UNIVERSITY OF AMSTERDAM

MSc Artificial Intelligence
Master Thesis

Optimal lossy compression for differentially private data release

by

Jonas Köhler

11400358

November 26, 2018

36 ECTS
Jan 2018 – Oct 2018

Supervisor:
Mijung Park

Assessor:
Efstratios Gavves

Conducted during a research internship at the Max-Planck-Institute for Intelligent Systems in Tuebingen, Germany.

Abstract

Differential privacy (DP) is the mathematical guarantee that the output of an algorithm operating on a data set has only a small dependency on any single individual in that set. A particular application of DP is the task of releasing privatized representations of data sets which should not be leaked to the public in their original form. Finding the right transformation of a data set, such that it is provably privatized but still preserves utility for the analysis task at hand, is a difficult task in theory and also practically unsolved for many applications. During this thesis we present an information theoretic approach to design DP data set release mechanisms, by reducing the problem to optimally compressing the data with respect to a measure of utility. As the optimal compression problem is inherently difficult to solve by itself, we analyze this approach for two linear instances of optimal compression for which an analytic solution exists and thus the analysis and sampling of privatized data sets is tractable. We further show in experiments that both methods cannot yield privacy/utility trade-offs that allow them to be used in practical tasks. We finish this research by proposing more complex approaches following this framework that are based on approximations and sampling methods and discuss their strengths and weaknesses.

Declaration

I declare to take full responsibility for the contents of this document.

I declare that the text and the work presented in this document are original and that no sources other than those mentioned in the text and its references have been used in creating it.

Jonas Köhler, Berlin, 20th November 2018.

Contents

1. Introduction

In this thesis, we will discuss the problem of preparing a data set such that the privacy of individuals in the set is not compromised and such that the prepared data set still possesses usefulness for some task. In particular we will study a private data release mechanism that is based on the idea of optimal compression.

Within this introduction we will motivate the need for a formal privacy guarantee, introduce differential privacy (DP) as a very strong guarantee, briefly present common facts about DP and further present three common use cases. We will then explain the problem of private data release as one of these use cases, discuss related work that has been conducted in this area of research and finish with listing the contributions of this thesis.

The next chapter will present our framework for private data release which is based on rate-distortion theory. The following two chapters will show two tractable instances of this framework. In chapter 5 we will analyze their privacy guarantees and how these relate to information theoretic quantities. Chapter 6 will present experiments in which we study the practical feasibility of using the derived methods for real-world data release. In the final chapter, we will propose three more advanced ways to approach privacy by compression, discuss their benefits and disadvantages and, for practical purposes, leave a feasibility test as an open question for future work.

1.1 Privacy in machine learning

Recent successes of machine learning in application domains such as computer assisted diagnostics or credit card scoring are based on the analysis of large amounts of data which might contain highly sensitive information about individuals. Especially in domains where scarcity of data is the major challenge for real-world applications (e.g. as frequently encountered when applying machine learning to medical domains), this can be a crucial hindrance to success. While many privacy concerns are justified, they make collaboration, data sharing and the statistical analysis of data across e.g. hospitals, corporations, governmental institutions and individuals difficult.

1.1.1 An utilitarian dilemma

This can be exemplified by the case of early-stage diagnosis of rare diseases. Many hospitals might not observe more than few instances of this disease and thus can hardly build any predictive modeling tools based on this knowledge. However, taken all hospitals together this amount of data might be sufficient in order to fine-tune a pre-trained model. It is reasonable to assume that such a diagnosis model should be developed by an independent provider who specialized in applying machine learning to medical domains, and potentially has access to a much stronger pool of resources and specialists than each individual hospital. However, simply

sharing data across medical institutions can already be a regulatory difficulty, thus it is unlikely to easily gain access to this data for any third party. As a consequence we are faced with a moral dilemma: we either might need to compromise the privacy of individual patients or we might not be able to create diagnosis models that could help to save the life of future patients. A possible remedy to this dilemma is given by *privatizing* of the data before handing it over to the external party. This means that we want to preserve the statistical information in the data required for a working diagnosis model, but still not to disclose information by which e.g. a malicious user would be able to infer that a specific person has been affected by the disease.

1.1.2 Privacy is not straightforward

Achieving privatization of a data set is not an easy task. While it is clear that full privacy can always be achieved by removing *all* information, it is more interesting how we will not compromise an individual and still provide some statistical value. In the next two paragraphs we will show why straightforward approaches to circumvent the dilemma might not help to achieve privacy at all.

Masking identifiers is not enough A simple approach to privatizing data is given by just removing all identifiers, e.g. the individual's name. This method, also called *pseudonymization*, is still used in practice, even though it can be shown that it does not imply strong privacy guarantees for an affected individual. As records contain more and more features, they become sufficient to identify individuals e.g. by matching the attributes to another database which contains the true identifier. As an example, knowing that a person is of German-Indonesian descent, living in Amsterdam, playing the Harpsichord and is part of a Rugby team might already be sufficient to reduce the number of individuals in question to a very small number. This breach of privacy by re-identification is called the *linkage attack*[17] and there are already a couple of examples where this became prominent in the past: in 1997 the MIT graduate student Latanya Sweeney has been able to re-identify the medical records of Massachusett's governor William Weld by linking public data to the pseudonymized records [7].

Complex models might learn too much Another frequently proposed approach is training the model at the client's site and only handing over inferred model parameters for the online task. However, this might imply another privacy breach when the trained models consist of a tremendous amount of parameters. As recent work shows, deep learning models containing billions of parameters are able to over-fit to anything, even to completely random labels [55]. It is thus a reasonable question to ask how much information about individuals is contained in the weights if such a model has been trained on sensitive information. In fact, other recent research showed the possibility to attack such models. You can infer whether an instance has been part of the training data [46] and - even worse - reconstruct individuals from the training set [21].

1.1.3 Requirements for a formal privacy model

These examples show that a strict notion of privacy is required that is even hiding whether an individual was present in the data set or not. Furthermore, we need a quantitative measure that allows to compare two methods by their privacy and how this privacy affects usefulness. Besides that there are further requirements which emerged to be useful for practical privacy. The two most important ones are *composability* and *post-processing invariance*.

Composability The analysis of complex algorithms requires breaking them down to simple building blocks. If the privacy guarantee of parts can be composed to an overall privacy guarantee, we can prove privacy for each component and then aggregate it to prove privacy for the whole algorithm.

Post-processing invariance Additionally, we want to ensure that any post-processing of something private cannot disclose more information as given by the privacy guarantee. This can be exemplified with an encrypted medical record. Without decryption key such a record would be useless. However, once the key gets into the hand of a malicious user, it would reveal all information and compromise all privacy. A procedure that possesses post-processing invariance would rule out such edge-case.

1.2 Differential privacy and its approximation

There are many notions of privacy (e.g. see [49] for an up to date survey). The most prominent however is given by *differential privacy* [17] (DP). This guarantee requires that algorithms operating on data sets consisting of nearly the same individuals should yield similar results with a high probability. In this case the dependency of the algorithm's output on a particular individual is low and does not leak information. This can be made precise by the following definition.

Definition 1.2.1 (Differential Privacy [17])**.** Let $\mathcal{A}$ be a randomized algorithm that takes a data set $\mathbf{X} \in \mathcal{X}^n$ from a domain $\mathcal{X}$ and produces an output $\mathcal{A}(\mathbf{X}) \in \mathcal{Y}$ from an output set $\mathcal{Y}$. Let $P_{\mathcal{A}}$ denote the probability measure induced on $\mathcal{Y}$ by the randomness of $\mathcal{A}$. Then $\mathcal{A}$ is *ϵ-differentially-private* (ϵ-DP) iff for all $\mathbf{y} \in \mathcal{Y}$, for all possible data sets $\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{X}^n$ which differ only in one entry (i.e. written as sets $|\mathbf{X}_1 \Delta \mathbf{X}_2| = 1$) and all possible $P_{\mathcal{A}}$-measurable subsets $B \subset \mathcal{Y}$

$$\left| \log \left(\frac{P_{\mathcal{A}}\left(\mathcal{A}(\mathbf{X}_1) \in B\right)}{P_{\mathcal{A}}\left(\mathcal{A}(\mathbf{X}_2) \in B\right)} \right) \right| \leq \epsilon \tag{1.1}$$

First, note that for $\epsilon \to 0$ both output distributions must be identical. This directly implies that $\mathcal{A}$ must be insensitive to the input at all. For $\epsilon \gg 1$ this definition becomes loose very quickly. We will illustrate this abstract definition by a small example.

Example 1: Privacy on a scale Assume we have the little village Mediocristan. Most of the villagers are in good shape and possess a body weight of around 70kg. The only exception is Rudy, who suffers from a rare metabolic disorder and weighs twice of that. For a national health survey a doctor visits the village in order to measure the average weight of its population. He will select 100 villagers randomly, measure their weight and finally report that average. If Rudy was part of a sample, the resulting average would equate 70.7kg. If he was absent it would have become 70kg. Thus reporting that average discloses the information, whether Rudy was part of the sample or not. We conclude that the algorithm `average_simple` should not be private according to definition 1.2.1.

1.2.1 Achieving ϵ-DP: Laplace mechanism

A common method to achieve ϵ-DP for counting queries is given by the *Laplace mechanism*[17]. Let $\mathcal{A}'$ be the original, non private, algorithm. We first analyze its *sensitivity* according to the

L_1-norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^{D} |x_i|$. The sensitivity of a function $f\colon \mathcal{X}^N \to \mathbb{R}^D$ with respect to a norm $\|\cdot\|$ is given by

$$\Delta f = \sup_{\mathbf{X},\ \mathbf{X}'\ :\ |\mathbf{X}_1 \Delta \mathbf{X}_2|} \|f(\mathbf{X}) - f(\mathbf{X}')\|. \tag{1.2}$$

Then we sample Laplacian noise according to $\xi \sim \text{Lap}\left(0, \frac{\Delta \mathcal{A}'}{\epsilon}\right)$ and define a new algorithm by setting

$$\mathcal{A}(\mathbf{X}) = \mathcal{A}'(\mathbf{X}) + \xi. \tag{1.3}$$

As shown in [17], theorem 3.6 this noised-up version $\mathcal{A}$ satisfies definition 1.2.1. The magnitude of Laplacian noise reveals a fundamental *privacy/utility* trade-off: for privacy we want to keep ϵ low. However, for utility we require only a small amount of perturbation.

Example 2: back to Mediocristan The national health service decided to rerun the survey using DP techniques. The doctor will again sample 100 random villagers, measure their weight and compute the average using `average_simple`. However, now there is a prior assumption that people weighing more than 200kg are unlikely. So measured weights beyond that value are clipped to 200kg. As a result `average_simple` has sensitivity of 2. By adding Laplacian noise with magnitude 2 to the output of `average_simple` we obtain the 1-DP algorithm `average_laplace`. When the doctor reports the noisy average 70.7kg there are now two possible explanations. Either Rudy was part of the sample, or the additional 0.7 are a result of the Laplacian noise. The likelihood of that value would increase by a moderate factor of $\exp(1) \approx 2.7$ if he was in the sample. Thus, with only one observation, it becomes difficult for an attacker to guess his participation. To see the privacy/utility tradeoff at work, consider a sample size of 1. In this case `average_simple` has sensitivity 200. Using the Laplace mechanism to achieve 1-DP requires Laplacian noise addition with magnitude 200. This clearly destroys all utility of taking the average. The example illustrates that controlling sensitivity is crucial for DP to work in practice.

1.2.2 Relaxation of differential privacy

Definition 1.2.1 is difficult to handle in practice, due to the strong supremum bound which has to hold for each possible pair of adjacent data sets. Thus many relaxations have been proposed to weaken the definition. This can either be done by demanding the guarantee to hold with high probability, either over the algorithm's output or over the data source itself, or by weakening the strong point-wise bound into much looser average bounds. Additionally, the Laplace mechanism complicates the privacy/utility analysis for many applications and leads to strong perturbation especially in high dimensions. This motivated the search for an approximation of DP that includes the addition of Gaussian noise, while still providing strong guarantees. This so-called (ϵ, δ)-*approximate differential privacy* is given by

Definition 1.2.2 (Approximate differential privacy [17])**.** Let $\mathcal{A}$ be a randomized algorithm that takes a data set $\mathbf{X} \in \mathcal{X}^n$ from a domain $\mathcal{X}$ and produces an output $\mathcal{A}(\mathbf{X}) \in \mathcal{Y}$ from an output set $\mathcal{Y}$. Let $P_{\mathcal{A}}$ denote the probability measure induced on $\mathcal{Y}$ by the randomness of $\mathcal{A}$. Then $\mathcal{A}$ is (ϵ, δ)-*differentially-private* ((ϵ, δ)-DP) iff for all $\mathbf{y} \in \mathcal{Y}$, for all possible data sets $\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{X}^n$ with $|\mathbf{X}_1 \Delta \mathbf{X}_2| = 1$ and all possible $P_{\mathcal{A}}$-measurable subsets $B \subset \mathcal{Y}$

$$\left| \log \left(\frac{P_{\mathcal{A}}\left(\mathcal{A}(\mathbf{X}_1) \in B\right) - \delta}{P_{\mathcal{A}}\left(\mathcal{A}(\mathbf{X}_2) \in B\right)} \right) \right| \leq \epsilon \tag{1.4}$$

For $\epsilon > 0$ and $\delta \ll 1$ this guarantee is very similar to 1.2.1. In fact, for $\delta \to 0$ both definitions become the same. However, if δ gets too big, the definition includes algorithms with very poor privacy, even if $\epsilon = 0$:

1. sample a number $u \in (0, 1)$ uniformly
2. if $u > \delta$ release a dummy symbol
3. if $u < \delta$ release the raw patient record.

This algorithm releases *any data* into the wild with a probability of δ. If δ is very small it cannot be useful, however if δ is not very small it strictly violates privacy.

1.2.3 Achieving (ϵ, δ)-DP: the Gaussian mechanism

Approximate differential privacy can be achieved by adding Gaussian noise to the output of a non-private algorithm. This procedure is called *Gaussian mechanism*[17] and is given by

Theorem 1.2.1 (Theorem 3.8. in [17]). *Let $f\colon \mathcal{X}^N \to \mathcal{Y}$ be a function with L_2 sensitivity Δf. Let $\epsilon \in (0, 1), \delta \in [0, 1]$. For a constant $c^2 > 2\log(1.25/\delta)$ let*

$$\mathcal{A}(\mathbf{X}) = f(\mathbf{X}) + \boldsymbol{\xi} \tag{1.5}$$

be an algorithm where the perturbation is sampled $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \nu^2\mathbf{I})$ with noise magnitude

$$\nu \geq \frac{c\Delta f}{\epsilon}. \tag{1.6}$$

Then $\mathcal{A}$ is (ϵ, δ)-DP.

This original result has been extended to weaker guarantees with $\epsilon \geq 1$. This so-called *improved Gaussian mechanism* [6] has been shown to be optimal in the sense of

Theorem 1.2.2 (Theorem 8 in [6]). *Let $f\colon \mathcal{X}^N \to \mathcal{Y}$ a function with L_2 sensitivity Δf. For any $\epsilon \geq 0$ and $\delta \in [0, 1]$ the algorithm $\mathcal{A}(\mathbf{x}) = f(\mathbf{x}) + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \nu^2\mathbf{I})$ is (ϵ, δ)-DP if and only if*

$$\Phi\left(\frac{\eta - \epsilon}{\sqrt{2\eta}}\right) - \exp(\epsilon)\Phi\left(\frac{-\eta - \epsilon}{\sqrt{2\eta}}\right) \leq \delta, \tag{1.7}$$

where $\eta = \frac{(\Delta f)^2}{2\nu^2}$ and $\Phi(\cdot)$ is the CDF of the standard normal distribution. .

Theorem 1.2.2 allows to find tight lower bounds for methods where the DP guarantee relies on Gaussian perturbation. The data set release method that we discuss in this thesis utilizes another privacy mechanism for its motivation. However, the two instances that we will present can be interpreted as instances of the Gaussian mechanism. Using theorem 1.2.2 allows a tighter analysis of their privacy/utility trade-off as we will see later.

1.2.4 Post-processing invariance

Post-processing invariance has ben shown for both DP and approximate DP: if $\mathcal{A}$ is an algorithm satisfying ϵ-DP ((ϵ, δ)-DP) and $g\colon \mathcal{Y} \to \mathcal{Z}$ is any probabilistic function of its output, then the algorithm $\mathcal{A}' = g \circ \mathcal{A}$ is ϵ-DP ((ϵ, δ)-DP) private as well (Proposition 2.1. in [17]). Without any further access to the original data we cannot obtain more information helping to undo the noise addition.

Example 3 If the doctor computed `average_laplace` 50 times using the same sample, he could average the results to approximate the true mean with very high accuracy and thus again compromise privacy. If he evaluates `average_laplace` just once, he could only guess the participation of Rudy with a chance of success as provided by ϵ.

1.2.5 Composition theorems

DP further possesses composition theorems. A simple composition theorem is *additive composition* [17]. If $\mathcal{A}_1$ and $\mathcal{A}_2$ are (ϵ_1, δ_1)-DP, (ϵ_2, δ_2)-DP algorithms and $\mathcal{A}$ be an algorithm that runs $\mathcal{A}_1$ and $\mathcal{A}_2$ as a subroutine its output will be $(\epsilon_1 + \epsilon_1, \delta_1 + \delta_2)$-DP. In followup work the composition has been tightened using *advanced composition* [18] and the recently proposed moment's accountant [1].

Example 4 The composition theorem explains, why multiple evaluations of `average_laplace` will again impose a high privacy risk. Using the additive composition we obtain $\epsilon = 50$ for the iterated evaluation. Thus the likelihood to detect Rudy's participation is increased by a factor of $\exp(50) \approx 5,184705529 \cdot 10^{21}$.

1.2.6 Applying DP in practice

In this section, we present three major applications of DP in practice.

Release of aggregate statistics

The classic application is the release of *summary statistics* of the data, e.g. the mean, the count of occurrence of a specific attribute or higher order statistics like variance. In the sense of definition 1.2.1 the algorithm $\mathcal{A}$ takes a data set and estimates the wanted statistics. There has been plenty of literature about releasing private aggregate statistics (see e.g. [16] for a survey) and we will not further deepen this application within this thesis.

Parameter estimation

The second application is privately estimating model parameters. Here we assume a parameterized family of functions $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, and a loss function that describes how well a hypothesis performs for a task (e.g. the cross entropy for a classification task, or the negative log likelihood in the case of generative modeling). Given a sample $\mathbf{X}$ of the data source, we aim to infer a $\boldsymbol{\theta}$, such that f_θ will also perform well for an unseen sample of the same source (in the context of supervised learning) or that f_θ allows to formulating a generative model such that samples from this model and a fresh sample from the source are distributed as similar as possible. Any algorithm $\mathcal{A}$ that performs the inference of $\boldsymbol{\theta} = \mathcal{A}(\mathbf{X})$ is traditionally called a *learner*. In the case of training deep neural networks via error back-propagation this learner is a gradient based method. If $\mathcal{A}$ satisfies DP, it implies that the distribution over the possibly inferred model parameters $\boldsymbol{\theta}$ will change only slightly if only one sample in $\mathbf{X}$ has been changed. While it is possible to achieve this guarantee for classic learning algorithms (e.g. see [26] for a survey), it is still a major challenge to obtain similar guarantees for learners that infer the parameters of deep neural networks. Recent work derived (ϵ, δ)-DP learning algorithms that perform moderately well for small image recognition tasks [1, 42] and language models [36]. Achieving high-quality private learners that estimate neural network parameters with quality similar to non-private counterparts still poses as a major challenge and remains an open question.

Releasing private data sets

The third class of applications of differential privacy deals with privatized feature representations of the full data set or even fully synthesized proxy data sets. This will be the major focus in this thesis and we will discuss in more detail in the following section.

1.3 Private data release

We look at the problem of an owner of private data who aims to prepare it for public release in a way that important statistical properties stay preserved while keeping the privacy risk for individuals low. This problem can split into two major approaches: learning a generative model in a private way, that allows to sample proxy data sets or transforming the data into a privatized representation. The latter perspective will be the major focus of this thesis. In both settings we assume that the data stems from a domain $\mathcal{X}$ and is distributed according to a source distribution $P_{\mathbf{X}}$. Assuming N different iid samples of this source, every data set can be seen as a sample $\mathbf{X} \sim P_{\mathbf{X}}^N$ of the product distribution. We further assume that we want to release either the transformed original data set or the generated proxy data set by representing it in a feature domain $\mathcal{T}$ (note that we could also have $\mathcal{T} = \mathcal{X}$). Without privacy in mind, we could think of this as a (potentially probabilistic) feature transformation $F(\mathbf{X}) = \mathbf{T}$. This map allows to define a distribution $P_{\mathbf{T}}$ over the feature representations $\mathbf{T}$.

1.3.1 Releasing proxy data from generative models

In this framework we aim to model $P_{\mathbf{T}}$ using a surrogate distribution $Q_{\mathbf{T}|\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ and then release a proxy data set $\mathbf{T} \sim Q_{\mathbf{T}|\boldsymbol{\theta}}^M$ (note that we could have $M \neq N$). The parameters $\boldsymbol{\theta}$ will have to be estimated from the sample $\mathbf{X}$ using a randomized algorithm $\mathcal{A}$. We can express this procedure as the Markov chain

$$\mathbf{X} \underset{\boldsymbol{\theta} \sim \mathcal{A}(\boldsymbol{\theta}|\mathbf{X})}{\longrightarrow} \theta \underset{\mathbf{t} \sim Q_{\mathbf{T}|\boldsymbol{\theta}}}{\longrightarrow} \mathbf{T}. \tag{1.8}$$

The post-processing invariance of DP implies, that we can sample an arbitrarily big proxy data set in feature representation $\mathcal{T}$ guaranteeing DP as long as $\mathcal{A}$ is a DP learning algorithm. However, we also see that the privacy/utility trade-off of such models entirely depends on the trade-off induced by the estimator $\mathcal{A}$. Thus, we can classify this approach to data release as instance of private parameter estimation sharing the same benefits and drawbacks.

1.3.2 Transforming data sets into privatized representations

In this framework we aim to find a noisy feature transformation F such that the release of $\mathbf{T} = F(\mathbf{X})$ is DP itself. During this thesis we will concentrate on feature transformations with trivial joint distributions $P_{\mathbf{X},\mathbf{T}}$. This allows us to analyze F as a point-wise noisy transformation of single data points and thus the conditional distribution $P_{\mathbf{T}|\mathbf{X}}$ factorizes over the transformed data points. If the induced point-wise conditional distribution $P_{\mathbf{T}|\mathbf{X}}$ does not vary too much for different $\mathbf{x} \sim P_{\mathbf{X}}$ then releasing $F(\mathbf{X})$ is private, as can quickly be shown by

Lemma 1.3.1. *Let $P_{\mathbf{X},\mathbf{T}}$ be a joint distribution over $\mathcal{X} \times \mathcal{T}$. If for each $S \subset \mathcal{T}$ and each $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$*

$$\log \frac{P_{\mathbf{T}|\mathbf{X}=\mathbf{x}}[S]}{P_{\mathbf{T}|\mathbf{X}=\mathbf{x}'}[S]} \leq \epsilon \tag{1.9}$$

then releasing $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_N)$ where $\mathbf{t}_i \sim P_{\mathbf{T},\mathbf{X}=\mathbf{x}_i}$ is ϵ-DP.

