκ των ιδιοτήτων και θεωρημάτων συγκέντρωσης που ισχύουν για αυτές, τις οποίες θα εκμεταλλευτούμε στη συνέχεια για τις δικές μας αποδείξεις.

Definition 1.3.1 (Sub-gaussian τυχαία μεταβλητή)**.** Μια τυχαία μεταβλητή X καλείται sub-gaussian εάν υπάρχει $K > 0$ τέτοιο ώστε, για κάθε $\lambda : |\lambda| \leq 1/K$,

$$\mathbb{E}\left[\exp\left(\lambda^2 X^2\right)\right] \leq \exp(\lambda^2 K^2) \, .$$

Το μικρότερο K για το οποίο ισχύει η ανωτέρω ιδιότητα ονομάζεται η *sub-gaussian norm* του X και συμβολίζεται με $\|X\|_{\psi_2}$.

Definition 1.3.2 (Sub-gaussian τυχαίο διάνυσμα)**.** Ένα τυχαίο διάνυσμα $\boldsymbol{X} \in \mathbb{R}^d$ ονομάζεται sub-gaussian αν για κάθε διάνυσμα $\boldsymbol{u} \in \mathbb{R}^d$, το εσωτερικό γινόμενο $\langle \boldsymbol{X}, \boldsymbol{u} \rangle$ είναι μια sub-gaussian τυχαία μεταβλητή. Η "sub-gaussian norm" ενός sub-gaussian τυχαίου διανύσματος ορίζεται ως εξής:

$$\|\boldsymbol{X}\|_{\psi_2} = \sup_{\boldsymbol{u} \in S^{d-1}} \|\langle \boldsymbol{X}, \boldsymbol{u} \rangle\|_{\psi_2} \ ,$$

όπου $S^{d-1} = \{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_2 = 1\}$ είναι η επιφάνεια της d-διάστατης μοναδιαίας σφαίρας.

Παραδείγματα sub-gaussian κατανομών.

- Κάθε Κανονική τυχαία μεταβλητή $X \sim \mathcal{N}(0, \sigma^2)$ είναι sub-gaussian με sub-gaussian norm $\|X\|_{\psi_2} \leq C\sigma$, για κάποια καθολική σταθερά $C > 0$.
- Όλες οι φραγμένες τυχαίες μεταβλητές X τέτοιες ώστε $|X| \leq \eta$ είναι sub-gaussian με sub-gaussian norm $\|X\|_{\psi_2} \leq \frac{1}{\sqrt{\ln 2}} \cdot \eta$.
- Κάθε Rademacher μεταβλητή (συμμετρική Bernoulli μεταβλητή που παίρνει τις τιμές $\{\pm 1\}$ με πιθανότητα $1/2$ στην τύχη) είναι sub-gaussian με sub-gaussian norm $\|X\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}$.

Lemma 1.3.1 (Ιδιότητες της sub-gaussian norm)**.** *Έστω $\boldsymbol{X}$ να είναι ένα sub-gaussian τυχαίο διάνυσμα. Τότε, ισχύουν τα ακόλουθα:*

- *Για κάθε σταθερά $c > 0$, το τυχαίο διάνυσμα $c\boldsymbol{X}$ είναι sub-gaussian, με $\|c\boldsymbol{X}\|_{\psi_2} = c\,\|\boldsymbol{X}\|_{\psi_2}$.*
- *Εάν $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu_X}$, τότε το τυχαίο διάνυσμα $\boldsymbol{X} - \boldsymbol{\mu_X}$ είναι και αυτό sub-gaussian, με*

$$\|\boldsymbol{X} - \boldsymbol{\mu_X}\|_{\psi_2} \leq C\,\|\boldsymbol{X}\|_{\psi_2}\ ,$$

για κάποια καθολική σταθερά $C > 0$.

Theorem 1.3.2 (Γενική ανισότητα του Hoeffding (Vershynin, 2018))**.** *Έστω ότι $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες, sub-gaussian τυχαίες μεταβλητές με μέση τιμή μηδέν. Τότε, υπάρχει μια σταθερά $c > 0$, ώστε για κάθε $t > 0$ να ισχύει ότι*

$$\mathbf{Pr}\left[\left|\sum_{i=1}^{n} X_i\right| > t\right] \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^{n}\|X_i\|_{\psi_2}^2}\right)\ .$$

Για περαιτέρω ενδιαφέροντα αποτελέσματα σχετικά με τις sub-gaussian τυχαίες μεταβλητές, παραπέμπουμε τον αναγνώστη στους Vershynin (2018); Rivasplata (2012); Hsu et al. (2012).

1.4 Περιβάλλοντα παλινδρόμησης

Σε αυτή την ενότητα, ορίζουμε αυστηρά τα περιβάλλοντα που μας ενδιαφέρουν, δηλαδή αυτά της Προσαρμογής Ελαχίστων Τετραγώνων, Γραμμικής Παλινδρόμησης και Δυαδικής Παλινδρόμησης, μαζί με τις όποιες τεχνικές υποθέσεις κάνουμε.

Στο απλό μοντέλο Γραμμικής Παλινδρόμησης, παρατηρούμε επισημειωμένα παραδείγματα $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, όπου οι επιγραφές y_i υποθέτουμε ότι δημιουργούνται από ένα υποκρυπτόμενο γραμμικό μοντέλο $\boldsymbol{\beta}^T\boldsymbol{X}_i$ με ένα σφάλμα που ακολουθεί την κανονική κατανομή, για κάποιους συντελεστές της παλινδρόμησης $\boldsymbol{\beta} \in \mathbb{R}^d$. Ο στόχος μας είναι να εκτιμήσουμε αυτό το "πραγματικό" υποκρυπτόμενο $\boldsymbol{\beta}$ με έναν αφορικά ιδιωτικό τρόπο. Συγκεκριμένα, έχουμε το ακόλουθο:

Assumption 1.4.1. *Τα επισημειωμένα παραδείγματα $(\boldsymbol{X}_i, y_i)$, $i = 1, \ldots, n$, είναι ανεξάρτητες και ισόνομα κατανεμημένες τυχαίες μεταβλητές. Επίσης, τα διανύσματα $\boldsymbol{X}_i \in \mathbb{R}^d$ λαμβάνονται από μια Κανονική κατανομή $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ που ικανοποιεί την σχέση $\mathbb{I}_d \preceq \Sigma \preceq \kappa\mathbb{I}_d$ για κάποια καθολική παράμετρο $\kappa > 0$. Το μοντέλο γέννησης των επιγραφών y_i έχει ως εξής: υπάρχει ένα διάνυσμα $\boldsymbol{\beta} \in \mathbb{R}^d$ τέτοιο ώστε, δοθέντος ενός διανύσματος $\boldsymbol{X}_i \in \mathbb{R}^d$,*

$$y_i = \boldsymbol{\beta}^T\boldsymbol{X}_i + \epsilon_i, \quad \text{για κάθε } i = 1, \ldots, n\,,$$

όπου οι τυχαίες μεταβλητές ϵ_i είναι ανεξάρτητες και ισόνομα κατανεμημένες, δειγματοληπτημένες από μια κανονική κατανομή μηδενικής μέσης τιμής $\mathcal{N}(0, \sigma_\epsilon^2)$.

Στο πρόβλημα Προσαρμογής Ελαχίστων Τετραγώνων, παρατηρούμε επισημειωμένα παραδείγματα $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ και επιθυμούμε να παραγάγουμε μια (ϵ, δ)-διαφορικά ιδιωτική έκδοση του Εκτιμητή Ελαχίστων Τετραγώνων (LSE):

$$\boldsymbol{\beta}^\star = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}^T\boldsymbol{X}_i\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\boldsymbol{X}_i\right) = \left(\frac{1}{n}X^TX\right)^{-1}\frac{1}{n}X^T\boldsymbol{y},$$

όπου $X = [\boldsymbol{X}_i]_{i=1}^n \in \mathbb{R}^{n\times d}$ είναι ο πίνακας των συμμεταβλητών και $\boldsymbol{y} = [y_i]_{i=1}^n \in \mathbb{R}^d$ είναι το διάνυσμα των επιγραφών αντίστοιχα. Σημειώνουμε ότι, αντίθετα με το παρακάτω πρόβλημα της Δυαδικής Παλινδρόμησης, σε αυτό το πρόβλημα δεν κάνουμε καμία πρότερη υπόθεση στον τρόπο σύνδεσης των φραγμένων επιγραφών y_i με τα διανύσματα των συμμεταβλητών $\boldsymbol{X}_i$.

Συγκεκριμένα, για το πρόβλημα της Προσαρμογής Ελαχίστων Τετραγώνων, κάνουμε τις ακόλουθες τεχνικές υποθέσεις:

Assumption 1.4.2. *Τα επισημειωμένα παραδείγματα* $(\boldsymbol{X}_i, y_i)$, $i = 1, \ldots, n$ *είναι ανεξάρτητα και ισόνομα κατανεμημένα. Επίσης, τα τυχαία διανύσματα* $\boldsymbol{X}_i \in \mathbb{R}^d$ *προέρχονται από μια κανονική κατανομή* $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ *που ικανοποιεί τις παρακάτω υποθέσεις:*

$$\|\boldsymbol{\mu}\|_2 \leq R \quad \textit{και} \quad \mathbb{I}_d \preceq \Sigma \preceq \kappa \mathbb{I}_d \,,$$

ενώ οι επιγραφές ικανοποιούν $\frac{1}{\rho} \leq |y_i| \leq c$ *για κάποιες καθολικές παραμέτρους του προβλήματος* $\rho, c, \kappa, R > 0$.

Τέλος, στο περιβάλλον της Δυαδικής Παλινδρόμησης, επιπροσθέτως των ανωτέρω υποθέσεων, υποθέτουμε επίσης ότι οι επιγραφές y_i είναι δυαδικές (δηλαδή $y_i \in \{-1, +1\}$) και παράγονται από ένα Γενικευμένο Γραμμικό Μοντέλο που συνδέει τις επιγραφές με τις συμμεταβλητές. Σε αντίθεση με το περιβάλλον της Προσαρμογής Ελαχίστων Τετραγώνων, αυτό το Γενικευμένο Γραμμικό Μοντέλο παραμετροποιείται από ένα "αληθινό" $\boldsymbol{\beta} \in \mathbb{R}^d$. Ο στόχος μας εδώ είναι να δώσουμε έναν εκτιμητή αυτού του $\boldsymbol{\beta}$ ξανά μέσω *της ίδιας* (ϵ, δ)-διαφορικά ιδιωτικής έκδοσης του LSE που χρησιμοποιήσαμε ανωτέρω. Πιο συγκεκριμένα, κάνουμε την ακόλουθη επιπρόσθετη υπόθεση:

Assumption 1.4.3. *Υπάρχει ένα διάνυσμα* $\boldsymbol{\beta} \in \mathbb{R}^d$ *τέτοιο ώστε, δοθέντος ενός διανύσματος* $\boldsymbol{X}_i \in \mathbb{R}^d$,

$$\mathbf{Pr}[y_i = +1 | \boldsymbol{X}_i] = f(\boldsymbol{\beta}^T \boldsymbol{X}_i), \quad \textit{για κάθε } i = 1, \ldots, n,$$

όπου $f : \mathbb{R} \to [0, 1]$ *είναι μια μη-φθίνουσα, συνεχώς διαφορίσιμη συνάρτηση που ικανοποιεί τις εξής σχέσεις:*

$$\lim_{x \to -\infty} f(x) = 0 \quad \textit{και} \quad \lim_{x \to \infty} f(x) = 1 \,.$$

Επιπροσθέτως, οι συμμεταβλητές $\boldsymbol{X}_i$ *είναι μηδενικής μέσης τιμής, δηλαδή* $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_i] = \mathbf{0}$.

Το ανωτέρω ορισμένο πιθανοτικό μοντέλο ισχύει σε πολλά σημαίνοντα πρακτικά περιβάλλοντα. Για παράδειγμα, ισχύει για την Λογιστική Παλινδρόμηση, όπου η συνάρτηση του μοντέλου είναι $f(x) = 1/\left(1 + e^{-x}\right)$. Επιπρσθέτως, ισχύει και για γραμμικά διαχωρίσιμες Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ).

Τέλος, η υπόθεσή μας ότι $\boldsymbol{\mu} = \mathbf{0}$ είναι συνήθης (δείτε, για παράδειγμα, τους Kulkarni et al. (2021); Cai et al. (2020); Daskalakis et al. (2020); Bernstein and Sheldon (2019); Sheffet (2017); Erdogdu (2016)) και λογική στο περιβάλλον της Δυαδικής Παλινδρόμησης: ακόμη και αν αγνοήσουμε τους περιορισμούς ιδιωτικότητας, οι εγγυήσεις δειγματικής πολυπλοκότητας για οποιονδήποτε πιθανό εκτιμητή θα επιδεινώνονται ταχέως καθώς το $\boldsymbol{\mu}$ κινείται μακριά από την αρχή των αξόνων. Αυτό θα συμβαίνει ακριβώς διότι, υπό Κανονικές μεταβλητές, το ποσοστό των δειγμάτων της μίας κλάσης θα μειώνεται εκθετικά όσο η απόσταση του $\boldsymbol{\mu}$ από το διαχωριστικό υπερεπίπεδο (το οποίο περνάει από την αρχή των αξόνων) αυξάνεται.

1.4.1 Το λήμμα του Stein και τα Γενικευμένα Γραμμικά Μοντέλα

Διατυπώνουμε εδώ μια πολυδιάστατη εκδοχή του Λήμματος του Stein (Stein, 1981) όπως αναφέρεται από τον Liu (1994).

Lemma 1.4.1 (Stein's Lemma (Liu, 1994))**.** *Έστω* $\boldsymbol{Z} \in \mathbb{R}^p, \boldsymbol{W} \in \mathbb{R}^q$ *να είναι από κοινού κανονικά (jointly Gaussian) τυχαία διανύσματα και μια συνάρτηση* $f : \mathbb{R}^q \to \mathbb{R}$ *που είναι διαφορίσιμη σχεδόν παντού, με* $\mathbb{E}_{\boldsymbol{W}}\left[|\partial f(\boldsymbol{W})/\partial W_i|\right] < \infty$ *για κάθε* $i \in [q]$. *Τότε, ισχύει ότι* $\mathbf{Cov}[\boldsymbol{Z}, f(\boldsymbol{W})] = \mathbf{Cov}[\boldsymbol{Z}, \boldsymbol{W}]\, \mathbb{E}[\nabla f(\boldsymbol{W})]$.

Το λήμμα αυτό βρίσκει απευθείας εφαρμογή στα Γενικευμένα Γραμμικά Μοντέλα με Κανονικές (Gaussian) συμμεταβλητές (Erdogdu, 2016; Brillinger, 2012a,b): για το Γενικευμένο Γραμμικό Μοντέλο του Brillinger (2012a) με Κανονικές (Gaussian) συμμεταβλητές, του οποίου η συνάρτηση του μοντέλου ικανοποιεί το ανωτέρω λήμμα, μπορεί να δειχθεί (Brillinger, 2012a,b) ότι ο Εκτιμητής Ελαχίστων Τετραγώνων συγκλίνει ασυμπτωτικά στο διάνυσμα των αληθινών παραμέτρων $\boldsymbol{\beta}$ του Γενικευμένου Γραμμικού Μοντέλου με πιθανότητα 1, ως προς έναν πολλαπλασιαστικό παράγοντα k. Αυτός ο παράγοντας είναι ανάλογος του παράγοντα k που εμφανίζεται στη συνέχεια της ανάλυσής μας.

1.5 Επέκταση εκτίμησης αναμενόμενης τιμής και συνδιασποράς για sub-gaussian τυχαίες μεταβλητές

Σε τεχνικό επίπεδο, η εργασία μας εξαρτάται από (και επεκτείνει) τα εργαλεία που αναπτύχθηκαν από τους Kamath et al. (2019) για την ιδιωτική εκτίμηση της μέσης τιμής $\boldsymbol{\mu}$ και της συνδιασποράς Σ μιας d-διάστατης Κανονικής κατανομής.

Ο αλγόριθμος που προτείνουν, τον οποίο ονομάζουμε LearnGaussian-hd, ικανοποιεί την ακόλουθη εγγύηση:

Theorem 1.5.1 (Πολυδιάστατη εκτίμηση κανονικής κατανομής (Kamath et al., 2019))**.** *Υπάρχει ένας $(\epsilon^2/2 + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP αλγόριθμος πολυωνυμικού χρόνου* LearnGaussian-hd *ο οποίος λαμβάνει τουλάχιστον*

$$n = \widetilde{O}\left(\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\epsilon} + \frac{d^{3/2}\log^{1/2}(\kappa) + d^{1/2}\log^{1/2}(R)}{\epsilon}\right)$$

ανεξάρτητα και ισόνομα κατανεμημένα δείγματα $\boldsymbol{X}_i$, $i \in [n]$ μιας d-διάστατης κανονικής κατανομής $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ με άγνωστη μέση τιμή $\boldsymbol{\mu} \in \mathbb{R}^d$ και άγνωστο πίνακα συνδιασποράς $\Sigma \in \mathbb{R}^{d\times d}$ που ικανοποιούν τις σχέσεις $\|\boldsymbol{\mu}\|_2 \leq R$ και $\mathbb{I}_d \preceq \Sigma \preceq \kappa\mathbb{I}_d$, και εξάγει εκτιμητές $\widehat{\boldsymbol{\mu}}, \widehat{\Sigma}$ τέτοιους ώστε με μεγάλη πιθανότητα $\mathrm{TV}(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}(\widehat{\boldsymbol{\mu}}, \widehat{\Sigma})) \leq \alpha$.

1.5.1 Συνοπτική ανασκόπηση του αλγορίθμου

Εδώ δίνουμε μια συνοπτική ανασκόπηση του αλγορίθμου LearnSubGaussian-hd, δηλαδή της επέκτασής μας για τον αλγόριθμο των Kamath et al. (2019), την οποία επέκταση χρησιμοποιούμε στον Αλγόριθμο 1.

Το κύριο συστατικό στοιχείο του αλγορίθμου εκτίμησης του πίνακα συνδιασποράς είναι ο αλγόριθμος NaivePCE (Αλγόριθμος 1 των Kamath et al. (2019)), ο οποίος διαισθητικά θα ήταν η πρώτη προσπάθειά μας να προσδώσουμε ιδιωτικότητα στην διαδικασία εκτίμησης συνδιασποράς. Πιο συγκεκριμένα, αρχικά περιορίζουμε τα δείγματα εισόδου, έπειτα προσθέτουμε έναν τυχαίο Gaussian πίνακα στην εμπειρική συνδιασπορά που προκύπτει από αυτά τα δείγματα και τέλος προβάλλουμε στον κώνο των θετικά ημι-ορισμένων πινάκων. Ωστόσο, αυτός ο πρώτος αλγόριθμος έχει γραμμική εξάρτηση της ακρίβειάς του στην μέγιστη ιδιοτιμή κ του πίνακα συνδιασποράς Σ, ενώ εμείς στοχεύουμε σε λογαριθμική εξάρτηση. Επομένως, σημειώνοντας ότι η ανωτέρω εξάρτηση είναι βέλτιστη όταν η μέγιστη ιδιοτιμή είναι σταθερής τάξης, αναζητούμε μήπως θα μπορούσαμε να μετασχηματίσουμε τα δείγματα εισόδου $\boldsymbol{X}_i$ σε $A\boldsymbol{X}_i$ τέτοια ώστε η μέγιστη ιδιοτιμή του πίνακα συνδιασποράς των $A\boldsymbol{X}_i$ (ο οποίος είναι $A\Sigma A$ για συμμετρικούς πίνακες A) να ικανοποιεί την ζητούμενη συνθήκη.

Ο αλγόριθμος εκτίμησης συνδιασποράς, επομένως, ξεκινάει με την αποδοτική εύρεση ενός τέτοιου πίνακα A (ο οποίος καλείται και "preconditioner") σύμφωνα με τον αλγόριθμο PPC (Αλγόριθμος 3 των Kamath et al. (2019)) ο οποίος, εν συντομία, κάνει τα ακόλουθα: χρησιμοποιεί $O(\log\kappa)$ διαδοχικούς γύρους του NaivePCE αλγορίθμου έτσι ώστε κάθε γύρος να "απαλείφει" τις ιδιοδιευθύνσεις με μεγάλη διασπορά, μετασχηματίζοντας κάθε διαδοχικό κ_j (για

$1 \le j \le O(\log \kappa)$ όπου j είναι ο αριθμός του τρέχοντος γύρου) σε $\kappa_{j+1} = 0.7\kappa_j$. Έπειτα από $O(\log \kappa)$ γύρους, η τελική μέγιστη ιδιοτιμή του $A\Sigma A$ θα είναι σταθερής τάξης, όπως ήταν επιθυμητό. Έπειτα από αυτή τη διαδικασία που προσδιορίζει αυτό τον A, εκτελούμε τον NaivePCE στα δείγματα $A\boldsymbol{X}_i$ με ένα αποτέλεσμα $\widetilde{\Sigma}$ και ο συνολικός αλγόριθμος εκτίμησης συνδιασποράς επιστρέφει το $\widehat{\Sigma} = A^{-1}\widetilde{\Sigma}A^{-1}$. Για λεπτομέρειες σε αυτές τις διαδικασίες, παραπέμπουμε τον ενδιαφερόμενο αναγνώστη στους Kamath et al. (2019).

Εάν κάποιος ήθελε επιπροσθέτως να εκτιμήσει και την μέση τιμή, θα έπρεπε πρώτα να βρει έναν πίνακα A όπως παραπάνω μέσω $2n$ δειγμάτων $\frac{1}{\sqrt{2}}\left(\boldsymbol{X}_{2i} - \boldsymbol{X}_{2i-1}\right), 1 \le i \le n$ τα οποία είναι ανεξάρτητα και ισόνομα κατανεμημένα με ίδιο πίνακα συνδιασποράς Σ και έπειτα να λάβει επιπρόσθετα n δείγματα $\boldsymbol{X}_i$ και να εφαρμόσει τον αλγόριθμο εκτίμησης μέσης τιμής των Karwa and Vadhan (2018) σε κάθε συντεταγμένη των $A\boldsymbol{X}_i$ ξεχωριστά. Η διαφορά του αλγορίθμου μας με τους Kamath et al. (2019) είναι ότι καλούμε τον αλγόριθμο των Karwa and Vadhan (2018) με $R = \infty$, το οποίο είναι επιτρεπτό και αποδοτικό στο περιβάλλον μας διότι έχουμε $\delta > 0$. Για τις πλήρεις περιγραφές παραπέμπουμε στους Αλγορίθμους 1 και 4 των Karwa and Vadhan (2018).

1.5.2 Επέκταση των εγγυήσεων των Kamath et al. (2019) και Karwa and Vadhan (2018) σε περιβάλλον sub-gaussian τυχαίων μεταβλητών

Οι αλγόριθμοι ελαφρώς τροποποιημένοι όπως δείξαμε στην προηγούμενη ενότητα έχουν τις ακόλουθες εγγυήσεις.

Lemma 1.5.2 (Ιδιωτική Εκτίμηση Συνδιασποράς, παραλλαγή των Kamath et al. (2019))**.** *Για κάθε $\epsilon, \delta, \gamma, \kappa, \alpha > 0$, υπάρχει ένας $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP αλγόριθμος, ο οποίος δοθέντων n ανεξάρτητων και ισόνομα κατανεμημένων δειγμάτων $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ από μια sub-gaussian πολυδιάστατη κατανομή με μέση τιμή $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{0}$ και πίνακα συνδιασποράς $\mathbb{E}[\boldsymbol{X}_i\boldsymbol{X}_i^T] = \Sigma$ με $\mathbb{I}_d \preceq \Sigma \preceq \kappa\mathbb{I}_d$ και*

$$n = O\left(\frac{d + \log(1/\gamma)}{\alpha^2} + \frac{d^{3/2}\mathrm{polylog}\left(\frac{d}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\mathrm{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right),$$

εξάγει $\widehat{\Sigma}$ ώστε $\left\|\Sigma^{-1/2}\left(\widehat{\Sigma} - \Sigma\right)\Sigma^{-1/2}\right\|_2 \le O(\alpha)$ με πιθανότητα $1 - O(\gamma)$.

Lemma 1.5.3 (Ιδιωτική Εκτίμηση Μέσης Τιμής, παραλλαγή των Karwa and Vadhan (2018); Kamath et al. (2019))**.** *Για κάθε $\epsilon, \delta, \gamma, \kappa, \alpha > 0$, υπάρχει ένας $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP αλγόριθμος, ο οποίος δοθέντων n ανεξάρτητων και ισόνομα κατανεμημένων δειγμάτων $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ από μια sub-gaussian πολυδιάστατη κατανομή με μέση τιμή $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{\mu}$ και πίνακα συνδιασποράς $\mathbb{E}[\boldsymbol{X}_i\boldsymbol{X}_i^T] = \Sigma$ με $\mathbb{I}_d \preceq \Sigma \preceq \kappa\mathbb{I}_d$ και*

$$n = O\left(\frac{d\log(\frac{d}{\gamma})}{\alpha^2} + \frac{d\mathrm{polylog}(\frac{d\log(1/\delta)}{\alpha\gamma\epsilon})}{\alpha\epsilon} + \frac{\sqrt{d}\log(\frac{d}{\gamma\delta})}{\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\mathrm{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right),$$

εξάγει έναν (συμμετρικό) πίνακα A και ένα διάνυσμα $\widehat{\boldsymbol{\mu}}$ τέτοια ώστε $\mathbb{I}_d \preceq A\Sigma A \preceq 1000\mathbb{I}_d$ και $\|A(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|_2 \le \alpha$ με πιθανότητα $1 - O(\gamma)$.

1.6 Ιδιωτική εκτίμηση σε απλά γραμμικά μοντέλα

Σε αυτή την ενότητα, σκοπεύουμε να δείξουμε ότι η εκτίμηση των συντελεστών παλινδρόμησης $\boldsymbol{\beta}$ υπό Gaussian συμμεταβλητές μπορεί, στην περίπτωση της Γραμμικής Παλινδρόμησης,

να γίνει αρκετά ευκολότερα από την κύρια μέθοδο που μελετούμε στην επόμενη ενότητα. Η τεχνική εκτίμησης συνίσταται στην χρήση των κλασσικών αλγορίθμων εκτίμησης συνδιασποράς μέσω μιας αναγωγής της εκτίμησης των συντελεστών συνδιασποράς στην εκτίμηση πινάκων συνδιασποράς για κανονικά κατανεμημένα (Gaussian) τυχαία διανύσματα.

Από τις υποθέσεις του μοντέλου της Γραμμικής Παλινδρόμησης που δείξαμε σε προηγούμενη ενότητα, μπορούμε να συμπεράνουμε ότι η κατανομή των επιγραφών y_i είναι η κανονική (Gaussian) κατανομή $\mathcal{N}(\boldsymbol{\beta}^T\boldsymbol{\mu}, \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \sigma_\epsilon^2)$, επομένως ορίζοντας τα διανύσματα $\boldsymbol{Z}_i \in \mathbb{R}^{d+1}$ ως

$$\boldsymbol{Z}_i = \begin{bmatrix} \boldsymbol{X}_i \\ y_i \end{bmatrix},$$

μπορούμε να δούμε ότι ακολουθούν και αυτά την κανονική κατανομή, με πίνακα συνδιασποράς Σ' ο οποίος μπορεί να γραφεί σε μορφή υποπινάκων ως εξής:

$$\Sigma' = \mathbb{E}\left[\boldsymbol{Z}_i\boldsymbol{Z}_i^T\right] = \begin{bmatrix} \Sigma & \Sigma\boldsymbol{\beta} \\ \boldsymbol{\beta}^T\Sigma & \sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} \end{bmatrix}.$$

Η ανωτέρω εξίσωση δείχνει ότι ένας φυσιολογικός τρόπος να εκτιμήσουμε το $\boldsymbol{\beta}$ θα ήταν να εκτιμήσουμε τους πίνακες Σ, Σ' και έπειτα, εξάγοντας την τελευταία στήλη του πίνακα Σ', η οποία (χωρίς το τελευταίο στοιχείο) είναι $\Sigma\boldsymbol{\beta}$, να πολλαπλασιάσουμε από αριστερά με το αντίστροφο της εκτίμησης του Σ που έχουμε. Πράγματι, αποδεικνύουμε ότι αυτή η προσέγγιση δουλεύει και, όταν η εκτίμηση των παραπάνω πινάκων συνδιασποράς γίνει με τους διαφορικά ιδιωτικούς αλγορίθμους των Kamath et al. (2019), το τελικό αποτέλεσμα είναι ένα διαφορικά ιδιωτικός εκτιμητής του $\boldsymbol{\beta}$ για το μοντέλο Γραμμικής Παλινδρόμησης με τις υποθέσεις που αναφέραμε σε ανωτέρω ενότητα.

