# VARIATIONAL AUTOENCODERS IN BAYESIAN LINEAR INVERSE PROBLEMS

Lappeenranta-Lahti University of Technology LUT

Master's Program in Computational Engineering, Master's Thesis

2022

Anastasiia Kashina

Examiner: Professor Tapio Helin
Professor Lassi Roininen

# ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering

Anastasiia Kashina

**Variational Autoencoders in Bayesian linear inverse problems**

Master's thesis
2022
53 pages, 3 figures
Examiners: Professor Tapio Helin and Professor Lassi Roininen

Keywords: Bayesian inference, variational autoencoder, inverse problem, dimensionality reduction

The Bayesian approach to inverse problems updates the prior distribution of an unknown quantity into a non-trivial posterior distribution conditioned on indirect observations. This thesis studies a methodology for constructing such a prior distribution through variational autoencoder learning. Variational autoencoder parameterizes the prior distribution using neural networks. The advantage of such an approach is that the Bayesian linear inverse problem can be formulated and solved in a latent space with a dimension lower than the initial space dimension using the VAE.

# ACKNOWLEDGEMENTS

I am very grateful to my supervisor, Prof. Tapio Helin, for his mentoring and assistance with my thesis.

Also, I want to thank all my teachers from LUT University for the interesting and advanced courses that helped me expand and deepen my knowledge in mathematics.

In the end, many thanks to my family and friends from all over the world, your love and support are always with me.

Lappeenranta, May 31, 2022

*Anastasiia Kashina*

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ANN | artificial neural network |
| CM | conditional mean |
| DAG | directed acyclic graph |
| DLVM | deep latent variable model |
| ELBO | evidence lower bound |
| i.i.d | independent identically distributed |
| KL | Kullback-Liebler |
| MAP | maximum a posteriori |
| PDF | probability density function |
| SGB | stochastic gradient-based |
| SVD | singular value decomposition |
| VAE | variational autoencoder |
| w.r.t. | with respect to |

# CONTENTS

**1 INTRODUCTION** **7**
1.1 Background . . . 7
1.2 Objectives and delimitations . . . 8
1.3 Structure of the thesis . . . 8

**2 INVERSE PROBLEMS** **10**
2.1 Problem statement . . . 10
2.2 Regularized solution in a vector form . . . 11

**3 BAYESIAN INFERENCE FOR INVERSE PROBLEMS** **14**
3.1 Introduction to probability theory . . . 14
3.2 Bayesian approach . . . 17
3.3 Bayes' formula . . . 19
3.4 Well-posedness of a Bayesian inverse problem . . . 21
3.5 Point estimation of the posterior distribution . . . 26
3.5.1 Maximum a posteriori estimator . . . 26
3.5.2 Conditional mean estimator . . . 26

**4 VARIATIONAL AUTOENCODERS** **28**
4.1 Encoders and autoencoders . . . 28
4.1.1 Encoder-decoder system . . . 28
4.1.2 Autoencoders . . . 30
4.2 Variational autoencoders preliminaries . . . 31
4.2.1 Probabilistic models . . . 31
4.2.2 Bayesian networks . . . 32
4.2.3 Deep latent variable models . . . 34
4.3 Definition of a variational autoencoder . . . 35
4.4 Evidence lower bound . . . 37
4.5 Stochastic gradient-based optimization . . . 40

**5 INVERSE PROBLEM SOLUTION USING VAE** **44**
5.1 Inverse problem statement in a VAE perspective . . . 44
5.2 Construction of a prior density . . . 45
5.3 Construction of a posterior density over a latent variable . . . 46
5.4 Algorithm of solving a linear inverse problem with a VAE . . . 47

**6 DISCUSSION** **49**
6.1 Current study . . . 49

6.2 Future work . . . . . 49

**7 CONCLUSION** **50**

**REFERENCES** **51**

# 1 INTRODUCTION

## 1.1 Background

Inverse problems form a wide class of scientific problems in different fields: from imaging science to physics. There are lots of established examples of these problems, such as deconvolution [1], tomography [2], and inverse Sturm-Liouville problems [3]. The main challenge that arises when solving them is their ill-posedness, which means that the existence, uniqueness, and continuity of the solution depending on the observed data are not guaranteed.