The feature transformation F can now be *data independent*, e.g. if $\mathbf{x}$ is transformed to $\mathbf{t}$ using a random projection, or it can be *data dependent*, which involves parameter estimation using an estimator $\mathcal{A}_{\boldsymbol{\theta}|\mathbf{X}}$. If the parameters of the feature transformation can be estimated on an independent, (possibly smaller) public sample $\mathbf{X}' \sim P_{\mathbf{X}}$, the privacy guarantees of this release mechanism solely depend on the point-wise transformation f. Thus $\mathcal{A}$ could be any non-private estimator:

$$\begin{aligned} \mathbf{X}' &\underset{\boldsymbol{\theta}\sim\mathcal{A}_{\boldsymbol{\theta}|\mathbf{X}}}{\longrightarrow} \boldsymbol{\theta} \\ \mathbf{X}, \boldsymbol{\theta} &\underset{\mathbf{T}\sim F_{\mathbf{X},\boldsymbol{\theta}}}{\longrightarrow} \mathbf{T} \qquad \Rightarrow (\epsilon_f, \delta_f)\text{-DP}. \end{aligned}$$

If however, only private data is available, we must utilize private parameter estimation technique prior to transforming the data. Thus we would need to combine the result of lemma 1.3.1 with the privacy guarantee of this estimation using one of the composition methods:

$$\begin{aligned} \mathbf{X}' &\underset{\boldsymbol{\theta}\sim\mathcal{A}_{\boldsymbol{\theta}|\mathbf{X}}}{\longrightarrow} \boldsymbol{\theta} \qquad \Rightarrow (\epsilon_{\mathcal{A}}, \delta_{\mathcal{A}})\text{-DP} \\ \mathbf{X}, \boldsymbol{\theta} &\underset{\mathbf{T}\sim F_{\mathbf{X},\boldsymbol{\theta}}}{\longrightarrow} \mathbf{T} \qquad \Rightarrow ((\epsilon_f + \epsilon_{\mathcal{A}}), (\delta_f + \delta_{\mathcal{A}}))\text{-DP}. \end{aligned}$$

1.3.3 Related work

After presenting this coarse taxonomy of approaches to private data release, we will finish this chapter discussing some of the recent work that has been done in transforming data sets into privatized representation.

A first line of work tried to find feature representations, that preserve useful information for downstream tasks by projecting the data onto low dimensional spaces, before perturbing it. This is motivated by the idea that low dimensional representations might possess a smaller sensitivity than the raw data itself and thus the amount of perturbation noise could be significantly lower. In [33] it is assumed that the data is sparsely represented and that data privatization can be considered as a compressed sampling approach. After projecting the data matrix randomly onto a low dimensional space, compressed sensing is applied to reconstruct the high dimensional, sparse representation as much as possible. It can then be shown that this reconstruction satisfies ϵ-DP while still maintaining statistical usefulness for downstream tasks like regression problems. A similar line of work using random projections onto low dimensional sub-spaces is used in [57, 10, 30]. The first work shows that random sub-space projections can maintain useful statistics about the data while still providing (ϵ, δ)-DP. The latter two works show that this further preserves metrical properties: e.g. closeness in the representation space stays correlated with closeness in the original high-dimensional representation. All these methods have in common, that they use random representations that are not adjusted to the specific data source and further do not provide an explicit notion of statistical utility that they aim to preserve.

Another line of work tries to model that data source implicitly before transforming the data set. In [28] the authors developed a noisy version of PCA and LDA to project the data on a linear subspace preserving enough statistical information e.g. to discriminate among two classes. A similar non-linear approach is given by [53] which uses a Wavelet transform of the data to identify useful representations before perturbing them to achieve (ϵ, δ)-DP.

If the set of queries is already known a-priori, it is possible to optimize a privatized representation with respect to them [54, 25]. This does not necessarily need to involve real feature

space projections: e.g. in [40], the data is represented according to noisy taxonomy trees. Theoretical work is given by [18, 29, 11] who study the problem of private data release from the perspective of learning theory. They establish lower bounds and show that it is possible to learn useful representations. However [48] shows that privatizing data sets, which could outperform the Laplace mechanism with respect to arbitrary queries, is a hard problem. Thus for practical reasons a constrained notion of utility must be assumed. A broader and more detailed discussion of recent approaches as mentioned above can further be found in surveys such as [58].

1.4 Contributions of this thesis

We will close the introduction by listing what we see as the major contributions of this research.

- We introduce a framework for private data release using point-wise transformations by showing that this problem can be discussed from the perspective of optimal lossy compression (chapter 2). This perspective yields a private data release mechanism which preserves utility, either given explicitly as function of data and private representation or given implicitly utilizing side information provided by auxiliary data,
- We derive two tractable instances of this framework based on the strong assumptions of linear compression and Gaussian distributed data (chapters 3 and 4),
- We analyze the resulting privacy guarantees for both derived methods according to our framework and discuss how these guarantees relate to information theoretic properties if the privatization is again seen as lossy compression (chapter 5),
- We show experiments suggesting that both approaches will only deliver inferior results if used for private data release in real-world applications (chapter 6),
- We discuss extensions of the results to more complex scenarios in which the strong assumptions of Gaussian sources or linear compression schemes are broken (chapter 7).

2. Differentially private data release using compression

In this chapter we propose a private data release mechanism which is based on point-wise transformations of the sample. We show that we can find point-wise privacy transformations under a given notion of utility by finding the *optimal* lossy compression of the data. Here optimality is defined as the simplest representation of the data such that it can still be useful according to the measure of utility. We start with explaining rate-distortion (RD) theory as the formal model to study lossy compression and show how this theory allows to find a nearly optimal privatizing transformation. We further show how this can be connected to the Information Bottleneck Principle which utilizes side information (e.g. classification labels) given by auxiliary data rather than an explicit function of utility to guide the compression.

2.1 Rate-distortion theory

Rate-distortion theory [14] formalizes the principle of *lossy compression*: How much can we compress a signal to a possibly much simpler representation, such that in the presence of noise, we can still identify those features that we deem to be important for our task at hand? We define this relevance by introducing a *distortion measure* $d(\mathbf{X}, \mathbf{T}) \geq 0$ which expresses how strongly a compressed representation $\mathbf{T}$ deviates from our original data $\mathbf{X}$. We further require that, on average, $d(\mathbf{X}, \mathbf{T})$ should be not too bad. The amount of information shared between $\mathbf{X}$ and its representation $\mathbf{T}$ is given by the *mutual information* of both random variables. If $P_{\mathbf{X},\mathbf{Y}}$ denotes the joint distribution of $\mathbf{X}$ and $\mathbf{T}$ with density function $\rho_{\mathbf{X},\mathbf{Y}}$ this is defined by

$$I(\mathbf{X};\mathbf{T}) = \int d\mathbf{x}\rho_{\mathbf{X}}(\mathbf{x}) \int d\mathbf{y}\rho_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y})\ p(\mathbf{x},\mathbf{y}) \log\left(\frac{\rho_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{\rho_{\mathbf{X}}(\mathbf{x})\rho_{\mathbf{Y}}(\mathbf{y})}\right) \tag{2.1}$$

$$= h(\mathbf{T}) - h(\mathbf{T}|\mathbf{X}), \tag{2.2}$$

where

$$h(\mathbf{X}) = -\int d\mathbf{x}\ p(\mathbf{x}) \log p(\mathbf{x}) \tag{2.3}$$

is called the *differential entropy* of a random variable $\mathbf{X}$. This quantity describes how much information is contained in its distribution as measured in *nats*[1] and serves as the natural generalization of the entropy of a discrete random variable. The mutual information can thus be considered as the average number of nats that can be transmitted over a channel by encoding $\mathbf{X}$ to $\mathbf{T}$. Obviously, it is possible to always achieve perfect compression by just projecting everything on a fixed code word. In this situation $I(\mathbf{X};\mathbf{T}) = 0$, however we would expect $d(\mathbf{X}, \mathbf{T})$ to be high on average. In the other extreme, we could represent $\mathbf{X}$ by itself yielding

[1] A *nat* is the unit of information expressed in base $e = \exp(0)$ compared to the more commonly unit *bit*, which expresses it in base 2.

no compression at all. In such situation we end up with $I(\mathbf{X};\mathbf{T}) = h(\mathbf{X})$ as we have no other choice than trying to encode all the information about $\mathbf{X}$ without any loss. The choice of d allows us to quantify which information of $\mathbf{X}$ is really needed for our task, such that we can trade off a loss in usefulness by a gain in compression.

Example: the compression of cats and dogs Assume we aim to encode pictures of cats and dogs into a representation that allows us to still distinguish both classes from each other. It is sufficient to represent each by a bit representing the respective class. This representation is perfectly useful for our task, and still requires only a minimum amount of information. On the other hand we might be interested to rank those animals by their furriness. Assuming both cats and dogs can be more or less furry, this aspect of information is totally gone in the one-bit representation. If we encoded them by their furriness though, i.e. by encoding it as a value in $[0, 1]$ we lose all information about their species.

2.1.1 Formal statement of the RD problem

Expressing this trade-off as a formal statement, we are searching for the (probabilistic) transformation

$$P_{\mathbf{T}|\mathbf{X}} = \arg\min_{P_{\mathbf{T}|\mathbf{X}}} I(\mathbf{X};\mathbf{T}) \tag{2.4}$$
$$s.t. \qquad \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}},\ \mathbf{t}\sim P_{\mathbf{T}|\mathbf{X}=\mathbf{x}}}[d(\mathbf{X},\mathbf{T})] < D$$

where $D > 0$ is the maximum average distortion that we are willing to allow. This optimization problem is called the *Rate-Distortion-Problem* (RD-problem) and has been studied tremendously in the context of signal compression and quantization (e.g. see [14] ch. 10) since its discovery by Shannon [45, 44]. It can be shown, that this problem always has a unique solution, which allows to express the compression rate $\mathcal{R}(\mathcal{D})$ as a function of the distortion $\mathcal{D} = \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}},\ \mathbf{t}\sim P_{\mathbf{T}|\mathbf{X}=\mathbf{x}}}[d(\mathbf{X},\mathbf{T})]$, where $\mathcal{R}(\mathcal{D}) = I(\mathbf{X};\mathbf{T})$ is defined via the unique minimizer of (2.4) for the given distortion limit.

2.2 RD-theory and the exponential mechanism

In a more recent work [39] it has been shown that there is a formal connection between the optimal solution of (2.4) and differential privacy. The RD-problem can be solved, by reformulating the constrained optimization problem into an unconstrained optimization problem, using a Lagrange multiplier $\beta > 0$ that weighs the amount of compression against the average distortion. This allows us to reformulate (2.4) as

$$P_{\mathbf{T}|\mathbf{X}} = \arg\min_{P_{\mathbf{T}|\mathbf{X}}} \underbrace{I(\mathbf{X},\mathbf{T}) + \beta\mathbb{E}_{\mathbf{X}\sim P_{\mathbf{x}},\ \mathbf{t}\sim P_{\mathbf{T}|\mathbf{X}=\mathbf{x}}}[d(\mathbf{X},\mathbf{T})]}_{\mathcal{L}(P_{\mathbf{T}|\mathbf{X}})}, \tag{2.5}$$

where we call $\mathcal{L}\left(P_{\mathbf{T}|\mathbf{X}}\right)$ the Lagrange functional for this problem. A classic result (e.g. a derivation is given in [47]) shows that compression rate and expected distortion are related by

$$\frac{\delta\mathcal{R}}{\delta\mathcal{D}} = -\beta, \tag{2.6}$$

where $\frac{\delta\mathcal{R}}{\delta\mathcal{D}}$ denotes the variational derivative of the compression rate by the distortion [2]. Using this formulation, we can find the optimal compression $P_{\mathbf{T}|\mathbf{X}}$ by minimizing $\mathcal{L}$ over the space of

[2] If the function $\mathcal{R}(\mathcal{D})$ is known e.g. in analytic form, this is just the derivative with respect to $\mathcal{D}$. However, in the way we defined $\mathcal{R}(\mathcal{D})$ and $\mathcal{D}$ as functionals of the optimal compression $P_{\mathbf{T}|\mathbf{X}}$ this becomes a varational derivative obtained by small around the optimum $P_{\mathbf{T}|\mathbf{X}}$.

all possible distributions. This results in

Theorem 2.2.1. *The minimizer distribution of* $\mathcal{L}\left(P_{\mathbf{T}|\mathbf{X}}\right)$ *is given by*

$$P_{\mathbf{T}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) = \frac{\exp\left(-\beta d(\mathbf{x},\mathbf{t})\right) P_{\mathbf{T}}(\mathbf{t})}{\int d\mathbf{t} \exp\left(-\beta d(\mathbf{x},\mathbf{t})\right) P_{\mathbf{T}}(\mathbf{t})}. \tag{2.7}$$

where the marginal $P_T(\mathbf{t})$ *is implicitly given by*

$$P_{\mathbf{T}}(\mathbf{t}) = \int d\mathbf{x}\ P_{\mathbf{T}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) P_{\mathbf{X}}(\mathbf{x}). \tag{2.8}$$

If $\mathbf{x}$ in this context is a whole data set (i.e. matrix of N entries sampled from the source $P_{\mathbf{X}}$), this distribution is known in the context of differential privacy as *exponential mechanism* [37] and possesses the following privacy guarantee:

Theorem 2.2.2 (Theorem 6 in [37])**.** *Assume for the sensitivity of the distortion*

$$\Delta d = \sup_{\mathbf{t}} \sup_{|\mathbf{X}\Delta\mathbf{X}'|=1} |d(\mathbf{X},\mathbf{t}) - d(\mathbf{X}',\mathbf{t})| < \infty \tag{2.9}$$

then releasing a sample $\mathbf{t}$ *according to the distribution as given in theorem 2.2.1 is* $2\Delta d\beta$*-DP.*

Furthermore, there is a guarantee that this mechanism samples close to optimal results

Theorem 2.2.3 (Lemma 7 in [37])**.** *Denote* $g(\mathbf{x}) = \min_{\mathbf{t}} d(\mathbf{x},\mathbf{t})$ *the lowest achievable distortion for an input* $\mathbf{x}$ *and let* $A(\mu) = \{\mathbf{t} : d(\mathbf{x},\mathbf{t}) < g(\mathbf{x}) + \mu\}$ *denote the set of possibly released elements that are not more than* μ *worse than the achievable optimum. Then*

$$P_{\mathbf{T}}\left[\{\mathbf{t} : d(\mathbf{x},\mathbf{t}) > g(\mathbf{x}) + 2\mu\}\right] < \frac{\exp(-\beta\mu)}{P_{\mathbf{T}}\left[A(\mu)\right]}. \tag{2.10}$$

In section 2.4 we will explain, how this form of the optimal compression allows to formulate a private data release mechanism.

2.3 Compression via the Information Bottleneck

Another related approach to lossy compression was given by the *information bottleneck method* [47]. Assume we have side information $\mathbf{Y} \sim P_{\mathbf{Y}}$, i.e. labels in a regression or classification task, such that $\mathbf{X}$ and $\mathbf{Y}$ together possess a non-trivial joint distribution $P_{\mathbf{X},\mathbf{Y}}$. Now optimal compression of $\mathbf{X}$ to a representation $\mathbf{T}$ can be formulated as the task of minimizing $I(\mathbf{X},\mathbf{T})$, while keeping $I(\mathbf{T},\mathbf{Y})$ as high as possible. Similarly as before, by introducing a Lagrange multiplier $\beta > 1$ for this trade-off [3], we can express the problem as the variational minimization problem of finding

$$P^*_{\mathbf{T}|\mathbf{X}} = \arg\min_{P_{\mathbf{X},\mathbf{T}}} \mathcal{L}(P_{\mathbf{X},\mathbf{T}}) = I(\mathbf{X},\mathbf{T}) - \beta I(\mathbf{T},\mathbf{Y}). \tag{2.11}$$

[3] As shown in [13] the problem is degenerated for $\beta \leq 0$ yielding the trivial solution $I(\mathbf{X},\mathbf{T}) = I(\mathbf{Y},\mathbf{T}) = 0$.

We will now show how the Information Bottleneck and the rate-distortion problem are related following a line of reasoning as given in [23]. Let $\mathbf{X} \in \mathcal{X}, \mathbf{Y} \in \mathcal{Y}$ with joint distribution $P_{\mathbf{X},\mathbf{Y}}$, having joint density $p_{\mathbf{X},\mathbf{Y}}$. Let $\mathcal{T}$ be another domain set. Now let

$$\mathbb{P} = \left\{ \rho_{\mathbf{T},\mathbf{X},\mathbf{Y}} \ : \ \mathbf{T} \to \mathbf{X} \to \mathbf{Y} \text{ is a Markov chain and } p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = \int d\mathbf{t} \ \rho_{\mathbf{T},\mathbf{X},\mathbf{Y}}(\mathbf{t},\mathbf{x},\mathbf{y}) \right\} \tag{2.12}$$

be the set of all possible joint distributions over $\mathcal{T} \times \mathcal{X} \times \mathcal{Y}$, such that $\mathbf{T} \to \mathbf{X} \to \mathbf{Y}$ is a Markov chain, i.e. that $\mathbf{Y}$ is independent of $\mathbf{T}$ for a given $\mathbf{X}$ and such that the marginalization over $\mathbf{T}$ equals $p_{\mathbf{X},\mathbf{Y}}$. Now for each $\rho \in \mathbb{P}$ we can define the expected distortion $D(\rho)$ between $\mathbf{X}$ and representation $\mathbf{T}$ induced by the choice of ρ by setting

$$D(\rho) = \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}},\ \mathbf{t}\sim\rho_{\mathbf{T}|\mathbf{X}=\mathbf{x}}} \left[D_{KL} \left[P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \| \rho_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} \right] \right] \tag{2.13}$$

$$= \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}},\ \mathbf{t}\sim\rho_{\mathbf{T}|\mathbf{X}=\mathbf{x}}} \left[\int d\mathbf{y} \ p_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}) \log \frac{p_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y})}{\rho_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}(\mathbf{y})} \right]. \tag{2.14}$$

Here

$$d_\rho(\mathbf{x},\mathbf{t}) = D_{KL} \left[P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \| \rho_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} \right] \tag{2.15}$$

serves as the local distortion measure between samples $\mathbf{x}$ and $\mathbf{t}$. Now we can formulate

Theorem 2.3.1. *Let*

$$\rho_{IB} = \arg\min_{\rho\in\mathbb{P}} I(\mathbf{X};\mathbf{T}) - \beta I(\mathbf{T};\mathbf{Y}) \tag{2.16}$$

be the optimal joint distribution $\rho_{IB} \in \mathbb{P}$ obtained from minimizing the information bottleneck functional and

$$\rho_{RD} = \arg\min_{\rho\in\mathbb{P}} I(\mathbf{X};\mathbf{Y}) + \beta D(\rho) \tag{2.17}$$

be the optimal joint distribution $\rho_{RD} \in \mathbb{P}$ obtained from minimizing the rate-distortion functional, with respect to $D(\rho)$. Then

$$\rho_{IB} = \rho_{RD}. \tag{2.18}$$

This means, the optimizing ρ of the information bottleneck coincides with with finding the exponential mechanism solution with respect to the distortion measure $d_\rho(\mathbf{x},\mathbf{t})$. Indeed, the minimizing distribution of the IB functional possesses the solution

Theorem 2.3.2. *The IB functional has the optimal solution*

$$P^*_{\mathbf{T}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) = \frac{\exp\left(-\beta D_{KL}\left[P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \| P^*_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}\right]\right) P^*_{\mathbf{T}}(\mathbf{t})}{Z(\mathbf{X},\beta)} \tag{2.19}$$

where

$$P^*_{\mathbf{T}}(\mathbf{t}) = \int d\mathbf{x} \ P_{\mathbf{X}}(\mathbf{x}) P^*_{\mathbf{T}|\mathbf{X}=\mathbf{x}}(\mathbf{t})$$

$$P^*_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}(\mathbf{y}) = \frac{\int d\mathbf{x} \ P^*_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}(\mathbf{y}) P^*_{\mathbf{T}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) P_{\mathbf{X}}(\mathbf{x})}{P^*_{\mathbf{T}}(\mathbf{t})}$$

$$Z(\mathbf{X},\beta) = \int d\mathbf{t} \ \exp\left(-\beta D_{KL}\left[P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \| P^*_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}\right]\right) P^*_{\mathbf{T}}(\mathbf{t}).$$

This implies, that once we found the optimal solution of the IB functional, we can again interpret it as an instance of the exponential mechanism according to the distortion function d_ρ. While before, we had to hand-design d a-priori according to our task, by solving (2.11) we simultaneously obtain the distortion measure, that naturally preserves as much of our auxiliary information as possible given the compression level.

2.4 Releasing private data by optimal compression

We can utilize these results to formulate a private data release mechanism that explicitly optimizes its privatized output according to a utility measure. We assume that the utility can be represented by a *factorizing* distortion function d, i.e. if $\mathbf{X} \sim P_{\mathbf{X}}^N$ is our private data sample and $\mathbf{T}$ a privatized representation, we assume that we can write

$$d(\mathbf{X}, \mathbf{T}) = \sum_{n=1}^{N} d(\mathbf{x}_n, \mathbf{t}_n). \tag{2.20}$$

In such a scenario we can formulate and solve the rate distortion problem with respect to $P_{\mathbf{X}}$ and d to obtain a point-wise compression function. By compressing each single data point $\mathbf{x}_i$ individually onto its compressed representation $\mathbf{t}_i$, we obtain a compressed data set $\mathbf{T}$. For a fixed upper bound on the expected distortion (or lower bound on the expected utility) $\mathbf{T}$ shares the least amount of mutual information with $\mathbf{X}$. The following theorem tells us, that such a compressed representation of $\mathbf{X}$ is private.

Theorem 2.4.1. *Let $P_{\mathbf{T}|\mathbf{X}}$ be the optimal solution of the the RD-problem given the point-wise distortion d with sensitivity Δd and a source distribution $P_{\mathbf{X}}$. Let $\mathbf{X} \sim P^N(\mathbf{X})$ be a data set. Then the release of $\mathbf{T} \sim P_{\mathbf{T}|\mathbf{X}}^N$ obtained by point-wise compressing $\mathbf{X}$ is $2\Delta d\beta$-DP.*

If we are in a situation, where we cannot explicitly formulate such a factorizing d in order to capture our notion of utility, we can still approximate it implicitly using auxiliary side information $\mathbf{Y}$ and then encoding the data according to the IB principle. As a result, we obtain a private data release mechanism, optimized with respect to an arbitrary measure of utility as long as its describing distortion function d factorizes over samples and possesses bounded sensitivity.