Αλγόριθμος. Έχοντας πρόσβαση σε n ανεξάρτητα και ισόνομα κατανεμημένα δείγματα $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, όπου $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma), i \in [n]$, ο αλγορίθμος αρχικά υποογίζει μία διαφορικά ιδιωτική εκτίμηση $\widehat{\Sigma}$ του πίνακα συνδιασποράς της d-διάστατης κανονικής κατανομής $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, χρησιμοποιώντας τον αλγόριθμο LearnGaussian-hd. Έπειτα, ο αλγόριθμος σχηματίζει τα διανύσματα $\boldsymbol{Z}_i \in \mathbb{R}^{d+1}$ όπως παραπάνω και υπολογίζει μια διαφορικά ιδιωτική εκτίμηση $\widehat{\Sigma}'$ του πίνακα συνδιασποράς της $(d+1)$-διάστατης κανονικής κατανομής με πίνακα συνδιασποράς της παραπάνω μορφής υποπινάκων, ξανά χρησιμοποιώντας τον αλγόριθμο LearnGaussian-hd. Από τον πίνακα $\widehat{\Sigma}'$, ο αλγόριθμος λαμβάνει μόνο τα πρώτα d στοιχεία της τελευταίας στήλης αυτού του πίνακα, τα οποία ονομάζουμε ως $\widehat{\Sigma\boldsymbol{\beta}} \in \mathbb{R}^d$. Μέσω αυτών των εκτιμήσεων, η διαφορικά ιδιωτική εκτίμηση του $\boldsymbol{\beta}$ δίδεται τελικά από την σχέση:

$$\widehat{\boldsymbol{\beta}} = \widehat{\Sigma}^{-1}\widehat{\Sigma\boldsymbol{\beta}},$$

της οποίας η ιδιωτικότητα προκύπτει από τα κατάλληλα θεωρήματα σύνθεσης διαφορικά ιδιωτικών μηχανισμών. Προχωρούμε τώρα στην διατύπωση της εγγύησης ακρίβειας αυτού του εκτιμητή.

Theorem 1.6.1 (Ακρίβεια του $\widehat{\boldsymbol{\beta}}$ στην Ιδιωτική Εκτίμηση Γραμμικής Παλινδρόμησης)**.** *Κάτω από την υπόθεση 7.0.1 με παραμέτρους (κ, Σ'), για κάθε παραμέτρους ιδιωτικότητας $\epsilon, \delta > 0$, ακρίβειας $\alpha, \eta > 0$ και εμπιστοσύνης $\gamma \in (0,1)$, ο ιδιωτικός αλγόριθμος που αντιστοιχεί στην παραπάνω σχέση είναι $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-διαφορικά ιδιωτικός και εάν ο αριθμός των επισημειωμένων παραδειγμάτων είναι τουλάχιστον:*

$$n = O\left(\frac{d + \log(1/\gamma)}{\eta^2} + \frac{d^{3/2}\mathrm{polylog}\left(\frac{d}{\eta\gamma\epsilon}\right)}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log\left(\kappa(\Sigma')\right)}\mathrm{polylog}\left(\frac{d\log(\kappa(\Sigma'))}{\gamma\epsilon}\right)}{\epsilon}\right)$$
$$+ O\left(\frac{d + \log(1/\gamma)}{\alpha^2} + \frac{d^{3/2}\mathrm{polylog}\left(\frac{d}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon}\right),$$

τότε με πιθανότητα τουλάχιστον $1 - O(\gamma)$ *η έξοδος* $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ *και οι "αληθινοί" συντελεστές παλινδρόμησης* $\boldsymbol{\beta}$ *ικανοποιούν τη σχέση:*

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2 \leq \|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_2^2 \leq O\left(\alpha^2\right) \cdot \|\boldsymbol{w}\|_2^2 + O\left(\eta^2\right) \cdot \lambda_{max}^2(\Sigma'),$$

όπου $\kappa(\Sigma') = \frac{\lambda_{max}(\Sigma')}{\lambda_{min}(\Sigma')}$ *είναι ο δείκτης κατάστασης (condition number) του πίνακα* Σ', $\widehat{\boldsymbol{w}} = \Sigma^{1/2}\widehat{\boldsymbol{\beta}}$ *και* $\boldsymbol{w} = \Sigma^{1/2}\boldsymbol{\beta}$.

1.7 Ιδιωτική εκτίμηση για την Προσαρμογή Ελαχίστων Τετραγώνων και την Δυαδική Παλινδρόμηση

1.7.1 Προσαρμογή Ελαχίστων Τετραγώνων

Ο διαφορικά ιδιωτικός LSE για το περιβάλλον Προσαρμογής Ελαχίστων Τετραγώνων συνοψίζεται στον Αλγόριθμο 1. Εν ολίγοις, ο αλγόριθμος υπολογίζει διαφορικά ιδιωτικούς εκτιμητές των ποσοτήτων

$$(X^T X/n)^{-1} \quad \text{και} \quad X^T \boldsymbol{y}/n,$$

των οποίων το γινόμενο δίδει τον εκτιμητή $\boldsymbol{\beta}^\star$.

Η εκτίμηση της πρώτης ποσότητας συμβαίνει ως ακολούθως. Έχοντας πρόσβαση σε n δείγματα $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, όπου $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma), i \in [n]$, ο αλγόριθμος στην αρχή υπολογίζει διαφορικά ιδιωτικές εκτιμήσεις $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}, \widehat{\Sigma}_{\boldsymbol{X}})$ της αναμενόμενης τιμής και του πίνακα συνδιασποράς της d-διάστατης κανονικής κατανομής $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, χρησιμοποιώντας τον αλγόριθμο LearnGaussian-hd. Αυτοί οι εκτιμητές μπορούν να χρησιμοποιηθούν για την εκτίμηση της ποσότητας $(X^T X/n)^{-1}$ μέσω της σχέσης:

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T \approx \mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T] \approx \widehat{\Sigma}_{\boldsymbol{X}} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T.$$

Η δεύτερη ποσότητα, δηλαδή ο όρος $X^T \boldsymbol{y}/n$, είναι κάπως δυσκολότερο να εκτιμηθεί με διαφορικά ιδιωτικό τρόπο, επειδή οι όροι $y_i \boldsymbol{X}_i$ που τον απαρτίζουν *δεν είναι* κανονικοί (Gaussian). Ωστόσο, λόγω των φραγμένων τυχαίων μεταβλητών y_i, μπορούμε να διασφαλίσουμε ότι οι εν λόγω όροι είναι sub-gaussian. Ως μια σημαντική τεχνική συνεισφορά, επεκτείνουμε την ανάλυση των Kamath et al. (2019) και Karwa and Vadhan (2018) σε sub-gaussian τυχαίες μεταβλητές (δείτε τον αλγόριθμο LearnSubGaussian-hd) και λαμβάνουμε μια διαφορικά ιδιωτική εκτίμηση της μέσης τιμής $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y}$ για τα sub-gaussian τυχαία διανύσματα $y_i \boldsymbol{X}_i$.

Έχοντας αυτούς τους εκτιμητές, ο διαφορικά ιδιωτικός LSE δίνεται τελικά από τη σχέση:

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\Sigma}_{\boldsymbol{X}} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T\right)^{-1} \widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y},$$

του οποίου η ιδιωτικότητα προκύπτει από κατάλληλα θεωρήματα σύνθεσης διαφορικά ιδιωτικών αλγορίθμων. Το κύριο αποτέλεσμά μας όσον αφορά την ιδιωτικότητα και την ακρίβεια του εν λόγω εκτιμητή έχει ως εξής:

Theorem 1.7.1 (Ιδιωτικότητα και Ακρίβεια του $\widehat{\boldsymbol{\beta}}$ στην Ιδιωτική Προσαρμογή Ελαχίστων Τετραγώνων)**.** *Κάτω από την υπόθεση 5.2.1 με παραμέτρους* (κ, c, ρ, R)*, για όλες τις παραμέτρους ιδιωτικότητας* $\epsilon, \delta > 0$*, ακρίβειας* $\alpha, \eta > 0$ *και εμπιστοσύνης* $\gamma \in (0,1)$*, ο αλγόριθμος* PrivateLearnLSE *(που ορίζεται στον Αλγόριθμο 1) είναι* $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$*-διαφορικά*

Algorithm 1 Ιδιωτική Εκτίμηση του Εκτιμητή Ελαχίστων Τετραγώνων.

Είσοδος: $(X, \boldsymbol{y}) = (\boldsymbol{X}_i, y_i)_{i\in[n]}$ with $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}, \Sigma$ are unknown.
Παράμετροι: Privacy $\epsilon, \delta > 0$, accuracy $\alpha, \eta > 0$, confidence $\gamma \in (0,1)$, covariance spectral norm bound κ, upper bound of labels c.
Έξοδος: Estimate $\widehat{\boldsymbol{\beta}}$ that approaches the LSE $\boldsymbol{\beta}^\star$ in ℓ_2 norm with high probability.
procedure PrivateLearnLSE($(X, \boldsymbol{y}), \epsilon, \delta, \alpha, \eta, \gamma, \kappa$)
    $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}, \widehat{\Sigma}_{\boldsymbol{X}}) \leftarrow$ LearnGaussian-HD$(\{\boldsymbol{X}_i\}_{i\in[n]}, O(\epsilon), O(\delta), O(\alpha), \gamma, \kappa)$.
    $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y} \leftarrow$ LearnSubGaussian-HD$(\{y_i\boldsymbol{X}_i\}_{i\in[n]}, O(\epsilon), O(\delta), O(\eta), \gamma, c^2\kappa)$.
    Δώσε ως αποτέλεσμα την ιδιωτική εκτίμηση $\widehat{\boldsymbol{\beta}} = \left(\widehat{\Sigma}_{\boldsymbol{X}} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T\right)^{-1}\widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y}$.

ιδιωτικός. Επιπροσθέτως, εάν ο αριθμός των επισημειωμένων παραδειγμάτων είναι τουλάχιστον:

$$n = O\left(\frac{d\log(\frac{d}{\gamma})}{\eta^2} + \frac{d\text{polylog}(\frac{d\log(1/\delta)}{\eta\gamma\epsilon})}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log(\kappa\rho c)}\text{polylog}\left(\frac{d\log(\kappa\rho c)}{\gamma\epsilon\delta}\right)}{\epsilon}\right)$$
$$+ O\left((1+R)\left(\frac{d\log(\frac{d}{\gamma})}{\alpha^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d\log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\text{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right)\right),$$

τότε με πιθανότητα τουλάχιστον $1 - O(\gamma)$ *η έξοδος* $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ *και ο LSE* $\boldsymbol{\beta}^*$ *ικανοποιούν τη σχέση:*

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right\|_2^2 \le O\left(\alpha^2\right)\cdot\left\|\Sigma^{1/2}\boldsymbol{\beta}^\star\right\|_2^2 + O\left(\eta^2\right)\cdot c^2\,.$$

1.7.2 Δυαδική Παλινδρόμηση

Έπειτα, στρέφουμε την προσοχή μας στο περιβάλλον Δυαδικής Παλινδρόμησης, στην οποία ισχύουν τόσο η υπόθεση 5.2.1 όσο και η υπόθεση 5.3.1. Μελετούμε τις ιδιότητες του αλγορίθμου PrivateLearnLSE υπό την ανωτέρω επιπρόσθετη υπόθεση: η μόνη μεταβολή στον Αλγόριθμο 1, εν συγκρίσει με την ανωτέρω ενότητα, είναι ότι δεν απαιτείται η εκτίμηση $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$, εφόσον από υπόθεση έχουμε ότι $\boldsymbol{\mu} = \mathbf{0}$. Επομένως, θέτουμε $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} = \mathbf{0}$ στον Αλγόριθμο, με τους υπόλοιπους όρους να υπολογίζονται όπως στην προαναφερθείσα ενότητα. Αποδεικνύεται ότι ο αλγόριθμος παρουσιάζει τις ακόλουθες εγγυήσεις.

Theorem 1.7.2 (Ιδιωτικότητα και Ακρίβεια του $\widehat{\boldsymbol{\beta}}$ στην Ιδιωτική Δυαδική Παλινδρόμηση)**.** *Κάτω από την υπόθεση 5.2.1 με παράμετρο συνδιασποράς* κ *και την υπόθεση 5.3.1 με αληθινό συντελεστή παλινδρόμησης* $\boldsymbol{\beta} \in \mathbb{R}^d$*, για κάθε παραμέτρους ιδιωτικότητας* $\epsilon, \delta > 0$*, ακρίβειας* $\alpha, \eta > 0$ *και εμπιστοσύνης* $\gamma \in (0,1)$*, ο αλγόριθμος* PrivateLearnLSE *(οριζόμενος ως Αλγόριθμος 1) με* $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} = \mathbf{0}$ *είναι* $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$*-διαφορικά ιδιωτικός. Επιπροσθέτως, εάν ο αριθμός των επισημειωμένων παραδειγμάτων είναι τουλάχιστον:*

$$n = O\left(\frac{d\log(\frac{d}{\gamma})}{\eta^2} + \frac{d\text{polylog}(\frac{d\log(1/\delta)}{\eta\gamma\epsilon})}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\text{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon\delta}\right)}{\epsilon}\right)$$
$$+ O\left(\frac{d\log(\frac{d}{\gamma})}{\alpha^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d\log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon}\right),$$

τότε με πιθανότητα τουλάχιστον $1 - O(\gamma)$ *η έξοδος* $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ *ικανοποιεί τη σχέση:*

$$\|\widehat{\boldsymbol{\beta}} - k\boldsymbol{\beta}\|_2^2 \le O\left(\alpha^2\right)\cdot\left(1 + \left\|k\Sigma^{1/2}\boldsymbol{\beta}\right\|_2^2\right) + O\left(\eta^2\right),$$

όπου $k = \frac{2n}{n-d-1}\,\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right)\right]$ *.*

1.8 Επίλογος

Στην εποχή της έκρηξης των πληροφοριών, ο ρόλος της ιδιωτικότητας και το δικαίωμα του κάθε ατόμου (του οποίου τα δεδομένα γίνονται αντικείμενο επεξεργασίας) να είναι βέβαιο ότι δεν θα υπάρξει κατάχρησή τους λαμβάνουν κεντρική θέση στον σχεδιασμό των αλγορίθμων. Η Διαφορική Ιδιωτικότητα είναι το κύριο εργαλείο που έχει ανακαλυφθεί για να παρέχει προστασία με αυστηρές, μαθηματικά επαληθεύσιμες εγγυήσεις. Ωστόσο, εξαιτίας της αυστηρής υφής των περιορισμών που είναι καθοριστικοί στον ορισμό της Διαφορικής Ιδιωτικότητας, είναι συνήθως δύσκολη η ενασχόληση με μη-φραγμένες ποσότητες και πώς θα μπορέσαμε να τις προστατεύσουμε. Σε αυτή την εργασία, δίνουμε και αναλύουμε εκτιμητές για να βγάλουμε συμπεράσματα σε πολλά ενδιαφέροντα περιβάλλοντα παλινδρόμησης με μη-φραγμένες συμμεταβλητές, αποδεικνύοντας επίσημα ότι είναι ιδιωτικοί και αποδοτικοί.

Πιστεύουμε ότι μια σειρά εργασιών σε μη-φραγμένες μεταβλητές είναι εξαιρετικού ενδιαφέροντος τόσο ως προς το θεωρητικό όσο και ως προς το πρακτικό τους μέρος. Πιθανές μελλοντικές ερευνητικές προεκτάσεις βασισμένες σε αυτή την εργασία θα μπορούσαν να αναδυθούν ώστε να ωθήσουν τα όρια της χρήσης μη-φραγμένων μεταβλητών, για παράδειγμα μέσω συμπερίληψης πιθανών εξαρτήσεων ανάμεσα στα δεδομένα εισόδου. Επιπροσθέτως, είναι γνωστό ότι τα κάτω όρια σε διαφορικά ιδιωτικούς μηχανισμούς εκτίμησης για περιβάλλοντα παλινδρόμησης είναι δύσκολο να εξαχθούν και απέχουν από το να χαρακτηριστούν ως βέλτιστα (δείτε, π.χ., την δουλειά του Wang (2018)) και συνεπώς η εξέταση πιθανών κάτω ορίων σε μη-φραγμένα περιβάλλοντα θα ήταν ακόμη μια πολλά υποσχόμενη ερευνητική κατεύθυνση. Αυτό θα καταλήξει εντέλει σε πολύ ενδιαφέρουσες πρακτικές υλοποιήσεις εφαρμόσιμες σε πραγματικά συστήματα Μηχανικής Μάθησης, με την δυνατότητα να παρέχουν εγγυήσεις ιδιωτικότητας καλύτερες από τις τρέχουσες υλοποιήσεις αντίστοιχων συστημάτων, ακόμη και στην περίπτωση δεδομένων χρηστών που απέχουν αυθαίρετα πολύ από το αναμενόμενο.

Chapter 2

Introduction

Modern algorithms interface intricately with databases that contain users' data. In certain cases such as healthcare and finance, or even more generally, users may consider their data to be sensitive, or simply desire not to be identified through the (publicly released) result of an algorithm; for example, a machine learning model trying to predict a certain type of cancer should not be able to reveal some individual as participating in the dataset from which the algorithm has learned the model.

Ensuring the private handling of every individual's data, while at the same time making inferences about a population's behavior are two targets that largely are at odds with each other. On the one hand, guaranteeing perfect privacy means that no useful statistical information may be extracted, whilst, on the other hand, ensuring a perfectly accurate result based on the data at hand necessitates the curtailing (or even worst, abolition) of privacy. Therefore, it is an exciting topic to figure out the tradeoffs that would enable to both guarantee privacy for the individuals contributing their data and to estimate a statistical quantity of interest with adequate accuracy.

To this end, there have been in the past efforts to *anonymize* datasets in order for them to be released for some common purpose, which eventually led to grand fiascoes through the re-identification of specific users from those "anonymized" datasets (see, e.g., Narayanan and Shmatikov (2008) for a case in point). Clearly, having a non-rigorous claim that "all personal data got erased before release" is grossly inadequate; there needs to be a mathematically rigorous *privacy guarantee* which the maker of the algorithm would be able to definitively assert that their algorithm satisfies.

Differential Privacy (Dwork et al., 2006) is a concept developed as a response to the above delineated context: it offers a strict privacy guarantee with highly favorable, strong properties that guard the privacy of the individuals whose data is contained in a dataset used by an algorithm. It is based on the general principle that, ultimately, privacy for an individual is about protecting them from being identified through the (publicly released) result of some algorithm. Intuitively, Differential Privacy (notated as DP) *guarantees* that including or excluding any specific individual in the dataset (through a modification of one of the entries of the dataset) should not be able to affect "a lot" the result of the algorithm.

More specifically, consider a dataset (collection of n samples) $X = (X_1, X_2, \ldots, X_n) \in \mathcal{X}^n$. We define two datasets $X, X' \in \mathcal{X}^n$ as adjacent, if they differ on at most one individual sample (e.g., sample $X_i \neq X'_i$ for some i and all other samples are the same). A *randomized* algorithm M (also called a "mechanism" for historical reasons) is *differentially private*, if the distributions of $M(X)$ and $M(X')$ are "very similar" for every pair of adjacent datasets X, X'. We provide the mathematically precise definition of DP in Section 3.1, along with what we mean by distribution "similarity."

As mentioned above, Differential Privacy is popular due to its strong, highly favorable properties that show that this type of privacy guarantee is *preserved* through a wide variety of operations, either on the algorithm's result ("post-processing") or by simultaneously executing many (differentially private) algorithms on the private dataset ("composition"). We formalize these properties in Section 3.2.

In order to offer a differentially private estimation algorithm, the standard course of action is as follows: we begin with a non-private algorithm (that computes some function $f(X)$ of the input dataset X) and add random noise η of an adjustable level depending on the level of privacy that we desire to offer. The output of the (private) algorithm is now $f(X) + \eta$. The random noise η may follow a wide variety of distributions, with by far the most prominent being the Laplace and the Gaussian distributions. We analyze these basic privacy-inducing mechanisms in Section 3.3.

Early work in Differential Privacy focused on estimating quantities of a dataset (such as the median, or the most frequent sample), instead of an underlying population (where we would, for instance, be interested in estimating the mean of a population). The connection of Differential Privacy to statistical inference came much later in the computer science literature (see, e.g., Dwork et al. (2010)). For statistical inference, but also more generally, there are two targets in which we are interested: 1. guaranteeing that the estimation algorithm is *private* (differentially private, in particular), and 2. ensuring that our algorithm is *accurate* with respect to the underlying population quantity that we wish to estimate. It is apparent that when we are dealing with an underlying population, we seek our privacy guarantee to be "worst-case", i.e., including any and all outliers (data points far from the expected values) that might be present in the dataset, since we would want to protect every individual's data from identification, regardless of their aberration from the mean. On the contrary, the private statistical estimation algorithm at hand can offer a meaningful accuracy guarantee in an "average-case" regime, i.e., when the dataset that the algorithm receives is by and large "good enough."

However, due to the worst-case nature of the Differential Privacy guarantee, it is in general very hard to construct algorithms for unbounded input, i.e., when the samples may be arbitrarily far from an "expected value." This is why early-era literature is blithe with boundedness assumptions on the samples of the dataset (see, e.g., Dwork and Roth (2014) for a review). Even for the problem of Mean Estimation, when we have samples that follow a multivariate Gaussian distribution (considered to be one of the most fundamental problems of statistics and incidentally, one where there may be *outliers* arbitrarily far away from the mean, even though most samples are around a well-defined mean), it was not until very recently that Kamath et al. (2019) managed to construct a differentially private algorithm for efficiently estimating the mean and covariance matrix of a multidimensional Gaussian distribution. Their work built upon research from Karwa and Vadhan (2018), who showed how to estimate the mean of (univariate) Gaussian random variables in an efficient way.

In this work, we make a step further and examine private and efficient statistical estimation in common regression settings, where we have unbounded input in the dataset. Regression is a set of processes fundamental to statistical inference, preliminary forms of which date back to the early days of Legendre (1806) and Gauss (see, e.g., Plackett (1949)). There is an abundance of regression settings: hereby, we examine many of those settings (such as finding the Least Squares Estimate, and estimating the regression coefficients of an underlying linear model from labelled data) under Gaussian marginals. The input dataset consists of n labelled points $(\boldsymbol{X}_i, y_i)$ with covariates $\boldsymbol{X}_i$ and labels y_i, where crucially, the covariates $\boldsymbol{X}_i$ are allowed to be *unbounded*.

The work on learning linear models alone is vast (see Cai et al. (2020); Wang (2018) for two recent reviews). Nevertheless, despite the intense interest on this topic (see the works of Iyengar et al. (2019); Zhang et al. (2017); Jain and Thakurta (2014), to name just a few), *all of the existing work on regression provides differential-privacy guarantees assuming bounded covariates.* Intuitively, this can be explained by inspecting even the simple least squares estimator used in linear regression. It is easy to see that estimator's *sensitivity*, i.e., its variability under changes on a single sample, is determined by the design matrix (i.e., the matrix of samples). As sensitivity has a direct effect on differential privacy guarantees, bounding the

design matrix's eigenvalues is the prevalent approach for bounding the sensitivity. For this reason, assuming bounded covariates is a ubiquitous assumption in DP literature on both linear regression and learning generalized linear models.

This assumption is quite restrictive, and is frequently identified as a deficiency of DP regression algorithms from a practical standpoint (Anonymous, 2019). It is also a significant drawback from a theoretical standpoint, as it precludes studying DP-estimators on data sampled from distributions of *unbounded support*. Even the Gaussian distribution, perhaps the most commonly used generative distribution in statistical machine learning literature (Deng et al., 2021; Daskalakis et al., 2020; Kini and Thrampoulidis, 2020; Diakonikolas et al., 2019b; Nakkiran, 2019; Kreidler et al., 2018), cannot be used in conjunction with the existing DP regression algorithms and maintain DP guarantees.

Our work aims to directly address this, by providing DP algorithms for regression assuming (unbounded) Gaussian covariates. In doing so, we leverage and extend the recent work of Kamath et al. (2019), who proposed differentially private mechanisms for estimating the mean and the covariance matrix of high-dimensional Gaussian random vectors.

2.1 Our contributions

Our first major contribution is to answer the following question in the affirmative:

Question 2.1.1. *Is private regression analysis with unbounded covariates possible?*

We study this problem in the context of three scenarios: Linear Regression, Least Squares Fitting, and Binary Regression. In all of these settings we assume (unbounded) Gaussian covariates.

In the Least Squares Fitting setting (see Section 5.2), given a training set $\{(\boldsymbol{X}_i, y_i)\}$, our goal is to efficiently and privately compute an estimate that is close to the Least Squares Estimate (LSE), i.e., the coefficients of the best-fitting linear function. In this problem, we assume that labels y_i are bounded, but make no further assumptions on how they relate to the covariates $\boldsymbol{X}_i \in \mathbb{R}^d$. Our main result is the following:

Informal Theorem 3. *For accuracy $\alpha > 0$ and privacy guarantees $\epsilon, \delta > 0$, there exists an efficient (ϵ, δ)-DP algorithm that, with high probability, approximates arbitrarily α-closely the Least Squares Estimate using $n = \widetilde{O}\left(d/\alpha^2 + d^{3/2}\log(1/\delta)/(\alpha\epsilon)\right)$ samples.*

In our second setting, that of simple Linear Regression (see Section 5.1), we show that using a variant of the methods of Kamath et al. (2019) and Karwa and Vadhan (2018) is enough to provide a private and accurate estimate of the underlying regression coefficient.

Finally, in the setting of Binary Regression (see Section 5.3), we further assume that labels are binary (i.e., $y_i = \pm 1$) and that covariates are zero mean. Moreover, labels are generated by a generalized linear model of the form $\mathbf{Pr}[y_i = +1|\boldsymbol{X}_i] = f(\boldsymbol{\beta}^T\boldsymbol{X}_i)$, where $f : \mathbb{R} \to [0, 1]$ is the model function and $\boldsymbol{\beta} \in \mathbb{R}^d$ is the true regression coefficient. This setting captures some of the most fundamental machine learning tasks, such as logistic regression and learning linearly-separable Support Vector Machines (SVMs). Our second main result is that the same differentially private estimator we used in Least Squares Fitting scenario can applied to the Binary Regression to obtain the following guarantees:

Informal Theorem 4. *For accuracy $\alpha > 0$ and privacy guarantees $\epsilon, \delta > 0$, there exists an efficient (ϵ, δ)-DP algorithm that, with high probability, approximates arbitrarily α-closely the true regression coefficient up to a multiplicative factor using $n = \widetilde{O}\left(d/\alpha^2 + d^{3/2}\log(1/\delta)/(\alpha\epsilon)\right)$ samples.*

To the best of our knowledge, these results constitute the first efficient and private algorithm for regression analysis with unbounded feature vectors. From a technical standpoint,

our analysis relies on the fact that the LSE requires the calculation of the inverse of a moment matrix, as well as the expectation of a central random quantity $y_i \boldsymbol{X}_i$. The former can be estimated in a private way using recent techniques developed by Kamath et al. (2019) and Karwa and Vadhan (2018); we show that the second can also be estimated similarly by extending the work Kamath et al. (2019) and Karwa and Vadhan (2018) to sub-gaussian vectors. We also show that the sample complexity dependence on ambient dimension d can be improved in both of our regression settings (from $\widetilde{O}(d^2)$ to $\widetilde{O}(d^{3/2})$) compared to covariance estimation.