Methods for the elimination of ill-posedness are presented by a family of so-called regularization methods, which are based on the idea of obtaining a solution for a well-posed approximation of an initial deterministic model of an inverse problem, assuming strict norm bounds to the measurement noise. From a computational point of view, that leads to optimization tasks in multi-dimensional spaces.

A more pliant approach for noise modeling is offered by Bayesian inference. From its perspective, the model uncertainty expresses the assumption that all the variables, involved in solving process, are random quantities, sampled according to their (known or unknown) distributions. Based on that, the solution to an inverse problem is sought in a form of a conditional distribution, which characterizes what information is available regarding the unknown based on the given observation. One of the advantages and at the same time the difficulties of this method is the need to describe what statistical information about the unknown is available prior to the observed data. Traditionally, the prior distribution of the unknown quantity has a form of a simple parameterized function, which is set up by an observer. This method, however, can lead to unrealistically simple models and couldn't reflect some properties of unknown quantity. To avoid that, the predefined examples of the unknown variable, collected as a dataset, might be used to construct the prior distribution. In that case, the tool for solving inverse problems is an artificial neural network.

The variational autoencoder (VAE) offers a technically efficient method to define the distribution of the unknown variable through a generative model in a space of a lower dimension. Using the predefined set of training examples with a statistical model of an inverse problem, its learning process allows obtaining a solution as a conditional distribution in a latent space, from where samples can be obtained effectively. The methodology of a VAE is based on a dimensionality reduction principle, which implicitly leads to the opportu-

nity of choosing the distribution for the latent variables to be relatively simple in order to increase the speed of calculations. The ability to obtain a decoding mapping after the learning allows to transfer of all sampling actions to a latent space since all samples could be decoded with relatively high precision.

The subject of this work is to study the process of implementation of a VAE techniques into a solving process of a linear inverse problem. The research is theoretical in nature and consists in studying the main features of the variational autoencoder, such as building an optimization function for learning, the principles of task modeling within its framework, as well as studying the inverse problem and its solution in a Bayesian sense.

## 1.2 Objectives and delimitations

The main objective of the thesis is to provide an incorporation of a VAE framework into a solving process of a Bayesian linear inverse problem. For that purpose the following goals are formulated:

- studying inverse problems and their properties in a classical vector form;
- describing features of Bayesian inverse problems;
- describing the theoretical basis of the VAE;
- combining of these studies into a methodology of solving a linear inverse problem using a VAE.

The main delimitation of the work is it's theoretical devoting, which leaves the numerical experiments out of scope.

## 1.3 Structure of the thesis

The thesis consists 7 of chapters. Chapter 1 presents the background information about the study, its objectives, and delimitations. Chapter 2 introduces a concept of a linear inverse problem, its statement, and a solution in a vector form. Chapter 3 establishes the Bayesian approach to solving linear inverse problems. Chapter 4 is devoted to studying the basic principle, structure, and optimization objective of a variational autoencoder. Chapter 5

presents the algorithm of solving a linear inverse problem using a VAE. Chapters 6 and 7 contains the conclusions made during the study and summarize its results, respectively.

# 2 INVERSE PROBLEMS

This chapter is devoted to introducing inverse problems and their solution in a deterministic setting. The idea of regularization of inverse problems for stabilizing computation aiming to obtain the solution is presented here also. The content of the chapter is based mostly on [4, 5].

## 2.1 Problem statement

In this subsection the introduction of the mathematical notion of an *inverse problem* in a multi-dimensional real-valued vector space is presented and extended with the statement of a *linear inverse problem*.

Let us consider a vector $\mathbf{x} \in \mathbb{R}^m$ as unknown variable and vector $\mathbf{y} \in \mathbb{R}^n$ as its indirect *measurement* (or observation). The relation between these two entities is expressed as a *forward* mapping $F : \mathbb{R}^m \mapsto \mathbb{R}^n$ and defined as a model

$$\mathbf{y} = F(\mathbf{x}) \tag{1}$$

which is a called an *inverse problem* when finding a vector $\mathbf{x} \in \mathbb{R}^m$ for given $\mathbf{y} \in \mathbb{R}^n$.

The problem described by equation (1) represents a *noise-free* inverse problem, which means that there were no errors in the measurement.