2.5 Challenges of the method

Using this mechanism for practical applications requires to overcome two non-trivial challenges: dealing with potentially unbounded sensitivities and finding a tractable form of the minimizer of either the RD or the IB problem.

2.5.1 Unbounded sensitivities

In real problems we are frequently faced with distortion measures that might attain very high or even unbounded sensitivity. Still, if we can guarantee that the probability of obtaining representations leading to high sensitivity of the distortion is low, will still have (ϵ, δ)-DP.

Theorem 2.5.1 (Following theorem 5.2.2 on p. 86 of [38])**.** *Let $\mathbf{t} \sim p_{\mathbf{T}}(\mathbf{t})$ have a tail-bound, such that for each possible $\delta > 0$, there is an $\epsilon > 0$ with*

$$P\left[\mathbf{t} \in \left\{ \sup_{\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}} |d(\mathbf{x}_1, \mathbf{t}) - d(\mathbf{x}_2, \mathbf{t})| > \frac{\epsilon}{2\beta} \right\}\right] < \delta. \tag{2.21}$$

Then releasing a sample from the exponential mechanism with respect to distortion d is (ϵ, δ)-DP.

2.5.2 Finding the optimal minimizer

A more difficult problem is finding a computable solution to (2.4) or (2.11) that goes beyond the presented formal equations.

General rate-distortion problem

While the solution as presented in theorem 2.2.1 has a seemingly simple form, the major challenge is that the normalizer of the sampler density can be very difficult to compute if it does not possess an analytic form. Furthermore, there is a complex coupling between the sampler and the implicit prior $P_{\mathbf{T}}\mathbf{t}$ which both rely on each other. There has been a tremendous amount of research on different ways of approximating this sampler (see e.g. [8] for a survey). Most notable is the *Blahut-Arimoto* (BA) algorithm [9, 5], which allows approximating the optimal solution for discrete sources in an iterative fashion and will converge to the true optimum in the infinite time limit. In some simple cases however, e.g. in the case of a Gaussian source and assuming L_2 distortion, it is possible to find an analytic solution. We derive this instance in chapter 3 and analyze its privacy guarantees according to our framework in chapter 5.

Information Bottleneck Principle

Due to the dependencies of $P^*_{\mathbf{T}|\mathbf{X}=\mathbf{x}}, P^*_{\mathbf{T}}$ and $P^*_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}$ this solution stays formal in the general case and its solution would involve solving the coupled equations, which is intractable in the general case. For discrete variables [47] give an adaptation of the BA algorithm which converges to a local optimum of (2.11). However, in contrast to the original BA algorithm, which will eventually converge to the unique minimum, it is not guaranteed, that such a local optimum has a form similar to (2.19). If the joint distribution $P_{\mathbf{X},\mathbf{Y}}$ is restricted to stem from a set of "well-behaved" distributions, such as multivariate Gaussians, an analytic solution can be found. We present this instance in chapter 4 and analyze its privacy guarantees according to our framework in chapter 5.

3. Optimal linear compression of Gaussian data under L_2 distortion.

In this chapter we derive the optimal compression for the L_2 distortion, under the strong assumption that the source distribution is Gaussian and that the compression scheme is represented by a linear projection with additive Gaussian noise. This model has been well studied in the RD literature and the solution of this model has a close relation to the principal component analysis of the covariance matrix of the signal source. We start with one-dimensional signals and further extend this result to the high-dimensional case. In both analyses it is assumed that the covariance matrix $\mathbf{\Sigma}$ of the source signal is already known or has already been estimated a-priori. If we do not have access to a public sample to estimate it, this involves using a private parameter estimation technique (see section 5.1).

3.1 One dimensional Gaussian signals

Let our source distribution be a zero centered Gaussian signal $x \sim P_X(x) = \mathcal{N}(x|0, \sigma^2)$ and assume that we have the L_2 distortion

$$d(x,t) = \tfrac{1}{2}(x-t)^2. \tag{3.1}$$

We further assume the noisy linear transformation $t = ax + \xi$, where $a \in \mathbb{R}$ and $\xi \sim \mathcal{N}(0, \lambda^2)$ is an independent noise term. This gives us

$$\begin{aligned} t|x &\sim P_{T|X=x}(t) = \mathcal{N}(t|ax, \lambda^2) \\ t &\sim P_T(t) = \mathcal{N}(t|0, a^2\sigma^2 + \lambda^2). \end{aligned} \tag{3.2}$$

Rewriting the distortion to

$$d(x,t) = \tfrac{1}{2}((1-a)x - \xi)^2, \tag{3.3}$$

gives us

$$\begin{aligned} \mathbb{E}_{x,t}[d(x,t)] &= \tfrac{1}{2}(1-a)^2\mathbb{E}_{x,\xi}[x^2] - (1-a)\mathbb{E}_{x,\xi}[x\xi] + \tfrac{1}{2}\mathbb{E}_{x,\xi}[\xi^2] \\ &= \tfrac{1}{2}(1-a)^2\sigma^2 - (1-a)\mathbb{E}_x[x]\mathbb{E}_\xi[\xi] + \tfrac{1}{2}\lambda^2 \\ &= \tfrac{1}{2}\left((1-a)^2\sigma^2 + \lambda^2\right). \end{aligned} \tag{3.4}$$

The mutual information between source and representation is given by

$$I(x;t) = h(t) - h(t|x) \tag{3.5}$$

$$= \frac{1}{2}\left(\log(a^2\sigma^2 + \lambda^2) - \log\lambda^2\right). \tag{3.6}$$

From this we can formulate the RD-Lagrangian with respect to parameters a and λ^2 and a chosen Lagrange multiplier $\beta > 0$

$$\mathcal{L}(a, \lambda^2) = \frac{1}{2}\left(\log(a^2\sigma^2 + \lambda^2) - \log\lambda^2 + \beta\left((1-a)^2\sigma^2 + \lambda^2\right)\right).$$

which gives us the minimization problem

$$a_*, \lambda_*^2 = \arg\min_{a,\lambda^2} \mathcal{L}(a, \lambda^2). \tag{3.7}$$

The solution of this minimization problem is given in the following

Theorem 3.1.1. *The optimal solution pair of* (3.7) *given signal variance* σ^2 *exists iff* β *is chosen such that* $\beta\sigma > 1$*. In this case*

$$a = \frac{\beta\sigma - 1}{\beta\sigma} \tag{3.8}$$

$$\lambda^2 = \frac{\beta\sigma - 1}{\beta^2\sigma} \tag{3.9}$$

While this solution is a necessary condition to be the global solution of the RD functional we also need to give the sufficient condition, that it indeed is an instance of the exponential mechanisms with respect to d. This is given in

Theorem 3.1.2. *Given signal variance* σ^2*, let* a *and* λ^2 *the optimal solution pair as obtained in 3.1.1 for a chosen* β *such that* $\beta\sigma^2 > 1$*. Then*

$$p_{T|X=x}(t) \propto \exp(-\beta d(x,t)) \cdot p(t) \tag{3.10}$$

where

$$p_T(t) = \int dx\ p_{T|X=x}(t) \cdot p_X(x) = \mathcal{N}(t|0, a^2\sigma^2 + \lambda^2). \tag{3.11}$$

Thus transforming $t = ax + \xi$ *for* $\xi \sim \mathcal{N}(0, \lambda^2)$ *corresponds to sampling from the exponential mechanism.*

3.2 Multivariate Gaussian signals

Now we will turn to multivariate signals $\mathbf{x} \sim P_{\mathbf{X}} = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Similar as in the one-dimensional case, we use L_2-distortion in $\mathbb{R}^D$

$$d(\mathbf{x}, \mathbf{t}) = \frac{1}{2}\|\mathbf{x} - \mathbf{t}\|_2^2. \tag{3.12}$$

Again we assume a linear noisy transformation

$$\mathbf{t} = \mathbf{A}\mathbf{x} + \boldsymbol{\xi}, \tag{3.13}$$

for a projection matrix $\mathbf{A} \in \mathbb{R}^{D\times D}$ and a noise source $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$. This directly gives us

$$\begin{aligned} \mathbf{t}|\mathbf{x} &\sim P_{\mathbf{T}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) = \mathcal{N}(\mathbf{A}\mathbf{x}, \mathbf{\Lambda}) \\ \mathbf{t} &\sim P_{\mathbf{T}}(\mathbf{t}) = \mathcal{N}\left(\mathbf{0}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{\Lambda}\right). \end{aligned} \tag{3.14}$$

By expressing

$$d(\mathbf{x}, \mathbf{t}) = \frac{1}{2}\|(\mathbf{I} - \mathbf{A})\mathbf{x} - \boldsymbol{\xi}\|_2^2 \tag{3.15}$$

we get (see appendix)

$$\mathbb{E}_{\mathbf{x},\boldsymbol{\xi}}\left[d(\mathbf{x}, \mathbf{t})\right] = \tfrac{1}{2}\left(\mathbf{tr}\left((\mathbf{I} - \mathbf{A})\,\boldsymbol{\Sigma}\,(\mathbf{I} - \mathbf{A})^T\right) + \mathbf{tr}\left(\boldsymbol{\Lambda}\right)\right). \tag{3.16}$$

As source and prior are both multivariate Gaussian distributions the mutual information is given by

$$\begin{aligned} I(\mathbf{X}, \mathbf{T}) &= h(\mathbf{T}) - h(\mathbf{T}|\mathbf{X}) && (3.17)\\ &= \tfrac{1}{2}\left(\log\det\left|\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \boldsymbol{\Lambda}\right| - \log\det\left|\boldsymbol{\Lambda}\right|\right), && (3.18)\end{aligned}$$

such that for $\beta > 0$ we end up with the rate-distortion Lagrangian

$$\mathcal{L}(\mathbf{A}, \boldsymbol{\Lambda}) = \tfrac{1}{2}\left(\log\det\left|\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \boldsymbol{\Lambda}\right| - \log\det\left|\boldsymbol{\Lambda}\right| + \beta\mathbf{tr}\left((\mathbf{I} - \mathbf{A})\,\boldsymbol{\Sigma}\,(\mathbf{I} - \mathbf{A})^T\right) + \mathbf{tr}\left(\boldsymbol{\Lambda}\right)\right). \tag{3.19}$$

Its solution is given by the following

Theorem 3.2.1. *Given signal covariance $\boldsymbol{\Sigma}$, let σ_i denote the i-th eigenvalue of $\boldsymbol{\Sigma}$ and $\mathbf{v}_i$ the corresponding eigenvector. The optimal pair $\mathbf{A}_*, \boldsymbol{\Lambda}_*$ of (3.19) exists iff β is chosen such that $\beta\sigma_i > 0$ for each i. In this case*

$$\begin{aligned} \mathbf{A}_* &= \mathbf{V}\mathbf{D}_{\mathbf{A}_*}\mathbf{V}^T,\ \boldsymbol{\Lambda}_* = \mathbf{V}\mathbf{D}_{\boldsymbol{\Lambda}_*}\mathbf{V}^T && (3.20)\\ \mathbf{D}_{\mathbf{A}_*} &= diag\left(\frac{\beta\sigma_1 - 1}{\beta\sigma_1}, \ldots, \frac{\beta\sigma_D - 1}{\beta\sigma_D}\right) && (3.21)\\ \mathbf{D}_{\boldsymbol{\Lambda}_*} &= diag\left(\frac{\beta\sigma_1 - 1}{\beta^2\sigma_1}, \ldots, \frac{\beta\sigma_D - 1}{\beta^2\sigma_D}\right) && (3.22)\end{aligned}$$

As before we have to show that this solution indeed follows the exponential mechanism. This is given by

Theorem 3.2.2. *Given signal covariance $\boldsymbol{\Sigma}$, let $\mathbf{A}$ and $\boldsymbol{\Lambda}$ be the optimal pair as obtained in theorem 3.2.1 for a chosen β such that $\beta\sigma_i > 1$ for each eigenvalue σ_i of $\boldsymbol{\Sigma}$. Then*

$$P_{\mathbf{T}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) \propto \exp(-\beta d(\mathbf{x}, \mathbf{t})) \cdot P_{\mathbf{T}}(\mathbf{t}) \tag{3.23}$$

where

$$P_{\mathbf{T}}(\mathbf{t}) = \int dx\ P_{\mathbf{T}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) \cdot P_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \boldsymbol{\Lambda}). \tag{3.24}$$

Thus transforming $\mathbf{t} = \mathbf{A}\mathbf{x} + \boldsymbol{\xi}$ for $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$ corresponds to sampling from the exponential mechanism.

3.3 Pruning low variance components

In realistic settings and high dimensional data the eigenvalues $\sigma_1 > \cdots > \sigma_D$ of $\mathbf{\Sigma}$ will not be very uniform. On the contrary, in many cases only a few eigenvalues will have significant power, whereas the majority of them will be very close to zero. Applying theorem 3.2.1 directly in such a scenario, would require to set $\beta > \frac{1}{\sigma_D}$. This implies that the noise for the strongest first components becomes very small, and thus imposes an increased privacy risk for the overall release mechanism. By first pruning away the $D - k$ smallest dimensions of $\mathbf{\Sigma}$ we can still use the theorem. This can be beneficial if $\frac{\sigma_1}{\sigma_k}$ is not too. In this case β can become smaller and as a consequence we will obtain more noise to privatize the strongest components. This projection is given by the operator $\mathbf{Q} = \mathbf{V}_k \mathbf{P}_k \mathbf{V}^T$, where $\mathbf{V}_k$ denotes the matrix of eigenvectors corresponding the first k biggest eigenvalues and $\mathbf{P}_k$ is the $k \times D$ matrix with ones on the diagonal and zeros elsewhere. The privacy implications of this pruning method is discussed in section 5.6.

3.4 Relation to classic results

We close this chapter by connecting our derived solution to classic results from RD theory and dimension reduction using PCA.

3.4.1 Relation to RD theory for memoryless Gaussian channels

Linear compression of Gaussian sources under the L_2 distortion are a classic result in RD theory (see e.g. [14], chapter 10.3.2). In this context, we are interested in finding the rate distortion function $\mathcal{R}(\mathcal{D})$ to understand how much we can compress (and quantize) Gaussian sources for a fixed average distortion. We can reproduce the classic result now by plugging in our optimal solution 3.1.1 to obtain

$$\begin{aligned}
\mathcal{D} &= \mathbb{E}\left[(x-t)^2\right] && (3.25)\\
&= ((1-a)^2\sigma^2 + \lambda^2) && (3.26)\\
&= \left(\frac{1}{\beta^2\sigma} + \frac{\beta\sigma - 1}{\beta^2\sigma}\right) && (3.27)\\
&= \frac{1}{\beta}. && (3.28)
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{R} &= I(x;t) && (3.29)\\
&= \frac{1}{2}\log\left(\frac{a^2\sigma^2 + \lambda^2}{\lambda^2}\right) && (3.30)\\
&= \frac{1}{2}\log\left(\frac{a^2\sigma^2}{\lambda^2} + 1\right) && (3.31)\\
&= \frac{1}{2}\log\left(\beta\sigma\right) && (3.32)\\
&= \frac{1}{2}\log\left(\frac{\sigma}{\mathcal{D}}\right). && (3.33)
\end{aligned}$$

if and only iff $\sigma > \mathcal{D}$, otherwise $\mathcal{R} = 0$. However, this classic discussion did not take into account that the obtained distribution is also an instance of the exponential mechanism. Analyzing the privacy guarantees obtained from the mechanism further requires an explicit solution of

the projection matrix and noise covariance as functions of the privacy determining Lagrange parameter β. As we will see later in chapter 5, the optimal privacy guarantees obtained by compressing Gaussian sources under L_2 distortion is only depending on information theoretic quantities. Thus they can directly be predicted using classic approaches from RD theory.

3.4.2 Relation to principal component analysis

The result of theorem 3.2.1 together with the idea of pruning low variance components has furthermore an interesting connection to *principal component analysis* (PCA). In PCA, we are interested in finding an (orthogonal) projection $\mathbf{U} \in \mathbb{R}^{k \times D}$ of a data set $\mathbf{X} \in \mathbb{R}^{N \times D}$ to a k-dimensional linear subspace, such that we preserve most of its variance along each axis. By pruning those dimensions with low variance, we can preserve most of the *relevant* information, while often being able to reduce dimensionality significantly. Similar to the presented compression model *relevance* is expressed by requiring that the L_2 distance between reconstructed data and source is as small as possible. This is can be formulated as finding

$$\mathbf{U}_* = \arg\min_{\mathbf{U}} \|\mathbf{X} - \mathbf{X}\mathbf{U}^T\mathbf{U}\|_2^2 \tag{3.34}$$

$$s.t. \qquad \mathbf{U}\mathbf{U}^T = \mathbf{I}_{D'}. \tag{3.35}$$

The optimal solution $\mathbf{U}_*$ contains those k eigenvectors corresponding to the k biggest eigenvalues of $\mathbf{\Sigma}$. While in classic PCA, we are not interested in scaling eigenvalues, the projection in 3.2.1 scales them down according to the set privacy level. For $\beta \to \frac{1}{\sigma_i}$ the i-th component contains less and less information until it eventually gets pruned out. As $\beta \to \infty$ we will keep more components and scale down kept eigenvalues less.

4. Optimal linear compression of Gaussian data using side information.

In this chapter we present the optimal compression according to the Information Bottleneck principle, again under the same strong assumptions that the source distribution is Gaussian and that the compression scheme is a linear projection with additive noise. This model has found its way to the literature as *Gaussian Information Bottleneck*[13] and similarly as in the last chapter has an analytic solution. Interestingly, it is not trivial that the linear Bottleneck solution is even the unique minimizer of (2.11) [22]. Only this last result allows to discuss it as another instance of privacy by optimal compression. As derived before, the distortion measure is not given a-priori like it has been in the L_2 case. Instead, it is implicitly determined by preserving relevance with respect to auxiliary information $\mathbf{Y}$. Similar as before, we assume that we already have access to the joint covariance $\mathbf{\Sigma_{X,Y}}$. Either we could have estimated it on a public sample, or we would need to utilize a DP covariance estimator (see 5.1) prior to compression and compose the privacy guarantees.

4.1 The Gaussian Information Bottleneck (GIB)

In [13] problem (2.11) was solved analytically for a restricted class of transformation. Here it is assumed that $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian distributed having zero mean and a joint full-rank co-variance matrix

$$\mathbf{\Sigma_{X,Y}} = \begin{bmatrix} \mathbf{\Sigma_X} & \mathbf{\Sigma_{XY}} \\ \mathbf{\Sigma_{YX}} & \mathbf{\Sigma_Y} \end{bmatrix}. \tag{4.1}$$

Further the search space is restricted to compressed representations $\mathbf{T}$ where $P_{\mathbf{T}|\mathbf{X}}$ is Gaussian as well. That way $\mathbf{T}$ can be expressed by

$$\mathbf{T} = \mathbf{A}\mathbf{X} + \boldsymbol{\xi} \tag{4.2}$$

for a transformation matrix $\mathbf{A}$ and independent noise $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\xi}|\mathbf{0}, \mathbf{\Sigma_\xi})$.

4.1.1 Analytic solution for the constrained search space

Equation (4.2) directly gives us

$$\mathbf{\Sigma_{T|X}} = \mathbf{\Sigma_\xi}. \tag{4.3}$$

Using common matrix algebra for co-variance matrices we compute

$$\begin{aligned} \mathbf{\Sigma_T} &= \mathbf{A}\mathbf{\Sigma_X}\mathbf{A}^T + \mathbf{\Sigma_\xi} \\ \mathbf{\Sigma_{TX}} &= \mathbf{A}\mathbf{\Sigma_X} \\ \mathbf{\Sigma_{TY}} &= \mathbf{A}\mathbf{\Sigma_{XY}} \end{aligned} \tag{4.4}$$

and by building the Schur complement we get the conditional matrix

$$\mathbf{\Sigma_{T|Y}} = \mathbf{A}\mathbf{\Sigma_{X|Y}}\mathbf{A}^T + \mathbf{\Sigma_{\xi}}. \tag{4.5}$$

By using the analytic form for the mutual information between two Gaussian random variables we can re-express (2.11) as a minimization problem with respect to parameters $\mathbf{A}$ and $\mathbf{\Sigma_{\xi}}$:

$$\mathbf{A}^*, \mathbf{\Sigma}^*_{\boldsymbol{\xi}} = \arg\min_{\mathbf{A}, \mathbf{\Sigma_{\xi}}} \left(\log|\mathbf{\Sigma_T}| - \log\left|\mathbf{\Sigma_{T|X}}\right| - \beta\left(\log|\mathbf{\Sigma_T}| - \log\left|\mathbf{\Sigma_{T|Y}}\right|\right)\right). \tag{4.6}$$

As was shown in [13] $\mathbf{A}$ and $\mathbf{\Sigma_{\xi}}$ depend on each other, such that by fixing $\mathbf{\Sigma_{\xi}} = \mathbf{I}$ this can be simplified to

$$\mathbf{A}^* = \arg\min_{\mathbf{A}}(1-\beta)\log\left|\mathbf{A}\mathbf{\Sigma_X}\mathbf{A}^T + \mathbf{I}\right| - \beta\log\left|\mathbf{A}\mathbf{\Sigma_{X|Y}}\mathbf{A}^T + \mathbf{I}\right|. \tag{4.7}$$

Theorem 3.1 of [13] solve (4.7) using standard calculus and show that it yields an eigenvalue problem. This eigenvalue problem produces the solution in form of

Theorem 4.1.1 (Theorem 3.1 of [13])**.** *Given a joint covariance matrix $\mathbf{\Sigma_{X,Y}}$ and a chosen $\beta > 1$, let $\mathbf{v}_1, \ldots, \mathbf{v}_d$ be the left eigenvectors of $\mathbf{\Sigma_{X|Y}}\mathbf{\Sigma_X^{-1}}$ sorted by ascending eigenvalues $\lambda_1, \ldots, \lambda_d$. Then the optimal $\mathbf{A}$ solving (4.7) is given by the projection matrix*

$$\mathbf{A} = \left[\mathbf{u}_1^T, \ldots, \mathbf{u}_d^T\right] \tag{4.8}$$

where the vectors $\mathbf{u}_i$ are defined as

$$\mathbf{u}_i = \begin{cases} \alpha_i \mathbf{v}_i & \beta > \frac{1}{1-\lambda_i} \\ \mathbf{0} & o.w. \end{cases} \tag{4.9}$$

α_i are given by

$$\alpha_i = \sqrt{\frac{\beta\left(1-\lambda_i\right) - 1}{\lambda_i \mathbf{v}_i^T \mathbf{\Sigma_X} \mathbf{v}_i}}. \tag{4.10}$$

This solution is again just a necessary condition and further solved 2.11 only for the subset of linear projection. Thus, so far we cannot apply 2.3.1 to analyze the privacy guarantee of the produced compression.