2.2 Related work

Generic work on Differential Privacy has been vast so far, making it impracticable to cite in detailed form the evolution of the field. Instead, and as usual in these cases, for a thorough overview of the phases that Differential Privacy underwent before culminating in its current form, as well as detailed citations of previous work, we refer to the chapters of Dwork and Roth (2014) and the references therein. We focus the rest of this section on research that is immediately relevant to our work, not merely on the broader DP spectrum.

Differentially Private Regression and GLMs with Bounded Covariates. Linear regression is of course a true workhorse of statistics, and there has been a significant body of work on the design of computationally and statistically efficient differentially private regression algorithms (see e.g., the recent surveys of Cai et al. (2020); Wang (2018) and the references therein). Approaches include objective perturbation (Iyengar et al., 2019; Kifer et al., 2012; Zhang et al., 2012; Chaudhuri et al., 2011), output perturbation (Asi and Duchi, 2020; Iyengar et al., 2019; Zhang et al., 2017; Jain and Thakurta, 2014), gradient perturbation (Abadi et al., 2016; Bassily et al., 2014), subsample-and-aggregate (Barrientos et al., 2019; Dwork and Smith, 2010), and sufficient statistics perturbation (Alabi et al., 2020; Wang, 2018; McSherry and Mironov, 2009). Additionally, several works study generalizations of such mechanisms to Generalized Linear Models (GLMs) (Kulkarni et al., 2021; Iyengar et al., 2019; Jain and Thakurta, 2014; Kifer et al., 2012). All above works, however, either operate under a random setting with bounded covariates, or use a fixed design matrix X with bounded minimum eigenvalue on $X^T X$. Such strong assumptions on the boundedness of feature vectors are precisely the kind of assumptions that our work aims to mend.

Mean and Covariance Estimation. The study of differentially private mechanisms for mean and covariance estimation under bounded covariates is classical (see, e.g., Amin et al. (2019); Dwork et al. (2014); McSherry and Mironov (2009)). Sheffet (2017) studies covariance estimation under Gaussian samples, also applying it to the Least Squares Fitting problem we study here; nevertheless, their differential privacy guarantee assumes an upper bound on the covariates. Karwa and Vadhan (2018) resolve, for the first time, the problem of differentially private (unbounded) univariate Gaussian mean estimation with almost optimal dependence on problem parameters. Also in the univariate setting, Bun et al. (2015) learn more general distributions w.r.t. Kolmogorov distance, which is weaker than the total variation considered by Karwa and Vadhan (2018); Diakonikolas et al. (2015) extend this work to total variation distance, again for univariate distributions.

Kamath et al. (2019) extend the work of Karwa and Vadhan (2018) to multivariate mean and covariance estimation for high-dimensional Gaussian random vectors – see Section 6.1 for a detailed description of their guarantees. Related to our setting, Cai et al. (2020) provide lower bounds for the sample complexity of differentially-private learning the mean of Gaussian random vectors, though the estimation algorithms they propose operate over bounded covariates. Our work relies on (and extends) Kamath et al. (2019), applying it both to sub-gaussian vectors as well as to Least Squares Fitting, which Karwa and Vadhan (2018) identify

as an *important* possible extension of their work. Recently, Aden-Ali et al. (2021) obtained nearly optimal sample complexity upper bounds for agnostically learning multivariate Gaussians in an approximate differentially private setting; however, they exhibit an existential proof, commenting that no computational method is known to match this upper bound. The latter indicates the difficulties that arise when agnostic unbounded settings are considered.

LSE for GLMs. The differentially private algorithm we propose applies Least Squares Estimation (LSE) to learn the parameters of a Binary Generalized Linear Model (GLM) (see Theorem 8.1.2) and, more generally, to perform Least Squares Fitting over bounded labels (c.f. Theorem 8.1.1). It is well known that, under Gaussian marginals, LSE is an unbiased estimator of the parameter vector of an affine GLM, up to a scaling factor (Erdogdu, 2016; Sun et al., 2014; Brillinger, 2012a). This is a consequence of Stein's Lemma (Liu, 1994; Stein, 1981) – see also Section 5.3.2. In the binary setting, LSE can also be seen as a special case of the so-called Linear Discriminant Analysis (LDA) classification algorithm (Hastie et al., 2009). As a result, our Theorem 8.1.2 can also be seen as a differentially private implementation of LDA.

Chapter 3

Background on Differential Privacy (DP)

As argued in the introduction, privacy is a highly sought-after attribute of data-handling algorithms, especially nowadays, with the vast wealth of information readily available to both companies and potentially malicious actors. Should the results of a (statistical) procedure run on private information become public in some way, we would want to ensure that the released results will be future-proof against any possible future privacy violation. This is due to the fact that the released data will be retained indefinitely exactly in the form that they were originally released, without the ability for them to amended, shall a privacy violation occur. Additionally, precisely because any future "adversary" might access data dated back in time, it is simply insufficient to account only for present-day challenges with present-day computing resources.

In 2006, the notion of Differential Privacy was mathematically formulated (Dwork et al., 2006). It is the rigorous form of data privacy that stood the test of time, and has become the most used notion of statistical privacy nowadays. Differential Privacy solves the challenge posed in the previous paragraph: any adversary, no matter how much computational power they have, will not be able to distinguish the participation of one specific individual in a (private) dataset through the results of a differentially private algorithm run on this dataset. Due to its properties that we will explore, Differential Privacy is widely used, with the most recent prominent example being the 2020 U.S. Decennial Census (US Census Bureau, 2020).

3.1 Definition of (central) DP

Differential Privacy is a property of a randomized algorithm/mechanism, denoted by M. Intuitively, what Differential Privacy states is that no matter which (one) individual is altered from a large collection of individuals' data, the output of the algorithm will remain "approximately the same;" to be more precise, the probability distribution of the possible outputs of the algorithm will remain approximately the same. In order to formulate that precisely, privacy parameters $\epsilon, \delta > 0$ are utilized, which one would want to be as close to zero as possible for "more privacy."

The key thought in privacy is that we need the privacy guarantee to hold for *any* individual, i.e., be a worst-case guarantee. This is contrary to the popular statistical assumption of some particular distributional model, where it is assumed that data comes from a particular family of distributions and the provided guarantees are average-case. Else, it does not really matter for every individual if they might not be protected, had the privacy guarantee been "for most individuals" or had some other notion of averaging amongst individuals' privacy compromise occurred. This is why we ultimately need a mathematical formulation that will provide its guarantee for every dataset the algorithm (that we wish to claim as "private") might receive.

We now move forward towards the definition of DP. We say that a randomized algorithm/mechanism M that takes as input a dataset X is (ϵ, δ) (centrally/globally) differentially private (DP), if for all datasets X, X' that are adjacent (i.e., only a single element/sample is different, which might be denoted as $\|X - X'\| = 1$ for some appropriately defined norm)

and all possible subsets of the range of the algorithm's output $Y \subseteq \text{Im}(M)$, it holds that the probability that the output of the algorithm (when run on dataset X) belongs to the set Y is close to the probability that the output of the algorithm (when run on the *adjacent* dataset X') belongs to the same set Y. In particular, we have the following formal definition.

Definition 3.1.1 (Differential Privacy (Dwork et al., 2006))**.** A randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ϵ, δ)-differential privacy (equivalently, is said to be (ϵ, δ)-DP) if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$ that differ on at most one element,

$$\mathbf{Pr}[M(X) \in Y] \leq \exp(\epsilon)\, \mathbf{Pr}[M(X') \in Y] + \delta\,, \forall Y \subseteq \mathcal{Y}\,.$$

Where did δ come from? The above definition of DP is sometimes called "approximate DP," whereas in the case that $\delta = 0$, i.e., $\mathbf{Pr}[M(X) \in Y] \leq \exp(\epsilon)\, \mathbf{Pr}[M(X') \in Y]$, it is called "pure DP." Someone looking at Definition 3.1.1 might wonder what the significance of δ is, since one might quickly realize that this definition does not preclude "catastrophic" failures of privacy (Canonne, 2021). This fact may be illustrated by the following example algorithm:

- Get a database D of n secret data from n individuals.
- Select at random $\delta \cdot n$ rows of this database D.
- Release those $\delta \cdot n$ rows directly at the output.

It is evident that the result of this algorithm (the last step) is a "catastrophic failure of privacy," since a portion –albeit very small– of the database got released out in the clear, thus leaving those individuals *completely* unprotected. Interestingly, this algorithm is (ϵ, δ)-DP for any $\epsilon > 0$ (and even for $\epsilon = 0$!). This example indicated that caution should be exercised with setting δ; in particular, it is suggested that $\delta \ll 1/n$ and more specifically, be cryptographically small, i.e., smaller than the reverse of any polynomial of n (e.g., be $1/2^n$). This is usually feasible in most settings and algorithms, since they exhibit some kind of $O(\log(1/\delta))$ dependence on δ.

Clearly, the approximate DP definition breaks the comparability/similarity between the outputs in the worst case. However, by utilizing the following definition of the "privacy loss random variable," one can see that δ is something like the probability that the "pure" ϵ-DP guarantee might fail to hold, so one might argue that "pure" privacy occurs with "high" probability $1 - \delta$. We mention, though, that in fact, the converse of does not hold exactly, and what holds is quite more nuanced, referenced in Lemma 9 of Canonne et al. (2020).

Definition 3.1.2 (Privacy Loss random variable)**.** Given a randomized algorithm M and two datasets D, D' differing on a single element, define the following *deterministic* function for every possible output y of the algorithm M:

$$f(y) = \log\left(\frac{\mathbf{Pr}\left[M(D) = y\right]}{\mathbf{Pr}\left[M(D') = y\right]}\right).$$

The Privacy Loss random variable (intuitively, how much the output of the algorithm helps in distinguishing D and D') is $Z = f(M(D))$, inheriting the randomness of M. As f states, Z also implicitly depends on D and D'.

Lemma 3.1.1 (ϵ-bounded Privacy Loss with probability $1 - \delta \Rightarrow (\epsilon, \delta)$-DP)**.** *Given an algorithm M, if for any two neighboring datasets D, D' it holds that the Privacy Loss random variable is bounded by ϵ with probability at least $1 - \delta$, i.e., that*

$$\mathbf{Pr}\left[Z(M, D, D') \leq \epsilon\right] \geq 1 - \delta\,,$$

then algorithm M is (ϵ, δ)-DP.

3.2 Properties of DP algorithms

We will hereby examine some properties of differentially private algorithms that make the provided privacy guarantee *robust to various changes* in the algorithm itself, in the group of individuals to whom the privacy is offered, and when combining the results of a series of algorithms run on (private) input.

Immunity to post-processing. The first property is the "post-processing" property, which intuitively states that it is impossible to degrade the privacy guarantee given by the DP definition without additional knowledge about the *private* dataset used to compute the differentially private statistic. In other words, no matter what other "outside information" an adversary might use, they cannot combine that information with the DP result to abridge the guaranteed level of privacy, so long as the private dataset used for the DP computation remains private on its entirety, i.e., the adversary has no knowledge of information correlated in some way with the (private) elements of the dataset used.

Lemma 3.2.1 (Post-processing)**.** *Given an (ϵ, δ)-DP randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$, let $f : \mathcal{Y} \to \mathcal{R}$ be an arbitrary randomized mapping. Then, $f \circ M : \mathcal{X}^n \to \mathcal{R}$ is (ϵ, δ)-DP.*

Any post-processing algorithm applied on the result of a (ϵ, δ)-DP algorithm is going to have absolutely no degradation of its privacy. Essentially, what this says is that we may obtain some intermediate private result (in a potentially easy-to-implement way) and then release this result, allowing anyone to perform any further (statistical or other) calculations on their own, without having to worry whether they might process this result in a certain way so as to violate the (ϵ, δ) privacy guarantee. The DP guarantee cannot be violated under post-processing, as is formally said.

Group privacy. The second desirable property is about what happens when we desire to guarantee privacy for groups of k individuals (and not just one, as the DP definition explicitly states). In this case, we *can* guarantee a similar fact for groups of k individuals, which is extremely useful, e.g., for members of a household. In the following Lemma, we refer only to the case of "pure" $(\epsilon, 0)$-DP, since the guarantees given for (ϵ, δ)-DP are quite intricate.

Lemma 3.2.2 (Group privacy)**.** *Given an $(\epsilon, 0)$-DP algorithm $M : \mathcal{X}^n \to \mathcal{Y}$, for every pair of datasets $X, X' \in \mathcal{X}^n$ that differ on at most k elements (i.e., k out of the original n elements have been altered), it holds that*

$$\mathbf{Pr}[M(X) \in Y] \leq \exp(k\epsilon)\, \mathbf{Pr}[M(X') \in Y]\,, \forall Y \subseteq \mathcal{Y}\,.$$

The proofs of Lemmata 3.2.1 and 3.2.2 are omitted, since they are not used in this thesis, and may be found in Dwork and Roth (2014).

Composition. We now move on to the composition theorems of DP, which state how Differentially Private algorithms may be combined together and analyze precisely how their privacy guarantees degrade from this "composition." The first (simple) composition theorem states that, if we apply two DP algorithms at the same input and release both results, the privacy guarantee degrades gracefully. In particular, if $\mathcal{A}(D)$ and $\mathcal{B}(D)$ are two $(\epsilon, 0)$-DP algorithms operating on a dataset D, then releasing the tuple $(\mathcal{A}(D), \mathcal{B}(D))$ is a $(2\epsilon, 0)$-DP algorithm.

Why is composition desirable? The greatest threat to privacy comes from a combination of different results, each of which might (at a first glance) seem private on their own. Let's reiterate the illuminating example of Steinke and Ullman (2020): consider an employer that

employs 1000 individuals and has insured them in a custom insurance plan of an insurance provider. The employer might want aggregated data over how many of their total employees have a certain condition. Let's suppose that the query is answered and is 42 of the current employees. This number on its own does not consitute a violation of privacy, since it is not possible for the employer to know whether some particular employee has or does not have the said condition. The employer only has the aggregated data, which would (at a first glance) seem safe. However, what happens if another individual gets hired at the employer's company? Then, the employer might be able to ask the same aggregating question again, and say the result was 43 this time. Immediately, the employer knows that the specific individual that they hired right now has the said condition. This constitutes an extreme violation of privacy. Observe that the employer was able to obtain this critical information by *combining* seemingly innocuous, aggregated data. This is why composition, one example of which is the use of the same algorithm (e.g., the aggregated data here) multiple times, is crucial for proving some sense of conservation of privacy. Combining information from multiple sources released without rigorous guarantees might prove to be detrimental for individuals, as Narayanan and Shmatikov (2008) eloquently illustrated in practice, utilizing public data from IMDb to break the anonymity of a Netflix dataset released for a competition.

Lemma 3.2.3 (Simple composition of $(\epsilon, 0)$-DP mechanisms)**.** *Let $M_1 : \mathcal{X}^n \to \mathcal{R}_1$ be an $(\epsilon_1, 0)$-DP algorithm, and $M_2 : \mathcal{X}^n \to \mathcal{R}_2$ be an $(\epsilon_2, 0)$-DP algorithm, whose internal randomizations are independent of each other. Define algorithm $M : \mathcal{X}^n \to \mathcal{R}_1 \times \mathcal{R}_2$ as the mapping $M(X) = (M_1(X), M_2(X))$. Then, algorithm M is $(\epsilon_1 + \epsilon_2, 0)$-DP.*

Since this result is rather easy to prove (the extension to the (ϵ, δ)-DP case requires measure-theoretic definitions and is, thus, omitted), we provide the standard proof hereby, which can extended by induction to the case of k mechanisms, whose results on the same dataset are to be released, to obtain a combined $\left(\sum_{i=1}^{k} \epsilon_i, 0\right)$-DP algorithm.

Proof. Let X, X' be adjacent datasets, and any $R_1 \subseteq \mathcal{R}_1, R_2 \subseteq \mathcal{R}_2$ such that

$$R' = \{(r_1, r_2) : r_1 \in R_1, r_2 \in R_2\} \subseteq \mathcal{R}_1 \times \mathcal{R}_2 \,.$$

Note that any subset of $\mathcal{R}_1 \times \mathcal{R}_2$ may be constructed in this way. Then, due to the individual privacy guarantees of each algorithm, it holds that

$$\mathbf{Pr}\left[M_1(X) \in R_1\right] \leq e^{\epsilon_1} \mathbf{Pr}\left[M_1(X') \in R_1\right] \text{ and } \mathbf{Pr}\left[M_2(X) \in R_2\right] \leq e^{\epsilon_2} \mathbf{Pr}\left[M_2(X') \in R_2\right] .$$

Finally, since the internal randomizations of the algorithms are independent of each other, we have the desired differential privacy guarantee for the combined algorithm M, as follows:

$$\begin{aligned} \mathbf{Pr}\left[M(X) \in R'\right] &= \mathbf{Pr}\left[M_1(X) \in R_1\right] \cdot \mathbf{Pr}\left[M_2(X) \in R_2\right] \\ &\leq e^{\epsilon_1 + \epsilon_2} \cdot \mathbf{Pr}\left[M_1(X') \in R_1\right] \cdot \mathbf{Pr}\left[M_2(X') \in R_2\right] \\ &= e^{\epsilon_1 + \epsilon_2} \cdot \mathbf{Pr}\left[M(X') \in R'\right] . \end{aligned}$$

□

However, the guarantee of $\left(\sum_{i=1}^{k} \epsilon_i, \sum_{i=1}^{k} \delta_i\right)$-DP for k mechanisms by Dwork and Lei (2009) is rather restrictive, since the algorithms are bound to be independent of each other, i.e., one may not take the output of another and some extra information from the (private) dataset and output a new result. Fortunately, there are sharper composition theorems, and in particular, the following *advanced composition theorem* that deals with the stronger environment that we seek: that of adaptive composition. The name "adaptive" comes from the fact that each query (algorithm, in fact) is submitted one-by-one and might be parameterized / might depend on the results of the previous queries / algorithms.

More specifically, Theorem 3.2.4 characterizes the privacy properties of a sequence of algorithms $M_1(X), M_2(X), \ldots, M_N(X)$, where the i-th algorithm may depend on the outcomes of the algorithms $M_1(X), M_2(X), \ldots, M_{i-1}(X)$, for $i \in [N]$.

Theorem 3.2.4 (Advanced Composition Theorem (Dwork and Roth, 2014))**.** *If M is an adaptive composition of differentially private algorithms $M_1, \ldots, M_N$, where M_i is (ϵ, δ_i)-DP for any $i \in [N]$, then it holds that M is $(\epsilon N, \sum_{i=1}^{N} \delta_i)$-DP and additionally, for every $\delta > 0$, M is $(\epsilon\sqrt{6N\log(1/\delta)}, \delta + \sum_{i=1}^{N} \delta_i)$-DP.*

3.3 Basic DP-inducing mechanisms

One popular way of constructing differentially private algorithms is to add some kind of noise to the "perfect" result that one would have expected from a non-private algorithm. This general method encompasses a number of techniques, and we will examine here some versions for numerical functions, that are classical and we will apply in our algorithms.

Before we move on to the first such mechanism (the Laplacian mechanism), we introduce the concept of ℓ_1 sensitivity of a vector-valued function, which intuitively captures the maximum (worst-case) change that one individual's data may cause to the underlying function whose value on the dataset we wish to estimate. Therefore, it is reasonable that we will need to add noise proportional to this sensitivity, in order to mask the participation of any single individual in the dataset.

Definition 3.3.1 (ℓ_1-sensitivity of function)**.** The ℓ_1 sensitivity of a function $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^d$ is the maximum perturbation $\Delta_1 \boldsymbol{f}$ that an adjacent dataset change (which is denoted by $\|x - y\| = 1$) may cause to the function in ℓ_1-norm:

$$\Delta_1 \boldsymbol{f} = \max_{\|x-y\|=1} \|\boldsymbol{f}(x) - \boldsymbol{f}(y)\|_1 \,.$$

Definition 3.3.2 (Laplace distribution)**.** We say that a random variable Z is distributed as per the Laplace distribution $\mathcal{L}(\mu, b)$ centered at μ with scale parameter b if its probability density function is:

$$g(z) = \frac{1}{2b} \exp\left(-\frac{|z-\mu|}{b}\right), \ \forall z \in \mathbb{R}\,.$$

Lemma 3.3.1 (Laplacian Mechanism)**.** *For a function $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^d$ and any $\epsilon > 0$, outputting the value of $\boldsymbol{f}(x) + \boldsymbol{\eta}$ is a $(\epsilon, 0)$-DP algorithm, where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_d)$ and the $\{\eta_i\}_{1 \le i \le d}$ are i.i.d. Laplacian random variables such that $\eta_i \sim \mathcal{L}(0, \Delta_1 \boldsymbol{f}/\epsilon)$.*

Proof. Let X, X' be adjacent datasets, i.e., such that $\|X - X'\| = 1$. Then, according to Definition 3.3.1, we have that

$$\left\|\boldsymbol{f}(X) - \boldsymbol{f}(X')\right\|_1 \le \Delta_1 \boldsymbol{f}\,. \tag{3.1}$$

Therefore, we may now immediately prove that the Laplacian Mechanism satisfies the definition of Differential Privacy (specifically, "pure" $(\epsilon, 0)$-DP), since the ratios of the probability density functions of the output $\boldsymbol{y} \in \mathbb{R}^d$ of the mechanism satisfy (by the subscript

notation, we notate the mechanisms where the input datasets are X and X' respectively):

$$\begin{aligned}\frac{p_X(\boldsymbol{y})}{p_{X'}(\boldsymbol{y})} &= \prod_{i=1}^{d} \frac{\exp\left(-\frac{\epsilon|y_i - f_i(X)|}{\Delta_1 \boldsymbol{f}}\right)}{\exp\left(-\frac{\epsilon|y_i - f_i(X')|}{\Delta_1 \boldsymbol{f}}\right)} \\ &= \exp\left(\sum_{i=1}^{d} \frac{\epsilon\left(|y_i - f_i(X)| - |y_i - f_i(X')|\right)}{\Delta_1 \boldsymbol{f}}\right) \\ &\leq \exp\left(\sum_{i=1}^{d} \frac{\epsilon|f_i(X') - f_i(X)|}{\Delta_1 \boldsymbol{f}}\right) \\ &= \exp\left(\frac{\epsilon\left\|\boldsymbol{f}(X') - \boldsymbol{f}(X)\right\|_1}{\Delta_1 \boldsymbol{f}}\right) \\ &\leq e^{\epsilon}\,,\end{aligned}$$

by the triangle inequality and Equation (3.1), as desired. □

It may also be proven (Dwork and Roth, 2014) that the accuracy of the Laplacian mechanism is as follows: if we name the result of the mechanism in Lemma 3.3.1 as $\boldsymbol{y}$ and reminding that the "perfect" result on the dataset x is $\boldsymbol{f}(x)$, then for every $\gamma \in (0,1)$, it follows by the tail of the Laplace distribution and a union bound that

$$\mathbf{Pr}\left[\|\boldsymbol{y} - \boldsymbol{f}(x)\|_\infty \leq \ln\left(\frac{d}{\gamma}\right) \cdot \frac{\Delta_1 \boldsymbol{f}}{\epsilon}\right] \geq 1 - \gamma\,,$$

where $\|\boldsymbol{a}\|_\infty = \max_{i \in [d]} |a_i|$ is the max-norm of a vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_d) \in \mathbb{R}^d$.

To continue, we mention the Gaussian mechanism, which varies correspondingly to the ℓ_2 sensitivity of a function, instead of the ℓ_1 and is one of the most significant mechanisms for inducing (ϵ, δ)-DP with $\delta > 0$. Notice that it is a significant difference of the Laplacian and the Gaussian mechanisms that the latter cannot be used to induce "pure" ($\delta = 0$) differential privacy, but instead requires that we set $\delta > 0$. This issue is inherent to the Gaussian mechanism and the way it adds noise proportional to the ℓ_2 sensitivity of a function (Dwork and Roth, 2014).

Definition 3.3.3 (ℓ_2-sensitivity of function)**.** The ℓ_2 sensitivity of a function $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^d$ is the maximum perturbation $\Delta_2 \boldsymbol{f}$ that an adjacent dataset change (which is denoted by $\|x - y\| = 1$) may cause to the function in ℓ_2-norm:

$$\Delta_2 \boldsymbol{f} = \max_{\|x-y\|=1} \|\boldsymbol{f}(x) - \boldsymbol{f}(y)\|_2\,.$$

Lemma 3.3.2 (Gaussian Mechanism)**.** *For a function $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^d$ and any $\epsilon \in (0,1), \delta > 0$, outputting the value of $\boldsymbol{f}(x) + \boldsymbol{\eta}$ is a (ϵ, δ)-DP algorithm, where $\boldsymbol{\eta} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbb{I}_d\right)$ consists of d independent, zero-mean Gaussian random variables with deviation $\sigma = \sqrt{2\ln(1.25/\delta)}\Delta_2 \boldsymbol{f}/\epsilon$.*

The proof is quite involved (in stark contrast to the pithy proof of the Laplacian mechanism), therefore we do not repeat it here, but refer the interested reader to the Appendix of Dwork and Roth (2014) for the full details.

Finally, we make a note on a variation of these results for matrices. When we desire to induce privacy on a function that outputs a matrix $A \in \mathbb{R}^{d \times d}$ (instead of a vector, as in Definitions 3.3.1 and 3.3.3), for example if the algorithm is responsible to perform covariance estimation given a dataset of sub-gaussian inputs, then the equivalent notion of sensitivity is the appropriate entry-wise norm corresponding to either Definition 3.3.1 or Definition 3.3.3. Specifically, for the ℓ_2 sensitivity, the appropriate entry-wise norm would be the Frobenius norm (see Section 4.2). The reason for this change is that outputting a $d \times d$ matrix is like outputting a d^2 vector (where all matrix entries are serialized) for the scope of the DP definition.

3.4 Differential Privacy Variants & Extensions

3.4.1 Zero-concentrated DP

Due to the extreme stringency of the "pure" $(\epsilon, 0)$-DP definition and conversely the caution required with the (ϵ, δ) one, much recent research has been devoted to alternative definitions that can act as relaxations of the "pure" definition, but that are stronger than approximate (ϵ, δ)-DP. One very successful such alternative is that of zero-concentrated DP. It is a slightly different definition of differential privacy that is roughly equivalent with the classical definition, according to Lemma 3.4.1 by Bun and Steinke (2016).

Definition 3.4.1 (Zero-concentrated DP (zCDP))**.** A randomized mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies ρ-zCDP if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$ that differ on at most one element, and for every $\alpha \geq 1$,

$$D_\alpha \left(M(X) || M(X') \right) \leq \rho\alpha \,,$$

where $D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \left(\mathbb{E}_{x \sim Q} \left[\left(\frac{P(x)}{Q(x)} \right)^\alpha \right] \right)$ is the α-Rényi divergence between the probability distributions P and Q.

Lemma 3.4.1 (An equivalence between zero-concentrated DP and "classical" DP)**.** *Let $M : \mathcal{X}^n \to \mathcal{Y}$ be a randomized mechanism. Then, the following results hold:*

- *If M is $\frac{\epsilon^2}{2}$-zCDP, then, for all $\delta > 0$, M is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP.*
- *If M is $(\epsilon, 0)$-DP, then M is $\frac{\epsilon^2}{2}$-zCDP.*

3.4.2 Local Differential Privacy

The question of whether the (central) DP model is the most appropriate one for every setting has been perpetual. In fact, in certain cases, it makes sense to enforce a stronger (or different altogether) model. For example, what if the user who is submitting their data to the central entity for estimation and release of a (private) statistic does not trust that this central aggregator will keep its promise for privacy? That is when another significant model, adjacent to (central) DP, arises: that of "Local Differential Privacy" (LDP). In this model, we are interested in assuming no trust at any centralized curator, and instead desire to enforce privacy on our own, rather than rely on the goodwill of another entity. We rigorously define the model below, which is stronger than (central) DP in the sense that any LDP algorithm can be transformed to provide the DP guarantee, and thus more stringent in its requirements and effect on the resulting accuracy of any private procedure.