A more sensible approach is to assume that the observation $\mathbf{y} \in \mathbb{R}^n$ is contains *noise*. Then, instead of (1) the approximate equation is given by

$$\mathbf{y}^\epsilon \approx F(\mathbf{x}) \tag{2}$$

where $\mathbf{y}^\epsilon$ is a *noisy observation*, such that

$$\|\mathbf{y}^\epsilon - \mathbf{y}\| \leq \epsilon \tag{3}$$

where $\epsilon > 0$ is the *noise level*.

A *linear inverse problem* then can be defined under the following assumptions.

**Assumption 1.** *The forward mapping $F : \mathbb{R}^m \mapsto \mathbb{R}^n$ is linear [6] and has its matrix*

*representation* $A \in \mathbb{R}^{n \times m}$.

**Assumption 2.** *The noise in the observation* $\mathbf{y}^{\epsilon} \in \mathbb{R}^n$ *can be expressed as an additive term given by*

$$\boldsymbol{\eta} = \mathbf{y} - \mathbf{y}^{\epsilon}$$

*where* $\boldsymbol{\eta} \in \mathbb{R}^n$.

**Definition 1.** *The inverse problem* (2) *under the assumptions 1 and 2 is given by*

$$\mathbf{y} = A\mathbf{x} + \boldsymbol{\eta} \tag{4}$$

*and called a* linear inverse problem.

Since the linear inverse problems are the topic of this thesis, the further work is referred mostly to the model (4) when it comes to the notion of an inverse problem.

## 2.2 Regularized solution in a vector form

In the previous subsection, the statement of an inverse problem was formulated for general and linear cases. This subsection is devoted to acquaintance with the idea of *regularization* and its using for obtaining a *deterministic solution* as a vector $\mathbf{x}^* \in \mathbb{R}^m$.

A naive way to obtain a vector solution $\mathbf{x}^* \in \mathbb{R}^m$ for the inverse problem (4) is given by following equality

$$\mathbf{x}^* = A^{-1}(\mathbf{y} - \boldsymbol{\eta}) \tag{5}$$

where $A^{-1}$ is an inverse of a matrix $A$.

However, the solution (5) does not exist if $m \neq n$ or if $n = m$, but matrix $A$ is singular.

To avoid these obstacles, a solution $\mathbf{x}^* \in \mathbb{R}^m$ can be obtained as a *minimum norm solution* [7], which is given by:

$$\mathbf{x}^* = A^{\dagger}(\mathbf{y} - \boldsymbol{\eta}) = VD^{\dagger}U^{\top}(\mathbf{y} - \boldsymbol{\eta}) \tag{6}$$

where exists a singular value decomposition (SVD) [8] of a matrix $A = UDV^{\top}$ and its Moore-Penrose inverse [9] $A^{\dagger} = VD^{\dagger}U^{\top}$.

Depending on relations between spaces dimensions there are three possible instances,

when trying to compute a solution (6) according to [10]:

1. When the problem is *overdetermined*, i.e. $n < m$, then if $A : \mathbb{R}^m \mapsto R(A) \subset \mathbb{R}^n$ (and the $R(A)$ is a range of a matrix $A$), the Moore-Penrose inverse matrix $A^\dagger : R(A) \mapsto \mathbb{R}^n$ exists. But, the instance of a noise vector $\boldsymbol{\eta} \in \mathbb{R}^n$ in unknown and cannot be substracted out, therefore, $\mathbf{y} \notin R(A)$.

2. When the problem is *underdetermined*, i.e. $n > m$ there exists an infinite number of solutions.

3. When $n = m$ there exists a Moore-Penrose inverse matrix $A^\dagger : \mathbb{R}^m \mapsto \mathbb{R}^n$. When the biggest singular value of the matrix $A$ is $\lambda_1$ and the smallest singular value is $\lambda_m$, the condition number $K = \frac{\lambda_1}{\lambda_m}$ is arbitrary large, the minimum norm solution $\mathbf{x}^* = A^\dagger(\mathbf{y}) - A^\dagger(\boldsymbol{\eta})$ is dominated by the second term $A^\dagger(\boldsymbol{\eta})$ and becomes meaningless. The error in a worst case is determined as $\|A^\dagger(\boldsymbol{\eta})\| \approx \frac{\|\boldsymbol{\eta}\|}{\lambda_k}$.

The case 3 is the most difficult to resolve for inverse problems. However, a stabilization of the vector form solution $\mathbf{x}^*$ can be provided using the *regularization*.