4.1.2 Global solution and relation to exponential mechanism

The sufficient condition that this solution is also the unique global minimizer of 2.11 was given in [22].

Theorem 4.1.2 (Theorem 1 of [22])**.** *Given $P_{\mathbf{T},\mathbf{X}} = \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma_{X,Y}}\right)$. Let $\mathbf{A}$ be the optimal solution of (4.7) as given in theorem 4.1.1. Then the joint distribution $P_{\mathbf{T},\mathbf{X}}$ as induced by*

$$P_{\mathbf{T}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) = \mathcal{N}\left(\mathbf{t}|\mathbf{A}\mathbf{x}, \mathbf{I}\right) \tag{4.11}$$

is the unique minimizer of 2.11.

This implies, that releasing $\mathbf{t} = \mathbf{A}\mathbf{x} + \boldsymbol{\xi}$ is equivalent to sampling from the exponential mechanism distribution with respect to the distortion

$$d(\mathbf{x}, \mathbf{t}) = D_{KL}\left[P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \| P_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}\right]. \tag{4.12}$$

Both $P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ and $P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ are multivariate normal distributions with means $\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$, $\mu_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ and covariance matrices $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$, $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ respectively. Given the optimal projection $\mathbf{A}$, we can compute all those distribution parameters analytically using Schur's lemma. Further the KL divergence between these two Gaussian distribution possesses the tractable analytic form

$$\begin{aligned} D_{KL}\left[P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \| P_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}\right] = \frac{1}{2}\big(&\mathbf{tr}\left(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-1}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\right) \\ &+ \left(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\right)^T \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-1}\left(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\right) \\ &- D + \log\left|\det\left(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\right)\right|\big). \end{aligned} \tag{4.13}$$

This allows us to both, sample from the mechanism efficently and analyze its privacy guarantees according to the derivation from the exponential mechanism.

5. Privacy analysis

In this chapter we will derive the privacy guarantees for both linear compression schemes. Both distortions possess unbounded sensitivity. However, we can achieve (ϵ, δ)-DP by establishing a tail-bound on the sensitivity according to theorem 2.5.1. Due to the special form of the presented a second analysis can be conducted that follows the *improved Gaussian mechanism*. As shown before, this mechanism has an exact lower bound on the privacy parameter ϵ given a fixed δ, the function sensitivity and the noise magnitude. This bound can only be computed numerically. Thus, we will discuss a close approximation to analytically express how the privacy guarantee depends on the information theoretic quantities of the compression.

5.1 Estimation of signal covariance

In the discussion so far, we assumed that the signal covariance was given a-priori. However, in practice we only observe a sampled data set $\mathbf{X} \sim P_{\mathbf{X}}^N$ and have to estimate $\mathbf{\Sigma}$ from it, before continuing with our analysis.

5.1.1 Estimation using public data

In a scenario, where we have access to public sample $\tilde{\mathbf{X}} \sim P_{\mathbf{X}}^M$ of the same source, we can estimate $\mathbf{\Sigma}$ on this sample and then this covariance to compress private data into the privatized representation. As estimating $\mathbf{\Sigma}$ would not involve any access to $\mathbf{X}$, the following privacy guarantees would stay unaffected.

5.1.2 DP estimation on private data

Having only access to private data we need to privately estimate the covariance first before continuing with deriving the optimal compression from it. An efficient estimator for the covariance of a dataset with ϵ-DP guarantee was given in [27]. This estimator requires adding a noise matrix with a spectral norm in the order of $O\left(\frac{D \log D}{N\epsilon}\right)$ to the covariance matrix. For the weakened guarantee of (ϵ, δ)-DP, [19] derived an optimal estimator, which only requires noise with a spectral norm in the order of $O\left(\frac{\sqrt{D}}{N\epsilon}\right)$. As both of compression schemes directly depend on the spectrum of the covariance, the Weyl inequality [51] gives an estimate of the utility that is preserved, after estimating $\mathbf{\Sigma}$ using one of both methods:

Theorem 5.1.1 ([51]). *Given a covariance matrix $\mathbf{\Sigma}$ having eigenvalues $\sigma_1, \ldots, \sigma_D$ and a noise matrix $\mathbf{P}$, let $\tilde{\mathbf{\Sigma}} = \mathbf{\Sigma} + \mathbf{P}$ be the perturbed covariance with eigenvalues $\tilde{\sigma_1}, \ldots, \tilde{\sigma_D}$. Then we have*

$$|\tilde{\sigma_i} - \sigma_i| < \|\mathbf{P}\|_2. \tag{5.1}$$

The privacy budget needed to estimate $\mathbf{\Sigma}$ has to be composed with the privacy guarantees obtained in the next sections using one of the composition methods.

5.2 Bounding sensitivity by clipping inputs

Achieving (ϵ, δ)-DP is unfeasible for both mechanisms, as long as the data comes from an unbounded domain. In this case the sensitivity of both distortions cannot be bounded. In practice however, we can assume, that the data is contained in a ball of radius R. If we estimate the covariance from such data, this implies

$$\|\mathbf{\Sigma}\|_2 = \max_{\|\mathbf{v}\|_2=1} \mathbf{v}^T \mathbf{\Sigma} \mathbf{v} = \max_{\|\mathbf{v}\|_2=1} \frac{1}{N} \sum_{i=1}^{N} \mathbf{v}^T \mathbf{x}\mathbf{x}^T \mathbf{v} \leq R^2. \tag{5.2}$$

Given this assumption, we can establish tail-bounds on the sensitivity and thus derive the privacy guarantees according to the framework.

5.3 Derivation from the exponential mechanism

Both derived compression schemes have been shown to be instances of the exponential mechanism. Using our assumption of the last section this allows us to tail-bound the sensitivity of the involved distortion measures. By further using theorem 2.5.1 we can establish an (ϵ, δ)-DP guarantee for both mechanisms.

5.3.1 Analysis for the compression of Gaussian data under L_2 distortion.

A tail-bound of the L_2 distortion is given by

Theorem 5.3.1. *Let* $d(\mathbf{x}, \mathbf{t}) = \frac{1}{2}\|\mathbf{x} - \mathbf{t}\|_2^2$ *for* $\mathbf{t} = \mathbf{A}\mathbf{x} + \boldsymbol{\xi}$, $\mathbf{x} \in B_R(\mathbf{0}) = \{\mathbf{x} \in \mathcal{R}^D : \|\mathbf{x}\|_2 = R\}$ *and* $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$. *Then*

$$P\left[\left\{\mathbf{t} : \sup_{\mathbf{x},\mathbf{x}' \in B_R(\mathbf{0})} |d(\mathbf{x}, \mathbf{t}) - d(\mathbf{x}', \mathbf{t})| > \frac{1}{2}\left(R + H(\mathbf{Q}, \delta)\right)^2\right\}\right] < \delta. \tag{5.3}$$

where

$$H(\mathbf{Q}, \delta) = \sqrt{\mathbf{tr}\left(\mathbf{Q}\right) + \sqrt{-\mathbf{tr}\left(\mathbf{Q}^T\mathbf{Q}\right)\log(1/\delta)} + \|\mathbf{Q}\|_2 \log(1/\delta)} \tag{5.4}$$

$$\mathbf{Q} = \mathbf{A}^T \mathbf{\Sigma} \mathbf{A}^T + \mathbf{\Lambda} \tag{5.5}$$

.

Using this result we can apply theorem (2.5.1) to obtain

Corollary 1. *Let* $\mathbf{A}, \mathbf{\Lambda}$ *be the optimal solution for solving the Gaussian RD-functional with respect to* L_2 *distortion. Then releasing a sample from the mechanism is* (ϵ, δ) *private as long as*

$$\epsilon > \beta\left(R + H(\mathbf{Q}, \delta)\right)^2. \tag{5.6}$$

5.3.2 Analysis for Gaussian Information Bottleneck

As similar privacy guarantee can be obtained for the GIB

Theorem 5.3.2. *Assume* $\mathbf{x} \in B_R(\mathbf{0}) = \{\mathbf{x} \in \mathcal{R}^D : \|\mathbf{x}\|_2 = R\}$. *Let* $\mathbf{A}$, *be the optimal GIB solution,* $\mathbf{t} = \mathbf{A}\mathbf{x} + \boldsymbol{\xi}$ *for* $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ *and let* $d(\mathbf{x}, \mathbf{t}) = D_{KL}\left[P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \| P_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}\right]$. *Then we have*

$$P\left[\left\{\mathbf{t} : \sup_{\mathbf{x},\mathbf{x}' \in B_R(\mathbf{0})} |d(\mathbf{x}, \mathbf{t}) - d(\mathbf{x}', \mathbf{t})| > \frac{1}{2}\left(\|\mathbf{P}\|_2 R + H(\mathbf{Q}, \delta)\right)^2\right\}\right] < \delta. \tag{5.7}$$

where

$$H(\mathbf{Q}, \delta) = \sqrt{\mathbf{tr}\left(\mathbf{Q}\right) + \sqrt{-\mathbf{tr}\left(\mathbf{Q}^T\mathbf{Q}\right)\log(1/\delta)} + \|\mathbf{Q}\|_2 \log(1/\delta)} \tag{5.8}$$

$$\mathbf{Q} = \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-1}\boldsymbol{\Sigma}_{\mathbf{YT}}\boldsymbol{\Sigma}_{\mathbf{T}}^{-1}\boldsymbol{\Sigma}_{\mathbf{TY}} \tag{5.9}$$

$$\mathbf{P} = \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \tag{5.10}$$

.

Again we can now apply theorem (2.5.1) to obtain

Corollary 2. *Let* $\mathbf{A}$ *be the optimal solution of the GIB functional. The releasing a sample from this mechanism is* (ϵ, δ)*-DP as long as*

$$\epsilon > \beta\left(\|\mathbf{P}\|_2 R + H(\mathbf{Q}, \delta)\right)^2 \tag{5.11}$$

5.4 Derivation from the improved Gaussian mechanism

In both methods we release $\mathbf{t} = f(\mathbf{x}) + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$. If we have for every $\mathbf{x}, \mathbf{x}' \in B_R(\mathbf{0}) = \{\mathbf{x} \in \mathcal{R}^D : \|\mathbf{x}\|_2 = R\}$ it implies

$$\|\mathbf{A}(\mathbf{x} - \mathbf{x}')\|_2 \leq \|\mathbf{A}\|_2 R. \tag{5.12}$$

Thus, if we release a data set by point-wise transforming $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ the overall sensitivity is completely determined by the sensitivity $\Delta f = \|\mathbf{A}\|_2 R$ for one point. By adding noise to the result of f, the representation $\mathbf{t}$ is at least as perturbed as $\tilde{\mathbf{t}} = f(\mathbf{x}) + \tilde{\boldsymbol{\xi}}$ where $\tilde{\boldsymbol{\xi}} \sim \mathcal{N}(\mathbf{0}, \lambda_{\min}^2 \mathbf{I})$ and $\lambda_{\min}^2$ denotes the smallest Eigenvalue of $\boldsymbol{\Lambda}$. This way, we can analyze both mechanisms according to theorem 1.2.2 using sensitivity $\Delta f = \|\mathbf{A}\|_2 R$ and Gaussian noise magnitude $\nu^2 = \lambda_{\min}^2$. The result as given in theorem 1.2.2 does not provide an easy way to analytically see the relation between compression and its effect on privacy. However, we can derive a sufficient condition for (ϵ, δ)-DP that is close to the optimal necessary condition:

Theorem 5.4.1. *Releasing* $\mathbf{t} = \mathbf{A}\mathbf{x} + \boldsymbol{\xi}$, *where* $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$ *is* (ϵ, δ)*-DP whenever*

$$\epsilon \geq \Phi^{-1}\left(1 - \delta\right)\sqrt{2\eta} + \eta \tag{5.13}$$

where $\eta = \frac{\|\mathbf{A}\|_2^2 R^2}{2\lambda_{\min}}$.

Accuracy of approximation We numerically evaluate how close the sufficient condition ϵ_{approx} as given in theorem 5.4.1 is to the optimal ϵ_{opt} as given in theorem 1.2.2. For $\delta = 10^{-4}, 10^{-5}, 10^{-6}$ and $\nu \in (0, 20]$ we numerically approximate the exact lower bound using a binary search and compare it to the lower bound as obtained from approximation. As ν grows the absolute error in ϵ approximately converges to 1, while the relative error vanishes quickly. As this becomes apparent mainly in the very low privacy area, we assume that the approximation serves as a good proxy for the true lower bound of the achievable privacy guarantee given a fixed ν and δ (see figure 5.1).

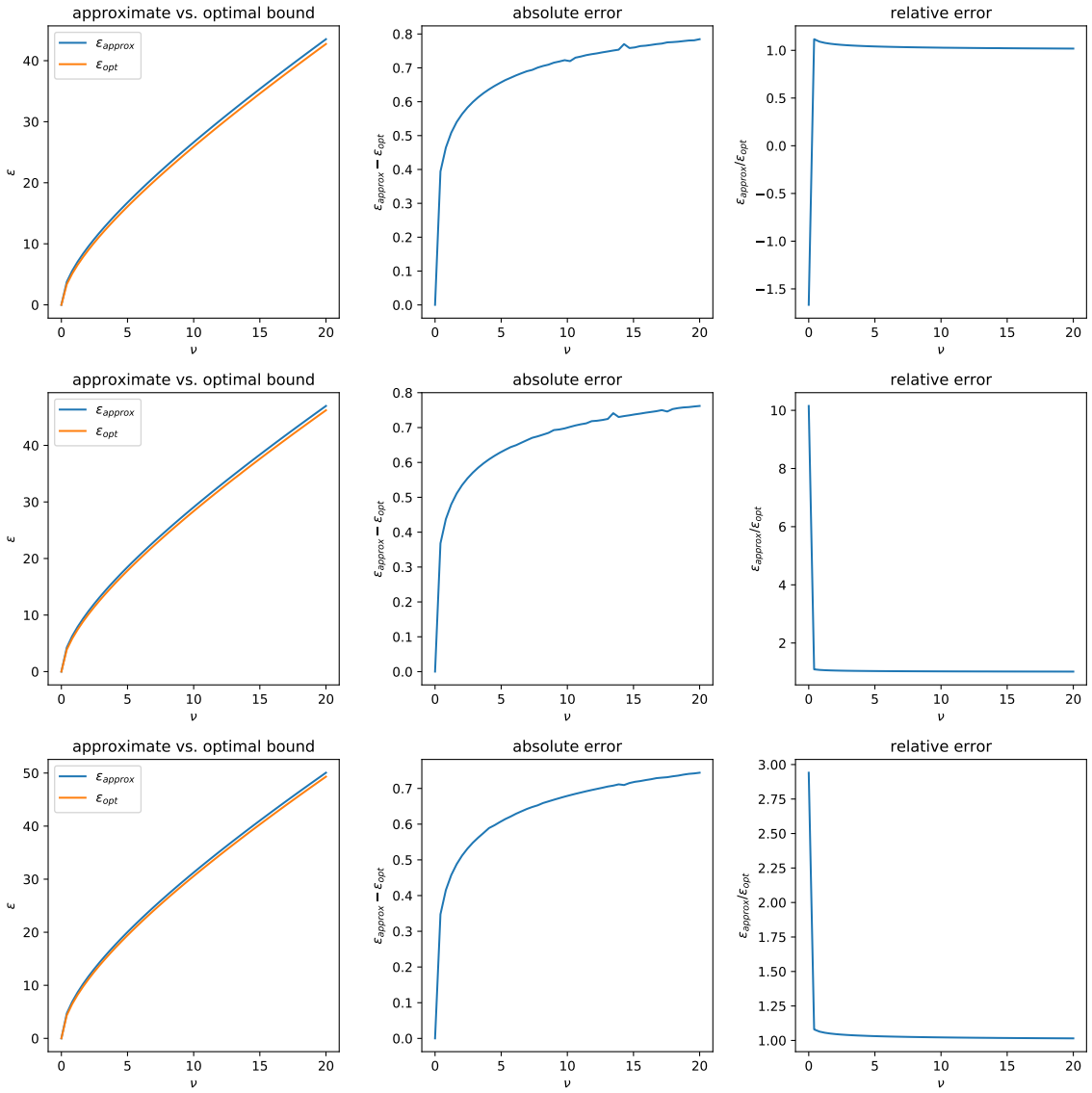

Figure 5.1: The error induced approximating the optimal ϵ which is achievable for a given δ and ν using the improved Gaussian mechanism by the sufficient condition as obtained in theorem 5.4.1. Columns show estimated bound, absolute error and relative error. Rows show $\delta = 10^{-4}$, $\delta = 10^{-5}$, $\delta = 10^{-6}$.

5.5 Relation of privacy bounds to information theoretic quantities.

We will now relate the obtained privacy bounds to the information theoretic quantities that drive the compression. If we study one-dimensional signals the Gaussian mechanism analysis for both compression problems depends on

$$\eta = \frac{a^2 R^2}{2\lambda^2} \geq \frac{a^2 \sigma^2}{2\lambda^2} = \frac{\mathrm{SNR}}{2}, \tag{5.14}$$

where SNR denotes the *transformed signal to noise ratio* of the channel. This quantity emerges as a natural lower bound to the achievable privacy risk using our derivation, as we have

$$\epsilon \geq \Phi^{-1}(1-\delta)\sqrt{\mathrm{SNR}} + \frac{\mathrm{SNR}}{2}. \tag{5.15}$$

Alternatively, we can study how the channel rate relates to the privacy level. For both channels we have the rate

$$I(x;t) = \frac{1}{2}\log\left(\frac{a^2\sigma^2 + \lambda^2}{\lambda^2}\right) \tag{5.16}$$

$$= \frac{1}{2}\log\left(\mathrm{SNR} + 1\right) \tag{5.17}$$

$$\tag{5.18}$$

from which we get

$$\mathrm{SNR} = \exp\left(2I(x;t)\right) - 1. \tag{5.19}$$

Thus if

$$n_{bits} = \frac{I(x;t)}{\log(2)} \tag{5.20}$$

denotes the number of bits that we require to encode $\mathbf{X}$ to $\mathbf{T}$ we have that

$$\epsilon \geq O(\exp(n_{bits})). \tag{5.21}$$

The numerically computed achievable rate in bits for a given guarantee of (ϵ, δ)-DP is plotted it in figure 5.2.

5.5.1 Compression of Gaussian data under $L2$-distortion

The signal to noise ratio of the optimal channel computes as

$$\mathrm{SNR}_{L_2} = \frac{a^2\sigma^2}{\lambda^2} = \beta\sigma - 1, \tag{5.22}$$

while its rate is given by

$$I(x;t) = \frac{1}{2}\log\left(\beta\sigma\right). \tag{5.23}$$

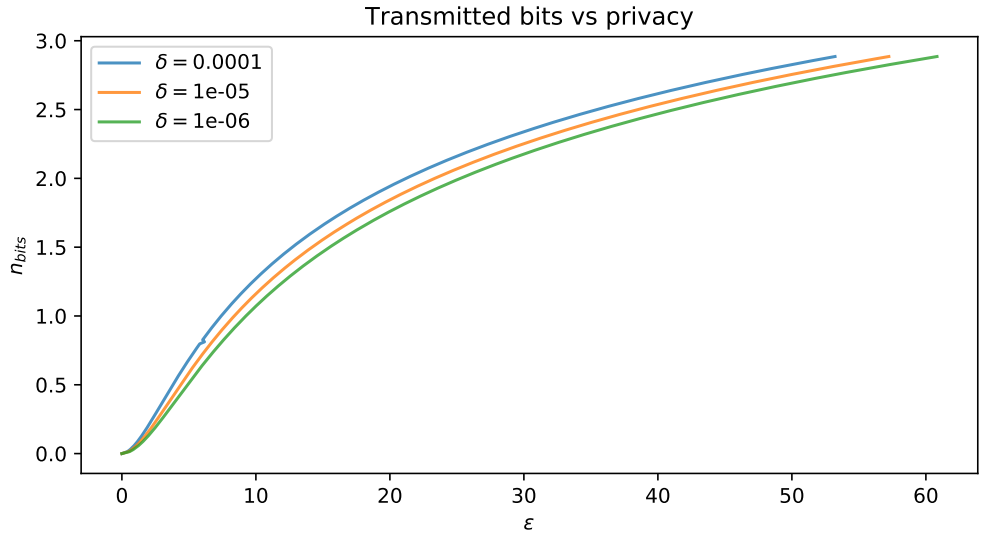

Figure 5.2: The achievable bit rate that can be transmitted with linear compression for a given approximate DP level of (ϵ, δ). We see that privacy risk grows as an exponential function of the number of preserved bits.

5.5.2 Gaussian Information Bottleneck

For the GIB where both x and y are one-dimensional we can write the joint covariance matrix as

$$\mathbf{\Sigma}_{x,y} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}. \tag{5.24}$$

This gives us

$$\sigma_{x|y}\sigma_x^{-2} = 1 - \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2} \tag{5.25}$$

which is the only Eigenvalue. The critical value β_c is now

$$\beta_c = \frac{1}{1 - \left(1 - \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}\right)} = \frac{\sigma_x^2\sigma_y^2}{\sigma_{xy}^2} = \frac{1}{\rho_{x,y}^2}, \tag{5.26}$$

where $\rho_{x,y}$ is the Pearson correlation coefficient between x and y. So for $\beta > \frac{1}{\rho_{x,y}^2}$ we get $\beta\rho_{x,y}^2 > 1$ and we have the optimal projection

$$a = \sqrt{\frac{\beta\rho_{x,y}^2 - 1}{\left(1 - \rho_{x,y}^2\right)\sigma_x^2}}. \tag{5.27}$$

Now $\lambda^2 = 1$ and thus the signal to noise ratio is given by

$$\mathrm{SNR}_{GIB} = \frac{a^2\sigma^2}{\lambda^2} = \frac{\beta\rho_{x,y}^2 - 1}{1 - \rho_{x,y}^2} \tag{5.28}$$

while the channel rate is

$$I(x;t) = \log \frac{\beta\rho_{x,y}^2 - 1}{1-\rho_{x,y}^2} + 1 \tag{5.29}$$

$$= \log \frac{(\beta-1)\rho_{x,y}^2}{1-\rho_{x,y}^2} \tag{5.30}$$

$$= \log \frac{\beta-1}{\beta_c-1}. \tag{5.31}$$

Here we observe, that the privacy/utility trade-off is determined by the Pearson correlation between data and auxiliary information. For a fixed Pearson correlation ρ_{xy}^2 the privacy risk only depends on the ratio $\frac{\beta-1}{\beta_c-1}$.