In the LDP model, each individual manages their private data and discloses a privatized form to a server after processing in such a way as to guarantee that the server is not able to discern any one individual's data, while still retaining global properties of the data throughout the entire population collected. This is in contrast to the (central) DP model, where data is gathered by a trusted central entity which is responsible for releasing the results after some kind of DP data analysis.

Due to its usefulness in software for users' devices, LDP algorithms have become widespread lately in practical settings, where it is of utmost importance to protect users without trusting a company's servers with private user data: local differential privacy has recently been integrated in statistics' collection from users' devices (e.g., for finding out the most used feature of a software, the most used emoji, the most used words in a sentence, or even new words that have become "viral", i.e., are most frequently used), such as on Google's Chrome browser (Google, 2013), and Apple's iOS (Apple, 2016).

Definition 3.4.2 ((ϵ, δ)-LDP Local Randomizer)**.** A randomized algorithm $Q_i : \mathcal{X} \to \mathcal{Y}$ (which takes the input of only the i-th user –notice that it is for **only one** user– and is being applied on each user's private input separately in the simple *non-interactive* LDP scheme) satisfies (ϵ, δ) local differential privacy (equivalently, is said to be (ϵ, δ)-LDP) if for every pair of possible (private) inputs $x, x' \in \mathcal{X}$,

$$\mathbf{Pr}[Q_i(x) \in Y] \leq \exp(\epsilon)\, \mathbf{Pr}[Q_i(x') \in Y] + \delta\,, \forall Y \subseteq \mathcal{Y}\,.$$

Algorithm Q_i is referred to as the "local randomizer," and the interactive LDP scheme allows for each Q_i to additionally depend on the (privatized) outputs $z_1, z_2, \ldots, z_{i-1}$ of the previous local randomizers, i.e., on the values $z_1 = Q_1(x_1), z_2 = Q_2(x_2, z_1), \ldots, z_{i-1} = Q_{i-1}(x_{i-1}, z_1, z_2, \ldots, z_{i-2})$. According to that definition, for the "sequentially interactive" LDP scheme, it would be the case that $z_i = Q_i(x_i, z_1, z_2, \ldots, z_{i-1})$ and for Definition 3.4.2 we would need to introduce the additional quantifiers $\forall z_1, z_2, \ldots, z_{i-1}$.

Randomized Response. The most basic mechanism used to induce LDP is Randomized Response, which is in fact inspired by social science research. In the social sciences, in order to study social trends on illicit or otherwise tabooed behavior, the researcher's questionnaire would ask for the yes-no question in consideration to be answered in a special way: the participant would first flip a coin (secretly from the researcher), and if the result was heads, then they would respond honestly, else if the result was tails, they would respond completely at random (with no regard to what their true answer would be). In this way, the participant has the option of "plausible deniability," i.e., they can easily claim that a "yes" answer (that would constitute illicit or otherwise tabooed behavior) was the result of the randomness from the coin flips and not in fact their true response. Therefore, the person interested in knowing any specific participant's response has no way of knowing whether the answer is real or not, and the individual is thus protected. Even though it might seem to one that this technique would deteriorate the usefulness of the aggregated statistic, this is not true, and there is a way to counter-balance the introduced randomness in aggregate and obtain a statistically accurate estimator. A rigorous description of the Randomized Response mechanism follows and for further results we refer the reader to Dwork and Roth (2014).

Consider one user's data x derived from a finite underlying domain $\mathcal{X}$ with size $|\mathcal{X}|$. The user would provide the (privatized) output z as sampled at random from the below probability distribution, where α is the desired privacy parameter for the mechanism to be $(\alpha, 0)$-LDP:

$$\forall y \in \mathcal{X} : \mathbf{Pr}\left[z = y\right] = \begin{cases} \frac{e^\alpha}{e^\alpha + |\mathcal{X}| - 1}, & \text{if } y = x \\ \frac{1}{e^\alpha + |\mathcal{X}| - 1}, & \text{else} \end{cases}.$$

That is, with a larger probability (the upper of the above cases, which we name p) the user will output $z = x$ (their true value), but with a probability (the lower of the above cases, which we name q) dependent on the privacy α desired, they will lie about it uniformly at random. When we have gathered (privatized) data $\{z_i\}_{1 \leq i \leq n}$ from n users, frequency estimation for the item $v \in \mathcal{X}$ from the underlying domain $\mathcal{X}$ may be done in the following way to re-adjust for the random noise introduced by the privatizing process:

$$\hat{f}_v = \frac{\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{z_i = v\} - q}{p - q}\,.$$

Along with LDP Frequency Oracles (which provide a LDP way to extract approximate frequencies of known and unknown items that are used by a set of users, e.g., emojis or words that are used by smartphone users), for a survey of which we refer to Dwork and Roth (2014), these two techniques are the most commonly used techniques to construct LDP algorithms.

Chapter 4

Mathematical Background

4.1 Notation

We use bold fonts for (column) vectors (e.g., $\boldsymbol{\beta}, \boldsymbol{y}$) and denote the set $\{1, \ldots, n\}$ as $[n]$. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we denote its ℓ_2 norm as $\|\boldsymbol{x}\|_2$. The indicator function of an event E is denoted as $\mathbb{1}\{E\}$, which is one when the event is realized, else zero.

When $\boldsymbol{X}_i \in \mathbb{R}^d$ for $i \in [n]$ are the (random) feature vectors and $y_i \in \mathbb{R}$ for $i \in [n]$ are the (random) labels of a regression setting (see Chapter 5 for the details), the matrix $X = [\boldsymbol{X}_1 \ \boldsymbol{X}_2 \ \ldots \ \boldsymbol{X}_n]^T \in \mathbb{R}^{n \times d}$ is called the (random) design matrix and the vector $\boldsymbol{y} = [y_1 \ y_2 \ \ldots \ y_n]^T \in \mathbb{R}^n$ is called the (random) response vector.

4.2 Matrix norms and PSD matrices

This section is going to be about reviewing some fundamental matrix theory definitions, properties and facts that we will use throughout our proofs and techniques in the following chapters. We begin with a characterization of positive semi-definite (PSD) matrices.

Definition 4.2.1. A symmetric matrix $A \in \mathbb{R}^{d \times d}$ is PSD, if any of the following equivalent conditions hold:

- For every vector $\boldsymbol{v} \in \mathbb{R}^d : \boldsymbol{v}^T A \boldsymbol{v} \geq 0$.
- All the eigenvalues of A are non-negative.
- All the principal minors of A are non-negative.
- There exists a matrix M such that $A = M^T M$.

Loewner order. The Loewner order (with symbols $\preceq$ and $\succeq$) is defined as the partial order induced by the convex cone of PSD matrices. In particular, we say that $A \preceq B$ for two symmetric matrices $A, B \in \mathbb{R}^{d \times d}$ if and only if $B - A$ is a PSD matrix. An interesting special case that we will utilize extensively in our proofs is that $A \preceq \kappa \mathbb{I}_d$ holds if and only if the largest eigenvalue of A is smaller than κ, i.e., $\lambda_{\max}(A) \leq \kappa$.

Matrix norms. For a matrix $A \in \mathbb{R}^{m \times n}$, we define its spectral norm as

$$\|A\|_2 = \max_{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 = 1} \|A\boldsymbol{x}\|_2 \,,$$

and its Frobenius norm as $\|A\|_F = \sqrt{\sum_{i \in [m], j \in [n]} a_{ij}^2}$, where a_{ij} is the entry on the i-th row and j-th column of matrix A. The two norms are roughly interchangeable, since it is true for any square matrix $A \in \mathbb{R}^{d \times d}$ that

$$\sqrt{d}\|A\|_F \leq \|A\|_2 \leq \|A\|_F \,.$$

Singular values. We continue by providing another characterization of the spectral norm as the largest singular value through the Singular Value Decomposition (SVD). SVD states that, for every matrix $A \in \mathbb{R}^{m\times n}$, there exist unitary (in fact, real orthogonal, since A is real) matrices $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$ such that $A = U\Sigma V^T$ for some rectangular diagonal matrix $\Sigma \in \mathbb{R}^{m\times n}$ whose diagonal entries are named as the "singular values" of A. More specifically, by standard linear algebra, it is true that the singular values of a square matrix $A \in \mathbb{R}^{d\times d}$ are the square roots of the eigenvalues of AA^T (which are the same as the square roots of the eigenvalues of A^TA). In this case, using the Cauchy-Schwartz inequality, it can be proven that the spectral norm of A is equal to its largest singular value, i.e.,

$$\|A\|_2 = \sigma_{\max}(A)\,.$$

Additionally, if matrix A is PSD, then its singular values are the same as its eigenvalues (since for PSD matrices A it holds that the eigenvalues of AA^T are the squares of the eigenvalues of A), hence $\|A\|_2 = \lambda_{\max}(A)$ when A is PSD.

Condition number. Finally, we define the condition number of a PSD, real symmetric matrix $A \in \mathbb{R}^{d\times d}$ as the ratio of its largest singular value to its smallest singular value, i.e.,

$$\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \geq 1\,.$$

Intuitively, the condition number of a matrix determines how "close" that matrix is to being non-invertible; more specifically, how much error is introduced when we compute the inverse of the said matrix. A singular/non-invertible matrix has an infinite condition number. The condition number also uncovers the error that will be made in trying to solve the linear system $A\boldsymbol{x} = \boldsymbol{b}$ when either an error in A or $\boldsymbol{b}$ exists (this error will roughly be multiplied by the condition number $\kappa(A)$ in the solution $\boldsymbol{x}$ of the system). If the condition number is large, we say that the problem $A\boldsymbol{x} = \boldsymbol{b}$ is "poorly-conditioned" and large errors will arise when approximately computing its solution.

4.3 Sub-gaussian random variables and vectors

In probability theory and its applications, one extremely useful and wide class of distributions is that of *sub-gaussian* distributions. This class contains many pivotal distributions, such as Gaussian, Bernoulli and *all* bounded distributions. Its usefulness arises mainly due to a Hoeffding-like (concentration) result that applies –almost by definition– to random variables that follow a sub-gaussian distribution.

We begin with definitions of the corresponding sub-gaussian distributions and the "sub-gaussian norm" for univariate and multivariate random variables, and move on to some of their properties that we shall use later on in our detailed proofs.

Definition 4.3.1 (Sub-gaussian random variable)**.** A random variable X is called a sub-gaussian random variable if there exists $K > 0$ such that, for all $\lambda : |\lambda| \leq 1/K$,

$$\mathbb{E}\left[\exp\left(\lambda^2 X^2\right)\right] \leq \exp(\lambda^2 K^2)\,.$$

The smallest K for which the above property holds is called the *sub-gaussian norm* of X, and is denoted as $\|X\|_{\psi_2}$.

Definition 4.3.2 (Sub-gaussian random vector)**.** A random vector $\boldsymbol{X} \in \mathbb{R}^d$ is called a sub-gaussian random vector if for all $\boldsymbol{u} \in \mathbb{R}^d$, the inner product $\langle \boldsymbol{X}, \boldsymbol{u}\rangle$ is a sub-gaussian random variable. The sub-gaussian norm of a sub-gaussian random vector is defined as follows:

$$\|\boldsymbol{X}\|_{\psi_2} = \sup_{\boldsymbol{u}\in S^{d-1}} \|\langle \boldsymbol{X}, \boldsymbol{u}\rangle\|_{\psi_2}\,,$$

where $S^{d-1} = \{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_2 = 1\}$ is the d-dimensional unit sphere.

There is a wealth of resources on sub-gaussian random variables and a useful generalization of them: sub-exponential random variables, which generalize the MGF (moment generating function) bound of Definition 4.3.1 to an exponential function of $|X|$ (instead of X^2) and for which most concentration inequalities continue to hold with appropriate modifications. A significant fact is that X is sub-gaussian if and only if X^2 is sub-exponential. Many interesting results have been aggregated by Vershynin (2018); Rivasplata (2012); Hsu et al. (2012), to which we refer the interested reader for further thorough treatment of mentioned lemmas hereby.

We continue with some examples of sub-gaussian distributions, along with properties of the sub-gaussian norm.

Examples of sub-gaussian distributions.

- Every centered Gaussian random variable $X \sim \mathcal{N}(0, \sigma^2)$ is sub-gaussian with sub-gaussian norm $\|X\|_{\psi_2} \leq C\sigma$, for some universal constant $C > 0$.
- All bounded random variables X such that $|X| \leq \eta$ are sub-gaussian with sub-gaussian norm $\|X\|_{\psi_2} \leq \frac{1}{\sqrt{\ln 2}} \cdot \eta$.
- Every Rademacher variable (symmetric Bernoulli variable that takes the values $\{\pm 1\}$ with probability $1/2$ at random) is sub-gaussian with sub-gaussian norm $\|X\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}$.

Lemma 4.3.1 (Properties of the sub-gaussian norm)**.** *Let $\boldsymbol{X}$ be a sub-gaussian random vector. Then, the following hold:*

- *For every constant $c > 0$, $c\boldsymbol{X}$ is a sub-gaussian random vector, with $\|c\boldsymbol{X}\|_{\psi_2} = c\,\|\boldsymbol{X}\|_{\psi_2}$.*
- *If $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu_X}$, then $\boldsymbol{X} - \boldsymbol{\mu_X}$ is a sub-gaussian random vector, with*

$$\|\boldsymbol{X} - \boldsymbol{\mu_X}\|_{\psi_2} \leq C\,\|\boldsymbol{X}\|_{\psi_2}\ ,$$

  for a universal constant $C > 0$.

The second property in Lemma 4.3.1 is sometimes called the "centering" property and gives us a way to reason about the centered counterparts of non-centered ($\boldsymbol{\mu_X} \neq \boldsymbol{0}$) random variables.

4.4 Concentration bounds due to sub-gaussianity

In this section, we will discuss some very useful concentration bounds on empirical quantities arising from sub-gaussian random variables and vectors. We begin by stating the General Hoeffding's Inequality for sub-gaussian random variables and continue with further results on various empirical quantities that concentrate "rapidly" to their expected values, due to the sub-gaussianity of the underlying distributions.

Theorem 4.4.1 (General Hoeffding's Inequality (Vershynin, 2018))**.** *Let $X_1, X_2, \ldots, X_n$ be independent sub-gaussian random variables with mean zero. Then, there exists a constant $c > 0$, for every $t > 0$, such that*

$$\mathbf{Pr}\left[\left|\sum_{i=1}^{n} X_i\right| > t\right] \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^{n}\|X_i\|_{\psi_2}^2}\right)\ .$$

Theorem 4.4.2 (Khintchine's inequality (Vershynin, 2018)). *Let $X_1, X_2, \ldots, X_n$ be independent sub-gaussian random variables with mean zero and unit variance. Also, consider a weight vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$. Then, there exists a constant $c > 0$, such that for every $p \geq 2$ it holds that*

$$\|\boldsymbol{a}\|_2 \leq \left(\mathbb{E}\left[\left| \sum_{i=1}^{n} a_i X_i \right|^p \right] \right)^{1/p} \leq c\sqrt{p} \|\boldsymbol{a}\|_2 \max_i \|X_i\|_{\psi_2} .$$

Theorem 4.4.3 (Sub-gaussian vector concentration bound: see, e.g., Lemma 2.21 of Diakonikolas et al. (2019a)). *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be i.i.d. sub-gaussian random vectors with mean $\mathbf{0}$ and covariance matrix $\mathbb{I}_d$. Then, there exist universal constants $A, B > 0$ such that, for every $t > 0$,*

$$\mathbf{Pr}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \right\|_2 > t \right] \leq 4 \exp\left(Ad - Bnt^2 \right) .$$

Theorem 4.4.4 (Empirical covariance estimation concentration bound: see, e.g., Lemma 2.22 of Diakonikolas et al. (2019a)). *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be i.i.d. sub-gaussian random vectors with mean $\mathbf{0}$ and covariance matrix $\mathbb{I}_d$. Then, there exist universal constants $A, B > 0$ such that, for every $t > 0$,*

$$\mathbf{Pr}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T - \mathbb{I}_d \right\|_2 > t \right] \leq 4 \exp\left(Ad - Bn \min(t, t^2) \right) .$$

Chapter 5

Regression Settings

In this chapter, we formally define the regression settings we are interested in, namely, the Least Squares Fitting, Linear Regression and the Binary Regression problems, as well as the associated technical assumptions we make.

5.1 Linear Regression

One widely studied regression model in the Differential Privacy literature (see, e.g., Amin et al. (2019); Sheffet (2017)) is the so-called "simple linear regression" model. In this model, we observe labeled examples $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, where the labels y_i are assumed to be generated as a deviation (with a Gaussian error) from an underlying "true" linear model $\boldsymbol{\beta}^T \boldsymbol{X}_i$, for some regression coefficient $\boldsymbol{\beta} \in \mathbb{R}^d$. Our goal is to estimate this "true" underlying $\boldsymbol{\beta}$ in a differentially private way. In particular, the full set of assumptions that we use follows:

Assumption 5.1.1. *Labeled examples $(\boldsymbol{X}_i, y_i)$, $i = 1, \ldots, n$, are i.i.d. Moreover, $\boldsymbol{X}_i \in \mathbb{R}^d$ are sampled from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ satisfying $\mathbb{I}_d \preceq \Sigma \preceq \kappa \mathbb{I}_d$ for some universal parameter $\kappa > 0$. The generative model of the labels y_i is as follows: there exists a $\boldsymbol{\beta} \in \mathbb{R}^d$ such that, given $\boldsymbol{X}_i \in \mathbb{R}^d$,*

$$y_i = \boldsymbol{\beta}^T \boldsymbol{X}_i + \epsilon_i, \quad \text{for all } i = 1, \ldots, n\,,$$

where ϵ_i are i.i.d. samples from a zero-mean Gaussian distribution $\mathcal{N}(0, \sigma_\epsilon^2)$.

5.2 Least Squares Fitting

In the Least Squares Fitting problem, we observe labeled examples $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, and wish to produce an (ϵ, δ)-differentially private version of the Least Squares Estimator (LSE):

$$\boldsymbol{\beta}^\star = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^d} \sum_{i=1}^n \left(y_i - \boldsymbol{\beta}^T \boldsymbol{X}_i\right)^2 = \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^T\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n y_i \boldsymbol{X}_i\right) = \left(\frac{1}{n} X^T X\right)^{-1} \frac{1}{n} X^T \boldsymbol{y}, \tag{5.1}$$

where $X = [\boldsymbol{X}_i]_{i=1}^n \in \mathbb{R}^{n \times d}$ is the matrix with feature vectors as rows and $\boldsymbol{y} = [y_i]_{i=1}^n \in \mathbb{R}^d$ is the vector of labels, respectively. We note that, in contrast to the Binary Regression problem below, we make no prior assumption on how the bounded labels y_i are linked to features $\boldsymbol{X}_i$; crucially, our differentially private algorithm *must not rely* on any presumed boundedness of features $\boldsymbol{X}_i$.

In particular, we make the following technical assumption:

Assumption 5.2.1. *Labeled examples $(\boldsymbol{X}_i, y_i)$, $i = 1, \ldots, n$, are i.i.d. Moreover, $\boldsymbol{X}_i \in \mathbb{R}^d$ are sampled from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ satisfying the following conditions:*

$$\|\boldsymbol{\mu}\|_2 \leq R \quad \text{and} \quad \mathbb{I}_d \preceq \Sigma \preceq \kappa \mathbb{I}_d\,, \tag{5.2}$$

while the labels satisfy $\frac{1}{\rho} \leq |y_i| \leq c$ for some universal parameters $\rho, c, \kappa, R > 0$.

The boundedness assumptions in Equation (5.2) are precisely the assumptions made by Kamath et al. (2019) in the context of Gaussian estimation. As discussed in Section 6.1, both upper bounds are natural, while the lower bound on the covariance comes without any loss of generality. Crucially, in contrast to the majority of prior works on regression, samples $\boldsymbol{X}_i$ are indeed unbounded, as they are sampled from $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Finally, the boundedness of the outputs y_i, $i \in [n]$, is a requirement we share with other works (e.g., Alabi et al. (2020); Wang (2018); Kifer et al. (2012); Zhang et al. (2012)), and is motivated from our focus on classification for this part, rather than standard linear regression, which is analyzed on the scope of the previous setting of Section 5.1.

5.3 Binary Regression

In the Binary Regression setting, in addition to the assumptions from the generic setting of Section 5.2, we assume that the labels y_i are binary (i.e., $y_i \in \{-1, +1\}$), and are produced by a Generalized Linear Model (GLM) linking these binary labels to features. In contrast to the previous setting, this GLM is parameterized by a "true" $\boldsymbol{\beta} \in \mathbb{R}^d$ (see Assumption 5.3.1 below). Our goal is to give an estimate of this $\boldsymbol{\beta}$ again via *the same* (ϵ, δ)-differentially private version of the LSE given by Equation (5.1).

In particular, in addition to Assumption 5.2.1, we make the following assumption in the Binary Regression setting:

Assumption 5.3.1. *There exists a $\boldsymbol{\beta} \in \mathbb{R}^d$ such that, given $\boldsymbol{X}_i \in \mathbb{R}^d$,*

$$\mathbf{Pr}[y_i = +1 | \boldsymbol{X}_i] = f(\boldsymbol{\beta}^T \boldsymbol{X}_i), \quad \text{for all } i = 1, \ldots, n, \tag{5.1}$$

where $f : \mathbb{R} \to [0, 1]$ is a non-decreasing, continuously differentiable function satisfying

$$\lim_{x \to -\infty} f(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} f(x) = 1\,. \tag{5.2}$$

Moreover, the features $\boldsymbol{X}_i$ are zero-mean, i.e., $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_i] = \mathbf{0}$.

The probabilistic model defined by Equation (5.1) and Equation (5.2) holds for many important practical settings. For instance, it holds for logistic regression, where the link function is $f(x) = 1/(1 + e^{-x})$. It also holds for Support Vector Machines (SVMs) with linearly separable data. For a discussion on these reductions, see Section 5.3.1.

Finally, our assumption that $\boldsymbol{\mu} = \mathbf{0}$ is common (see, e.g., Kulkarni et al. (2021); Cai et al. (2020); Daskalakis et al. (2020); Bernstein and Sheldon (2019); Sheffet (2017); Erdogdu (2016)) and well-motivated in the context of our Binary Regression setting: even ignoring privacy considerations, the sample complexity guarantees of any estimator will degrade rapidly as $\boldsymbol{\mu}$ gets farther away from the origin. This is precisely because, under Gaussian covariates, the fraction of samples of one class will decrease exponentially as the distance of $\boldsymbol{\mu}$ from the separating hyperplane (that passes through the origin) increases.

5.3.1 Logistic Regression and SVMs as models satisfying Assumption 5.3.1

We will show here how the models of Logistic Regression and linearly-separable SVMs fit into our probabilistic model of Binary Regression (see Assumption 5.3.1). For the Logistic Regression model, applying the model function $f(x) = 1/(1 + e^{-x})$ directly yields the conditional probabilistic model of $\mathbf{Pr}[Y = 1 | \boldsymbol{X}] = \frac{1}{1+e^{-\boldsymbol{\beta}^T \boldsymbol{X}}}$ for the regression coefficients $\boldsymbol{\beta}$, which is precisely the desired Logistic Regression model.

In the second case, we consider linearly-separable SVMs, where the data $(\boldsymbol{X}_i, y_i)$ are completely separated by an underlying hyperplane that we are trying to uncover. That is,

the model function would be $\operatorname{sgn}(\boldsymbol{\beta}^T \boldsymbol{X})$. However, such a function is neither smooth nor continuously differentiable near the origin, therefore we will apply the following trick: after receiving the perfectly linearly separable data, we will induce a minuscule amount of noise through a noisy model function f that is continuously differentiable and smooth everywhere (but crucially, near the origin). Intuitively, we smooth out the sign function. We can make infinitely good approximations of the sign function, and therefore passing the data through one of those before we apply our algorithm is sufficient to recover the true underlying $\boldsymbol{\beta}$ up to a scaling factor depending on the noise that we artificially introduced.

5.3.2 Stein's Lemma and Generalized Linear Models

We state a multivariate version of Stein's Lemma (Stein, 1981) due to Liu (1994).

Lemma 5.3.1 (Stein's Lemma (Liu, 1994))**.** *Let $\boldsymbol{Z} \in \mathbb{R}^p, \boldsymbol{W} \in \mathbb{R}^q$ be jointly Gaussian random vectors and let $f : \mathbb{R}^q \to \mathbb{R}$ be differentiable almost everywhere with $\mathbb{E}_{\boldsymbol{W}}\left[|\partial f(\boldsymbol{W})/\partial W_i|\right] < \infty$ for any $i \in [q]$. Then,* $\mathbf{Cov}[\boldsymbol{Z}, f(\boldsymbol{W})] = \mathbf{Cov}[\boldsymbol{Z}, \boldsymbol{W}]\, \mathbb{E}[\nabla f(\boldsymbol{W})]$.

The lemma has a direct application on Generalized Linear Models with Gaussian covariates (Erdogdu, 2016; Brillinger, 2012a,b): for the GLM of Brillinger (2012a) with Gaussian covariates, whose model function satisfies the conditions of Lemma 5.3.1, one can show (Brillinger, 2012a,b) that the Ordinary Least Squares Estimator asymptotically converges to the true parameter vector $\boldsymbol{\beta}$ of the GLM with probability 1, up to a scaling factor k. This is analogous to the scaling factor k that appears later in our analysis (see Theorem 8.1.2).

Chapter 6

Private Mean and Covariance Estimation of Sub-Gaussian Random Vectors

6.1 Introduction to DP Gaussian Parameter Estimation

At a technical level, our work depends on (and extends) the tools developed by Kamath et al. (2019) to privately estimate the mean $\boldsymbol{\mu}$ and covariance Σ of a d-dimensional Gaussian distribution. The algorithm they propose, which we call LearnGaussian-hd, satisfies the following guarantee:

Theorem 6.1.1 (Multivariate Gaussian Estimation (Kamath et al., 2019))**.** *There exists a polynomial time* $(\epsilon^2/2 + \epsilon\sqrt{2\log(1/\delta)}, \delta)$*-DP algorithm* LearnGaussian-hd *that takes at least*

$$n = \widetilde{O}\left(\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\epsilon} + \frac{d^{3/2}\log^{1/2}(\kappa) + d^{1/2}\log^{1/2}(R)}{\epsilon}\right)$$

i.i.d. samples $\boldsymbol{X}_i$*,* $i \in [n]$*, from a* d*-dimensional Gaussian* $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ *with unknown mean* $\boldsymbol{\mu} \in \mathbb{R}^d$ *and unknown covariance* $\Sigma \in \mathbb{R}^{d\times d}$ *satisfying* $\|\boldsymbol{\mu}\|_2 \leq R$ *and* $\mathbb{I}_d \preceq \Sigma \preceq \kappa\mathbb{I}_d$*, and outputs estimates* $\widehat{\boldsymbol{\mu}}, \widehat{\Sigma}$ *such that, with high probability,* $\mathrm{TV}(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}(\widehat{\boldsymbol{\mu}}, \widehat{\Sigma})) \leq \alpha$*.*

In short, LearnGaussian-hd produces differentially private estimates of the distribution's parameters using only $\widetilde{O}(d^2)$ samples. It operates under the following boundedness assumptions for the distributional parameters:

$$\|\boldsymbol{\mu}\|_2 \leq R \quad \text{and} \quad \mathbb{I}_d \preceq \Sigma \preceq \kappa\mathbb{I}_d,$$

even though, crucially, *the samples* $\boldsymbol{X}_i$ *themselves are unbounded*. Moreover, both upper bounds (R and κ) are mild and well-motivated: even if LearnGaussian-hd is applied to a sequence of datasets where these grow sub-exponentially, the sample complexity remains polynomial. Additionally, notice that $\mathbb{I}_d \preceq \Sigma$ comes w.l.o.g.: as long as the smallest eigenvalue of Σ is non-zero, we can rescale the vectors $\boldsymbol{X}_i$ to ensure that this holds. If, on the other hand, an eigenvalue of Σ is zero, then the distribution is degenerate: we can then apply LearnGaussian-hd in the subspace spanned by the features (in which Σ will have full rank).