Since the regularization and the methods for its performing are not a subject of this work, this subsection gives only a brief explanation of this idea. For the deep study of the regularization methods for inverse problems solving, see [11].

In order to obtain a solution in a vector form as a vector $\mathbf{x}^* \in \mathbb{R}^m$ the model of an initial problem described by equation (4) has to be *well-posed*.

**Definition 2.** *According to Haramard's notion [12] the problem is considered as* well-posed *if the following conditions are satisfied:*

*1) The solution $\mathbf{x}^* \in \mathbb{R}^m$ exists.*

*2) The solution $\mathbf{x}^* \in \mathbb{R}^m$ is unique.*

*3) The solution $\mathbf{x}^* \in \mathbb{R}^m$ depends continuously on data.*

The regularization methods provides the way to obtain the solution $\mathbf{x}^* \in \mathbb{R}^m$ not for the original ill-posed problem (4) but for its best uniquely solvable and continuously depending on data approximation, which is well-posed.

A regularization method approximates the solution $\mathbf{x}^* \in \mathbb{R}^m$ by a family of the regularized solutions

$$\mathbf{x}_\alpha = R_\alpha(\mathbf{y}), \alpha > 0 \tag{7}$$

that are stable with respect to changes in $\mathbf{y} \in \mathbb{R}^n$ and where the solution $\mathbf{x}_\alpha$ converges to the true solution $\mathbf{x}^* \in \mathbb{R}^m$ as $\alpha \to 0$.

There are several regularization methods applied to solve the inverse problem by creating a family of solution in the form (7), such as regularization by singular value truncation [13] and Tikhonov regularization [7]. For the brief study of these methods, also see [4].

The family of the solutions $\mathbf{x}_\alpha$ is dependent on the regularization parameter and based on that, the unique solution $\mathbf{x}^* \in \mathbb{R}^m$ can be chosen. In that case, the *regularized solution in a vector form* of an inverse problem (4) is a solution of its approximation problem.

In this work, a formula for obtaining a solution to the inverse problem using *Tikhonov regularization* is presented as an example.

**Example 1.** *The vector $\mathbf{x}_\alpha \in \mathbb{R}^m$ is called a Tikhonov regularized solution if it gives minimum to the functional*

$$F_\alpha(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|_2^2 + \alpha\|\mathbf{x}\|_2^2 \tag{8}$$

*where the choice of the approximation parameter $\alpha > 0$ is performed based on the noise level of the measurement $\mathbf{y} \in \mathbb{R}^n$. The parameter $\alpha$ can be chosen, for example, using the Morozov discrepancy principle [14]. The regularized solution $\mathbf{x}_\alpha^* \in \mathbb{R}^m$ is minimizing* (8) *is given by*

$$\mathbf{x}_\alpha^* = (A^\top A + \alpha\mathbb{I})^{-1}A^\top\mathbf{y} \tag{9}$$

*where $\mathbb{I} \in \mathbb{R}^{m\times m}$ is the identity matrix and the $R_\alpha = (A^\tau A + \alpha\mathbb{I})^{-1}A^\top$ is a reconstruction matrix.*

According to the example 1 the solution in a vector form $\mathbf{x}^* \in \mathbb{R}^m$ of an inverse problem 1 can be obtained in a form of an approximate regularized solution $\mathbf{x}_\alpha^* \in \mathbb{R}^m$, which is obtaining by solving an optimization problem.

# 3 BAYESIAN INFERENCE FOR INVERSE PROBLEMS

In the previous chapter, the notion of an inverse problem was given with the idea of obtaining a solution in a vector form using the regularization methods. In this chapter, the alternative approach to solving inverse problems, called the Bayesian approach, is introduced and discussed after a brief introduction to probability theory.

## 3.1 Introduction to probability theory

The content of the following subsection is devoted to reminding the basic probability concepts, which are used in the following part of the thesis. The main references for this subsection are [15, 16].