High-dimensional GIB In the high-dimensional case, the whole privacy analysis is determined by the biggest factor $\alpha_{\max}$. If $\sigma_1^2, \ldots, \sigma_D^2$ denote the Eigenvalues of $\mathbf{\Sigma_X}$ in descending order we will always have $\mathbf{v}^T\mathbf{\Sigma_X}\mathbf{v} \leq \sigma_1^2 \leq R^2$. Since $\beta_c > \lambda_i$ for each $\alpha_i \neq 0$, we obtain that

$$\eta \geq R^2\|\mathbf{A}\|_2^2 \geq \frac{\frac{\beta}{\beta_c} - 1}{\frac{\beta_c-1}{\beta_c}} \tag{5.32}$$

$$= \frac{\beta-1}{\beta_c-1}. \tag{5.33}$$

From this we can conclude, that the overall privacy risk of the high-dimensional GIB will always be higher than the risk of the one-dimensional GIB induced by the one-dimensional correlation between any two components x_i, y_j as given by their Pearson correlation ρ_{x_i,y_j}.

5.6 Impact of pruning low variance dimensions on the privacy guarantee

We will now study how pruning away the smallest Eigenvalues before compressing a multivariate Gaussian source with respect to the L_2 distortion affects the resulting privacy. In this situation we have

$$\eta = \frac{\left(\frac{\sigma_{\max}\beta-1}{\sigma_{\max}\beta}\right)^2 R^2}{\frac{\sigma_{\min}\beta-1}{\sigma_{\min}\beta^2}} \tag{5.34}$$

$$\geq \frac{\left(\frac{\sigma_{\max}\beta-1}{\sigma_{\max}\beta}\right)^2 \sigma_{\max}}{\frac{\sigma_{\min}\beta-1}{\sigma_{\min}\beta^2}} \tag{5.35}$$

$$= (\sigma_{\max}\beta - 1)\left(\frac{\sigma_{\max}\beta-1}{\sigma_{\min}\beta-1}\right)\frac{\sigma_{\min}}{\sigma_{\max}} \tag{5.36}$$

$$= (\sigma_{\max}\beta - 1)\underbrace{\left(\frac{\sigma_{\min}\beta - \frac{\sigma_{\min}}{\sigma_{\max}}}{\sigma_{\min}\beta-1}\right)}_{\alpha:=}. \tag{5.37}$$

So if $\frac{\sigma_{\min}}{\sigma_{\max}} \to 1$, we will have $\alpha \to 1$ and we will end up with the same privacy guarantees as in the one-dimensional channel. However, if $\frac{\sigma_{\min}}{\sigma_{\max}} \ll 1$, both $\beta \gg 1$ (as $\beta > \sigma_{\min}$ still has to hold)

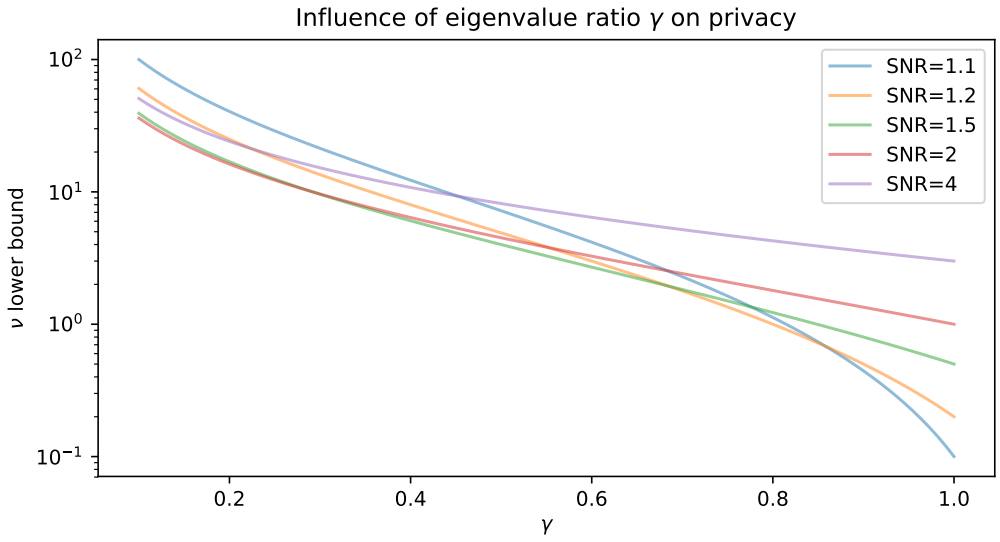

Figure 5.3: Here we plot the lower bound on ν (given in log coordinates) as derived in 5.6 for different signal to noise ratios as a function of $\gamma = \frac{\sigma_{\min}}{\sigma_{\max}}$. We see that ν grows quickly, as $\gamma \ll 1$. Thus, maintaining $\gamma \approx 1$ is crucial to maintain privacy.

and $\alpha \gg 1$ and thus the overall privacy risk will be big. By pruning away the $D - k$ smallest components first, such that $\frac{\sigma_{\min}}{\sigma_{\max}} = \gamma$ we obtain

$$\eta \geq \left(\frac{\sigma_{\min}\beta}{\gamma} - 1\right) \underbrace{\left(\frac{\sigma_{\min}\beta - \gamma}{\sigma_{\min}\beta - 1}\right)}_{\geq 1}. \tag{5.38}$$

Because $\sigma_{\min}\beta > 1$, this implies a privacy risk at least on level of the one-dimensional channel with $SNR = \frac{\sigma_{\min}\beta}{\gamma} \geq \frac{1}{\gamma}$. Thus keeping γ close to 1 is important to preserve privacy (see figure 5.3).

6. Experiments

We will now present experiments evaluating the practical feasibility of the derived methods for private data release. In a first experiment we compare the privacy guarantees as derived from the exponential mechanism to the approximated lower bound as derived from the improved Gaussian mechanism. In a second experiment, we study the privacy/utility trade-off using a toy classification task. Finally, we evaluate the privacy/utility trade-off of both methods for two real-world data sets involving sensitive data about individuals.

6.1 Comparison of derived privacy guarantees

In this experiment we compare how close the lower bound for ϵ as derived from the exponential mechanism is to the bound as derived from the Gaussian mechanism. We compare this for the Gaussian linear compression under L_2 distortion and the for the GIB for channels with varying Pearson correlation between the data and the auxiliary information. As derived in the last chapter, it suffices to study this relation in the one-dimensional case to get an estimate of the feasibility of higher dimensional instance.

Compression of Gaussian data under L_2-distortion As shown in figure 6.1 the derived bound from the exponential mechanism becomes loose very quickly, especially in the low privacy regime. Furthermore the constant offset of R shows a strict overestimation when $SNR \to 0$. If we compare the log ratios of both bounds, we see that it becomes constant in the low privacy regime.

GIB As shown in figure 6.2 a similar pattern emerges for the GIB. As the Pearson correlation $\rho_{x,y}$ between x and y grows, the bound obtained from the exponential mechanism approaches the bound obtained from the Gaussian mechanism. Especially when $SNR \to 0$ the exponential mechanism bound strictly overestimates the privacy risk due to the large offset around zero.

6.2 Toy experiment: 1D-Gaussians

We will now present a one-dimensional toy experiment to explore the privacy/utility trade-off for both methods, if privacy is bounded from below using the Gaussian mechanism.

Compression of Gaussian data under L_2-distortion We sample data $x \sim \mathcal{N}(0, 1)$ and assign labels

$$t = \begin{cases} 1 & x > 0 \\ 0 & o.w. \end{cases}. \tag{6.1}$$

Now we compress the data according to the derived optimal compression with respect to the L_2 distortion. We estimate the lower bound on the privacy risk by numerically computing the

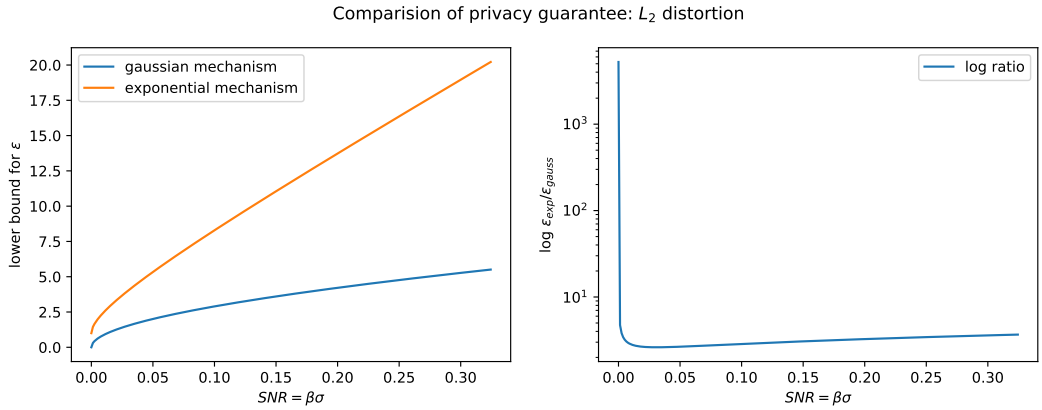

Figure 6.1: Comparison of the bounds on ϵ given $\delta = 10^{-5}$ for compressing Gaussian data with respect to L_2 distortion as derived from the exponential mechanism and the Gaussian mechanism respectively. Left: derived bounds by SNR. Right: log ratio of derived bounds by SNR.

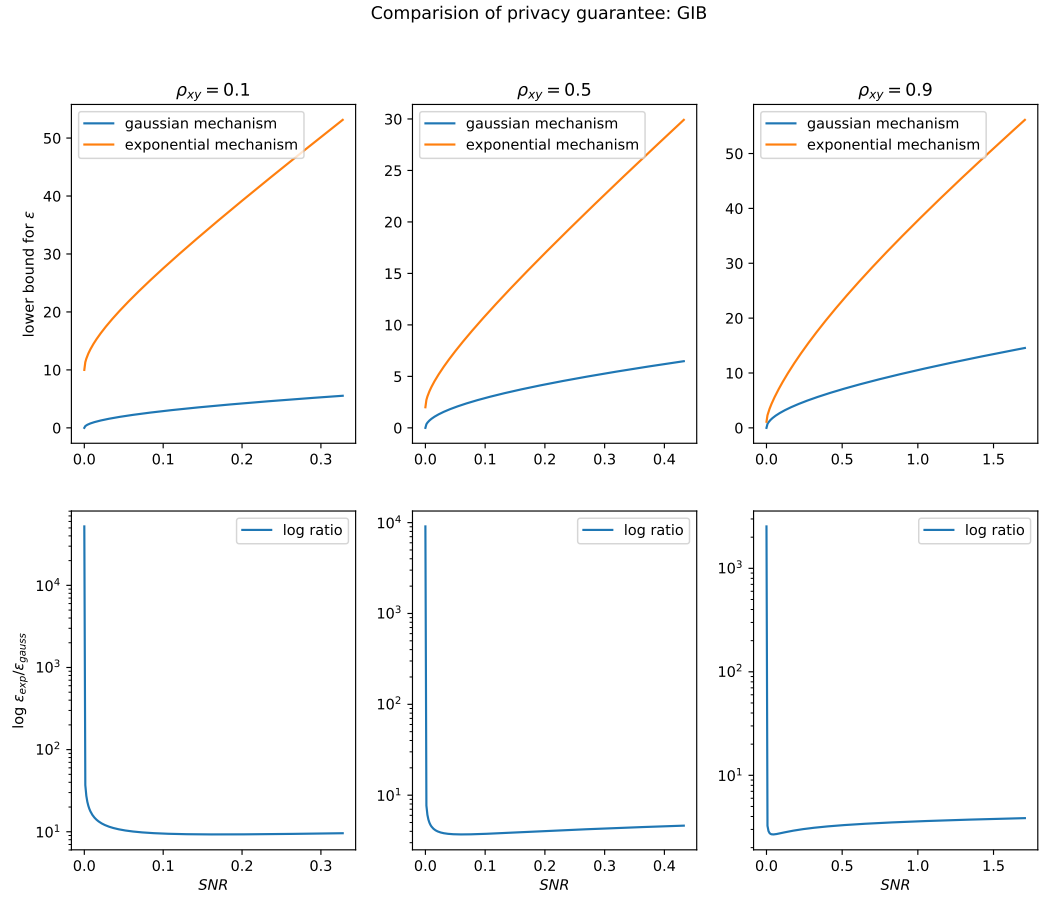

Figure 6.2: Comparison of the bounds on ϵ given $\delta = 10^{-5}$ for the GIB as derived from the exponential mechanism and the Gaussian mechanism respectively. Top row: derived bounds by SNR for Pearson correlation $0.1, 0.5, 0.9$. Bottom row: log ratio of derived bounds by SNR for Pearson correlation $0.1, 0.5, 0.9$.

result of 1.2.2 for $\delta = 10^{-5}$. Now, we evaluate the utility of the compressed representation by predicting the label

$$t' = \begin{cases} 1 & t > 0 \\ 0 & o.w. \end{cases} \tag{6.2}$$

Finally, we compute how well the sign of x can still be correctly predicted using the privatized representation t'.

GIB Here, we sample data

$$(x, y) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{x,y}) \tag{6.3}$$

where

$$\mathbf{\Sigma}_{x,y} = \begin{bmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{bmatrix}. \tag{6.4}$$

Again we assign the labels

$$t = \begin{cases} 1 & x > 0 \\ 0 & o.w. \end{cases}. \tag{6.5}$$

Now we compress the data according GIB using the auxiliary information y. Again, we estimate the lower bound on the privacy risk by numerically computing the result of 1.2.2 for $\delta = 10^{-5}$. As before, we evaluate the utility of the compressed representation by predicting the label

$$t' = \begin{cases} 1 & t > 0 \\ 0 & o.w. \end{cases} \tag{6.6}$$

Again, we compute how well the sign of x can still be correctly predicted using the privatized representation t'.

6.2.1 Results

As can be seen in figure 6.2.1 the best privacy/utility trade-off is obtained for compressing the data according to the L_2 distortion. This is a natural result, as the auxiliary data y is just a noisy encoding of this information. As $\rho_{\mathbf{X},\mathbf{Y}}$ grows the trade-off improves while the trade-off spanned by the L_2-compression is an upper bound. Even though we could obtain 100% accuracy for the non-compressed data, the prediction power of the privatized data is low. Especially in the high-privacy regime where $\epsilon \leq 1$, we barely beat a random guess. As we derived in the previous chapter, this effect will only get worse for high dimensions. Thus, we believe this toy experiment is already an indicator that both methods can not achieve a privacy/utility trade-off leading to an applicable private data release mechanism for real-world data.

6.3 Experiments on real data

In our last set of experiments we try to see whether both methods could be utilized to privatize real-world data while still preserving utility. For this purpose we chose two UCI classification data sets containing sensitive information about individuals and estimate to which extend linear compression is able to preserve information about the class.

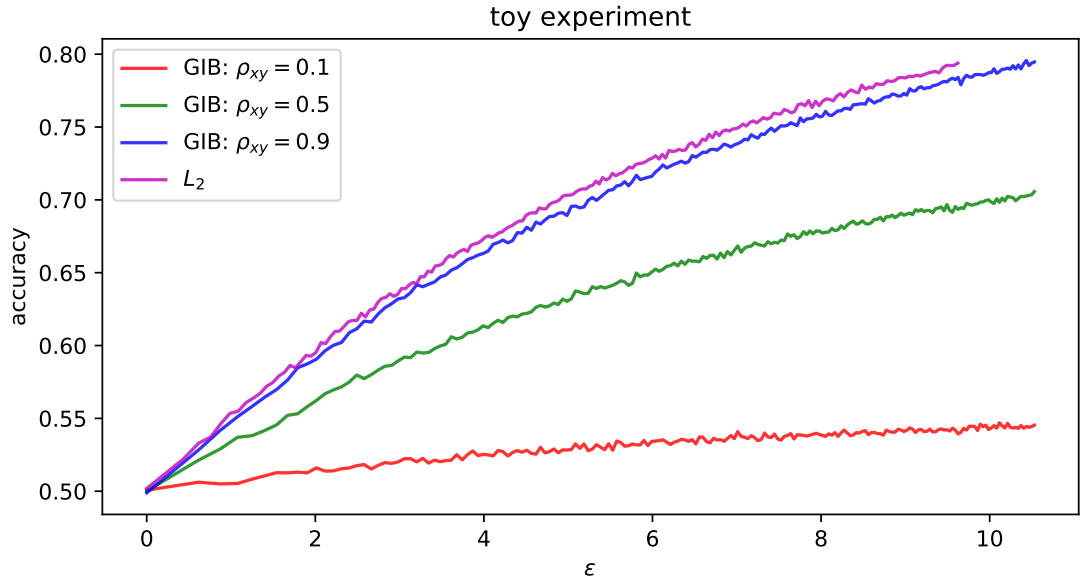

Figure 6.3: Shown the accuracy of predicting whether the compressed representation t stems from an x with $x > 0$. Across the four methods the compression based on L_2-distortion preserves most of this information in t. For an increase in correlation among x and y the privacy/utility trade-off of the GIB approaches this upper baseline.

6.3.1 Breast cancer data set

The Wisconsin breast cancer data set [52, 56] contains 699 patient samples with 10 attributes describing the result of a biopsy. The goal is to predict, whether the sample is benign (the patient is healthy) or whether it is malignant (the patient has developed breast cancer). The class distribution is given in figure 6.4. After removing NaN entries, we are left with 683 instances and 9 attributes. As all features are given by real values, we do not conduct any additional pre-processing except for zero-centering the data to satisfy the assumption that the data stems from a zero centered multivariate normal distribution.

6.3.2 Drug consumption data set

The data set[20] contains the answers of 1885 individuals who reported on their drug consumption. For a variety of different drugs they grade their behavior with a numerical scale from 1

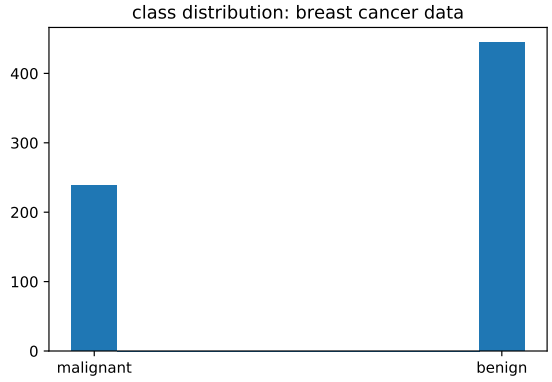

Figure 6.4: Class distribution of the breast cancer data set

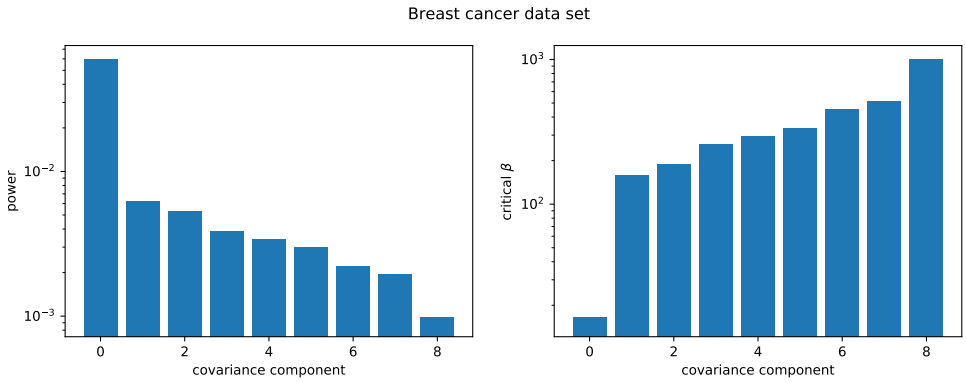

Figure 6.5: Distribution of eigenvalues of the covariance $\mathbf{\Sigma}$ (left) and the critical values of β determining when a component would be pruned out (right).

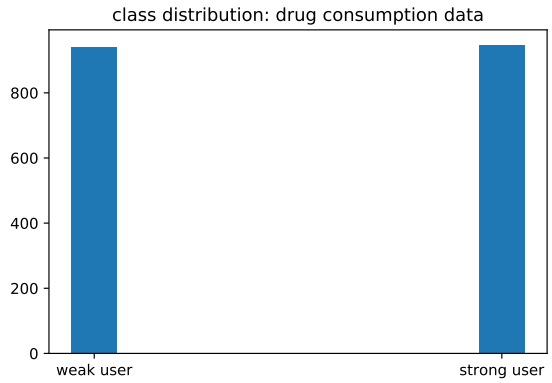

Figure 6.6: Class distribution of the drug consumption data set.

(no user) to 6 (heavy user). This is accompanied with some other features about the person, like their age or different psychometric scores. Together this data set contains 32 attributes. We now split the attributes into features that we want to protect, attributes that we want to predict and attributes that we consider as auxiliary information for the task. The first 13 attributes describe data about the participant that we consider to be the private part. The other attributes are their answers on drug consumption. A couple of drugs are legal and common drugs, like chocolate, coffee, alcohol or nicotine. All other drugs are considered to be hard drugs. By averaging the grade of the hard drugs, we obtain a single number that is an indicator whether a participant is a strong or a weak user of hard drugs. We now set a threshold on this indicator that splits the data set into two parts of roughly the same size (see figure 6.6). From this, we can formulate the prediction task as trying to identify a strong or a weak user given the provided features. Here we consider the vector of grades about the consumption of legal drugs as auxiliary information that might guide the compression in the GIB formulation. Besides zero-centering the features and the auxiliary information in order to satisfiy the mutual zero-centered Gaussian assumption we do not perform any further pre-processing.

6.3.3 Experimental setup

As we assume our data set to be Gaussian, we model the classification problem with a logistic regression. For both classification tasks, we first compute the baseline that always predicts one class. Furthermore, we used cross-validation to identify a non-private baseline that utilizes all original information about the individuals. We further assume, that we do not have to estimate

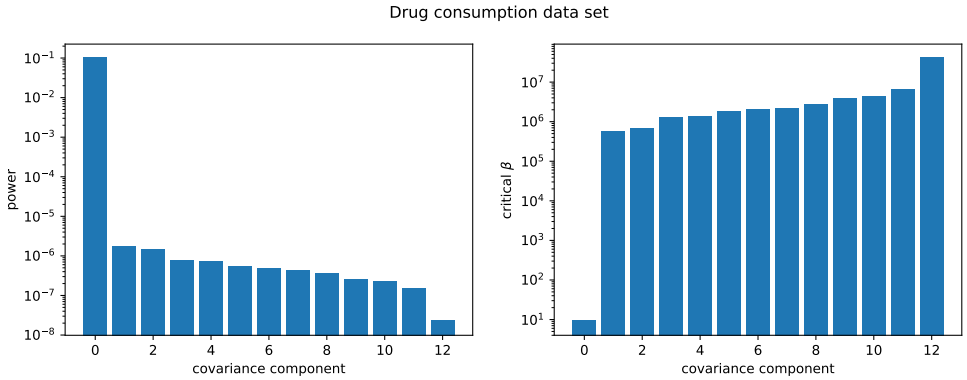

Figure 6.7: Distribution of eigenvalues of the covariance $\mathbf{\Sigma}$ (left) and the critical values of β determining when a component would be pruned out (right).