The main technical contribution of Kamath et al. (2019) is the recursive private preconditioning technique. This efficiently learns a symmetric matrix A, termed the *preconditioner* of the Gaussian distribution, that satisfies $\mathbb{I}_d \preceq A\Sigma A \preceq O(1)\mathbb{I}_d$. Multiplying the input samples with this preconditioner thus makes the Gaussian nearly spherical; this transformation allows Kamath et al. (2019) to efficiently reduce the estimation of $\boldsymbol{\mu}$ to the one-dimensional setting, previously studied by Karwa and Vadhan (2018), while A itself can be used to estimate Σ.

The original algorithm by Kamath et al. (2019) requires knowledge of both upper bounds κ and R. By switching the privacy guarantee from zero-concentrated DP to (ϵ, δ)-DP, we remove the requirement of prior knowledge of R by the algorithm, even though we still require κ as input. A more detailed description of this variant of LearnGaussian-hd and our adaptation of the guarantees by Kamath et al. (2019) can be found in Section 6.2.

6.2 Overview of Kamath et al. (2019) and Extension to Sub-Gaussian Regime

We first review the covariance estimation result of Kamath et al. (2019) (slightly modified), and present a generalization of their results to the sub-Gaussian case, whereupon we provide a discussion.

6.2.1 Equivalence of Privacy Guarantees of Kamath et al. (2019) to Classical DP

First of all, we note that the privacy guarantees of Kamath et al. (2019) hold for a variation of the Differential Privacy definition that is mentioned in Definition 3.4.1; specifically, $\frac{\epsilon^2}{2}$ "zero-concentrated DP." This variation is more lax than the classical (pure) ϵ-DP, but stronger than the (ϵ, δ)-DP that is commonly used in the privacy literature, as can be seen immediately from Lemma 3.4.1.

In our case, we prefer to keep the results of Theorem 8.1.1 and Theorem 8.1.2 in the classical (ϵ, δ)-DP definition, and therefore the respective algorithms are $(\frac{\epsilon^2}{2}+\epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP. In essence, this guarantee is equivalent to (ϵ, δ)-DP, but this equivalence is worse as ϵ gets larger, i.e., in the case that little privacy is desired. Additionally, the fact that $\delta > 0$ allows us to bypass the requirement of an upper bound on $\|\boldsymbol{\mu}\|_2$ on our variation, as will be analyzed in Lemma 6.2.2.

6.2.2 Algorithm Overview

Here, we provide a high-level description of the algorithm LearnSubGaussian-hd (used in Algorithm 2) that learns the mean and covariance matrix of a high-dimensional Gaussian distribution, which differs slightly from the LearnGaussian-hd considered by Kamath et al. (2019).

The building block of the covariance estimation algorithm is the NaivePCE algorithm (Algorithm 1 of Kamath et al. (2019)), which intuitively would be the first try at inducing privacy in the covariance estimation procedure. More specifically, it truncates the input samples, adds a random Gaussian matrix to the empirical covariance that arises from these samples, and outputs the projection of the final matrix to the PSD cone. However, this naive "first try" algorithm exhibits a linear dependence of the accuracy to the largest eigenvalue κ of the covariance matrix Σ (intuitively, the largest variance across any direction), whereas we aim for a $\log\kappa$ dependence. Therefore, noticing that the accuracy dependence is optimal when the aforementioned largest eigenvalue is of constant order, we seek to transform the samples $\boldsymbol{X}_i$ to $A\boldsymbol{X}_i$ such that the largest eigenvalue of the covariance matrix of $A\boldsymbol{X}_i$ (which is $A\Sigma A$ for symmetric matrices A) satisfies the above condition.

The covariance estimation algorithm, thus, begins by efficiently finding such a matrix A (the "preconditioner") according to the PPC algorithm (Algorithm 3 of Kamath et al. (2019)) which, in short, does the following: it uses $O(\log\kappa)$ successive rounds of the NaivePCE algorithm such that every round "eliminates" the eigendirections of largest variance (through an eigenvector decomposition and keeping intact for the next rounds only the eigenvalues that are smaller than half the current upper bound) hence transforming each successive κ_j (for $1 \le j \le O(\log\kappa)$ the number of the current round) to $\kappa_{j+1} = 0.7\kappa_j$. After $O(\log\kappa)$ rounds, the final largest eigenvalue of $A\Sigma A$ will be of constant order, as desired. After this procedure which finds A, NaivePCE is run on the samples $A\boldsymbol{X}_i$ with a result of $\widetilde{\Sigma}$, and the covariance estimation algorithm finally outputs $\widehat{\Sigma} = A^{-1}\widetilde{\Sigma}A^{-1}$. For details on those algorithms, we refer the interested reader to Kamath et al. (2019). Note that for these steps, the knowledge of κ, the upper bound on the largest eigenvalue of the covariance matrix Σ of the initial samples $\boldsymbol{X}_i$ is necessary for the calibrated truncation and noise addition to occur correctly.

Had someone wanted to also estimate the mean, they would first get a matrix A as above through $2n$ samples $\frac{1}{\sqrt{2}}\left(\boldsymbol{X}_{2i}-\boldsymbol{X}_{2i-1}\right)$, $1\leq i\leq n$ that are i.i.d. with the same covariance matrix Σ, and then draw n additional i.i.d. samples $\boldsymbol{X}_i$ (for a total of $3n$ samples) and apply the univariate mean estimation algorithm of Karwa and Vadhan (2018) to each coordinate of $A\boldsymbol{X}_i$ separately. Our algorithm's difference with Kamath et al. (2019) is that we call the algorithm of Karwa and Vadhan (2018) with $R=\infty$, which is allowable and efficient to do in our setting due to the privacy guarantee having $\delta>0$ (see the guarantees on Section 6.2.3). Once we are on a univariate sub-gaussian setting, and since we have a constant-order upper bound on the variance $\sigma^2=O(1)$ of $\left(A\boldsymbol{X}_i\right)_j$, which denotes the j-th coordinate of the random vector $A\boldsymbol{X}_i$, the algorithm for univariate mean estimation works as follows: First, we find a differentially private estimation of an upper bound B on the data with high probability in the following way: we split the whole range that the mean might be located $(-\infty,\infty)$ to bins of width $\sigma=O(1)$, and taking advantage of the concentration of sub-gaussian random variables around their mean, we use a differentially private histogram algorithm (Bun et al., 2016) to locate the most frequent bin, which (along with its neighboring bins) should contain all data points with high probability. Second, we truncate the input data $\left(A\boldsymbol{X}_i\right)_j$ to a range calculated according to the above estimated bound, such that all input samples fall within that range with high probability, and then add Laplacian noise (calibrated according to the differentially-private calculated bound B) to the empirical mean of the input samples. We output as the result of the univariate mean estimation algorithm this noisy empirical mean of the (truncated) input samples. For the full sub-algorithms that we discussed shortly here, we refer the interested reader to Algorithm 1 and Algorithm 4 of Karwa and Vadhan (2018) respectively.

Assuming the generic description of the algorithms above, we show why the proofs can be extended to the sub-gaussian case below. For the complete proofs that we modify, in light of conciseness in this appendix, we refer to the works of Kamath et al. (2019) and Karwa and Vadhan (2018).

6.2.3 Extension of the Guarantees of Kamath et al. (2019) and Karwa and Vadhan (2018) to the Sub-Gaussian Regime

The modified algorithms that we presented in Section 6.2.2 have the following guarantees, whereupon we will provide a proof sketch.

Lemma 6.2.1 (Private Covariance Estimation, variant of Kamath et al. (2019))**.** *For every $\epsilon,\delta,\gamma,\kappa,\alpha>0$, there exists an $(\frac{\epsilon^2}{2}+\epsilon\sqrt{2\log(1/\delta)},\delta)$-DP algorithm that, when given n i.i.d. samples $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ from a sub-gaussian multivariate distribution with mean $\mathbb{E}[\boldsymbol{X}_i]=\mathbf{0}$ and covariance matrix $\mathbb{E}[\boldsymbol{X}_i\boldsymbol{X}_i^T]=\Sigma$ with $\mathbb{I}_d\preceq\Sigma\preceq\kappa\mathbb{I}_d$ and*

$$n=O\left(\frac{d+\log(1/\gamma)}{\alpha^2}+\frac{d^{3/2}\mathrm{polylog}\left(\frac{d}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon}+\frac{d^{3/2}\sqrt{\log\kappa}\mathrm{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right),$$

outputs $\widehat{\Sigma}$ such that $\left\|\Sigma^{-1/2}\left(\widehat{\Sigma}-\Sigma\right)\Sigma^{-1/2}\right\|_2\leq O(\alpha)$ with probability $1-O(\gamma)$.

Lemma 6.2.2 (Private Mean Estimation, variant of Karwa and Vadhan (2018); Kamath et al. (2019))**.** *For every $\epsilon,\delta,\gamma,\kappa,\alpha>0$, there exists an $(\frac{\epsilon^2}{2}+\epsilon\sqrt{2\log(1/\delta)},\delta)$-DP algorithm that, when given n i.i.d. samples $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ from a sub-gaussian multivariate distribution with mean $\mathbb{E}[\boldsymbol{X}_i]=\boldsymbol{\mu}$ and covariance matrix $\mathbb{E}[\boldsymbol{X}_i\boldsymbol{X}_i^T]=\Sigma$ with $\mathbb{I}_d\preceq\Sigma\preceq\kappa\mathbb{I}_d$ and*

$$n=O\left(\frac{d\log(\frac{d}{\gamma})}{\alpha^2}+\frac{d\mathrm{polylog}(\frac{d\log(1/\delta)}{\alpha\gamma\epsilon})}{\alpha\epsilon}+\frac{\sqrt{d}\log(\frac{d}{\gamma\delta})}{\epsilon}+\frac{d^{3/2}\sqrt{\log\kappa}\mathrm{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right),$$

outputs a (symmetric) matrix A and a vector $\widehat{\boldsymbol{\mu}}$ such that $\mathbb{I}_d \preceq A\Sigma A \preceq 1000\mathbb{I}_d$ and $\|A(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|_2 \leq \alpha$ with probability $1 - O(\gamma)$.

First of all, the respective algorithms adumbrated in Section 6.2.2 hold for the case of sub-gaussian input random vectors too, because the concentration bounds that are utilized readily generalize to the sub-gaussian case. We provide here the variants of the concentration bounds that are needed for the algorithms of Kamath et al. (2019) and Karwa and Vadhan (2018), and, since their form is the same as the ones used in the respective proofs of the aforementioned works, we then show how the second modification with respect to the consideration of $R = \infty$ (see Section 6.2.2 for this modification) alters the guarantees provided.

By Diakonikolas et al. (2019a), we have the following generalizations of concentration bounds in the sub-gaussian regime:

Lemma 6.2.3. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be n i.i.d. samples from a sub-gaussian multivariate distribution with mean $\mathbb{E}[\boldsymbol{X}_i] = \mathbf{0}$ and covariance matrix $\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T] = \Sigma$. Then, with probability $1 - O(\gamma)$, it holds that*

$$\left\|\Sigma^{-1/2}\boldsymbol{X}_i\right\|_2^2 \leq d\log(n/\gamma),\ \forall i \in [n]\,.$$

Lemma 6.2.4 (Sub-gaussian covariance matrix estimation)**.** *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be n i.i.d. samples from a sub-gaussian multivariate distribution with mean $\mathbb{E}[\boldsymbol{X}_i] = \mathbf{0}$ and covariance matrix $\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T] = \Sigma$. Define $\boldsymbol{Z}_i = \Sigma^{-1/2}\boldsymbol{X}_i$ with covariance matrix $E[\boldsymbol{Z}_i \boldsymbol{Z}_i^T] = \mathbb{I}_d$. Then, with probability $1 - O(\gamma)$, all the following hold:*

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T - \mathbb{I}_d\right\|_2 \leq O\left(\sqrt{\frac{d+\log(1/\gamma)}{n}}\right)$$

$$\left(1 - O\left(\sqrt{\frac{d+\log(1/\gamma)}{n}}\right)\right)\cdot\mathbb{I}_d \preceq \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T \preceq \left(1 + O\left(\sqrt{\frac{d+\log(1/\gamma)}{n}}\right)\right)\cdot\mathbb{I}_d$$

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T - \mathbb{I}_d\right\|_F \leq O\left(\sqrt{\frac{d^2+\log(1/\gamma)}{n}}\right)$$

where $\|A\|_F$ is the Frobenius norm of a matrix A, defined as the square root of the sum of the squares of each of its entries.

Note that the necessary bounds for the univariate case are obtained simply by setting $d = 1$ in the above lemmata. These are required for adapting the proofs of Karwa and Vadhan (2018) to the sub-gaussian regime.

Additionally, we remark that we do not need to modify the concentration bound arising from the Hanson-Wright inequality for bounding the norm of the noise matrix that is added according to NaivePCE, since, even in the case that the inputs are sub-gaussian, the added noise is purely Gaussian.

Proof Sketch. We now provide a proof sketch for the modification of the proof on Kamath et al. (2019). In this sketch, we will use their notation to move forward with our slight variation of the lemmata.

First, we provide an alternative lemma from Karwa and Vadhan (2018) that removes the dependency on a prior bound for $\|\boldsymbol{\mu}\|_2$ when the DP guarantee that we desire to achieve is $\delta > 0$.

Lemma 6.2.5 (Variant of Karwa and Vadhan (2018))**.** *For every $\epsilon, \delta, \gamma, \kappa, \alpha > 0$, there exists an (ϵ, δ)-DP algorithm that, when given n i.i.d. samples $X_1, \ldots, X_n$ from a sub-gaussian*

univariate distribution with mean $\mathbb{E}[X_i] = \mu$ *and variance* $\mathbb{E}[(X_i - \mu)^2] = \sigma^2$ *with* $1 \le \sigma^2 \le \kappa$ *and*

$$n = O\left(\frac{\log(1/\gamma)}{\alpha^2} + \frac{\text{polylog}\left(\frac{\log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon} + \frac{\log(1/\delta) + \log(1/\gamma)}{\epsilon}\right),$$

outputs $\widehat{\mu}$ *such that* $|\widehat{\mu} - \mu| \le \alpha\kappa$ *with probability* $1 - \gamma$.

Generalizing this algorithm to the multivariate case, by following the algorithm NaivePME of Kamath et al. (2019), we obtain the following lemma, in the proof sketch of which we show only the modifications required in the proof of Kamath et al. (2019).

Lemma 6.2.6. *For every* $\epsilon, \delta, \gamma, \kappa, \alpha > 0$*, there exists an* $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$*-DP algorithm that, when given* n *i.i.d. samples* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ *from a sub-gaussian multivariate distribution with mean* $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{\mu}$ *and covariance matrix* $\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T] = \Sigma$ *with* $\mathbb{I}_d \preceq \Sigma \preceq \kappa\mathbb{I}_d$ *and*

$$n = O\left(\frac{\kappa^2 d\log(d/\gamma)}{\alpha^2} + \frac{\kappa d\,\text{polylog}\left(\frac{\kappa d\log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon} + \frac{\sqrt{d}\,(\log(1/\delta) + \log(d/\gamma))}{\epsilon}\right),$$

outputs $\widehat{\boldsymbol{\mu}}$ *such that* $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \le \alpha$ *with probability* $1 - \gamma$.

Proof Sketch. Following the procedure of Lemma 6.2.5 for each dimension of the multivariate vectors $\boldsymbol{X}_i$ with the appropriate parameter settings as described in NaivePME (Kamath et al., 2019), we obtain the following final result:

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \le \sqrt{d}\max_{1\le i\le d}|\widehat{\mu}_i - \mu_i| \le \sqrt{d}\left(\frac{\alpha}{\sqrt{d}}\right) = \alpha\,,$$

as desired. □

The final step (see Section 6.2.2) is to use the private "preconditioner" A in order to reduce the condition number of Σ (which is at most κ) to at most a constant, and obtain the final bound of the lemma that we seek. Again, we show the modification of the bound, hinging on Lemma 6.2.6 which is used to output the final estimate $\widehat{\boldsymbol{\mu}} = A^{-1}\widetilde{\boldsymbol{\mu}}$ from the estimate $\widetilde{\boldsymbol{\mu}}$ of the mean of the variables $A\boldsymbol{X}_i$, as follows by the lemma with $\kappa = O(1)$:

$$\|A(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|_2 = \|\widetilde{\boldsymbol{\mu}} - A\boldsymbol{\mu}\|_2 \le \alpha\,.$$

□

Chapter 7

Private Estimation on Simple Linear Models

In this chapter, we intend to show that estimation of the regression coefficient $\boldsymbol{\beta}$ under Gaussian covariates with unbounded feature vectors can, in the case of the Linear Regression model of Section 5.1, be done considerably easier than the main method we study in Chapter 8. The estimation technique consists of using only the classical differentially private covariance estimation algorithms by a reduction of regression coefficient estimation to Gaussian vector covariance estimation.

We remind to the reader the set of assumptions in this particular model, before continuing to state and analyze the estimation algorithm that we utilize.

Assumption 7.0.1. *Labeled examples $(\boldsymbol{X}_i, y_i)$, $i = 1, \ldots, n$, are i.i.d. Moreover, $\boldsymbol{X}_i \in \mathbb{R}^d$ are sampled from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ satisfying $\mathbb{I}_d \preceq \Sigma \preceq \kappa \mathbb{I}_d$ for some universal parameter $\kappa > 0$. The generative model of the labels y_i is as follows: there exists a $\boldsymbol{\beta} \in \mathbb{R}^d$ such that, given $\boldsymbol{X}_i \in \mathbb{R}^d$,*

$$y_i = \boldsymbol{\beta}^T \boldsymbol{X}_i + \epsilon_i, \quad \text{for all } i = 1, \ldots, n\,, \tag{7.1}$$

where ϵ_i are i.i.d. samples from a zero-mean Gaussian distribution $\mathcal{N}(0, \sigma_\epsilon^2)$.

We can deduce from Equation (7.1) that the (marginal) distribution of labels y_i is a Gaussian distribution $\mathcal{N}(\boldsymbol{\beta}^T \boldsymbol{\mu}, \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta} + \sigma_\epsilon^2)$, therefore, defining the vectors $\boldsymbol{Z}_i \in \mathbb{R}^{d+1}$ as

$$\boldsymbol{Z}_i = \begin{bmatrix} \boldsymbol{X}_i \\ y_i \end{bmatrix}, \tag{7.2}$$

we can see that they similarly follow a Gaussian distribution, with a covariance matrix Σ' that may be written in a block matrix form:

$$\Sigma' = \mathbb{E}\left[\boldsymbol{Z}_i \boldsymbol{Z}_i^T\right] = \begin{bmatrix} \Sigma & \Sigma\boldsymbol{\beta} \\ \boldsymbol{\beta}^T \Sigma & \sigma_\epsilon^2 + \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta} \end{bmatrix}. \tag{7.3}$$

Equation (7.3) indicates that a natural way to estimate $\boldsymbol{\beta}$ would be to estimate the covariance matrices Σ, Σ' and then, extracting the last column of Σ', which (without the last element) is $\Sigma\boldsymbol{\beta}$, left multiply by the inverse of the estimate of Σ that we have. Indeed, we show that this approach works and, when the estimation of the above covariance matrices is made according to the differentially private algorithms of Kamath et al. (2019), the end result is a differentially private estimator of $\boldsymbol{\beta}$ for the "simple linear regression" model of Assumption 7.0.1.

Algorithm. Having access to the n i.i.d. samples $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, where $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma), i \in [n]$, the algorithm initially computes a differentially private estimate $\widehat{\Sigma}$ of the covariance matrix of the d-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, using the algorithm LearnGaussian-hd, described in Section 6.1. Then, the algorithm forms the random vectors $\boldsymbol{Z}_i \in \mathbb{R}^{d+1}$ as in Equation (7.2) and computes a differentially private estimate $\widehat{\Sigma}'$ of the covariance matrix of the

$(d+1)$-dimensional Gaussian distribution with covariance matrix of the block form of Equation (7.3), again using the algorithm LearnGaussian-hd. From the matrix $\widehat{\Sigma}'$, the algorithm obtains only the first d elements of the last column of that matrix, naming them as $\widehat{\Sigma\boldsymbol{\beta}} \in \mathbb{R}^d$ (hinting at the form of Equation (7.3)). Armed with these estimates, the differentially private estimate of $\boldsymbol{\beta}$ is finally given by:

$$\widehat{\boldsymbol{\beta}} = \widehat{\Sigma}^{-1}\widehat{\Sigma\boldsymbol{\beta}}\,, \tag{7.4}$$

whose privacy follows from appropriate composition rules. We now proceed to prove the accuracy guarantee of this estimate.

Theorem 7.0.1 (Accuracy of $\widehat{\boldsymbol{\beta}}$ in Private Simple Linear Regression)**.** *Under Assumption 7.0.1 with parameters (κ, Σ') where Σ' is defined as in Equation (7.3), for all privacy parameters $\epsilon, \delta > 0$, accuracy parameters $\alpha, \eta > 0$ and confidence $\gamma \in (0,1)$, the private algorithm corresponding to Equation (7.4) is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-differentially private, and if the number of labeled examples is at least:*

$$n = O\left(\frac{d+\log(1/\gamma)}{\eta^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d}{\eta\gamma\epsilon}\right)}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log\left(\kappa(\Sigma')\right)}\text{polylog}\left(\frac{d\log(\kappa(\Sigma'))}{\gamma\epsilon}\right)}{\epsilon}\right)$$
$$+ O\left(\frac{d+\log(1/\gamma)}{\alpha^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon}\right),$$

then with probability at least $1 - O(\gamma)$ the output estimate $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ and the "true" regression coefficient $\boldsymbol{\beta}$ satisfy:

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2 \leq \|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_2^2 \leq O\left(\alpha^2\right) \cdot \|\boldsymbol{w}\|_2^2 + O\left(\eta^2\right) \cdot \lambda_{max}^2(\Sigma')\,, \tag{7.5}$$

where $\kappa(\Sigma') = \frac{\lambda_{max}(\Sigma')}{\lambda_{min}(\Sigma')}$ is the condition number of the block matrix Σ' as in Equation (7.3), $\widehat{\boldsymbol{w}} = \Sigma^{1/2}\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{w} = \Sigma^{1/2}\boldsymbol{\beta}$.

Proof. The privacy of the algorithm in this Theorem arises directly from the privacy of the differentially private covariance estimation algorithm and the composition theorems.

For the accuracy guarantee, we begin by adding and subtracting the quantities of each factor of $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$, as follows:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\widehat{\Sigma}^{-1} - \Sigma^{-1}\right)\Sigma\boldsymbol{\beta} + \widehat{\Sigma}^{-1}\left(\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right).$$

Then, substituting the quantities $\boldsymbol{w} = \Sigma^{1/2}\boldsymbol{\beta}$ and left multiplying both sides of the equation by $\Sigma^{1/2}$, we obtain that

$$\widehat{\boldsymbol{w}} - \boldsymbol{w} = \left(\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2} - \mathbb{I}_d\right)\boldsymbol{w} + \Sigma^{1/2}\widehat{\Sigma}^{-1}\left(\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right)$$
$$\Leftrightarrow \widehat{\boldsymbol{w}} - \boldsymbol{w} = \left(\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\right)\left[-\left(\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - \mathbb{I}_d\right)\boldsymbol{w} + \Sigma^{-1/2}\left(\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right)\right].$$

Using Cauchy-Schwartz and the sub-multiplicative property of the spectral norm, we establish the following inequality:

$$\|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_2^2 \leq 2\left\|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\right\|_2^2\left(\left\|\Sigma^{-1/2}\left(\widehat{\Sigma} - \Sigma\right)\Sigma^{-1/2}\right\|_2^2 \cdot \|\boldsymbol{w}\|_2^2 + \left\|\Sigma^{-1/2}\left(\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right)\right\|_2^2\right).$$

In order to bound the constituent terms of the right-hand-side of this inequality, we need a very similar claim to Claim 8.3.2 which we state below without proof (since it is almost the same as the main Section 8.3.2 that follows), Lemma 6.2.1 (which we remind to the reader that it is applied to the $2n$ sample differences $\frac{1}{\sqrt{2}}(\boldsymbol{X}_{2i} - \boldsymbol{X}_{2i-1})$, so that they have zero mean), and Claim 7.0.3, which is the main subject that we elaborate on in Section 7.1.

Claim 7.0.2 (Similar to Claim 8.3.2)**.** *When*

$$n = \Omega\left(d + \log(1/\gamma) + \frac{d^{3/2}\sqrt{\log\kappa}\text{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right),$$

the following inequality holds with probability $1 - O(\gamma)$*:*

$$\left\|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\right\|_2^2 \leq O\left(1\right).$$

Claim 7.0.3. *When*

$$n = \Omega\left(\frac{d + \log(1/\gamma)}{\eta^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d}{\eta\gamma\epsilon}\right)}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log\left(\kappa(\Sigma')\right)}\text{polylog}\left(\frac{d\log(\kappa(\Sigma'))}{\gamma\epsilon}\right)}{\epsilon}\right),$$

the following inequality holds with probability $1 - O(\gamma)$*:*

$$\left\|\Sigma^{-1/2}\left(\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right)\right\|_2^2 \leq O\left(\eta^2\right)\cdot\lambda_{max}^2(\Sigma').$$

Combining Claim 7.0.2, Lemma 6.2.1, and Claim 7.0.3 with a union bound of the respective events, we directly obtain Theorem 7.0.1, since

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2 = \left\|\Sigma^{-1/2}\Sigma^{1/2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\|_2^2 \leq \left\|\Sigma^{-1/2}\right\|_2^2\cdot\|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_2^2 \leq \|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_2^2,$$

by the sub-multiplicative property of the norm and because $\mathbb{I}_d \preceq \Sigma$. □

7.1 Proof of Claim 7.0.3

First of all, we note that, according to the first lines of the proof of Fact 8.3.11 which follows later, in order to apply the (accuracy) result of Lemma 6.2.1 to the covariance estimation of the random vectors $\boldsymbol{Z}_i \in \mathbb{R}^{d+1}$ as in Equation (7.2), a change of variables is needed, that affects solely the analysis of the algorithm (and more specifically, appears in a change of the sample complexity). Therefore, the specific result which applies in our case here is stated in the following fact.

Fact 7.1.1 (Covariance $\widehat{\Sigma}'$ estimation accuracy)**.** *For every* $\eta > 0$*, the output* $\widehat{\Sigma}'$ *of algorithm* LearnGaussian-hd *when given at least* n *samples* $\boldsymbol{Z}_i$ *with*

$$n = O\left(\frac{d + \log(1/\gamma)}{\eta^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d}{\eta\gamma\epsilon}\right)}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log\left(\kappa(\Sigma')\right)}\text{polylog}\left(\frac{d\log(\kappa(\Sigma'))}{\gamma\epsilon}\right)}{\epsilon}\right),$$

where $\kappa(\Sigma') = \frac{\lambda_{max}(\Sigma')}{\lambda_{min}(\Sigma')}$ *is the condition number of the block matrix* Σ' *as in Equation (7.3), satisfies the following accuracy guarantee with probability* $1 - O(\gamma)$*:*

$$\left\|\Sigma'^{-1/2}\left(\widehat{\Sigma}' - \Sigma'\right)\Sigma'^{-1/2}\right\|_2 \leq O(\eta). \tag{7.1}$$

We remind to the reader the form of the block matrix Σ' which is as follows:

$$\Sigma' = \begin{bmatrix}\Sigma & \Sigma\boldsymbol{\beta} \\ \boldsymbol{\beta}^T\Sigma & \sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}\end{bmatrix}. \tag{7.2}$$

By definition of the spectral norm, from Equation (7.1) we have that for every vector $\boldsymbol{u} \in \mathbb{R}^{d+1} : \|\boldsymbol{u}\|_2 \leq 1$, it holds that $\left\|\Sigma'^{-1/2}\left(\widehat{\Sigma}' - \Sigma'\right)\Sigma'^{-1/2}\boldsymbol{u}\right\|_2 \leq O(\eta)$. Taking advantage of the spectral decomposition of the (real symmetric, PSD) matrix $\Sigma' = U\Lambda U^T$ for some unitary orthogonal matrix U and diagonal matrix Λ, it is well-known that $\Sigma'^{-1/2} = \Lambda^{-1/2}U^T$, and because $\|U^T\boldsymbol{u}\|_2 = \|\boldsymbol{u}\|_2$ for every vector $\boldsymbol{u} \in \mathbb{R}^{d+1}$ (since U is an orthonormal matrix), and since $\sqrt{\lambda_{\max}(\Sigma')}\Lambda^{-1/2} \succeq \mathbb{I}_{d+1}$, we conclude that by choosing $\boldsymbol{u} \in \mathbb{R}^{d+1} : \sqrt{\lambda_{\max}(\Sigma')}\Lambda^{-1/2}\boldsymbol{u} = \boldsymbol{e}_{d+1}$ (where $\boldsymbol{e}_{d+1}$ is the unit vector that has only the $(d+1)$-th coordinate 1 and all other coordinates 0), it is true that $\left\|\Sigma'^{-1/2}\left(\widehat{\boldsymbol{v}}_{d+1} - \boldsymbol{v}_{d+1}\right)\right\|_2^2 \leq O(\eta^2) \cdot \lambda_{\max}(\Sigma')$, where $\widehat{\boldsymbol{v}}_{d+1}, \boldsymbol{v}_{d+1}$ are the last columns of the matrices $\widehat{\Sigma}'$ and Σ' respectively (see Equation (7.2) for what the last column looks like). Therefore, it is immediate that

$$\begin{aligned}\|\widehat{\boldsymbol{v}}_{d+1} - \boldsymbol{v}_{d+1}\|_2^2 &= \left\|\Sigma'^{1/2}\Sigma'^{-1/2}\left(\widehat{\boldsymbol{v}}_{d+1} - \boldsymbol{v}_{d+1}\right)\right\|_2^2 \\ &\leq \left\|\Sigma'^{1/2}\right\|_2^2 \cdot \left\|\Sigma'^{-1/2}\left(\widehat{\boldsymbol{v}}_{d+1} - \boldsymbol{v}_{d+1}\right)\right\|_2^2 \\ &\leq O(\eta^2) \cdot \lambda_{\max}^2(\Sigma')\,,\end{aligned}$$

and since the first d coordinates of $\widehat{\boldsymbol{v}}_{d+1} - \boldsymbol{v}_{d+1} \in \mathbb{R}^{d+1}$ are the vector $\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta} \in \mathbb{R}^d$ (see the structure of Equation (7.2)), we have that $\left\|\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right\|_2^2 \leq \|\widehat{\boldsymbol{v}}_{d+1} - \boldsymbol{v}_{d+1}\|_2^2 \leq O(\eta^2) \cdot \lambda_{\max}^2(\Sigma')$.

Claim 7.0.3 follows, since

$$\left\|\Sigma^{-1/2}\left(\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right)\right\|_2^2 \leq \left\|\Sigma^{-1/2}\right\|_2^2 \cdot \left\|\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right\|_2^2 \leq \left\|\widehat{\Sigma\boldsymbol{\beta}} - \Sigma\boldsymbol{\beta}\right\|_2^2 \leq O(\eta^2) \cdot \lambda_{\max}^2(\Sigma')\,,$$

by the sub-multiplicative property of the norm and because $\mathbb{I}_d \preceq \Sigma$.

The proof has now been completed. Of course, the condition number $\kappa(\Sigma')$ is an interesting quantity that merits consideration to examine what it depends upon. From Theorem 1 of Dembo (1988), one may deduce the following upper bound on the largest eigenvalue of Σ':

$$\lambda_{\max}(\Sigma') \leq 2\left(\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \max\left(\kappa, \sigma_\epsilon^2\right)\right)\,, \tag{7.3}$$

where we remind to the reader that $\kappa = \lambda_{\max}(\Sigma)$.

The lower bound on the smallest eigenvalue provided by the above work (Dembo, 1988) is non-optimal, since it may be negative at certain cases, while the matrix itself only ever exhibits non-negative eigenvalues (since it is PSD, by definition of being a covariance matrix). A better bound is given by Ma and Zarowski (1995), as follows:

$$\begin{aligned}\lambda_{\min}(\Sigma') &\geq \frac{\sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \lambda_{\min}(\Sigma)}{2} - \sqrt{\left(\frac{\sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \lambda_{\min}(\Sigma)}{2}\right)^2 - \sigma_\epsilon^2\lambda_{\min}(\Sigma)} \\ &= \frac{\sigma_\epsilon^2\lambda_{\min}(\Sigma)}{\frac{1}{2}\left(\sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \lambda_{\min}(\Sigma) + \sqrt{\left(\sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \lambda_{\min}(\Sigma)\right)^2 - 4\sigma_\epsilon^2\lambda_{\min}(\Sigma)}\right)} \\ &\geq \frac{\sigma_\epsilon^2\lambda_{\min}(\Sigma)}{\sigma_\epsilon^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} + \lambda_{\min}(\Sigma)}\,.\end{aligned} \tag{7.4}$$

The first form of the lower bound given above is tight, as may be noticed by examining the specific case of $\Sigma = \kappa\mathbb{I}_d$. In this case, one would have that:

$$\Sigma' = \begin{bmatrix}\kappa\mathbb{I}_d & \kappa\boldsymbol{\beta} \\ \kappa\boldsymbol{\beta}^T & \sigma_\epsilon^2 + \kappa\|\boldsymbol{\beta}\|_2^2\end{bmatrix} = \kappa\begin{bmatrix}\mathbb{I}_d & \boldsymbol{\beta} \\ \boldsymbol{\beta}^T & \frac{\sigma_\epsilon^2}{\kappa} + \|\boldsymbol{\beta}\|_2^2\end{bmatrix}\,,$$

reducing our calculations to the simple case of covariance matrix $\mathbb{I}_d$, for which the second matrix written in the above equation has eigenvalues t according to the roots of the equation

$$\begin{aligned} & (1-t)^d \left(\frac{\sigma_\epsilon^2}{\kappa} + \|\boldsymbol{\beta}\|_2^2 - t \right) - \sum_{i=1}^{d} \beta_i^2 (1-t)^{d-1} = 0 \\ \Leftrightarrow \quad & (1-t)^{d-1} \left(t^2 - \left(1 + \|\boldsymbol{\beta}\|_2^2 + \frac{\sigma_\epsilon^2}{\kappa} \right) t + \frac{\sigma_\epsilon^2}{\kappa} \right) = 0 \,, \end{aligned}$$

where β_i is the i-th coordinate of the vector $\boldsymbol{\beta}$. Therefore Σ' has the following smallest eigenvalue (the smallest of the two roots of the quadratic equation, which is guaranteed to be ≤ 1, i.e., smaller than the other eigenvalues):

$$\begin{aligned} \lambda_{\min}(\Sigma') &= \frac{\kappa}{2} \left(1 + \|\boldsymbol{\beta}\|_2^2 + \frac{\sigma_\epsilon^2}{\kappa} - \sqrt{\left(1 + \|\boldsymbol{\beta}\|_2^2 + \frac{\sigma_\epsilon^2}{\kappa} \right)^2 - 4\frac{\sigma_\epsilon^2}{\kappa}} \right) \\ &= \frac{\sigma_\epsilon^2 + \kappa\|\boldsymbol{\beta}\|_2^2 + \kappa}{2} - \sqrt{\left(\frac{\sigma_\epsilon^2 + \kappa\|\boldsymbol{\beta}\|_2^2 + \kappa}{2} \right)^2 - \kappa\sigma_\epsilon^2} \,, \end{aligned}$$

which neatly matches the first form of the lower bound given in Equation (7.4), since $\Sigma = \kappa\mathbb{I}_d$.

Therefore, the condition number $\kappa(\Sigma')$ (which is at most the ratio of the right-hand-sides of Equation (7.3) to Equation (7.4)) and the largest eigenvalue $\lambda_{\max}(\Sigma')$ depend on both $\boldsymbol{\beta}$ and σ_ϵ^2 besides the usual dependence on the smallest and largest eigenvalues of the feature vector covariance matrix Σ, respectively: $\lambda_{\max}(\Sigma) \leq \kappa$ and $\lambda_{\min}(\Sigma) \geq 1$. We note that this means that, when $\|\boldsymbol{\beta}\|_2$ is large, more samples will be necessary to achieve a fixed additive accuracy, as indicated by Equation (7.5).

Chapter 8

Private Estimation of Least Squares Fitting and Binary Regression

8.1 Main Results

We formally state our main results in this section. Our theorems provide $(\frac{\epsilon^2}{2}+\epsilon\sqrt{2\log(1/\delta)},\delta)$-DP guarantees for both the Least Squares Fitting and Binary Regression settings. This guarantee is, in essence, equivalent to (ϵ,δ)-DP. For a more detailed discussion on this issue, we refer the reader to Section 6.2.

8.1.1 Least Squares Fitting

Our differentially private LSE for the Least Squares Fitting setting is summarized in Algorithm 2. In short, the algorithm computes diffentially private estimates of the quantities

$$(X^TX/n)^{-1} \quad \text{and} \quad X^T\boldsymbol{y}/n, \tag{8.1}$$

whose product, by Equation (5.1) in Section 5.2, yields the LSE $\boldsymbol{\beta}^\star$.

The estimation of the first quantity proceeds as follows. Having access to the n i.i.d. samples $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, where $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma), i \in [n]$, Algorithm 2 initially privately computes differentially private estimates $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}, \widehat{\Sigma}_{\boldsymbol{X}})$ of the mean and covariance matrix of the d-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, using the algorithm LearnGaussian-hd, described in Section 6.1. These estimates, that satisfy the guarantees indicated in Theorem 6.1.1, can be used to estimate $(X^TX/n)^{-1}$ via the relationship:

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T \approx \mathbb{E}[\boldsymbol{X}_i\boldsymbol{X}_i^T] \approx \widehat{\Sigma}_{\boldsymbol{X}} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T\,.$$

The second quantity, i.e., the term $X^T\boldsymbol{y}/n$, is somewhat harder to estimate in a differentially private fashion, as constituent terms $y_i\boldsymbol{X}_i$ are *not* Gaussian. The boundedness of variables y_i, however, ensures that these terms are sub-gaussian. As an important technical contribution, we extend the analysis of Kamath et al. (2019) and Karwa and Vadhan (2018) to the sub-gaussian regime (see LearnSubGaussian-hd in Section 6.2.3), and obtain a private mean estimate $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y}$ for the sub-gaussian random vectors $y_i\boldsymbol{X}_i$.

Armed with these estimates, the differentially private LSE is finally given by:

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\Sigma}_{\boldsymbol{X}} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T\right)^{-1}\widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y}\,, \tag{8.2}$$

whose privacy follows from appropriate composition rules. We refer to the resulting algorithm, summarized in Algorithm 2, as PrivateLearnLSE. We remark that Algorithm 2 requires a priori knowledge of the covariance bound κ but not the mean bound R. We elaborate on this in Section 6.2.3. Our main result w.r.t. the privacy and accuracy of this estimator is as follows:

Algorithm 2 Private Estimation of Least Squares Estimate.

Input: $(X, \boldsymbol{y}) = (\boldsymbol{X}_i, y_i)_{i\in[n]}$ with $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}, \Sigma$ are unknown.
Parameters: Privacy $\epsilon, \delta > 0$, accuracy $\alpha, \eta > 0$, confidence $\gamma \in (0,1)$, covariance spectral norm bound κ, upper bound of labels c.
Output: Estimate $\widehat{\boldsymbol{\beta}}$ that approaches the LSE $\boldsymbol{\beta}^\star$ in ℓ_2 norm with high probability.
procedure PrivateLearnLSE$((X, \boldsymbol{y}), \epsilon, \delta, \alpha, \eta, \gamma, \kappa)$ ▷ *Sample size n satisfies Theorem 8.1.1.*
    $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}, \widehat{\Sigma}_{\boldsymbol{X}}) \leftarrow$ LearnGaussian-HD$(\{\boldsymbol{X}_i\}_{i\in[n]}, O(\epsilon), O(\delta), O(\alpha), \gamma, \kappa)$.
    $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y} \leftarrow$ LearnSubGaussian-HD$(\{y_i\boldsymbol{X}_i\}_{i\in[n]}, O(\epsilon), O(\delta), O(\eta), \gamma, c^2\kappa)$.
    Output the private estimate $\widehat{\boldsymbol{\beta}} = \left(\widehat{\Sigma}_{\boldsymbol{X}} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T\right)^{-1}\widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y}$.

Theorem 8.1.1 (Privacy and Accuracy of $\widehat{\boldsymbol{\beta}}$ in Private Least Squares Fitting)**.** *Under Assumption 5.2.1 with parameters (κ, c, ρ, R), for all privacy parameters $\epsilon, \delta > 0$, accuracy parameters $\alpha, \eta > 0$ and confidence $\gamma \in (0,1)$,* PrivateLearnLSE *(defined in Algorithm 2) is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-differentially private. Moreover, if the number of labeled examples is at least:*

$$n = O\left(\frac{d\log(\frac{d}{\gamma})}{\eta^2} + \frac{d\text{polylog}(\frac{d\log(1/\delta)}{\eta\gamma\epsilon})}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log(\kappa\rho c)}\text{polylog}\left(\frac{d\log(\kappa\rho c)}{\gamma\epsilon\delta}\right)}{\epsilon}\right)$$
$$+ O\left((1+R)\left(\frac{d\log(\frac{d}{\gamma})}{\alpha^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d\log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\text{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right)\right),$$

then with probability at least $1 - O(\gamma)$ the output estimate $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ and the LSE $\boldsymbol{\beta}^*$ satisfy:

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right\|_2^2 \le O\left(\alpha^2\right)\cdot\left\|\Sigma^{1/2}\boldsymbol{\beta}^\star\right\|_2^2 + O\left(\eta^2\right)\cdot c^2\,.$$

The proof can be found in Section 8.3. Intuitively, compared to Theorem 6.1.1, the number of samples we require grows as $\widetilde{O}(d^{3/2})$, slightly more favorably than the covariance estimation case. Moreover, the number of samples again grows polylogarithmically on κ (the bound on the covariance spectral norm) but linearly (rather than polylogarithmically) on R, the bound on the mean.

8.1.2 Binary Regression

We next turn our attention to the Binary Regression setting, in which *both* Assumption 5.2.1 and Assumption 5.3.1 apply. We study the properties of PrivateLearnLSE (Algorithm 2) under the above additional Assumption 5.3.1; the only modification of Algorithm 2, compared to the previous setting, is that we do not need the estimate $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$, as Assumption 5.3.1 states that $\boldsymbol{\mu} = \mathbf{0}$. As such, we set $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} = \mathbf{0}$ in Equation (8.2), with the remaining terms computed as in the previous section. We show that the resulting algorithm has the following guarantees.

Theorem 8.1.2 (Privacy and Accuracy of $\widehat{\boldsymbol{\beta}}$ in Private Binary Regression)**.** *Under Assumption 5.2.1 with covariance parameter κ and Assumption 5.3.1 with true parameter $\boldsymbol{\beta} \in \mathbb{R}^d$, for every privacy parameters $\epsilon, \delta > 0$, accuracy parameters $\alpha, \eta > 0$ and confidence $\gamma \in (0,1)$,* PrivateLearnLSE *(defined in Algorithm 2) with $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} = \mathbf{0}$ is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-differentially*

private. Moreover, if the number of labeled examples is at least:

$$n = O\left(\frac{d\log(\frac{d}{\gamma})}{\eta^2} + \frac{d\text{polylog}(\frac{d\log(1/\delta)}{\eta\gamma\epsilon})}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\text{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon\delta}\right)}{\epsilon}\right)$$
$$+ O\left(\frac{d\log(\frac{d}{\gamma})}{\alpha^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d\log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon}\right),$$

then with probability at least $1 - O(\gamma)$ *the output estimate* $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ *satisfies*

$$\|\widehat{\boldsymbol{\beta}} - k\boldsymbol{\beta}\|_2^2 \leq O\left(\alpha^2\right) \cdot \left(1 + \left\|k\Sigma^{1/2}\boldsymbol{\beta}\right\|_2^2\right) + O\left(\eta^2\right), \tag{8.3}$$

where $k = \frac{2n}{n-d-1}\,\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right)\right]$.

The theorem is proven in Section 8.4. As in Theorem 8.1.1, the sample complexity grows as $d^{3/2}$, and is merely polylogarithmic on κ. Moreover, as in classic (non-DP) work on binary regression via LSE (Brillinger, 2012a), our estimator learns the underlying "true" $\boldsymbol{\beta}$ up to a scaling factor k, that depends on the "sharpness" of the model function f (via its derivative f'). We note that, to discover the hyperplane separating positive from negative labels, it indeed suffices to learn only the direction of $\boldsymbol{\beta}$, not its magnitude.

To elaborate more on the effect of k: by Assumption 5.3.1, $\boldsymbol{\beta}^T\boldsymbol{X}_i$ is a zero mean Gaussian, while f' tends to zero as its argument reaches either $+\infty$ or $-\infty$. Hence, the expectation that determines k very much depends by the behavior of f' around 0. That is, if f is relatively flat (i.e., binary labels are "noisy"), k will be small, and more samples will be needed to achieve a better numerical accuracy in Equation (8.3); the converse is true when f is "sharp" (e.g., a sigmoid close to the sign function), and labels are less noisy. This dependence of the estimate accuracy on the noise inherent in the GLM (via the model function f) is natural.

8.2 Technical Overview

In this section, we provide a sketch of the technical highlights of the proofs of Theorem 8.1.1 and Theorem 8.1.2. We begin with Theorem 8.1.1, which deals with the problem of Least Squares Fitting. Our goal is to privatize the Least Squares Estimator (see Equation (5.1)) without significant accuracy loss. Hence, the differentially private algorithm (see Algorithm 2) computes a quantity $\widehat{\boldsymbol{\beta}}$ that is asymptotically the same as the Least Squares Estimate of Equation (5.1):

$$\boldsymbol{\beta}^\star = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{X}_i\right). \tag{8.1}$$

The structure of this estimate (product of two terms) hints at privatizing each term separately, and then Algorithm 2 would arise as a natural solution. To ensure that the desired privacy property holds, the key idea is to apply the composition of differentially private mechanisms (see Theorem 3.2.4), hence affording privacy to the whole algorithm. It thus suffices to consider privatized estimates of the individual terms. In order to achieve that, we utilize the private procedures for mean and covariance estimation of multivariate Gaussian random vectors, as presented in Kamath et al. (2019) and Karwa and Vadhan (2018).

The main conceptual observation for our main result is that the second term in Equation (8.1) consists, in fact, of sub-gaussian vectors. At a technical level, we have to expand the aforementioned works to the sub-gaussian regime. More to that, in order to reduce as

much as possible the dependence on the range of the mean value R of the feature vectors $\boldsymbol{X}_i$, we modify the multivariate mean estimation analysis of Kamath et al. (2019) to hold for unbounded mean feature vectors. As a technical tool, we use an alternative guarantee present in Karwa and Vadhan (2018) which allows us to disengage the concentration bounds from the bound on the mean, in the case that $\delta > 0$.

However, even using those variants of the algorithms, we still have to satisfy a stronger privacy desideratum: due to the privacy guarantees of Karwa and Vadhan (2018); Kamath et al. (2019) being on the entire sub-gaussian term $y_i\boldsymbol{X}_i$ being possibly altered, it is not straightforward to achieve the more general guarantee of altering the individual $(\boldsymbol{X}_i, y_i)$ pairs. We, thus, require a more nuanced version of privacy; that is, on the dataset of $\{(\boldsymbol{X}_i, y_i)\}_{i\in[n]}$. It turns out that we can tweak their version of privacy analysis to also satisfy this more subtle privacy definition.

For the desired accuracy guarantee on Algorithm 2, we have to control the quantity $\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right\|_2^2$. At a first sight, the above expression cannot be handled by standard concentration of measure phenomena. However, we provide a non-trivial decomposition:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star = \left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T\right)^{-1}\left(-Q_1\boldsymbol{\beta}^\star + \boldsymbol{Q}_2\right),$$

using the below quantities that we introduce:

$$Q_1 = \widehat{\Sigma} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}^T - \frac{1}{n}X^TX, \quad \text{and} \quad \boldsymbol{Q}_2 = \widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y} - \frac{1}{n}X^T\boldsymbol{y},$$

where $\widehat{\Sigma}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y}$ are the private outputs of the algorithms described in Algorithm 2, and $X, \boldsymbol{y}$ are the design matrix and the labels vector. This decomposition, when altered in geometry for normalization purposes by a transformation $\boldsymbol{w} = \Sigma^{-1/2}\boldsymbol{\beta}$ and $\widehat{\boldsymbol{w}} = \Sigma^{-1/2}\widehat{\boldsymbol{\beta}}$, enables us to control each term individually and obtain the desired bounds. The intuition behind this decomposition lies in the fact that each term of $Q_1, \boldsymbol{Q}_2$ vanishes asymptotically as the number of samples n increases.

The bounds on $Q_1, \boldsymbol{Q}_2$ are handled by further decomposing into the difference of private quantities and their actual values $(\Sigma, \boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\mu}_{\boldsymbol{X},y})$ and between empirical quantities and the actual values. To obtain tighter bounds on the individual terms of difference of private quantities and actual values, we use the private preconditioner matrix technique (Kamath et al. (2019)), that allows us to avoid a strict dependence on the largest eigenvalue κ of the covariance matrix Σ in our bounds (see Theorem 8.1.1). For a detailed proof of Theorem 8.1.1, see Section 8.3.

As far as our second main result (Theorem 8.1.2) is concerned, the key conceptual contribution is to introduce a new estimator $\boldsymbol{\beta}_s^\star$ that is defined with the help of n additional samples $(\boldsymbol{X}_i, y_i)$ (for a total of $2n$ samples) as follows:

$$\boldsymbol{\beta}_s^\star = \left(\frac{1}{n}\sum_{i=n+1}^{2n}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{X}_i\right).$$

This estimate resembles the Least Squares Estimate $\boldsymbol{\beta}^\star$ but *crucially introduces independence* between the two terms that constitute the Least Squares Estimate. This independence of the two terms is pivotal for proving that the estimate $\boldsymbol{\beta}_s^\star$ is an unbiased up to a multiplicative factor estimate of the true regression coefficient $\boldsymbol{\beta}$. In turn, this crucial observation is used to prove that our private estimate $\widehat{\boldsymbol{\beta}}$ (see Algorithm 2) is close to the true regression coefficient $\boldsymbol{\beta}$ up to a multiplicative factor, since the proof of Theorem 8.1.1 holds true even for the Least-Squares-resembling estimate $\boldsymbol{\beta}_s^\star$ (because of the independent handling of the aforementioned quantities $Q_1, \boldsymbol{Q}_2$). At a technical level, the above discussion is a result of probabilistic tools, such as the high-dimensional geometry of Wishart matrices. For a detailed proof of Theorem 8.1.2, see Section 8.4.

8.3 Proof of Theorem 8.1.1

We divide the proof of Theorem 8.1.1 in a series of claims. For convenience, we restate the (stronger) version of the Theorem that we will prove here:

Theorem 8.3.1 (Privacy and Accuracy of $\widehat{\boldsymbol{\beta}}$ in Private Least Squares Fitting)**.** *Under Assumption 5.2.1 with parameters (κ, c, ρ, R), for all privacy parameters $\epsilon, \delta > 0$, accuracy parameters $\alpha, \eta > 0$ and confidence $\gamma \in (0,1)$,* PrivateLearnLSE *(defined in Algorithm 2) is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-differentially private. Moreover, if the number of labeled examples is at least:*

$$n = O\left(\frac{d\log(\frac{d}{\gamma})}{\eta^2} + \frac{d\text{polylog}(\frac{d\log(1/\delta)}{\eta\gamma\epsilon})}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log(\kappa\rho c)}\text{polylog}\left(\frac{d\log(\kappa\rho c)}{\gamma\epsilon\delta}\right)}{\epsilon}\right) + O\left((1+R)\left(\frac{d\log(\frac{d}{\gamma})}{\alpha^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d\log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\text{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon}\right)}{\epsilon}\right)\right), \tag{8.1}$$

then with probability at least $1 - O(\gamma)$ the output estimate $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ and the LSE $\boldsymbol{\beta}^\star$ satisfy:

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right\|_2^2 \leq \|\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\|_2^2 \leq O\left(\alpha^2\right) \cdot \|\boldsymbol{w}^\star\|_2^2 + O\left(\eta^2\right) \cdot c^2\,, \tag{8.2}$$

where $\widehat{\boldsymbol{w}} = \Sigma^{1/2}\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{w}^\star = \Sigma^{1/2}\boldsymbol{\beta}^\star$.

The outline of the proof is as follows. First, we prove the privacy guarantee (see Section 8.3.1); the latter follows in a similar fashion as the privacy proof of Kamath et al. (2019), exploiting the extension to $\delta > 0$ we described in the previous section, and generalizing the obtained privacy for altering both y_i and $\boldsymbol{X}_i$ (see Section 8.2). In the case of accuracy (see Section 8.3.2), we first establish the following inequality:

$$\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\|_2^2 \leq 2\left\|\Sigma^{1/2}\left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right)^{-1}\Sigma^{1/2}\right\|_2^2 \left(\left\|\Sigma^{-1/2}Q_1\Sigma^{-1/2}\right\|_2^2 \cdot \|\boldsymbol{w}^\star\|_2^2 + \left\|\Sigma^{-1/2}\boldsymbol{Q}_2\right\|_2^2\right).$$

Via a series of claims (see Claim 8.3.2, Claim 8.3.3, and Claim 8.3.4) we bound each of the constituent terms of the right-hand-side of this inequality, finally yielding Equation (8.2) with as many samples as in Equation (8.1).

8.3.1 Proof of Privacy Guarantee

We show first that Algorithm 2, that computes $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ as in Theorem 8.1.1, is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP. Our dataset consists of n pairs $(\boldsymbol{X}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, therefore we will look into what happens if one of those pairs is altered: specifically, consider that the pair $(\boldsymbol{X}_i, y_i)$ becomes $(\boldsymbol{X}_i', y_i')$ for some (specific) i. The algorithm for $\widehat{\boldsymbol{\beta}}$ uses three sub-algorithms, which we claim will be $\left(\frac{1}{3}\left(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}\right), \frac{\delta}{3}\right)$-DP each (as described in the algorithm above, it suffices to consider $O(\epsilon)$ and $O(\delta)$ as parameters of each one), thereby giving us the final result of the claim by the advanced composition properties of differentially private mechanisms (Theorem 3.2.4).