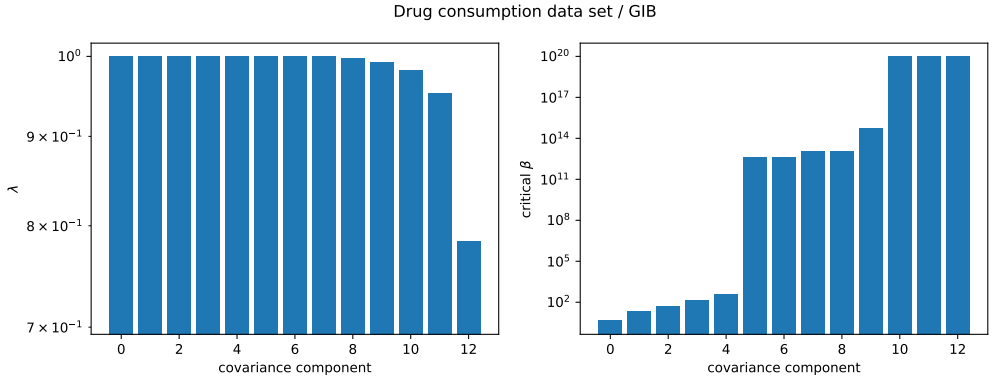

Figure 6.8: Distribution of eigenvalues λ_i of $\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{\Sigma}_{\mathbf{X}}^{-1}$ (left) and the critical values of β determining when a component would be pruned out (right).

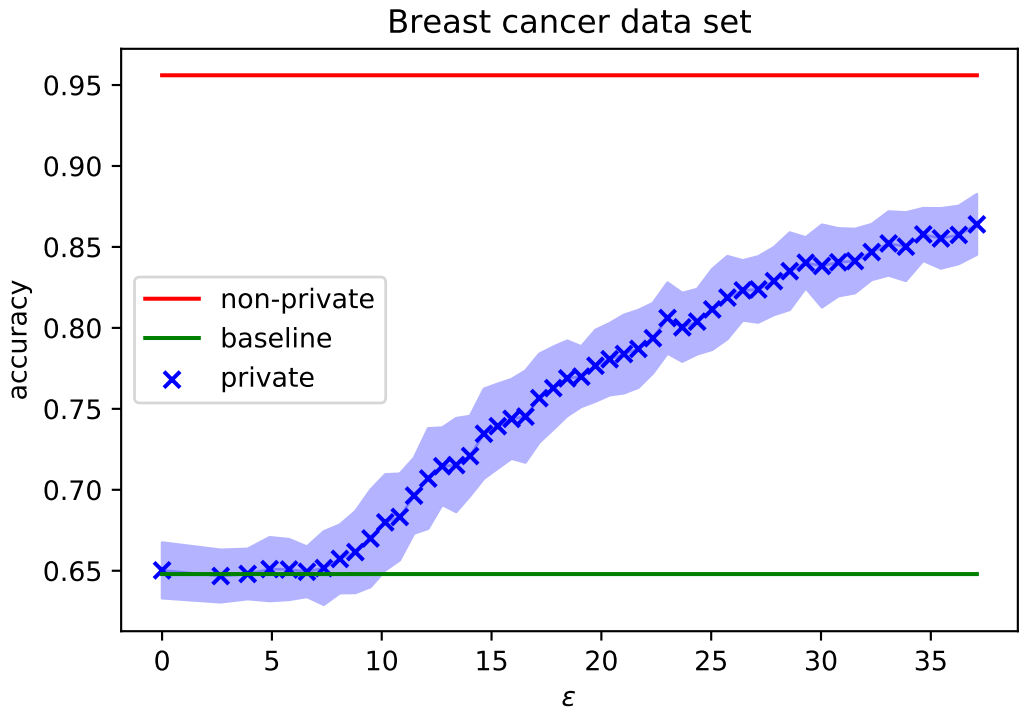

Figure 6.9: Accuracy for predicting breast cancer based on the features after having compressing them using the L_2 compression. Each point represents the mean of 100 cross-validations together with one standard deviation.

the covariance from the private data. This assumption would imply that this covariance was already estimated e.g. on public data. For both methods and $\delta = 10^{-5}$ we compute the optimal compression that can be achieved for a fixed lower bound of ϵ. Then we use cross-validation (100 random 50/50 splits of the data) to estimate the accuracy of a logistic regression classifier trained on the compressed data. As explained before, the privacy/utility trade-off in the high-dimensional setting is driven by the ratio of Eigenvalues (L_2 compression) or the critical values β_c (GIB). We report these values for both data sets in figures 6.5, 6.7 and 6.8.

6.3.4 Results

As we can see in figures 6.9, 6.10 and 6.11 in all three experiments (breast cancer data with L_2 compression, drug consumption data with L_2 and GIB compression) just significantly outperforming the baseline already requires a privacy risk of $\epsilon = 5$ and beyond. This means the probability to identify an individual within the data set is already increased by a factor of $\exp(5) \approx\sim 150$ and thus implies a serious privacy risk. Furthermore, we see that getting close to the non-private baseline requires a tremendously high privacy risk of $\epsilon \geq 20$ corresponding to a probability increase by a factor of $5 \cdot 10^7$. We conclude that a release of any of these data sets using any of the proposed compression methods would either yield a representation that only provides very inferior results for the actual task or implies a far too high privacy risk.

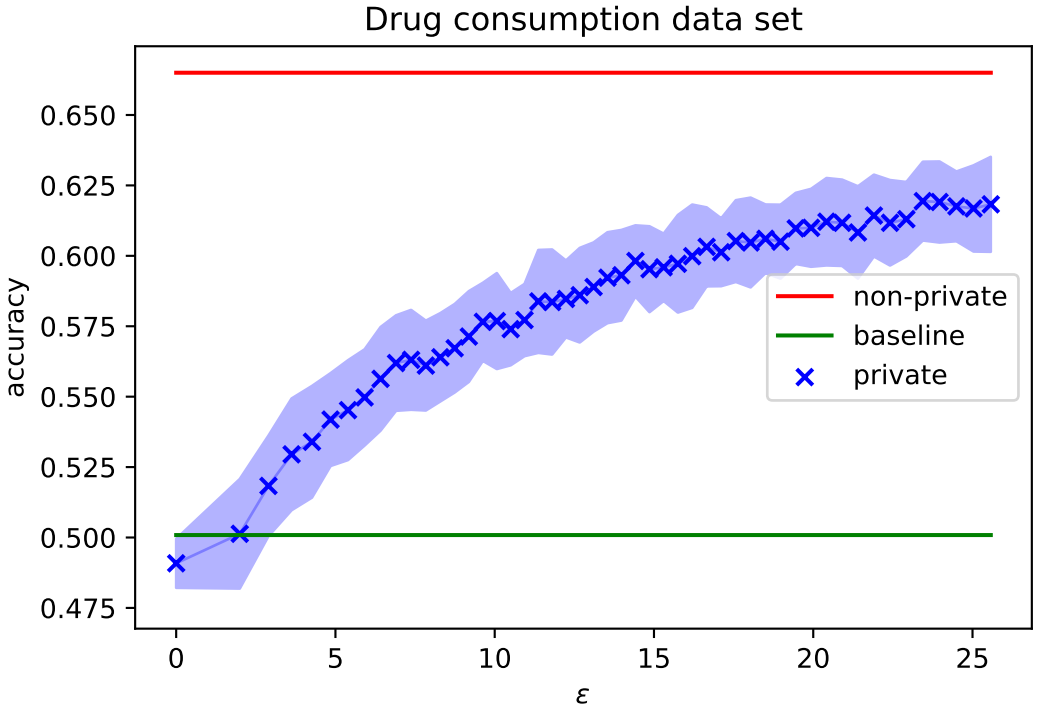

Figure 6.10: Accuracy for predicting drug usage based on the features after compressing them using the L_2 compression. Each point represents the mean of 100 cross-validations together with one standard deviation.

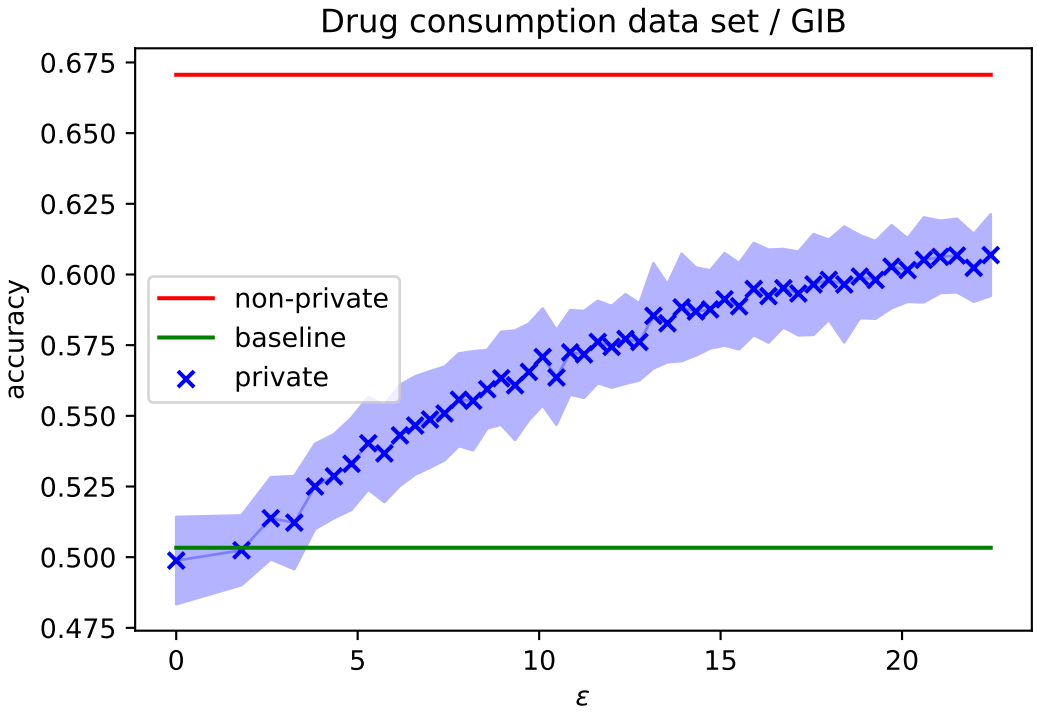

Figure 6.11: Accuracy for predicting drug usage based on the features after compressing them using the L_2 compression. Each point represents the mean of 100 cross-validations together with one standard deviation.

7. Going beyond Gaussianity

In this chapter we will first lay out why extending this work to non-Gaussian data is difficult. Then we discuss a couple of directions to incorporate more complex distributions (e.g. dropping the assumption of Gaussian data), to have fewer assumptions on the point-wise transformation (e.g. allowing non-linear transformations) and to allow a broader class of distortion measures. Due to the time constraint of this thesis project, we will leave any feasibility test or implementation of the proposed extensions as an open question for future research.

7.1 Difficulties of going beyond Gaussianity

In general we can structure the difficulties of a extending this work according to the following fundamental hardness results:

- Solving the RD or IB problem in general is NP hard [41]. Thus any practical solutions must involve relaxations or strong assumptions, e.g. as we proposed them for Gaussian data and linear compression schemes.
- Even sampling from the optimal solution of the RD problem as given in theorem 2.2.1 is NP hard as well [50]. This implies we need to guarantee that an approximate sampling scheme still preserves privacy guarantees.

As a result, any extension must introduce strong assumptions in order to be tractable or must try to approximate the true solution. This yields the following technical problems, that we see as major points for future work:

- Even *an infinitesimally close approximate solution to a true private solution might not be private*. Right now there are only few guarantees that relate closeness of an approximation to closeness in terms of privacy guarantees. E.g. in [50] it was shown, that if an approximation $Q_{\mathbf{T}|\mathbf{X}}$ is δ-close to the true, ϵ-DP distribution $P_{\mathbf{T}|\mathbf{X}}$ *for each possible adjacent* $\mathbf{X}$ as measured in L_1 norm, then $Q_{\mathbf{T}|\mathbf{X}}$ is also an $(\epsilon, g(\epsilon, \delta))$-DP distribution, where g is a function of ϵ and δ. The obvious challenge originates from the difficulty to prove this closeness for each possible pair of adjacent data sets.
- By assuming a specific source distribution $P_{\mathbf{X}}$ or a simplified but tractable notion of utility, we might be able to solve the RD problem analytically or sample from the true RD solution efficiently. However, in such cases for different source distributions or more involved utility measures such a solution would *not be optimal*. E.g. if we knew that our data stems from a low variance Gaussian mixture model where the mixture components are highly separated, the proposed linear compression would totally overestimate the necessary variance required to explain the data. While an optimal encoding for such a mixture distribution (with respect to L_2-distortion) would just encode each cluster independently, the linear projection approach must push everything to the origin and thus destroy most of the information about the cluster dependency.

- Complex models of the point-wise transformation might *require a high effort of private parameter estimation*. In the presented examples, we just needed to estimate the covariance of the source from either the private sample using a private estimator or by access to a public sample. Especially, with the use of general function approximators for approximating the RD solution, e.g. using deep neural networks, this becomes a severe issue on its own. It requires either to have access to a very big public data set, or to have very efficient DP parameter estimators for high-dimensional and very complex models. The first case might already make the need for a data set privatization method obsolete while the latter remains an open question and is still undergoing heavy research.

We will now propose approaches that could be used to either sample or approximate the optimal solution in future work and briefly discuss their benefits and challenges.

7.2 Optimization of the RD problem over convex sets

[32] gives an approximation of to the RD problem, by assuming a class of parameterized compression functions $\mathcal{Q} = \left\{Q_{\mathbf{T}|\mathbf{X},\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\right\}$, such that $\mathcal{Q}$ is a convex set over parameters $\boldsymbol{\theta}$. The original RD problem is now relaxed to the problem of finding

$$\boldsymbol{\theta}_* = \arg\min_{\theta\in\Theta} I_{\mathbf{x}\sim P_{\mathbf{X}},\ \mathbf{t}\sim Q_{\mathbf{T}|\mathbf{X},\boldsymbol{\theta}}}(\mathbf{X},\mathbf{T}) + \beta\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}},\ \mathbf{t}\sim Q_{\mathbf{T}|\mathbf{X},\boldsymbol{\theta}}}\left[d(\mathbf{x},\mathbf{t})\right]. \tag{7.1}$$

If the true minimizer distribution is already contained in $\mathcal{Q}$, both problems coincide. We can now see that our restriction to linear compression schemes is exactly such a convex class for which the constrained and the global optimum of the RD problem match if the data is Gaussian distributed. If the data is non-Gaussian, better compression schemes might exist, though those would leave the class of linear projections. Now [32] shows that for a convex class of compression functions finding the constrained optimum is feasible using a modification of the BA algorithm. Furthermore, this optimization gives strong guarantees on the rate of convergence. A convex class offering more modeling complexity with respect to the L_2 distortion compared to linear functions is given by Gaussian mixture models. As shown in [32] this class can efficiently be approximated using their modified BA algorithm and gives close approximations to the true RD-curve. A first challenge of this method will be getting tight privacy bounds. While it seems reasonable to bound the sensitivity of picking a mixture component $\mu(\mathbf{x})$ depending on the input $\mathbf{x}$ (e.g. by containing each cluster point within a unit ball) and then analyze the privacy induced by sampling $\mathbf{t}$ from a Gaussian around $\mu(\mathbf{x})$, possibly tighter bounds could be obtained using other means as we could also show in this thesis. A bigger challenge would be the private estimation of the mixture components and responsibilities as the algorithm as given in [32] would require multiple EM steps conducted over the whole data set. While Gaussian mixture models can also be estimated from small public samples (e.g. compared to deep neural networks), finding the convex minimum having only access to private data would require a modification of the algorithm e.g. by using techniques similar as presented in [43].

7.3 Approximations of the information bottleneck

Another route could be taken by approximating the information bottleneck using parameterized families $\mathcal{Q} = \left\{Q_{\mathbf{T}|\mathbf{X},\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\right\}$ as has been done in recent work [12, 3, 2] using variational inference. While we can only obtain privacy guarantees for the exact minimum of the IB problem, we can again analyze the privacy of intermediate results using complementary techniques. E.g.

the proposed model of [3] approximates $P_{\mathbf{T}|\mathbf{X}}$ with a deep neural network via the approximation

$$Q_{\mathbf{T}|\mathbf{X},\boldsymbol{\theta}} = \mathcal{N}\left(\boldsymbol{\mu}_\theta(\mathbf{x}), \sigma_{\boldsymbol{\theta}}^{cond}(\mathbf{x})\mathbf{I}\right) \tag{7.2}$$

$$Q_{\mathbf{T}} = \mathcal{N}\left(\mathbf{0}, \sigma_{\boldsymbol{\theta}}^{prior}\mathbf{I}\right), \tag{7.3}$$

where $\boldsymbol{\mu}_\cdot(\mathbf{x}), \sigma_{\boldsymbol{\theta}}^{cond}(\cdot)$ are the outputs of a multilayer perceptron. By controlling the sensitivity of $\boldsymbol{\mu}_\theta(\cdot)$ and enforcing a lower bound on $\sigma_{\boldsymbol{\theta}}^{cond}(\cdot)$ any point-wise release using $Q_{\mathbf{T}|\mathbf{X},\boldsymbol{\theta}}$ could again be seen as an instance of the improved Gaussian mechanism as we showed it for linear functions in section 5.4. Besides the major question, whether such hard constraints on the approximate posterior mean and variance still allow for effective learning the other big challenge lies in estimating $\boldsymbol{\theta}$ in a private way. [3] train their model using standard gradient based techniques such as SGD or Adam [31]. Whether it is possible to achieve a good privacy/utility tradeoff using private parameter estimation techniques, e.g. by using DP-SGD[1] remains an open question to this point that has to be evaluated in future work.

7.4 Particle based sampling of the minimizer distribution

In a final approach, we could think about sampling from 2.2.1 directly. As already mentioned the two major challenges are that the partition function in general stays intractable and for complicated distortion functions the re-weighting term $\exp(-\beta(\cdot,\cdot))$ might have a very complicated form. If d however is differentiable or possesses a sub-gradient, it is possible to apply recently proposed particle based sampling algorithms, such as [35, 4]. These methods iteratively update a set of initial particles only requiring access to $\nabla d(\cdot,\cdot)$. While this yields the least constrained approach to directly sample from 2.2.1 without learning a model in between this method again possesses two major limitations. Without clear bounds on the rate of convergence, we cannot analyze privacy according to the exponential mechanism derivation. The only guarantee of convergence for these particle based methods exist in the weak limit [34]. The other challenge lies in the iterative update procedure of the methods. Each update requires access to the original data. As we do not model a transformation or distribution but transform the private data iteratively, we also cannot use public data in this case. A possible solution could be updating the particles using DP gradient based estimators, such as DP-SGD[1] and accumulate the total privacy risk. Both questions, whether finite time convergence guarantees could be obtained for particle based sampling methods or whether an acceptable privacy/utility trade-off can be achieved by privatizing particle update steps are up to this point open questions for future research.

8. Summary & Outlook

During this thesis we presented differential privacy as a formal guarantee for algorithms to protect the privacy of individuals in a data set. We presented the problem of private data set release and how it could be approached by attacking it with tools from optimal compression. Due to the difficulty of solving optimal compression on its own, we studied two analytically tractable instances of this theory and derived the privacy bounds that can be predicted according to our framework. We then related the privacy guarantees to the information theoretic quantities driving the compression and evaluated the feasibility of the linear approaches for practical use. As our experiments showed, both approaches do not seem to provide a privacy/utility trade-off that could be used for sensitive data in real scenarios. Subsequently, we turned to discussing more complex instances of the framework using approximate approaches to solve the RD/IB problems. Due to the time constraints of this thesis and due to many open challenges that are still required to be solved first, we leave a practical evaluation of any of the proposed extensions for future work. Finally, we investigated the implication, that compression can yield data privatization even within the very strict notion of DP. While it has been broadly discussed before, that DP implies constraints on information theoretic quantities (e.g. [15]), this reverse direction is very unexplored so far. Our results do not yield a new method that could be used for solving the practical problem of private data release. However, we hope that the presented work might still provide new insights towards a better understanding of this reverse direction and maybe lead to useful private data release mechanisms in the future.

References

[1] Martin Abadi et al. "Deep learning with differential privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2016, pp. 308–318.

[2] Alessandro Achille and Stefano Soatto. "Information dropout: Learning optimal representations through noisy computation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[3] Alexander A Alemi et al. "Deep variational information bottleneck". In: *arXiv preprint arXiv:1612.00410* (2016).

[4] Luca Ambrogioni et al. "Wasserstein Variational Gradient Descent: From Semi-Discrete Optimal Transport to Ensemble Variational Inference". In: *arXiv preprint arXiv:1811.02827* (2018).

[5] Suguru Arimoto. "An algorithm for computing the capacity of arbitrary discrete memoryless channels". In: *IEEE Transactions on Information Theory* 18.1 (1972), pp. 14–20.

[6] Borja Balle and Yu-Xiang Wang. "Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising". In: *arXiv preprint arXiv:1805.06530* (2018).

[7] Daniel Barth-Jones. "The 're-identification' of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now". In: *SSRN* (2012). Available at SSRN: https://ssrn.com/abstract=2076397 or http://dx.doi.org/10.2139/ssrn.2076397.

[8] T. Berger and J. D. Gibson. "Lossy source coding". In: *IEEE Transactions on Information Theory* 44.6 (1998), pp. 2693–2723. ISSN: 0018-9448. DOI: 10.1109/18.720552.

[9] Richard Blahut. "Computation of channel capacity and rate-distortion functions". In: *IEEE transactions on Information Theory* 18.4 (1972), pp. 460–473.

[10] Jeremiah Blocki et al. "The Johnson-Lindenstrauss Transform Itself Preserves Differential Privacy". In: *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. FOCS '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 410–419. ISBN: 978-0-7695-4874-6. DOI: 10.1109/FOCS.2012.67. URL: https://doi.org/10.1109/FOCS.2012.67.

[11] Avrim Blum, Katrina Ligett, and Aaron Roth. "A learning theory approach to noninteractive database privacy". In: *Journal of the ACM (JACM)* 60.2 (2013), p. 12.

[12] Matthew Chalk, Olivier Marre, and Gasper Tkacik. "Relevant sparse codes with variational information bottleneck". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1957–1965.

[13] Gal Chechik et al. "Information bottleneck for Gaussian variables". In: *Journal of machine learning research* 6 (2005), pp. 165–188.

[14] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[15] Paul Cuff and Lanqing Yu. "Differential privacy as a mutual information constraint". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2016, pp. 43–54.

[16] Cynthia Dwork. "Differential privacy: A survey of results". In: *International Conference on Theory and Applications of Models of Computation*. Springer. 2008, pp. 1–19.

[17] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Found. Trends Theor. Comput. Sci.* 9 (2014), pp. 211–407. ISSN: 1551-305X. DOI: 10.1561/0400000042. URL: http://dx.doi.org/10.1561/0400000042.

[18] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. "Boosting and differential privacy". In: *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*. IEEE. 2010, pp. 51–60.

[19] Cynthia Dwork et al. "Analyze gauss: optimal bounds for privacy-preserving principal component analysis". In: *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. ACM. 2014, pp. 11–20.

[20] Elaine Fehrman et al. "The Five Factor Model of personality and evaluation of drug consumption risk". In: *Data Science*. Springer, 2017, pp. 231–242.

[21] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2015, pp. 1322–1333.

[22] Amir Globerson and Naftali Tishby. "On the Optimality of the Gaussian Information Bottleneck Curve". In: *Hebrew University Technical Report.* (2004). Retrieved from http://people.csail.mit.edu/~gamir/pubs/gopt_tr04.ps on 12/10/2018, 12:07 pm.

[23] Peter Harremoës and Naftali Tishby. "The information bottleneck revisited or how to choose a good distortion measure". In: *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*. IEEE. 2007, pp. 566–570.

[24] Daniel Hsu, Sham Kakade, Tong Zhang, et al. "A tail inequality for quadratic forms of subgaussian random vectors". In: *Electronic Communications in Probability* 17 (2012).

[25] Dong Huang et al. "Orthogonal mechanism for answering batch queries with differential privacy". In: *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*. ACM. 2015, p. 24.

[26] Zhanglong Ji, Zachary C Lipton, and Charles Elkan. "Differential privacy and machine learning: a survey and review". In: *arXiv preprint arXiv:1412.7584* (2014).

[27] Wuxuan Jiang, Cong Xie, and Zhihua Zhang. "Wishart Mechanism for Differentially Private Principal Components Analysis." In: *AAAI*. 2016, pp. 1730–1736.

[28] Xiaoqian Jiang et al. "Differential-Private Data Publishing Through Component Analysis". In: *Transactions on data privacy* 6.1 (Apr. 2013), pp. 19–34. URL: https://www.ncbi.nlm.nih.gov/pubmed/24409205.

[29] Shiva Prasad Kasiviswanathan et al. "What can we learn privately?" In: *SIAM Journal on Computing* 40.3 (2011), pp. 793–826.

[30] Krishnaram Kenthapadi et al. "Privacy via the johnson-lindenstrauss transform". In: *arXiv preprint arXiv:1204.2606* (2012).

[31] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[32] Yuval Kochman and Ram Zamir. "Computation of the Rate-Distortion Function Relative to a Parametric Class of Reproductions". In: *Proceedings of 41st annual Allerton conference on Communication, Control and Computing*. 2003, pp. 211–220.

[33] Yang D. Li et al. "Compressive Mechanism: Utilizing Sparse Representation in Differential Privacy". In: *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society*. WPES '11. Chicago, Illinois, USA: ACM, 2011, pp. 177–182. ISBN: 978-1-4503-1002-4. DOI: 10.1145/2046556.2046581. URL: http://doi.acm.org/10.1145/2046556.2046581.

[34] Qiang Liu. "Stein variational gradient descent as gradient flow". In: *Advances in neural information processing systems*. 2017, pp. 3115–3123.

[35] Qiang Liu and Dilin Wang. "Stein variational gradient descent: A general purpose bayesian inference algorithm". In: *Advances In Neural Information Processing Systems*. 2016, pp. 2378–2386.

[36] H Brendan McMahan et al. "Learning differentially private language models without losing accuracy". In: *arXiv preprint arXiv:1710.06963* (2017).

[37] Frank McSherry and Kunal Talwar. "Mechanism design via differential privacy". In: *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*. IEEE. 2007, pp. 94–103.

[38] Darakhshan J Mir. "Differential privacy: an exploration of the privacy-utility landscape". PhD thesis. Rutgers University-Graduate School-New Brunswick, 2013.

[39] Darakhshan J. Mir. "Information-Theoretic Foundations of Differential Privacy". In: *Proceedings of the 5th International Conference on Foundations and Practice of Security*. FPS'12. Montreal, QC, Canada: Springer-Verlag, 2013, pp. 374–381. ISBN: 978-3-642-37118-9. DOI: 10.1007/978-3-642-37119-6_25. URL: http://dx.doi.org/10.1007/978-3-642-37119-6_25.

[40] Noman Mohammed et al. "Differentially Private Data Release for Data Mining". In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '11. San Diego, California, USA: ACM, 2011, pp. 493–501. ISBN: 978-1-4503-0813-7. DOI: 10.1145/2020408.2020487. URL: http://doi.acm.org/10.1145/2020408.2020487.

[41] Brendan Mumey and Tomáš Gedeon. "Optimal mutual information quantization is NP-complete". In: *Proc. Neural Inf. Coding (NIC) Workshop*. 2003.

[42] Nicolas Papernot et al. "Semi-supervised knowledge transfer for deep learning from private training data". In: *arXiv preprint arXiv:1610.05755* (2016).

[43] Mijung Park et al. "DP-EM: Differentially Private Expectation Maximization". In: *arXiv preprint arXiv:1605.06995* (2016).

[44] Claude E Shannon. "A mathematical theory of communication". In: *Bell System Technical journal* 27.3 (1948), pp. 379–423.

[45] Claude E Shannon. "Probability of error for optimal codes in a Gaussian channel". In: *Bell System Technical Journal* 38.3 (1959), pp. 611–656.

[46] Reza Shokri et al. "Membership inference attacks against machine learning models". In: *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE. 2017, pp. 3–18.

[47] Naftali Tishby, Fernando C Pereira, and William Bialek. "The information bottleneck method". In: *Proceedings of 37th Annual Allerton Conference on Communication, Control and Computing* (1999), pp. 368–377.

[48] Jonathan Ullman. "Answering nˆ2+o(1) Counting Queries with Differential Privacy is Hard". In: *SIAM Journal on Computing* 45.2 (2016), pp. 473–496.

[49] Isabel Wagner and David Eckhoff. "Technical privacy metrics: a systematic survey". In: *ACM Computing Surveys (CSUR)* 51.3 (2018), p. 57.

[50] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. "Privacy for free: Posterior sampling and stochastic gradient monte carlo". In: *International Conference on Machine Learning*. 2015, pp. 2493–2502.

[51] Hermann Weyl. "Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)". In: *Mathematische Annalen* 71.4 (1912), pp. 441–479. ISSN: 1432-1807. DOI: 10.1007/BF01456804. URL: https://doi.org/10.1007/BF01456804.

[52] William H Wolberg and Olvi L Mangasarian. "Multisurface method of pattern separation for medical diagnosis applied to breast cytology." In: *Proceedings of the national academy of sciences* 87.23 (1990), pp. 9193–9196.

[53] Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. "Differential privacy via wavelet transforms". In: *IEEE Transactions on knowledge and data engineering* 23.8 (2011), pp. 1200–1214.

[54] Ganzhao Yuan et al. "Optimizing batch linear queries under exact and approximate differential privacy". In: *ACM Transactions on Database Systems (TODS)* 40.2 (2015), p. 11.

[55] Chiyuan Zhang et al. "Understanding deep learning requires rethinking generalization". In: *arXiv preprint arXiv:1611.03530* (2016).

[56] Jianping Zhang. "Selecting typical instances in instance-based learning". In: *Machine Learning Proceedings 1992*. Elsevier, 1992, pp. 470–479.

[57] S. Zhou, K. Ligett, and L. Wasserman. "Differential privacy with compression". In: *2009 IEEE International Symposium on Information Theory*. 2009, pp. 2718–2722. DOI: 10.1109/ISIT.2009.5205863.

[58] T. Zhu et al. "Differentially Private Data Publishing and Analysis: A Survey". In: *IEEE Transactions on Knowledge and Data Engineering* 29.8 (2017), pp. 1619–1638. ISSN: 1041-4347. DOI: 10.1109/TKDE.2017.2697856.

A. Appendix

A.1 Facts

A.1.1 Differential entropy of Gaussians

Theorem A.1.1. *Let* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. *Then*

$$h(\mathbf{x}) = \frac{1}{2} \log \left| \det 2\pi \mathbf{\Sigma} \right|. \tag{A.1}$$

A.1.2 Linear transformations of Gaussians

Theorem A.1.2. *Let* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}')$ *and* $\mathbf{A}, \mathbf{B}$ *be linear transformations. Then*

$$\mathbf{Ax} + \mathbf{By} \sim \mathcal{N}(\mathbf{0}, \mathbf{A\Sigma A}^T + \mathbf{B\Sigma' B}^T). \tag{A.2}$$

A.1.3 Expectation of squared norm

Theorem A.1.3. *Let* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. *Then* $\mathbb{E}[\|\mathbf{x}\|_2^2] = \mathbf{tr}\left(\mathbf{\Sigma}\right)$.

A.1.4 Tail-bound for squared norms

Theorem A.1.4 (Proposition 1.1 of [24])**.** *Let* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. *Then we have the tail-bound*

$$P\left[\|\mathbf{x}\|_2^2 > \mathbf{tr}\left(\mathbf{\Sigma}\right) + 2\sqrt{\mathbf{tr}\left(\mathbf{\Sigma}^2\right) t} + 2\|\mathbf{\Sigma}\|_2\right] < \exp(-t) \tag{A.3}$$

where $\|\mathbf{\Sigma}\|_2$ *denotes the spectral norm of the matrix* $\mathbf{\Sigma}$ *(i.e. its biggest eigenvalue).*

A.2 Proofs

A.2.1 Proof of lemma 1.3.1

Let $\mathbf{X}, \mathbf{X}' \sim P_{\mathbf{X}}^N$ be datasets such that $|\mathbf{X} \Delta \mathbf{X}'| = 1$. W.o.l.g. let $\mathbf{x}_n = \mathbf{x}'_n$ and $\mathbf{x}_N \neq \mathbf{x}'_N$. Let $S = S_1 \times \cdots \times S_N \subset \mathcal{T}^N$. Then we get

$$\log \frac{P^N_{\mathbf{T}|\mathbf{X}=f^N(\mathbf{X})}[S]}{P^N_{\mathbf{T}|\mathbf{X}=f^N(\mathbf{X}')}[S]} \overset{*}{=} \log \frac{\prod_{n=1}^N P_{\mathbf{T}|\mathbf{X}=f(\mathbf{x}_n)}[S_n]}{\prod_{n=1}^N P_{\mathbf{T}|\mathbf{X}=f(\mathbf{x}'_n)}[S_n]} \tag{A.4}$$

$$= \log \frac{P_{\mathbf{T}|\mathbf{X}=f(\mathbf{x}_n)}[S_N]}{P_{\mathbf{T}|\mathbf{X}=f(\mathbf{x}'_n)}[S_N]} \tag{A.5}$$

$$\leq \epsilon. \tag{A.6}$$

Here * follows from the fact, that the transformation f^N is defined point-wise.

A.2.2 Proof of theorem 2.3.1

We first see that for any $\rho \in \mathbb{P}$ and $\mathbf{X}, \mathbf{Y}, \mathbf{T} \sim \rho$ we get

$$
\begin{aligned}
I(\mathbf{X};\mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) && \text{(A.7)}\\
&\overset{*}{=} H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X},\mathbf{T}) && \text{(A.8)}\\
&= H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{T}) + H(\mathbf{Y}|\mathbf{T}) - H(\mathbf{Y}|\mathbf{X},\mathbf{T}) && \text{(A.9)}\\
&= I(\mathbf{Y};\mathbf{T}) + I(\mathbf{Y};\mathbf{X}|\mathbf{T}). && \text{(A.10)}
\end{aligned}
$$

where * follows from the Markov factorization of ρ. Now we have

$$
\begin{aligned}
I(\mathbf{Y};\mathbf{X}|\mathbf{T}) &= \int d\mathbf{t}\rho_{\mathbf{T}}(\mathbf{t}) \int d\mathbf{x}\; d\mathbf{y}\; \rho_{\mathbf{X},\mathbf{Y}|\mathbf{T}=\mathbf{t}}(\mathbf{x},\mathbf{y}) \log \frac{\rho_{\mathbf{X},\mathbf{Y}|\mathbf{T}=\mathbf{t}}(\mathbf{x},\mathbf{y})}{\rho_{\mathbf{X}|\mathbf{T}=\mathbf{t}}(\mathbf{x})\rho_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}(\mathbf{y})} && \text{(A.11)}\\
&= \int d\mathbf{t}\rho_{\mathbf{T}}(\mathbf{t}) \int d\mathbf{x}\; d\mathbf{y}\; \rho_{\mathbf{X},\mathbf{Y}|\mathbf{T}=\mathbf{t}}(\mathbf{x},\mathbf{y}) \log \frac{\rho_{\mathbf{Y}|\mathbf{X}=\mathbf{x},\mathbf{T}=\mathbf{t}}(\mathbf{y})\rho_{\mathbf{X}|\mathbf{T}=\mathbf{t}}(\mathbf{x})}{\rho_{\mathbf{X}|\mathbf{T}=\mathbf{t}}(\mathbf{x})\rho_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}(\mathbf{y})} && \text{(A.12)}\\
&= \int d\mathbf{x}\; d\mathbf{t}\; p_{\mathbf{X}}(\mathbf{x})\rho_{\mathbf{T}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) \int d\mathbf{y}\rho_{\mathbf{Y}|\mathbf{X}=\mathbf{x},\mathbf{T}=\mathbf{t}}(\mathbf{x}) \log \frac{\rho_{\mathbf{Y}|\mathbf{X}=\mathbf{x},\mathbf{T}=\mathbf{t}}(\mathbf{y})}{\rho_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}(\mathbf{y})} && \text{(A.13)}\\
&\overset{*}{=} \int d\mathbf{x}\; d\mathbf{t}\; p_{\mathbf{X}}(\mathbf{x})\rho_{\mathbf{T}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) \int d\mathbf{y}\rho_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{x}) \log \frac{\rho_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y})}{\rho_{\mathbf{Y}|\mathbf{T}=\mathbf{t}}(\mathbf{y})} && \text{(A.14)}\\
&= \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}},\ \mathbf{t}\sim\rho_{\mathbf{T}|\mathbf{X}=\mathbf{x}}} \left[D_{KL}\left[p_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \| \rho_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} \right] \right] && \text{(A.15)}\\
&= D(\rho), && \text{(A.16)}
\end{aligned}
$$

where again * follows from the Markov factorization of ρ. Combining these results, we obtain

$$
\begin{aligned}
\arg\min_{\rho\in\mathbb{P}} I(\mathbf{X};\mathbf{T}) - \beta I(\mathbf{T};\mathbf{Y}) &= \arg\min_{\rho\in\mathbb{P}} I(\mathbf{X};\mathbf{T}) - \beta I(\mathbf{X};\mathbf{Y}) + \beta D(\rho) && \text{(A.17)}\\
&= \arg\min_{\rho\in\mathbb{P}} I(\mathbf{X};\mathbf{T}) + \beta D(\rho), && \text{(A.18)}
\end{aligned}
$$

since $I(\mathbf{X};\mathbf{Y})$ is constant in ρ and will not affect the choice of the minimizer.

A.2.3 Proof of theorem 2.4.1

Due to our lemma (1.3.1) it suffices to show that the obtained feature transformation is not too sensitive with respect to our data. However, here we can utilize a proof that is similar to the privacy proof given for the exponential mechanism in [37]: let $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Then we get

$$
\begin{aligned}
\left| \log \frac{P_{\mathbf{T}|\mathbf{X}=\mathbf{x}}(\mathbf{t})}{P_{\mathbf{T}|\mathbf{X}=\mathbf{x}'}(\mathbf{t})} \right| &\leq \left| \beta \left(d(\mathbf{x},\mathbf{t}) - d(\mathbf{x}',\mathbf{t}) \right) \right| && \text{(A.19)}\\
&\quad + \left| \log \int d\mathbf{t} \exp\left(-\beta d(\mathbf{x},\mathbf{t})\right) P_{\mathbf{T}}(\mathbf{t}) - \log \int d\mathbf{t} \exp\left(-\beta d(\mathbf{x}',\mathbf{t})\right) P_{\mathbf{T}}(\mathbf{t}) \right|\\
&\leq |\beta\Delta d| + \left| \log \int d\mathbf{t} \exp\left(-|\beta d(\mathbf{x},\mathbf{t}) - \beta d(\mathbf{x}',\mathbf{t})|\right) P_{\mathbf{T}}(\mathbf{t}) \right|\\
&\leq |\beta\Delta d| + \left| \log \int d\mathbf{t} \exp\left(-|\beta\Delta d|\right) P_{\mathbf{T}}(\mathbf{t}) \right| && \text{(A.20)}\\
&\leq 2\beta\Delta d. && \text{(A.21)}
\end{aligned}
$$

.

A.2.4 Proof of theorem 2.5.1

Set $M := \left\{\sup_{\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}} |d(\mathbf{x}_1,\mathbf{t}) - d(\mathbf{x}_2,\mathbf{t})| > \epsilon\right\}$. Then we have for any measurable set $S \subset \mathcal{T}$

$$\begin{aligned} P[\mathbf{t}\in S|\mathbf{x}_1] &= P[\mathbf{t}\in S|\mathbf{x}_1,\mathbf{t}\in M]\cdot P[\mathbf{t}\in M] + P[\mathbf{t}\in S|\mathbf{x}_1,\mathbf{t}\notin M]\cdot P[\mathbf{t}\notin M] \\ &\leq \delta + P[\mathbf{t}\in S|\mathbf{x}_1,\mathbf{t}\notin M]\cdot P[\mathbf{t}\notin M] \\ &\leq \delta + \exp(\epsilon)P[\mathbf{t}\in S|\mathbf{x}_2,\mathbf{t}\notin M]\cdot P[\mathbf{t}\notin M] \\ &\leq \delta + \exp(\epsilon)\left(P[\mathbf{t}\in S|\mathbf{x}_2,\mathbf{t}\notin M]\cdot P[\mathbf{t}\notin M] + P[\mathbf{t}\in S|\mathbf{x}_2,\mathbf{t}\in M]\cdot P[\mathbf{t}\in M]\right) \\ &= \delta + \exp(\epsilon)P[\mathbf{t}\in S|\mathbf{x}_2] \end{aligned} \tag{A.22}$$

A.2.5 Proofs of theorems 3.1.1 and 3.1.2

Lemma A.2.1. *For a fixed $a \neq 0$ the minimum of $\mathcal{L}$ is attained at*

$$\lambda^2 = \frac{a^2\sigma^2}{2}\left(\sqrt{1+\frac{4}{a^2\sigma^2\beta}} - 1\right). \tag{A.23}$$

Proof. Solving $\frac{\partial\mathcal{L}}{\partial\lambda} = 0$ gives us

$$\frac{\partial\mathcal{L}}{\partial\lambda^2} = \left(a^2\sigma^2+\lambda^2\right)^{-1} - \left(\lambda^2\right)^{-1} + \beta = 0 \tag{A.24}$$

$$\implies \lambda^2 - a^2\sigma^2 - \lambda^2 + \beta\left(a^2\sigma^2+\lambda^2\right)\lambda^2 = 0 \tag{A.25}$$

$$\implies \lambda^4 + a^2\sigma^2\lambda^2 - \frac{a^2\sigma^2}{\beta} = 0. \tag{A.26}$$

Solving the square gives us

$$\lambda^2 \in \{-\frac{a^2\sigma^2}{2} \pm \sqrt{\frac{a^4\sigma^4}{4} + \frac{a^2\sigma^2}{\beta}}\}. \tag{A.27}$$

Using the constraint that $\lambda^2 > 0$ finally yields

$$\lambda^2 = \frac{a^2\sigma^2}{2}\left(\sqrt{1+\frac{4}{a^2\sigma^2\beta}} - 1\right). \tag{A.28}$$

□

Lemma A.2.2. *The optimal projection only exists iff $\sigma^2\beta > 1$. In this case*

$$a = \frac{\sigma^2\beta - 1}{\sigma^2\beta} \in (0,1). \tag{A.29}$$

Proof. Solving $\frac{\partial\mathcal{L}}{\partial a} = 0$ for a gives

$$\frac{\partial\mathcal{L}}{\partial a} = \left(a^2\sigma^2+\lambda^2\right)^{-1} a\sigma^2 - \beta(1-a) = 0 \tag{A.30}$$

Now we can reuse the result obtained in (A.2.1) to get

$$\implies 0 = \frac{a}{\frac{a^2\sigma^2}{2}\left(\sqrt{1+\frac{4}{a^2\sigma^2\beta}}+1\right)} - \beta + \alpha\beta \tag{A.31}$$

$$\implies \underbrace{\beta}_{>0} = a\left(\underbrace{\beta}_{>0} + \underbrace{\frac{1}{\frac{a^2\sigma^2}{2}\left(\sqrt{1+\frac{4}{a^2\sigma^2\beta}}+1\right)}}_{>0}\right). \tag{A.32}$$

From this equation, we can see that $a > 0$. We derive further

$$\implies \underbrace{\beta}_{>0}(1-a) = \underbrace{\frac{1}{\frac{a\sigma^2}{2}\left(\sqrt{1+\frac{4}{a^2\sigma^2\beta}}+1\right)}}_{>0} \tag{A.33}$$

from which we conclude $(1-a) > 0$ and thus $a \in (0,1)$. Now we can continue with

$$\implies \frac{1}{\beta(1-a)} = \frac{a^2\sigma^2}{2}\left(\sqrt{1+\frac{4}{a^2\sigma^2\beta}}+1\right) \tag{A.34}$$

$$\implies 1 = \frac{a(a-1)\sigma^2\beta}{2}\left(\sqrt{1+\frac{4}{a^2\sigma^2\beta}}+1\right) \tag{A.35}$$

$$\implies \left(\frac{2}{a(1-a)\sigma^2\beta}-1\right)^2 = 1+\frac{4}{a^2\sigma^2\beta} \tag{A.36}$$

$$\implies \frac{4}{a^2(1-a)^2\sigma^4\beta^2} - \frac{4}{a(1-a)\sigma^2\beta} + 1 = 1 + \frac{4}{a^2\sigma^2\beta} \tag{A.37}$$

$$\implies \frac{1-a(1-a)\sigma^2\beta}{a^2(1-a)^2\sigma^4\beta^2} = \frac{1}{a^2\sigma^2\beta} \tag{A.38}$$

$$\implies \frac{1-a(1-a)\sigma^2\beta}{(1-a)^2\sigma\beta} = 1 \tag{A.39}$$

$$\implies 1-a(1-a)\sigma^2\beta = (1-a)^2\sigma^2\beta \tag{A.40}$$

$$\implies 1-a\sigma^2\beta+a^2\sigma^2\beta = a^2\sigma^2\beta - 2a\sigma^2\beta + \sigma^2\beta \tag{A.41}$$

$$\implies a = 1-\frac{1}{\sigma^2\beta} = \frac{\sigma^2\beta-1}{\sigma^2\beta}. \tag{A.42}$$

□

Lemma A.2.3. *The optimal noise covariance only exists iff $\sigma^2\beta > 1$. In this case*

$$\lambda = \frac{a}{\beta}, \tag{A.43}$$

where a is given as in lemma (A.2.2).