For the covariance estimation algorithm (used both in the covariance estimation of the feature vectors $\boldsymbol{X}_i$ and in the mean estimation of the product $y_i\boldsymbol{X}_i$), it now suffices to show that the interface of the algorithms in Kamath et al. (2019), which is the NaivePCE algorithm (see Section 6.2.2 to read further on how this holds), is differentially private. Indeed,

this result arises by computing the sensitivity of the (truncated) empirical covariance (see Section 6.2.2) in the following way:

$$\left\| \frac{1}{n} \left(y_i^2 \boldsymbol{X}_i \boldsymbol{X}_i^T - y_i'^2 \boldsymbol{X}_i' \boldsymbol{X}_i'^T \right) \right\|_F \leq \frac{1}{n} \left(\|y_i \boldsymbol{X}_i\|_2^2 + \|y_i' \boldsymbol{X}_i'\|_2^2 \right) \leq O\left(\frac{d\kappa c^2 \log(n/\gamma)}{n} \right) ,$$

since the truncation happens in accordance with Lemma 6.2.3, thereby allowing us to add the Gaussian noise of the magnitude prescribed in Kamath et al. (2019) verbatim. Hence, the rest of the algorithms in Kamath et al. (2019) that hinge on NaivePCE and Differential Privacy composition theorems (like Theorem 3.2.4) are differentially private.

In a similar fashion, the algorithm for mean estimation in Karwa and Vadhan (2018) depends upon the stability-based histogram learner of Bun et al. (2016) which is in turn based on the idea of introducing Laplacian noise to the empirical histogram of a dataset. Denoting X_{ij} as the j-th coordinate of the feature vector $\boldsymbol{X}_i$, we notice that the sensitivity of the empirical counting function is $2/n$ regardless of the input variations (at most 2 bins could have their counts altered in the worst case, if some $y_i X_{ij}$ changed its location from the bin it was to another, whereas all the other products had the same locations in bins). This sensitivity is, thus, independent of whether both y_i and X_{ij} were changed in our model, thereby affording the desired level of privacy to the whole algorithm. □

8.3.2 Proof of Accuracy Guarantee

For simplicity in notation, in what follows we notate $\boldsymbol{\mu} = \boldsymbol{\mu}_{\boldsymbol{X}}$ and $\boldsymbol{\mu}' = \boldsymbol{\mu}_{\boldsymbol{X},y}$ (likewise, $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$ and $\widehat{\boldsymbol{\mu}}' = \widehat{\boldsymbol{\mu}}_{\boldsymbol{X},y}$).

We begin by adding and subtracting the quantities of each factor of $\boldsymbol{\beta}^\star \in \mathbb{R}^d$, as follows:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star = \left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}^T \right)^{-1} \left(-Q_1 \boldsymbol{\beta}^\star + \boldsymbol{Q}_2 \right) \Leftrightarrow \left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}^T \right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \right) = -Q_1 \boldsymbol{\beta}^\star + \boldsymbol{Q}_2 ,$$

where

$$Q_1 = \widehat{\Sigma} + \widehat{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}^T - \frac{1}{n} X^T X, \quad \text{and} \quad \boldsymbol{Q}_2 = \widehat{\boldsymbol{\mu}}' - \frac{1}{n} X^T \boldsymbol{y} .$$

Then, substituting the quantities $\boldsymbol{w} = \Sigma^{1/2} \boldsymbol{\beta}$ and moving terms from the left side to the right of the equation, it holds that

$$\begin{aligned}
& \left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}^T \right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \right) = -Q_1 \boldsymbol{\beta}^\star + \boldsymbol{Q}_2 \\
\Leftrightarrow \; & \left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}^T \right) \Sigma^{-1/2} \left(\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star \right) = -Q_1 \Sigma^{-1/2} \boldsymbol{w}^\star + \boldsymbol{Q}_2 \\
\Leftrightarrow \; & \Sigma^{-1/2} \left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}^T \right) \Sigma^{-1/2} \left(\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star \right) = -\Sigma^{-1/2} Q_1 \Sigma^{-1/2} \boldsymbol{w}^\star + \Sigma^{-1/2} \boldsymbol{Q}_2 \\
\Leftrightarrow \; & \widehat{\boldsymbol{w}} - \boldsymbol{w}^\star = \Sigma^{1/2} \left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}^T \right)^{-1} \Sigma^{1/2} \left(-\Sigma^{-1/2} Q_1 \Sigma^{-1/2} \boldsymbol{w}^\star + \Sigma^{-1/2} \boldsymbol{Q}_2 \right) .
\end{aligned}$$

Using Cauchy-Schwartz and the sub-multiplicative property of the spectral norm, we establish the following inequality:

$$\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star\|_2^2 \leq 2 \left\| \Sigma^{1/2} \left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}^T \right)^{-1} \Sigma^{1/2} \right\|_2^2 \left(\left\| \Sigma^{-1/2} Q_1 \Sigma^{-1/2} \right\|_2^2 \cdot \|\boldsymbol{w}^\star\|_2^2 + \left\| \Sigma^{-1/2} \boldsymbol{Q}_2 \right\|_2^2 \right) .$$

We state three claims that bound the constituent terms of the right-hand-side of this inequality:

Claim 8.3.2. *When*

$$n = \Omega \left(d + \log(1/\gamma) + \frac{d^{3/2} \sqrt{\log \kappa} \operatorname{polylog} \left(\frac{d \log \kappa}{\gamma \epsilon} \right)}{\epsilon} \right) ,$$

the following inequality holds with probability $1 - O(\gamma)$*:*

$$\left\| \Sigma^{1/2} \left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}^T \right)^{-1} \Sigma^{1/2} \right\|_2^2 \leq O\left(1\right) .$$

Claim 8.3.3. *For every* $\alpha > 0$*, when*

$$n = \Omega\left((1+R) \left(\frac{d \log(\frac{d}{\gamma})}{\alpha^2} + \frac{d^{3/2} \mathrm{polylog}\left(\frac{d \log(1/\delta)}{\alpha \gamma \epsilon} \right)}{\alpha \epsilon} + \frac{d^{3/2} \sqrt{\log \kappa} \mathrm{polylog}\left(\frac{d \log \kappa}{\gamma \epsilon} \right)}{\epsilon} \right) \right) ,$$

the following inequality holds with probability $1 - O(\gamma)$*:*

$$\left\| \Sigma^{-1/2} Q_1 \Sigma^{-1/2} \right\|_2^2 \leq O(\alpha^2) .$$

Claim 8.3.4. *For every* $\eta > 0$*, when*

$$n = \Omega\left(\frac{d \log(\frac{d}{\gamma})}{\eta^2} + \frac{d \mathrm{polylog}(\frac{d \log(1/\delta)}{\eta \gamma \epsilon})}{\eta \epsilon} + \frac{\sqrt{d} \log(\frac{d}{\gamma \delta})}{\epsilon} + \frac{d^{3/2} \sqrt{\log(\kappa \rho c)} \mathrm{polylog}\left(\frac{d \log(\kappa \rho c)}{\gamma \epsilon} \right)}{\epsilon} \right) ,$$

the following inequality holds with probability $1 - O(\gamma)$*:*

$$\left\| \Sigma^{-1/2} \boldsymbol{Q}_2 \right\|_2^2 \leq O(\eta^2) \cdot c^2 .$$

We prove each of these claims individually below (see Section 8.3.2–Section 8.3.2). When combined with a union bound of the respective probabilistic events, these claims give the desired Theorem 8.1.1. In particular, we directly obtain Theorem 8.1.1, since

$$\left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \right\|_2^2 = \left\| \Sigma^{-1/2} \Sigma^{1/2} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \right) \right\|_2^2 \leq \left\| \Sigma^{-1/2} \right\|_2^2 \cdot \| \widehat{\boldsymbol{w}} - \boldsymbol{w}^\star \|_2^2 \leq \| \widehat{\boldsymbol{w}} - \boldsymbol{w}^\star \|_2^2 ,$$

by the sub-multiplicative property of the norm and because $\mathbb{I}_d \preceq \Sigma$. □

Proof of Claim 8.3.2

Following the procedure from Kamath et al. (2019) for covariance estimation, we recall the "private preconditioner" matrix A that is used to reduce the effect of the condition number of Σ from at most $O(\kappa)$ to at most a constant order factor. More specifically, the following lemma holds:

Lemma 8.3.5 (Theorem 3.11 of Kamath et al. (2019))**.** *For every* $\epsilon, \delta, \gamma, \alpha, \kappa > 0$ *, there exists an algorithm that, when given* n *i.i.d. samples* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ *from a sub-gaussian multivariate distribution with mean* $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{\mu}$ *and covariance matrix* $\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T] = \Sigma$ *with* $\mathbb{I}_d \preceq \Sigma \preceq \kappa \mathbb{I}_d$ *and*

$$n = O\left(\frac{d^{3/2} \sqrt{\log \kappa} \mathrm{polylog}\left(\frac{d \log \kappa}{\gamma \epsilon} \right)}{\epsilon} \right) ,$$

outputs a (symmetric) matrix A *(the "private preconditioner") such that* $\mathbb{I}_d \preceq A \Sigma A \preceq 1000 \mathbb{I}_d$ *with probability* $1 - O(\gamma)$*.*

Afterwards, the naive private estimation by addition of Gaussian noise through a random Gaussian matrix perturbation to the sample covariance matrix estimate is run, taking as input the "normalized" samples $A\boldsymbol{X}_i$, and by denoting $\widetilde{\Sigma}$ this estimate (which is explained in

the proof of Fact 8.3.7) and also $\widetilde{\boldsymbol{\nu}} = A\widehat{\boldsymbol{\mu}}$, we have that $\widehat{\Sigma} = A^{-1}\widetilde{\Sigma}A^{-1}$ and therefore with probability $1 - O(\gamma)$,

$$\left\|\Sigma^{1/2}\left(\widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right)^{-1}\Sigma^{1/2}\right\|_2^2 \leq \left\|\Sigma^{1/2}A\right\|_2^2 \cdot \left\|\left(\widetilde{\Sigma} + \widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^T\right)^{-1}\right\|_2^2 \cdot \left\|A\Sigma^{1/2}\right\|_2^2,$$

which gives the desired result, due to the following two facts.

Fact 8.3.6. *With probability* $1 - O(\gamma)$, $\left\|\Sigma^{1/2}A\right\|_2^2 = \left\|A\Sigma^{1/2}\right\|_2^2 \leq O(1)$.

Proof. Since A is a symmetric square matrix, $\left\|\Sigma^{1/2}A\right\|_2^2 = \left\|A\Sigma^{1/2}\right\|_2^2$. By using the definition of the spectral norm, the fact's statement is immediately obtained:

$$\left\|\Sigma^{1/2}A\right\|_2^2 = \sigma_{\max}^2\left(\Sigma^{1/2}A\right) = \lambda_{\max}\left(\left(\Sigma^{1/2}A\right)^T\Sigma^{1/2}A\right) = \lambda_{\max}\left(A\Sigma A\right) \leq 1000,$$

since by Lemma 8.3.5, $A\Sigma A \preceq 1000\mathbb{I}_d$ with probability $1 - O(\gamma)$. □

Fact 8.3.7. *With probability* $1 - O(\gamma)$, *when*

$$n = \Omega\left(d + \log(1/\gamma) + \frac{\sqrt{d}\text{polylog}\left(\frac{d}{\epsilon\gamma}\right)}{\epsilon}\right),$$

it holds that

$$\left\|\left(\widetilde{\Sigma} + \widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^T\right)^{-1}\right\|_2^2 \leq 1 + O\left(\sqrt{\frac{d + \log(1/\gamma)}{n}} + \frac{\sqrt{d}\log(1/\gamma)\log(n/\gamma)}{n\epsilon}\right).$$

Proof. In order to prove this fact, we need to delve further into the procedure by which $\widetilde{\Sigma}$ is generated (see Section 6.2.2), i.e., NaivePCE of Kamath et al. (2019). In short, we will use that $\widetilde{\Sigma}$ is the projection into the PSD cone of the empirical covariance matrix of the inputs (which are the vectors $A\boldsymbol{X}_1, \ldots, A\boldsymbol{X}_n$ where A is the above "preconditioner" matrix) plus a symmetric random matrix N of small Gaussian perturbations (that serves to enforce the privacy guarantee). Hence, the following holds:

$$\widetilde{\Sigma} = \text{proj}_{\text{PSD}}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T + N\right),$$

where $\boldsymbol{Z}_i = A\boldsymbol{X}_i$, and the matrix N is a symmetric random matrix with dimension $d \times d$ whose entries $N_{ij}, j \geq i$ are i.i.d. Gaussian random variables with zero mean and standard deviation $\sigma = \frac{d\log(n/\gamma)}{n\epsilon}$.

By Weyl's inequality for matrices, it is true for two real, symmetric matrices A, B and their sum $A+B$ that $\lambda_{\min}(A+B) \geq \lambda_{\min}(A) + \lambda_{\min}(B)$. Because $\left\|\left(\widetilde{\Sigma} + \widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^T\right)^{-1}\right\|_2^2 = \frac{1}{\lambda_{\min}^2\left(\widetilde{\Sigma} + \widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^T\right)}$, we will prove a lower bound about $\lambda_{\min}\left(\widetilde{\Sigma} + \widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^T\right) \geq \lambda_{\min}\left(\widetilde{\Sigma}\right)$ (since $\widetilde{\boldsymbol{\nu}}\widetilde{\boldsymbol{\nu}}^T$ is a PSD matrix) in the following way: we will show a bound about the minimum eigenvalue of the inner sum $\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T + N$, and argue that it is positive with high probability $1 - O(\gamma)$. This means that the projection of this matrix into the PSD cone is the same as the matrix itself with high probability, therefore the bound on the minimum eigenvalue will hold verbatim.

We begin by tailoring a lemma from random matrix theory referenced in Tao (2012) to the random matrix N:

Lemma 8.3.8 (Concentration of symmetric random matrices with Gaussian entries)**.** *Suppose that the entries N_{ij} for $j \geq i$ of a symmetric matrix N with dimensions $d \times d$ are i.i.d. Gaussian random variables with zero mean and variance σ^2. Then, there exist universal constants $C, c > 0$ such that the largest singular value of N satisfies for all $A \geq C$:*

$$\mathbf{Pr}\left[s_{max}(N) > A\sigma\sqrt{d}\right] \leq C\exp(-cAd)\,.$$

The above lemma means that with probability at least $1-\gamma$, we have that the largest singular value of N is at most

$$s_{\max}(N) \leq O\left(\frac{\sigma\log(1/\gamma)}{\sqrt{d}}\right).$$

Due to the matrix N being square symmetric with dimensions $d \times d$, it holds that its singular values are the absolute values of its eigenvalues, therefore $|\lambda_{\min}(N)|$ is one of the singular values of N (note that it could even be the largest), hence $|\lambda_{\min}(N)| \leq s_{\max}(N)$ and thus

$$\lambda_{\min}(N) \geq -s_{\max}(N) \geq -O\left(\frac{\sigma\log(1/\gamma)}{\sqrt{d}}\right) = -O\left(\frac{\sqrt{d}\log(1/\gamma)\log(n/\gamma)}{n\epsilon}\right), \tag{8.3}$$

since we remind the reader that $\sigma = \frac{d\log(n/\gamma)}{n\epsilon}$.

Lower bounds on the minimum eigenvalue of sample covariance matrices of the form are well-known in the literature, and we use here a version that appears in Diakonikolas et al. (2019a). Note that the vectors $\boldsymbol{Z}_i = A\boldsymbol{X}_i$, whose covariance matrix we are interested in, have a "normalized" distribution with covariance matrix $A\Sigma A^T = A\Sigma A$ (by the symmetry of A) that has at most a constant eigenvalue, since $A\Sigma A \preceq 1000\mathbb{I}_d$ by construction of the "preconditioner" A with probability $1-O(\gamma)$. Therefore, by classical covariance matrix estimation inequalities for eigenvalues of sample covariance matrices from Diakonikolas et al. (2019a), it holds that with probability $1-O(\gamma)$,

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^T\right) \geq 1 - O\left(\sqrt{\frac{d+\log(1/\gamma)}{n}}\right). \tag{8.4}$$

To conclude, we combine the lower bounds in eigenvalues of Equation (8.3) and Equation (8.4). By Weyl's inequality, with probability $1-O(\gamma)$, we have that:

$$\begin{aligned}\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^T + N\right) &\geq \lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^T\right) + \lambda_{\min}(N)\\ &\geq 1 - O\left(\sqrt{\frac{d+\log(1/\gamma)}{n}} + \frac{\sqrt{d}\log(1/\gamma)\log(n/\gamma)}{n\epsilon}\right).\end{aligned}$$

Choosing n such that the above lower bound is positive, i.e., if

$$n = \Omega\left(d + \log(1/\gamma) + \frac{\sqrt{d}\mathrm{polylog}\left(\frac{d}{\epsilon\gamma}\right)}{\epsilon}\right),$$

then the projection of the matrix sum into the PSD cone is equal to the matrix sum itself, therefore directly arriving at the final result by noting the additional fact that:

$$\left(\frac{1}{1-O(x)}\right)^2 \leq 1 + O(x)\,,$$

obtainable by a Taylor expansion since we chose n to be at least such that the denominator of the fraction is positive. □

By combining Fact 8.3.6 and Fact 8.3.7, we arrive at the final result of Claim 8.3.2.

Proof of Claim 8.3.3

Initially, breaking Q_1 as

$$\begin{aligned}
Q_1 &= \widehat{\Sigma} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T - \frac{1}{n}X^T X \\
&= \left(\widehat{\Sigma} - \Sigma\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)^T - \Sigma\right) \\
&\quad + \left(\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{\mu}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i - \boldsymbol{\mu}\right)^T - \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i - \boldsymbol{\mu}\right)\boldsymbol{\mu}^T\right) \\
&= \left(\widehat{\Sigma} - \Sigma\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)^T - \Sigma\right) - \boldsymbol{\mu}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i - \boldsymbol{\mu}\right)^T - \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i - \boldsymbol{\mu}\right)\boldsymbol{\mu}^T \\
&\quad + \left(\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)^T + \left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)\boldsymbol{\mu}^T + \boldsymbol{\mu}\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)^T\right),
\end{aligned}$$

it follows that

$$\begin{aligned}
\left\|\Sigma^{-1/2}Q_1\Sigma^{-1/2}\right\|_2^2 \leq\ & 2\left\|\Sigma^{-1/2}\left(\widehat{\Sigma} - \Sigma\right)\Sigma^{-1/2}\right\|_2^2 \\
& + 2\left\|\Sigma^{-1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)^T - \Sigma\right)\Sigma^{-1/2}\right\|_2^2 \\
& + 2\left\|\boldsymbol{\mu}\right\|_2^2 \cdot \left(2\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_i\right\|_2^2 + O\left(\left\|\Sigma^{-1/2}\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)\right\|_2^2\right)\right),
\end{aligned}$$

where we have used the variance-normalized vectors $\boldsymbol{V}_i = \Sigma^{-1/2}\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)$ which have covariance matrix $\mathbb{I}_d$.

We state and prove the two below facts which, along with Lemma 6.2.1, which we remind to the reader that it is applied to the $2n$ sample differences $\frac{1}{\sqrt{2}}\left(\boldsymbol{X}_{2i} - \boldsymbol{X}_{2i-1}\right)$, so that they have zero mean, and with a similar procedure to Fact 8.3.6 and Lemma 6.2.2, leads to the desired result immediately.

Fact 8.3.9. *For every $\alpha > 0$, with probability $1 - O(\gamma)$, when*

$$n = \Omega\left(\frac{d + \log(1/\gamma)}{\alpha^2}\right),$$

it holds that

$$\left\|\Sigma^{-1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)^T - \Sigma\right)\Sigma^{-1/2}\right\|_2^2 \leq O(\alpha^2).$$

Proof. First of all, we restate the left hand side of the inequality in terms of the variance-normalized vectors $\boldsymbol{V}_i$, as follows:

$$\left\|\Sigma^{-1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)^T - \Sigma\right)\Sigma^{-1/2}\right\|_2^2 = \left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_i\boldsymbol{V}_i^T - \mathbb{I}_d\right\|_2^2,$$

which we can bound with high probability by the classical empirical covariance estimation concentration bounds in Diakonikolas et al. (2019a). More specifically, with probability $1 - O(\gamma)$, we have that:

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_i\boldsymbol{V}_i^T - \mathbb{I}_d\right\|_2 = \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_i\boldsymbol{V}_i^T - \mathbb{I}_d\right) \leq O\left(\sqrt{\frac{d + \log(1/\gamma)}{n}}\right),$$

which directly implies the desired fact. □

Fact 8.3.10. *For every $\alpha > 0$, with probability $1 - O(\gamma)$, when*

$$n = \Omega\left(\frac{d + \log(1/\gamma)}{\alpha^2}\right),$$

it holds that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{V}_i \right\|_2^2 \leq O(\alpha^2).$$

Proof. The vector sum inside the desired term has a multivariate Gaussian distribution, and we can bound its ℓ_2-norm with high probability by the classical sub-gaussian concentration bounds in Diakonikolas et al. (2019a). More specifically, with probability $1 - O(\gamma)$, we have that:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{V}_i \right\|_2 \leq O\left(\sqrt{\frac{d + \log(1/\gamma)}{n}}\right),$$

which directly implies the desired fact. □

Directly combining Lemma 6.2.1, Lemma 6.2.2 with Fact 8.3.9 and Fact 8.3.10, one obtains the stated Claim 8.3.3.

Proof of Claim 8.3.4

Initially, breaking $\boldsymbol{Q}_2$ as

$$\boldsymbol{Q}_2 = \widehat{\boldsymbol{\mu}}' - \frac{1}{n} X^T \boldsymbol{y} = \left(\widehat{\boldsymbol{\mu}}' - \boldsymbol{\mu}'\right) - \left(\frac{1}{n} X^T \boldsymbol{y} - \boldsymbol{\mu}'\right),$$

it follows that

$$\left\|\Sigma^{-1/2} \boldsymbol{Q}_2\right\|_2^2 \leq \left\|\Sigma^{-1/2} A'^{-1}\right\|_2^2 \cdot \left\|A'\left(\widehat{\boldsymbol{\mu}}' - \boldsymbol{\mu}'\right)\right\|_2^2 + \left\| \frac{1}{n} \sum_{i=1}^{n} y_i \boldsymbol{V}_i - \Sigma^{-1/2} \boldsymbol{\mu}' \right\|_2^2,$$

where we have again used the notation of the variance-normalized vectors $\boldsymbol{V}_i = \Sigma^{-1/2} \boldsymbol{X}_i$, and the "private preconditioner" matrix A' (in this case, obtained for the mean estimation of the random vectors $y_i \boldsymbol{X}_i$), due to Lemma 8.3.5, as detailed below (see the first lines of the proof of Fact 8.3.11).

Before stating and proving the facts which lead to the desired result, we need to prove the sub-gaussianity of the vectors $y_i \boldsymbol{X}_i$ that are crucial for the conditions of Lemma 8.3.5 and the rest of our proof.

Proposition 8.3.1 (Sub-gaussianity of $y_i \boldsymbol{X}_i$)**.** *Let $\boldsymbol{X}_i$ be a random vector sampled according to a multivariate Gaussian distribution with mean value $\boldsymbol{\mu}$ and covariance matrix Σ such that $\mathbb{I}_d \preceq \Sigma$, and y_i be a random variable such that $\frac{1}{\rho} \leq |y_i| \leq c$. Then, $y_i \boldsymbol{X}_i$ are sub-gaussian random vectors with covariance matrix Σ' such that*

$$\frac{1}{\rho^2} \mathbb{I}_d \preceq \Sigma' \preceq c^2 \Sigma,$$

and sub-gaussian norm

$$\|y_i \boldsymbol{X}_i\|_{\psi_2} \leq c \|\boldsymbol{X}_i\|_{\psi_2}.$$

Proof. First of all, we have that $\Sigma' = \mathbb{E}\left[y_i^2 \boldsymbol{X}_i \boldsymbol{X}_i^T\right] \preceq c^2 \Sigma$, since the eigenvalues of $\mathbb{E}\left[(c^2 - y_i^2) \boldsymbol{X}_i \boldsymbol{X}_i^T\right]$ are non-negative, because the quantity inside the expectation is always a positive semi-definite matrix, and expectation is a linear operator, thus the eigenvalues of the matrix in expectation are also non-negative, by the Courant-Fischer min-max theorem.

Similarly, it can be seen that $\frac{1}{\rho^2}\mathbb{I}_d \preceq \frac{1}{\rho^2}\Sigma \preceq \Sigma'$. We proceed to prove the second part of the proposition, which is the sub-gaussian norm inequality.

We prove the sub-gaussianity of the desired vectors, by Definition 4.3.2: consider a unit vector $\boldsymbol{u} \in S^{d-1}$, then it holds that

$$\mathbb{E}\left[\exp(\lambda^2 y_i^2 \langle \boldsymbol{X}_i, \boldsymbol{u}\rangle^2)\right] \leq \mathbb{E}\left[\exp((\lambda c)^2 \langle \boldsymbol{X}_i, \boldsymbol{u}\rangle^2)\right] \leq \exp(\lambda^2 (cK)^2)\,,$$

for all $\lambda : |\lambda| \leq 1/(Kc)$, where K is the sub-gaussian norm of $\langle \boldsymbol{X}_i, \boldsymbol{u}\rangle$. The sub-gaussian norm follows:

$$\|y_i \boldsymbol{X}_i\|_{\psi_2} \leq c\|\boldsymbol{X}_i\|_{\psi_2}\,.$$

□

We are now ready to present the facts that lead to the claim.

Fact 8.3.11. *With probability* $1 - O(\gamma)$*, when*

$$n = \Omega\left(\frac{d^{3/2}\sqrt{\log(\kappa\rho c)}\mathrm{polylog}\left(\frac{d\log(\kappa\rho c)}{\gamma\epsilon}\right)}{\epsilon}\right)\,,$$

it holds that

$$\left\|\Sigma^{-1/2}A'^{-1}\right\|_2^2 \leq O(c^2)\,.$$

Proof. According to Proposition 8.3.1, the conditions of Lemma 8.3.5 apply to the variables $\rho y_i \boldsymbol{X}_i$ by a change of variables in the sample complexity of $\kappa' = \rho^2 c^2 \kappa$, where we remind to the reader that κ is the largest eigenvalue of the covariance matrix of the feature vectors: $\Sigma \preceq \kappa \mathbb{I}_d$.

Therefore, with

$$n = \Omega\left(\frac{d^{3/2}\sqrt{\log(\kappa\rho c)}\mathrm{polylog}\left(\frac{d\log(\kappa\rho c)}{\gamma\epsilon}\right)}{\epsilon}\right)\,,$$

we obtain a matrix A' (which is ρ times the A given by the algorithm of Lemma 8.3.5 as stated in the previous paragraph) such that with probability $1 - O(\gamma)$,

$$\mathbb{I}_d \preceq A'\Sigma' A' \preceq 1000\mathbb{I}_d\,. \tag{8.5}$$

Note that the knowledge of ρ is *not* required for Algorithm 2, since the change of variables that we did only affects the analysis that we performed here (the privacy of the algorithm is solely based on the upper bound c of the labels, and *not* on ρ).