Proof. Plugging the result of lemma (A.2.2) into the result of lemma (A.2.1) yields

$$\lambda^2 = \frac{(\sigma^2\beta-1)^2\sigma^2}{2\sigma^4\beta^2}\left(\sqrt{1+\frac{4}{\frac{(\sigma^2\beta-1)^2}{\sigma^4\beta^2}\sigma^2\beta}}-1\right) \tag{A.44}$$

$$= \frac{(\sigma^2\beta-1)^2}{2\sigma^2\beta^2}\left(\sqrt{1+\frac{4\sigma^2\beta}{(\sigma^2\beta-1)^2}}-1\right) \tag{A.45}$$

$$= \frac{(\sigma^2\beta-1)^2}{2\sigma^2\beta^2}\left(\frac{1}{\sigma^2\beta-1}\sqrt{(\sigma^2\beta-1)^2+4\sigma^2\beta}-1\right) \tag{A.46}$$

$$= \frac{(\sigma^2\beta-1)^2}{2\sigma^2\beta^2}\left(\frac{1}{\sigma^2\beta-1}\sqrt{\sigma^4\beta^2-2\sigma^2\beta+1+4\sigma^2\beta}-1\right) \tag{A.47}$$

$$= \frac{(\sigma^2\beta-1)^2}{2\sigma^2\beta^2}\left(\frac{1}{\sigma^2\beta-1}(\sigma^2\beta+1)-1\right) \tag{A.48}$$

$$= \frac{(\sigma^2\beta-1)^2}{2\sigma^2\beta^2}\left(\frac{\sigma^2\beta+1-(\sigma^2\beta-1)}{\sigma^2\beta-1}\right) \tag{A.49}$$

$$= \frac{(\sigma^2\beta-1)^2}{2\sigma^2\beta^2}\left(\frac{2}{\sigma^2\beta-1}\right) \tag{A.50}$$

$$= \frac{\sigma^2\beta-1}{\sigma^2\beta^2}. \tag{A.51}$$

□

Lemma A.2.4. *For the optimal a, λ^2, we have the relation*

$$\beta + \frac{1}{a^2\sigma^2+\lambda^2} = \frac{1}{\lambda^2}. \tag{A.52}$$

Proof. Plugging in the results for λ^2 and a gives us

$$\beta + \frac{1}{a^2\sigma^2+\lambda^2} = \beta + \frac{1}{\left(\frac{\sigma^2\beta-1}{\sigma^2\beta}\right)^2\sigma^2+\frac{\sigma^2\beta-1}{\sigma^2\beta}} \tag{A.53}$$

$$= \beta + \frac{1}{\frac{(\sigma^2\beta-1)^2}{\sigma^2\beta^2}+\frac{\sigma^2\beta-1}{\sigma^2\beta}} \tag{A.54}$$

$$= \beta + \frac{\sigma^2\beta}{(\sigma^2\beta-1)^2+\sigma^2\beta-1} \tag{A.55}$$

$$= \frac{(\sigma^2\beta-1)^2\beta+\sigma^2\beta^2-\beta+\sigma^2\beta^2}{(\sigma^2\beta-1)^2+\sigma^2\beta-1} \tag{A.56}$$

$$= \frac{(\sigma^2\beta-1)^2\beta+\sigma^2\beta^2-\beta+\sigma^2\beta^2}{\sigma^4\beta^2-2\sigma^2\beta+1+\sigma^2\beta-1} \tag{A.57}$$

$$= \frac{\sigma^4\beta^3-2\sigma^2\beta^2+\beta+2\sigma^2\beta^2-\beta}{\sigma^4\beta^2-\sigma^2\beta} \tag{A.58}$$

$$= \frac{\sigma^4\beta^3}{\sigma^4\beta^2-\sigma^2\beta} \tag{A.59}$$

$$= \frac{\sigma^2\beta^2}{\sigma^2\beta-1} \tag{A.60}$$

$$= \frac{1}{\lambda^2} \tag{A.61}$$

□

Proof of theorem 3.1.1. Follows directly from lemmas (A.2.2) and (A.2.3). □

Proof of theorem 3.1.2. We have

$$
\begin{aligned}
-\log p_{T|X=x}(t) &= \frac{1}{2}\frac{(t-ax)^2}{\lambda_*^2} + const && \text{(A.62)}\\
&= \frac{1}{2}\frac{1}{\lambda_*^2}t^2 - \frac{a}{\lambda_*^2}tx + const && \text{(A.63)}\\
&= \frac{1}{2}\beta t^2 + \frac{1}{2}\frac{1}{a_*^2\sigma^2+\lambda_*^2}t^2 - \beta tx + const && \text{(A.64)}\\
&= \frac{1}{2}\beta(t-x)^2 + \frac{1}{2}\frac{1}{a_*^2\sigma^2+\lambda_*^2}t^2 + const && \text{(A.65)}\\
&= \beta d(x,t) - \log p_T(t) && \text{(A.66)}
\end{aligned}
$$

where eq. (A.64) uses lemma (A.2.4) and thus

$$p_{T|X=x}(t) \propto \exp(-\beta d(x,t)) \cdot p(t). \tag{A.67}$$

□

A.2.6 Proofs of theorems 3.2.1 and 3.2.2

Lemma A.2.5. *The expected distortion is*

$$\mathbb{E}_{\mathbf{x}\sim p_{\mathbf{X}},\ \mathbf{t}\sim p_{\mathbf{T}|\mathbf{X}=\mathbf{x}}}\left[d(\mathbf{x},\mathbf{t})\right] = \frac{1}{2}\left(\mathbf{tr}\left(\boldsymbol{\Sigma}\right) + \mathbf{tr}\left(\left(\mathbf{I}-\mathbf{A}\right)\boldsymbol{\Sigma}\left(\mathbf{I}-\mathbf{A}\right)^T\right)\right). \tag{A.68}$$

Proof.

$$
\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim p_{\mathbf{X}},\ \mathbf{t}\sim p_{\mathbf{T}|\mathbf{X}=\mathbf{x}}}\left[d(\mathbf{x},\mathbf{t})\right] &= \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}),\ \mathbf{t}\sim\mathcal{N}(\mathbf{A}\mathbf{x},\boldsymbol{\Lambda})}\left[\|\mathbf{x}-\mathbf{t}\|_2^2\right] && \text{(A.69)}\\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}),\ \boldsymbol{\xi}\sim\mathcal{N}(0,\boldsymbol{\Lambda})}\left[\|\mathbf{x}-\mathbf{A}\mathbf{x}-\boldsymbol{\xi}\|_2^2\right]\\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}),\ \boldsymbol{\xi}\sim\mathcal{N}(0,\boldsymbol{\Lambda})}\left[\|(\mathbf{I}-\mathbf{A})\,\mathbf{x}-\boldsymbol{\xi}\|_2^2\right].
\end{aligned}
$$

Further, by using theorem A.1.2 we get

$$(\mathbf{I}-\mathbf{A})\,\mathbf{x}-\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},(\mathbf{I}-\mathbf{A})\,\boldsymbol{\Sigma}\,(\mathbf{I}-\mathbf{A})^T+\boldsymbol{\Lambda}). \tag{A.70}$$

Finally, we can use theorem A.1.3 to conclude

$$\mathbb{E}_{(\mathbf{I}-\mathbf{A})\mathbf{x}-\boldsymbol{\xi}\sim\mathcal{N}(0,(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{A})^T+\boldsymbol{\Lambda})}\left[\|(\mathbf{I}-\mathbf{A})\,\mathbf{x}-\boldsymbol{\xi}\|_2^2\right] = \mathbf{tr}\left(\left(\mathbf{I}-\mathbf{A}\right)\boldsymbol{\Sigma}\left(\mathbf{I}-\mathbf{A}\right)^T+\boldsymbol{\Lambda}\right). \tag{A.71}$$

□

Lemma A.2.6. *The optimal perturbation covariance $\boldsymbol{\Lambda}_*$ is given by*

$$\boldsymbol{\Lambda}_* = \mathbf{V}\mathbf{D}_{\boldsymbol{\Lambda}}\mathbf{V}^T \tag{A.72}$$

where

$$\mathbf{A}_*\boldsymbol{\Sigma}\mathbf{A}_*^T = \mathbf{V}\mathbf{D}\mathbf{V}^T \tag{A.73}$$

is the eigendecomposition of $\mathbf{A}_*\boldsymbol{\Sigma}\mathbf{A}_*^T$ given the optimal projection $\mathbf{A}_*$ and the diagonal matrix $\mathbf{D}_{\boldsymbol{\Lambda}}$ is defined by

$$\mathbf{D}_{\boldsymbol{\Lambda}} = \frac{\mathbf{D}}{2}\left(\sqrt{\mathbf{I}+\frac{4}{\beta\mathbf{D}}}-\mathbf{I}\right). \tag{A.74}$$

Proof. We solve

$$\mathbf{0} = \frac{\partial \mathcal{L}}{\partial \mathbf{\Lambda}} \tag{A.75}$$

$$\implies \mathbf{0} = \left(\mathbf{A\Sigma A}^T + \mathbf{\Lambda}\right)^{-1} - \mathbf{\Lambda}^{-1} + \beta\mathbf{I} \tag{A.76}$$

$$\implies \mathbf{0} = \mathbf{\Lambda} - \mathbf{A\Sigma A}^T + \mathbf{\Lambda} + \beta\left(\mathbf{A\Sigma A}^T + \mathbf{\Lambda}\right)\mathbf{\Lambda} \tag{A.77}$$

$$\implies \mathbf{0} = \mathbf{\Lambda}^2 + \mathbf{A\Sigma A}^T\mathbf{\Lambda} - \frac{1}{\beta}\mathbf{A\Sigma A}^T \tag{A.78}$$

This matrix equation is quadratic in $\mathbf{\Lambda}$ and possesses a solution set

$$\mathbf{\Lambda} \in \left\{-\frac{1}{2}\mathbf{A\Sigma A}^T \pm \mathbf{B}\right\} \tag{A.79}$$

whenever there is a matrix $\mathbf{B}$ such that

$$\mathbf{BB}^T = \frac{1}{4}(\mathbf{A\Sigma A}^T)^2 + \frac{1}{\beta}\mathbf{A\Sigma A}^T. \tag{A.80}$$

As $\mathbf{A}$ has full rank the matrix $\mathbf{A\Sigma A}^T$ is positive definite and thus possesses an eigendecomposition

$$\mathbf{A\Sigma A}^T = \mathbf{VDV}^T. \tag{A.81}$$

Furthermore, due to positive definiteness we know for each diagonal element $d_i > 0$. Therefore, we can set

$$\mathbf{B} = \mathbf{V}\sqrt{\frac{1}{4}\mathbf{D}^2 + \frac{1}{\beta}\mathbf{D}}\mathbf{V}^T \tag{A.82}$$

where the square-root is considered element-wise over the diagonal elements. By plugging back into (A.79) and rewriting, we obtain

$$\mathbf{\Lambda} \in \left\{\mathbf{V}\frac{D}{2}\left(-\mathbf{I} \pm \sqrt{\mathbf{I} + \frac{4}{\beta\mathbf{D}}}\right)\mathbf{V}^T\right\}. \tag{A.83}$$

However, as we require $\mathbf{\Lambda}$ to be positive definite, this leads to the unique solution

$$\mathbf{\Lambda} = \mathbf{V}\frac{\mathbf{D}}{2}\left(\sqrt{\mathbf{I} + \frac{4}{\beta\mathbf{D}}} - \mathbf{I}\right)\mathbf{V}^T \tag{A.84}$$

$$= \mathbf{VD}_{\mathbf{\Lambda}}\mathbf{V}^T. \tag{A.85}$$

□

Lemma A.2.7. *The optimal projection matrix $\mathbf{A}_*$ is given by*

$$\mathbf{A}_* = \mathbf{V}\left(\mathbf{I} - \frac{1}{\beta}\left(\mathbf{D} + \mathbf{D}_{\mathbf{\Lambda}_*}\right)^{-1}\right)^{-1}\mathbf{V}^T \tag{A.86}$$

Proof. We solve

$$0 = \frac{\partial \mathcal{L}}{\partial \mathbf{A}} \tag{A.87}$$

$$\implies 0 = \left(\mathbf{A\Sigma A} + \mathbf{\Lambda}\right)^{-1}\mathbf{A\Sigma} - \beta\left(\mathbf{I} - \mathbf{A}\right)\mathbf{\Sigma} \tag{A.88}$$

$$\implies \left(\beta\mathbf{I} + \left(\mathbf{A\Sigma A} + \mathbf{\Lambda}\right)^{-1}\right)\mathbf{A} = \beta\mathbf{I} \tag{A.89}$$

Now reuse lemma (A.2.6) and to get

$$\mathbf{V}\left(\mathbf{I}+\frac{1}{\beta}\left(\mathbf{D}+\mathbf{D}_{\mathbf{\Lambda}_*}\right)^{-1}\right)\mathbf{V}^T\mathbf{A}=\mathbf{I} \tag{A.90}$$

$$\implies \mathbf{A}=\mathbf{V}\left(\mathbf{I}+\frac{1}{\beta}\left(\mathbf{D}+\mathbf{D}_{\mathbf{\Lambda}_*}\right)^{-1}\right)^{-1}\mathbf{V}^T \tag{A.91}$$

□

Lemma A.2.8. *The covariance* $\mathbf{\Sigma}$ *has the eigenspace decomposition*

$$\mathbf{\Sigma}=\mathbf{V}\mathbf{D}\left(\mathbf{I}+\frac{1}{\beta}\left(\mathbf{D}+\mathbf{D}_{\mathbf{\Lambda}_*}\right)^{-1}\right)^{2}\mathbf{V}^T. \tag{A.92}$$

Proof. This follows directly from lemma (A.2.7)

$$\mathbf{V}\mathbf{D}\mathbf{V}^T=\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T \tag{A.93}$$

$$\implies \mathbf{V}\mathbf{D}\mathbf{V}^T=\mathbf{V}\left(\mathbf{I}+\frac{1}{\beta}\left(\mathbf{D}+\mathbf{D}_{\mathbf{\Lambda}_*}\right)^{-1}\right)\mathbf{V}^T\mathbf{\Sigma}\mathbf{V}\left(\mathbf{I}+\frac{1}{\beta}\left(\mathbf{D}+\mathbf{D}_{\mathbf{\Lambda}_*}\right)^{-1}\right)\mathbf{V}^T \tag{A.94}$$

$$\implies \mathbf{\Sigma}=\mathbf{V}\mathbf{D}\left(\mathbf{I}+\frac{1}{\beta}\left(\mathbf{D}+\mathbf{D}_{\mathbf{\Lambda}_*}\right)^{-1}\right)^{2}\mathbf{V}^T. \tag{A.95}$$

□

Equipped with these lemmas, we can now proof theorem 3.2.1

Proof of theorem 3.2.1. Denote d_i the i-th diagonal element of $\mathbf{D}$ and let a_i, λ_i be the i-th eigenvalue of $\mathbf{A}_*, \mathbf{\Lambda}_*$ respectively. Then we have

$$d_i=a^2\sigma_i. \tag{A.96}$$

From this we get

$$\lambda_i=\frac{a_i^2\sigma_i}{2}\left(\sqrt{1+\frac{4}{a_i^2\beta\sigma_i}}-1\right) \tag{A.97}$$

and

$$a_i=\left(1+\frac{1}{\beta}\left(a_i^2\sigma_i+\lambda_i\right)^{-1}\right)^{-1} \tag{A.98}$$

$$\implies \frac{1}{a_i}=1+\frac{1}{\beta}\left(a_i^2\sigma_i+\lambda_i\right) \tag{A.99}$$

$$\implies a_i\beta-\beta+\frac{a_i}{a_i^2\sigma_i+\lambda_i}=0. \tag{A.100}$$

However, these are exactly the same equations as given in the one-dimensional case. □

Proof of theorem 3.2.2. As in the proof of theorem 3.2.1 we can reduce the problem to showing it for each one-dimensional component independently. However, then we can just reuse theorem 3.1.2. □

A.2.7 Proof of theorem 5.3.1

Fix $\mathbf{t}$. As $d(\cdot,\cdot) > 0$ we have

$$\sup_{\mathbf{x},\mathbf{x}' \in B_R(\mathbf{0})} |d(\mathbf{x},\mathbf{t}) - d(\mathbf{x}',\mathbf{t})| \leq \sup_{\mathbf{x} \in B_R(\mathbf{0})} d(\mathbf{x},\mathbf{t}) \tag{A.101}$$

$$\leq \sup_{\mathbf{x} \in B_R(\mathbf{0})} \frac{1}{2}(\|\mathbf{x}\|_2 + \|\mathbf{t}\|_2)^2. \tag{A.102}$$

Thus, it suffices to bound $\|\mathbf{t}\|_2$. Now $\mathbf{t} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, so we can use the bound given in theorem A.1.4 to obtain

$$P\left[\|\mathbf{t}\|_2^2 > (H(\mathbf{Q},\delta))^2\right] < \delta. \tag{A.103}$$

A.2.8 Proof of theorem 5.3.2

Both $\mathbf{Y}|\mathbf{X}$ and $\mathbf{Y}|\mathbf{T}$ are multivariate Gaussians. Thus, the KL-divergence has the form

$$d(\mathbf{x},\mathbf{t}) = D_{KL}\left[P_{\mathbf{Y}|\mathbf{X}} \| P_{\mathbf{Y}|\mathbf{T}}\right] \tag{A.104}$$

$$= \frac{1}{2}\Big(\mathbf{tr}\left(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-1}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\right) + \left(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\right)^T \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-1}\left(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\right) \tag{A.105}$$

$$- D + \log\left|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\right|\Big).$$

Now fix $\mathbf{t}$. We get

$$2 \sup_{\mathbf{x},\mathbf{x}' \in B_R(\mathbf{0})} |d(\mathbf{x},\mathbf{t}) - d(\mathbf{x}',\mathbf{t})| = \sup_{\mathbf{x},\mathbf{x}' \in B_R(\mathbf{0})} \Big| \left(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\right)^T \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-1}\left(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\right) \tag{A.106}$$

$$- \left(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}'}\right)^T \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-1}\left(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}'}\right) \Big|.$$

As $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}$ is positive definite, so is its inverse. Thus we have

$$2 \sup_{\mathbf{x},\mathbf{x}' \in B_R(\mathbf{0})} |d(\mathbf{x},\mathbf{t}) - d(\mathbf{x}',\mathbf{t})| \leq \sup_{\mathbf{x} \in B_R(\mathbf{0})} \left(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\right)^T \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-1}\left(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{T}=\mathbf{t}} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\right). \tag{A.107}$$

Now the conditional means are given by

$$\mu_{\mathbf{Y}|\mathbf{T}} = \boldsymbol{\Sigma}_{\mathbf{YT}}\boldsymbol{\Sigma}_{\mathbf{T}}^{-1}\mathbf{t} \tag{A.108}$$

$$\mu_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathbf{x}. \tag{A.109}$$

Thus we can rewrite

$$2 \sup_{\mathbf{x},\mathbf{x}' \in B_R(\mathbf{0})} |d(\mathbf{x},\mathbf{t}) - d(\mathbf{x}',\mathbf{t})| \leq \mathbf{z}^T\mathbf{z} \tag{A.110}$$

for

$$\mathbf{z} = \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathbf{x} + \boldsymbol{\Sigma}_{\mathbf{YT}}\boldsymbol{\Sigma}_{\mathbf{T}}^{-1}\mathbf{t}\right). \tag{A.111}$$

So we get

$$\|\mathbf{z}\|_2 \leq \|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\|_2 R + \|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{YT}}\boldsymbol{\Sigma}_{\mathbf{T}}^{-1}\mathbf{t}\|_2. \tag{A.112}$$

The random variable $\mathbf{c} = \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-\frac{1}{2}} \mathbf{\Sigma}_{\mathbf{YT}} \mathbf{\Sigma}_{\mathbf{T}}^{-1} \mathbf{t}$ is distributed as

$$\mathbf{c} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-\frac{1}{2}} \mathbf{\Sigma}_{\mathbf{YT}} \mathbf{\Sigma}_{\mathbf{T}}^{-1} \mathbf{\Sigma}_{\mathbf{TY}} \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-\frac{1}{2}}\right). \tag{A.113}$$

Thus can again use theorem A.1.4 to obtain the bound

$$P\left[\|\mathbf{c}\|_2 > H(\mathbf{Q}, \delta)\right] < \delta \tag{A.114}$$

where

$$\mathbf{Q} = \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{T}}^{-1} \mathbf{\Sigma}_{\mathbf{YT}} \mathbf{\Sigma}_{\mathbf{T}}^{-1} \mathbf{\Sigma}_{\mathbf{TY}}. \tag{A.115}$$

A.2.9 Proof of theorem 5.4.1

Let

$$\epsilon = \Phi^{-1}(1-\delta)\sqrt{2\eta} + \eta. \tag{A.116}$$

From

$$\underbrace{\exp(\epsilon)}_{>0} \cdot \underbrace{\Phi\left(\frac{-\eta-\epsilon}{\sqrt{2\eta}}\right)}_{>0} > 0 \tag{A.117}$$

we get that

$$\begin{aligned}
\Phi\left(\frac{\eta-\epsilon}{\sqrt{2\eta}}\right) - \exp(\epsilon)\Phi\left(\frac{-\eta-\epsilon}{\sqrt{2\eta}}\right) &\leq \Phi\left(\frac{\eta-\epsilon}{\sqrt{2\eta}}\right) && \text{(A.118)}\\
&= \Phi\left(\frac{\eta - \Phi^{-1}(1-\delta)\sqrt{2\eta}\lambda_{\min} - \eta}{\sqrt{2\eta}}\right) && \text{(A.119)}\\
&= \Phi\left(-\Phi^{-1}(1-\delta)\right) && \text{(A.120)}\\
&= \Phi\left(\Phi^{-1}(\delta)\right) && \text{(A.121)}\\
&= \delta. && \text{(A.122)}
\end{aligned}$$

Now we use theorem 1.2.2.