Finally, the desired result holds with probability $1 - O(\gamma)$:

$$\left\|\Sigma^{-1/2}A'^{-1}\right\|_2^2 \leq \left\|\Sigma^{-1/2}\Sigma'^{1/2}\right\|_2^2 \cdot \left\|\Sigma'^{-1/2}A'^{-1}\right\|_2^2 \leq c^2 \cdot 1 = c^2\,,$$

since for the two quantities of interest we have separately the following:

By Proposition 8.3.1 and properties of the positive semi-definite order, we have that

$$\begin{aligned}
&\Sigma' \preceq c^2\Sigma \\
\Rightarrow\ &c^2\Sigma'^{-1} \succeq \Sigma^{-1} \\
\Rightarrow\ &\Sigma^{-1} - c^2\Sigma'^{-1} \preceq O \\
\Rightarrow\ &\Sigma'^{1/2}\Sigma^{-1}\Sigma'^{1/2} \preceq c^2\mathbb{I}_d \\
\Rightarrow\ &\left\|\Sigma^{-1/2}\Sigma'^{1/2}\right\|_2^2 \leq c^2\,.
\end{aligned}$$

At the same time, by Equation (8.5), we obtain the final term:

$$\left\|\Sigma'^{-1/2}A'^{-1}\right\|_2^2 = \frac{1}{\sigma_{\min}^2\left(\Sigma'^{1/2}A'\right)} = \frac{1}{\lambda_{\min}\left(\left(\Sigma'^{1/2}A'\right)^T\Sigma'^{1/2}A'\right)} = \frac{1}{\lambda_{\min}\left(A'\Sigma'A'\right)} \leq 1\,.$$

□

Fact 8.3.12. *For every $\eta > 0$, with probability $1 - O(\gamma)$, when*

$$n = \Omega\left(\frac{d\log(\frac{d}{\gamma})}{\eta^2} + \frac{d\text{polylog}(\frac{d\log(1/\delta)}{\eta\gamma\epsilon})}{\eta\epsilon} + \frac{\sqrt{d}\log(\frac{d}{\gamma\delta})}{\epsilon} + \frac{d^{3/2}\sqrt{\log(\kappa\rho c)}\text{polylog}\left(\frac{d\log(\kappa\rho c)}{\gamma\epsilon}\right)}{\epsilon}\right),$$

it holds that

$$\left\|A'\left(\widehat{\boldsymbol{\mu}}' - \boldsymbol{\mu}'\right)\right\|_2^2 \leq O(\eta^2)\,.$$

Proof. This fact is a direct implication of Proposition 8.3.1 which guarantees the conditions for Lemma 6.2.2 to hold. □

Fact 8.3.13. *For every $\eta > 0$, with probability $1 - O(\gamma)$, when*

$$n = \Omega\left(\frac{d + \log(1/\gamma)}{\eta^2}\right),$$

it holds that

$$\left\|\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{V}_i - \Sigma^{-1/2}\boldsymbol{\mu}'\right\|_2^2 \leq O(\eta^2)\cdot c^2\,.$$

Proof. We will prove that, under the stated conditions,

$$\left\|\frac{1}{n}\sum_{i=1}^n \frac{y_i\boldsymbol{V}_i}{c} - \Sigma^{-1/2}\frac{\boldsymbol{\mu}'}{c}\right\|_2 \leq O(\eta)\,,$$

and the result will follow.

First of all, we prove that $\frac{y_i\boldsymbol{V}_i}{c}$ is sub-gaussian and calculate a bound on its sub-gaussian norm. Similarly to the sub-gaussianity of Proposition 8.3.1, and by Lemma 4.3.1, we have that

$$\left\|\frac{y_i\boldsymbol{V}_i}{c}\right\|_{\psi_2} = \frac{\|y_i\boldsymbol{V}_i\|_{\psi_2}}{c} \leq C_1\left\|\boldsymbol{V}_i\right\|_{\psi_2} \leq C_2\,,$$

for some universal constants $C_1, C_2 > 0$, since $\boldsymbol{V}_i = \Sigma^{-1/2}\boldsymbol{X}_i$ are variance-normalized random vectors.

Then, noting that

$$\mathbb{E}\left[\frac{y_i\boldsymbol{V}_i}{c}\right] = \Sigma^{-1/2}\frac{\boldsymbol{\mu}'}{c}\,,$$

and by Lemma 4.3.1, it holds that the (centered) quantity $\frac{y_i\boldsymbol{V}_i}{c} - \Sigma^{-1/2}\frac{\boldsymbol{\mu}'}{c}$ is also sub-gaussian with sub-gaussian norm at most a constant times the sub-gaussian norm of the non-centered random vector $\frac{y_i\boldsymbol{V}_i}{c}$. Notice that the covariance matrix of the centered quantity above is $\preceq \mathbb{I}_d$. We will leverage this relationship, alongside the sub-gaussianity of the quantity, to prove the final concentration inequality, from which the fact follows:

Lemma 8.3.14. *There exist universal constants $A, B > 0$ such that, for all $t > 0$,*

$$\mathbf{Pr}\left[\left\|\frac{1}{n}\sum_{i=1}^n \frac{y_i\boldsymbol{V}_i}{c} - \mathbb{E}\left[\frac{y_i\boldsymbol{V}_i}{c}\right]\right\|_2 > t\right] \leq 4\exp\left(Ad - Bnt^2\right)$$

Proof. We denote the covariance matrix of $\frac{y_i \boldsymbol{V}_i}{c}$ as Σ'', for which it holds that $\Sigma'' \preceq \mathbb{I}_d$, and we also name the variance-normalized random vectors $(\Sigma'')^{-1/2} \frac{y_i \boldsymbol{V}_i}{c}$ as $\boldsymbol{W}_i$ (therefore, $\mathbb{E}[\boldsymbol{W}_i \boldsymbol{W}_i^T] = \mathbb{I}_d$).

By a classical result of sub-gaussian concentration inequalities (see, e.g., Lemma 2.21 of Diakonikolas et al. (2019a)), we have that there exist universal constants $A, B > 0$ such that, for all $t > 0$,

$$\mathbf{Pr}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W}_i - \mathbb{E}\left[\boldsymbol{W}_i\right]\right\|_2 > t\right] \leq 4\exp\left(Ad - Bnt^2\right) .$$

Additionally, by definition of the spectral norm, and since $\Sigma'' \preceq \mathbb{I}_d$, we have that:

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W}_i - \mathbb{E}\left[\boldsymbol{W}_i\right]\right\|_2 &\geq \left\|\left(\Sigma''\right)^{1/2}\right\|_2 \left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W}_i - \mathbb{E}\left[\boldsymbol{W}_i\right]\right\|_2 \\ &\geq \left\|\left(\Sigma''\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W}_i - \mathbb{E}\left[\boldsymbol{W}_i\right]\right)\right\|_2 ,\end{aligned}$$

namely, that:

$$\begin{aligned}\mathbf{Pr}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \frac{y_i \boldsymbol{V}_i}{c} - \mathbb{E}\left[\frac{y_i \boldsymbol{V}_i}{c}\right]\right\|_2 > t\right] &= \mathbf{Pr}\left[\left\|\left(\Sigma''\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W}_i - \mathbb{E}\left[\boldsymbol{W}_i\right]\right)\right\|_2 > t\right] \\ &\leq \mathbf{Pr}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W}_i - \mathbb{E}\left[\boldsymbol{W}_i\right]\right\|_2 > t\right] \\ &\leq 4\exp\left(Ad - Bnt^2\right) ,\end{aligned}$$

as desired. □

Utilizing Lemma 8.3.14, Fact 8.3.13 follows. □

Directly combining Fact 8.3.11, Fact 8.3.12, and Fact 8.3.13, we obtain the stated Claim 8.3.4.

8.4 Proof of Theorem 8.1.2

Again, for convenience, we restate here the (stonger) version of Theorem 8.1.2 that we will prove here:

Theorem 8.4.1 (Privacy and Accuracy of $\widehat{\boldsymbol{\beta}}$ in Private Binary Regression)**.** *Under Assumption 5.2.1 with covariance parameter κ and Assumption 5.3.1 with true parameter $\boldsymbol{\beta} \in \mathbb{R}^d$, for every privacy parameters $\epsilon, \delta > 0$, accuracy parameters $\alpha, \eta > 0$ and confidence $\gamma \in (0,1)$,* PrivateLearnLSE *(defined in Algorithm 2) with $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} = \mathbf{0}$ is $(\frac{\epsilon^2}{2} + \epsilon\sqrt{2\log(1/\delta)}, \delta)$-differentially private. Moreover, if the number of labeled examples is at least:*

$$\begin{aligned}n = O&\left(\frac{d\log(\frac{d}{\gamma})}{\eta^2} + \frac{d\text{polylog}(\frac{d\log(1/\delta)}{\eta\gamma\epsilon})}{\eta\epsilon} + \frac{d^{3/2}\sqrt{\log\kappa}\text{polylog}\left(\frac{d\log\kappa}{\gamma\epsilon\delta}\right)}{\epsilon}\right) \\ &+ O\left(\frac{d\log(\frac{d}{\gamma})}{\alpha^2} + \frac{d^{3/2}\text{polylog}\left(\frac{d\log(1/\delta)}{\alpha\gamma\epsilon}\right)}{\alpha\epsilon}\right) ,\end{aligned}$$

then with probability at least $1 - O(\gamma)$ the output estimate $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ satisfies

$$\|\widehat{\boldsymbol{\beta}} - k\boldsymbol{\beta}\|_2^2 \leq \|\widehat{\boldsymbol{w}} - k\boldsymbol{w}\|_2^2 \leq O\left(\alpha^2\right) \cdot \left(1 + \|k\boldsymbol{w}\|_2^2\right) + O\left(\eta^2\right) ,$$

where $\widehat{\boldsymbol{w}} = \Sigma^{1/2}\widehat{\boldsymbol{\beta}}$, $\boldsymbol{w} = \Sigma^{1/2}\boldsymbol{\beta}$ and $k = \frac{2n}{n-d-1}\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T \boldsymbol{X}_i\right)\right]$.

Proof. The privacy of the algorithm in this Theorem arises directly from the privacy of Theorem 8.1.1, since the algorithm is the same.

For the accuracy guarantee, as discussed in the Technical overview (Section 8.2), we first break the norm of the vector difference $\left\|\widehat{\boldsymbol{\beta}} - k\boldsymbol{\beta}\right\|_2^2$ into the distance from the estimate $\boldsymbol{\beta}_s^\star$, for which we remind to the reader that we define as

$$\boldsymbol{\beta}_s^\star = \left(\frac{1}{n}\sum_{i=n+1}^{2n} \boldsymbol{X}_i \boldsymbol{X}_i^T\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{X}_i\right) , \tag{8.1}$$

that resembles the Least Squares Estimate but *crucially introduces independence* between the two terms that constitute the Least Squares Estimate (to which we can apply our result from Theorem 8.1.1), and the distance of the Least-Squares-resembling estimate from a constant multiplicative factor of the true regression coefficient $\boldsymbol{\beta}$ (respectively, of the estimate $\boldsymbol{w}_s^\star$ from $\boldsymbol{w} = \Sigma^{1/2}\boldsymbol{\beta}$), as follows:

$$\|\widehat{\boldsymbol{w}} - k\boldsymbol{w}\|_2^2 \leq \|\widehat{\boldsymbol{w}} - \boldsymbol{w}_s^\star\|_2^2 + \|\boldsymbol{w}_s^\star - k\boldsymbol{w}\|_2^2 .$$

The first term gets bounded by Theorem 8.1.1, since as we also noted in the Technical overview (Section 8.2), the independence between Q_1 and $\boldsymbol{Q}_2$ in our proof of Theorem 8.1.1 allows us to prove the same claim for $\boldsymbol{\beta}_s^\star$ as we did for $\boldsymbol{\beta}^\star$ above (see Section 8.3). In the remainder of the proof, we focus on bounding the second term.

We first supply the following central Lemma, which uncovers the (unbiased up to a multiplicative factor) relation between the Least-Squares-resembling estimate $\boldsymbol{\beta}_s^\star$ and the true regression coefficient $\boldsymbol{\beta}$, following from Stein's Lemma:

Lemma 8.4.2. *There exists a multiplicative factor $k \in \mathbb{R}_+$ that depends on the model function f, where f as defined in Assumption 5.3.1, such that the estimate $\boldsymbol{\beta}_s^\star$, as in Equation (8.1), is an unbiased up to a multiplicative factor estimate of the true parameter $\boldsymbol{\beta}$ of Assumption 5.3.1, i.e.,*

$$\mathbb{E}\left[\boldsymbol{\beta}_s^\star\right] = k\boldsymbol{\beta} .$$

Proof. First, we note the following equality following from the definitions of Assumption 5.3.1:

$$\mathbb{E}\left[y_i | \boldsymbol{X}_i\right] = 2f\left(\boldsymbol{\beta}^T \boldsymbol{X}_i\right) - 1 . \tag{8.2}$$

Also, by a classical result on Wishart matrices (for instance, see Anderson (2003)), it is true that the inverse sample covariance matrix is proportional to the true covariance matrix for multivariate Gaussian random vectors:

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=n+1}^{2n} \boldsymbol{X}_i \boldsymbol{X}_i^T\right)^{-1}\right] = \frac{n}{n-d-1}\Sigma^{-1} . \tag{8.3}$$

By the independence of the first n samples $(1 \dots n)$ from the next $(n+1 \dots 2n)$, the law of iterated expectations, using Equation (8.2), Equation (8.3) and the zero-mean property of the feature vectors $\boldsymbol{X}_i$, we have that:

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\beta}_s^\star\right] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=n+1}^{2n} \boldsymbol{X}_i \boldsymbol{X}_i^T\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{X}_i\right)\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=n+1}^{2n} \boldsymbol{X}_i \boldsymbol{X}_i^T\right)^{-1}\right] \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \, \mathbb{E}[y_i | \boldsymbol{X}_i]\right] \\
&= \frac{n}{n-d-1}\Sigma^{-1}\, \mathbb{E}\left[\boldsymbol{X}_i \left(2f\left(\boldsymbol{\beta}^T \boldsymbol{X}_i\right) - 1\right)\right] \\
&= \frac{n}{n-d-1}\Sigma^{-1}\, \mathbf{Cov}\left[\boldsymbol{X}_i,\ 2f\left(\boldsymbol{\beta}^T \boldsymbol{X}_i\right) - 1\right] ,
\end{aligned}$$

Now, an application of Stein's Lemma (see Lemma 5.3.1), since $\boldsymbol{X}_i$ and $\boldsymbol{\beta}^T\boldsymbol{X}_i$ are jointly Gaussian, suggests that

$$\mathbf{Cov}\left[\boldsymbol{X}_i,\ 2f\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right)-1\right]=2\,\mathbf{Cov}\left[\boldsymbol{X}_i,\boldsymbol{\beta}^T\boldsymbol{X}_i\right]\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right)\right]=2\,\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right)\right]\ \Sigma\,\boldsymbol{\beta}\,,$$

and combining with the above equality yields

$$\mathbb{E}\left[\boldsymbol{\beta}_s^\star\right]=k\boldsymbol{\beta}\,,$$

where

$$k=\frac{2n}{n-d-1}\,\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T\boldsymbol{X}_i\right)\right]\,.$$

□

Continuing to the proof of the result, we use the form as written with the expectation, breaking the term into three sub-terms by adding and subtracting the same quantities (defining the variance-normalized vectors $\boldsymbol{V}_i=\Sigma^{-1/2}\boldsymbol{X}_i$), to deduce that

$$\begin{aligned}
\|\boldsymbol{w}_s^\star-k\boldsymbol{w}\|_2^2&=\|\boldsymbol{w}_s^\star-\mathbb{E}\left[\boldsymbol{w}_s^\star\right]\|_2^2\\
&=\left\|\Sigma^{1/2}\left(\frac{1}{n}\sum_{i=n+1}^{2n}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{X}_i\right)-\frac{n}{n-d-1}\Sigma^{-1/2}\,\mathbb{E}\left[y_j\boldsymbol{X}_j\right]\right\|_2^2\\
&\leq 2\left\|\left(\frac{1}{n}\sum_{i=n+1}^{2n}\boldsymbol{V}_i\boldsymbol{V}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{V}_i\right)-\left(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{V}_i\right)\right\|_2^2\\
&+2\left\|\left(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{V}_i\right)-\mathbb{E}\left[y_j\boldsymbol{V}_j\right]\right\|_2^2+2\left\|\frac{d+1}{n-d-1}\,\mathbb{E}\left[y_j\boldsymbol{V}_j\right]\right\|_2^2\\
&=2\left\|\left(\frac{1}{n}\sum_{i=n+1}^{2n}\boldsymbol{V}_i\boldsymbol{V}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=n+1}^{2n}\boldsymbol{V}_i\boldsymbol{V}_i^T-\mathbb{I}_d\right)\left(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{V}_i\right)\right\|_2^2\\
&+2\left\|\frac{1}{n}\sum_{i=1}^{n}\left(y_i\boldsymbol{V}_i-\mathbb{E}\left[y_j\boldsymbol{V}_j\right]\right)\right\|_2^2+2\left\|\frac{d+1}{n-d-1}\,\mathbb{E}\left[y_j\boldsymbol{V}_j\right]\right\|_2^2\,.
\end{aligned}$$

When we have at least as many samples as required for Theorem 8.1.1, it follows from the sub-proofs presented in the proof of this Theorem above (specifically, Fact 8.3.7, Fact 8.3.9, Fact 8.3.10, and Fact 8.3.13 of Section 8.3) that the whole quantity is bounded as follows:

$$\|\boldsymbol{w}_s^\star-\mathbb{E}\left[\boldsymbol{w}_s^\star\right]\|_2^2\leq O\left(\alpha^2\right)+O\left(\eta^2\right)\,.$$

□

Chapter 9

Conclusion

In an era of information explosion, the role of privacy and the right of every individual whose data is being processed to be assured that it will not be mishandled is assuming a central place along the design of algorithms. Differential Privacy is the main tool that has arisen to provide protection with strict, mathematically sound guarantees. Yet, because of the stringent nature of the constraints that are ingrained in the definition of Differential Privacy, it is typically hard to reason about unbounded quantities and how we could protect them. In this work, we provide and analyze estimates for inference in many interesting regression settings with unbounded covariates, formally proving that they are private and efficient.

We believe that the line of work on unbounded covariates is of great interest with respect to both theory and practice. Potential future research based on this work could emerge that pushes the frontier of unboundedness, for instance by relaxing the i.i.d. assumptions on the provided data (to account for potential dependencies among feature vectors). In addition, lower bounds in differentially private regression regimes are known to be evasive and sub-optimal (see, e.g., Wang (2018)), therefore examining possible lower bounds in unbounded regimes for regression-like environments would be another promising future direction. This will eventually result in highly interesting practical schemes applied in actual Machine Learning systems, with the grand potential of providing stronger privacy guarantees than existent implementations, even in cases where some individual data points will be unboundedly farther away than expected.

Limitations. Finally, we would like to note that privacy is one of the essential prerequisites of a modern Machine Learning system. Other important objectives to consider include fairness with respect to the system's users and societal benefit as a whole. Notice that there has been research indicating that privacy could adversely affect, or at least interfere with, some of these additional desiderata (Jagielski et al., 2019; Bagdasaryan et al., 2019). Such considerations fall outside of the scope of this work and we refer the interested reader to ongoing literature on fair and responsible machine learning (Barocas et al., 2019).

Βιβλιογραφία

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

I. Aden-Ali, H. Ashtiani, and G. Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. In V. Feldman, K. Ligett, and S. Sabato, editors, *Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 185–216. PMLR, 2021.

D. Alabi, A. McMillan, J. Sarathy, A. Smith, and S. Vadhan. Differentially private simple linear regression, 2020.

K. Amin, T. Dick, A. Kulesza, A. Munoz, and S. Vassilvitskii. Differentially private covariance estimation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/4158f6d19559955bae372bb00f6204e4-Paper.pdf.

T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J, 3rd ed edition, 2003. ISBN 9780471360919.

Anonymous. Review #2 of "Differentially Private Covariance Estimation", 2019. URL https://papers.nips.cc/paper/2019/file/4158f6d19559955bae372bb00f6204e4-Reviews.html.

Apple. Engineering privacy for your users, 2016. URL https://asciiwwdc.com/2016/sessions/709.

H. Asi and J. C. Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14106–14117. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a267f936e54d7c10a2bb70dbe6ad7a89-Paper.pdf.

E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. Differential privacy has disparate impact on model accuracy. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/fc0de4e0396fff257ea362983c2dda5a-Paper.pdf.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

A. F. Barrientos, J. P. Reiter, A. Machanavajjhala, and Y. Chen. Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, 28(2):440–453, 2019.

R. Bassily, A. Smith, and A. Thakurta. Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473, Oct. 2014. doi: 10.1109/FOCS.2014.56. ISSN: 0272-5428.

G. Bernstein and D. R. Sheldon. Differentially private bayesian linear regression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/f90f2aca5c640289d0a29417bcb63a37-Paper.pdf.

D. R. Brillinger. *A Generalized Linear Model With "Gaussian" Regressor Variables*, pages 589–606. Springer New York, New York, NY, 2012a. ISBN 978-1-4614-1344-8. doi: 10.1007/978-1-4614-1344-8_34. URL https://doi.org/10.1007/978-1-4614-1344-8_34.

D. R. Brillinger. The identification of a particular nonlinear time series system. In *Selected Works of David Brillinger*, pages 607–613. Springer, 2012b.

M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In M. Hirt and A. Smith, editors, *Theory of Cryptography*, pages 635–658, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg. ISBN 978-3-662-53641-4.

M. Bun, K. Nissim, U. Stemmer, and S. Vadhan. Differentially private release and learning of threshold functions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 634–649. IEEE, 2015.

M. Bun, K. Nissim, and U. Stemmer. Simultaneous private learning of multiple concepts. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, page 369–380, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340571. doi: 10.1145/2840728.2840747. URL https://doi.org/10.1145/2840728.2840747.

T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy in generalized linear models: Algorithms and minimax lower bounds, 2020.

C. Canonne. What is δ, and what δifference does it make?, 2021. URL https://differentialprivacy.org/flavoursofdelta/.

C. L. Canonne, G. Kamath, and T. Steinke. The discrete gaussian for differential privacy. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15676–15688. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/b53b3a3d6ab90ce0268229151c9bde11-Paper.pdf.

K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011. URL http://jmlr.org/papers/v12/chaudhuri11a.html.

C. Daskalakis, D. Rohatgi, and E. Zampetakis. Truncated linear regression in high dimensions. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10338–10347. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/751f6b6b02bf39c41025f3bcfd9948ad-Paper.pdf.

A. Dembo. Bounds on the extreme eigenvalues of positive-definite toeplitz matrices. *IEEE Transactions on Information Theory*, 34(2):352–355, 1988. doi: 10.1109/18.2651.

Z. Deng, A. Kammoun, and C. Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, page iaab002, Apr. 2021. ISSN 2049-8772. doi: 10.1093/imaiai/iaab002. URL https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iaab002/6209694.

I. Diakonikolas, M. Hardt, and L. Schmidt. Differentially private learning of structured discrete distributions. In *NIPS*, pages 2566–2574, 2015.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Robust Estimators in High-Dimensions Without the Computational Intractability. *SIAM Journal on Computing*, 48(2):742–864, Jan. 2019a. ISSN 0097-5397, 1095-7111. doi: 10.1137/17M1126680. URL https://epubs.siam.org/doi/10.1137/17M1126680.

I. Diakonikolas, W. Kong, and A. Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '19, page 2745–2754, USA, 2019b. Society for Industrial and Applied Mathematics.

C. Dwork and J. Lei. Differential privacy and robust statistics. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, page 371–380, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585062. doi: 10.1145/1536414.1536466. URL https://doi.org/10.1145/1536414.1536466.

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, Aug. 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

C. Dwork and A. Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2), 2010.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum. Differential privacy under continual observation. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC '10, page 715–724, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300506. doi: 10.1145/1806689.1806787. URL https://doi.org/10.1145/1806689.1806787.

C. Dwork, K. Talwar, A. Thakurta, and L. Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014.

M. A. Erdogdu. Newton-stein method: An optimization method for glms via stein's lemma. *Journal of Machine Learning Research*, 17(215):1–52, 2016. URL http://jmlr.org/papers/v17/16-062.html.

Google. Rappor (randomized aggregatable privacy preserving ordinal responses), 2013. URL https://www.chromium.org/developers/design-documents/rappor.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer series in statistics. Springer, New York, NY, 2nd ed edition, 2009. ISBN 9780387848570 9780387848587.

D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(none):1 – 6, 2012. doi: 10.1214/ECP.v17-2079. URL https://doi.org/10.1214/ECP.v17-2079.

R. Iyengar, J. P. Near, D. Song, O. Thakkar, A. Thakurta, and L. Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316, 2019. doi: 10.1109/SP.2019.00001.

M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. S. Malvajerdi, and J. Ullman. Differentially private fair learning. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3000–3008. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/jagielski19a.html.

P. Jain and A. Thakurta. (near) dimension independent risk bounds for differentially private learning. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page I–476–I–484. JMLR.org, 2014.

G. Kamath, J. Li, V. Singhal, and J. Ullman. Privately learning high-dimensional distributions. In A. Beygelzimer and D. Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1853–1902, Phoenix, USA, 25–28 Jun 2019. PMLR. URL http://proceedings.mlr.press/v99/kamath19a.html.

V. Karwa and S. Vadhan. Finite Sample Differentially Private Confidence Intervals. In A. R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 44:1–44:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-060-6. doi: 10.4230/LIPIcs.ITCS.2018.44. URL http://drops.dagstuhl.de/opus/volltexte/2018/8344.

D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. In S. Mannor, N. Srebro, and R. C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 25.1–25.40, Edinburgh, Scotland, 25–27 Jun 2012. JMLR Workshop and Conference Proceedings. URL http://proceedings.mlr.press/v23/kifer12.html.

G. R. Kini and C. Thrampoulidis. Analytic Study of Double Descent in Binary Classification: The Impact of Loss. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2527–2532, Los Angeles, CA, USA, June 2020. IEEE. ISBN 9781728164328. doi: 10.1109/ISIT44484.2020.9174344. URL https://ieeexplore.ieee.org/document/9174344/.

S. M. Kreidler, B. M. Ringham, K. E. Muller, and D. H. Glueck. Calculating power for the general linear multivariate model with one or more Gaussian covariates. *Communications in Statistics - Theory and Methods*, Feb. 2018. ISSN 0361-0926. URL https://www.tandfonline.com/doi/full/10.1080/03610926.2018.1433849.

T. Kulkarni, J. Jälkö, A. Koskela, S. Kaski, and A. Honkela. Differentially private bayesian inference for generalized linear models, 2021.

A. M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes: avec un supplément contenant divers perfectionnemens de ces méthodes et leur application aux deux comètes de 1805*. Courcier, 1806.

J. S. Liu. Siegel's formula via stein's identities. *Statistics & Probability Letters*, 21(3):247–251, 1994.

E. Ma and C. Zarowski. On lower bounds for the smallest eigenvalue of a hermitian positive-definite matrix. *IEEE Transactions on Information Theory*, 41(2):539–540, 1995. doi: 10.1109/18.370166.

F. McSherry and I. Mironov. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–636, 2009.

P. Nakkiran. More data can hurt for linear regression: Sample-wise double descent, 2019.

A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, 2008. doi: 10.1109/SP.2008.33.

R. L. Plackett. A Historical Note on the Method of Least Squares. *Biometrika*, 36(3/4):458, Dec. 1949. ISSN 00063444. doi: 10.2307/2332682. URL https://www.jstor.org/stable/2332682?origin=crossref.

O. Rivasplata. Subgaussian random variables: An expository note. Technical report, University of Alberta, Dept. of Math. and Stat. Sciences, 11 2012.

O. Sheffet. Differentially private ordinary least squares. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3105–3114. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/sheffet17a.html.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

T. Steinke and J. Ullman. Why privacy needs composition, 2020. URL https://differentialprivacy.org/privacy-composition/.

Y. Sun, S. Ioannidis, and A. Montanari. Learning mixtures of linear classifiers. In *International Conference on Machine Learning*, pages 721–729. PMLR, 2014.

T. Tao. *Topics in random matrix theory*. Number v. 132 in Graduate studies in mathematics. American Mathematical Society, Providence, R.I, 2012. ISBN 9780821874301.

US Census Bureau. 2020 census data products: Disclosure avoidance modernization, 2020. URL https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html.

R. Vershynin. *High-dimensional probability: an introduction with applications in data science*. Number 47 in Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge ; New York, NY, 2018. ISBN 9781108415194.

Y. Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In A. Globerson and R. Silva, editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 93–103. AUAI Press, 2018. URL http://auai.org/uai2018/proceedings/papers/40.pdf.

J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett. Functional mechanism: Regression analysis under differential privacy. *Proc. VLDB Endow.*, 5(11):1364–1375, July 2012. ISSN 2150-8097. doi: 10.14778/2350229.2350253. URL https://doi.org/10.14778/2350229.2350253.

J. Zhang, K. Zheng, W. Mou, and L. Wang. Efficient private erm for smooth objectives. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3922–3928, 2017. doi: 10.24963/ijcai.2017/548. URL https://doi.org/10.24963/ijcai.2017/548.