**Definition 3.** *The $\sigma$-algebra $\mathcal{F}$ satisfies the following conditions:*

1. $\emptyset \in \mathcal{F}$.
2. *If the event $A \in \mathcal{F}$, then its complement $A^c \in \mathcal{F}$.*
3. *If $A_1, A_2, ... \in \mathcal{F}$ is a countable sequence of events, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.*

**Definition 4.** *The probability measure $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ satisfies the following conditions:*

1. *For all $A \in \mathcal{F}$ it holds that $\mathbb{P}(A) \geq 0$.*
2. $\mathbb{P}(\Omega) = 0$.
3. *If $A_1, A_2, ... \in \mathcal{F}$ are* pairwise disjoint, *i.e. $A_i \bigcap A_j = 0$ for all $i \neq j$, then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.*

**Definition 5.** *The* probability space *is defined as a triplet $(\Omega, \mathcal{F}, P)$, where $\Omega \neq \emptyset$ is a set of all possible outcomes, $\mathcal{F}$ is a $\sigma$-algebra on $\Omega$ is a set of all possible events and the measure $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ describes the* probability *of events in $\mathcal{F}$ .*

**Remark 1.** *In definition 5 $\sigma$-algebra $\mathcal{F}$ is a collection of sets $\Omega$, i.e. $\mathcal{F} \subseteq 2^{\Omega}$.*

**Definition 6.** *A measure is called $\sigma$-finite, if $\Omega$ is a countable union of measurable sets with finite measure.*

For the further work the probability space is specified with the $\Omega = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^n)$ (Borel $\sigma$-algebra) and with Lebesgue measure $\mathbb{P}$ on $\mathbb{R}^n$.

After determination of a probability space, a notion of a *random variable* must be given.

**Definition 7.** *A measurable mapping*

$$X : (\Omega, \mathcal{F}) \mapsto (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$$

*is called a* random variable. *It means that if* $A \in \mathcal{B}(\mathbb{R}^n)$*, then* $X^{-1} \in \mathcal{F}$.

Definition 7 allows to determine a *probability distribution function*.

**Definition 8.** *The function* $\mu(A) = P(X^{-1}(A))$ *is called a* probability distribution *of that* $X \in A$. *It is defined as a measure on* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

The shorthand for the definition 8 is $X \sim \mu$.

**Definition 9.** *There are two measures* $\mu$ *and* $\nu$ *on the same measure space* $(\Omega, \mathcal{F})$. *The measure* $\mu$ *is* absolutely continuous *with respect to (w.r.t.) to* $\nu$*, if*

$$\nu(A) = 0 \implies \mu(A) = 0$$

*and it is denoted as* $\mu \ll \nu$. *If* $\mu \ll \nu$ *and* $\nu \ll \mu$*, then they are called* equivalent.

For the absolutely continuous measures, the *Radon-Nikodym theorem* can be formulated as follows.

**Theorem 1.** *Let* $\mu$ *and* $\nu$ *be two measures on the same measure space* $(\Omega, \mathcal{F})$. *I the measure* $\nu$ *is* $\sigma$*-finite and* $\mu \ll \nu$*, then there exists* $f \in L^1(\Omega, \mathcal{F}, \nu)$ *such that*

$$\mu(A) = \int_A f(\mathbf{x}) d\nu(\mathbf{x})$$

*for all* $A \in \mathcal{F}$. *The function* $f(\mathbf{x}) = \frac{d\mu}{d\nu}(\mathbf{x})$ *is called the Radon-Nikodym derivative of* $\mu$ *w.r.t.* $\nu$.

*Proof.* For studying of the rigorous proof of that theorem, see [17]. □

The theorem 1 allows us to formulate the notion of the *probability density function (PDF)*.

**Definition 10.** *Let $\mu$ be a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and let $\mathcal{L}$ denote a Lebesgue measure on $\mathbb{R}^n$ and $\mu \ll \mathcal{L}$. Then, by theorem 1 there exists $\pi \in L^1(\mathbb{R}^n)$ such that*

$$\mu(A) = \int_A \pi(\mathbf{x}) d\mathbf{x}$$

*for all $A \in \mathcal{B}(\mathbb{R}^n))$ and the function $\pi(\mathbf{x})$ is called a probability density function (PDF) on $X$.*

**Definition 11.** *Let $X : \Omega \mapsto \mathbb{R}^m$ and $Y : \Omega \mapsto \mathbb{R}^n$ be random variables. The* joint distribution *of $X$ and $Y$ on $: \mathbb{R}^m \times \mathbb{R}^n$ is:*

$$\mu_{X,Y}(A, B) = P(X^{-1}(A) \cap Y^{-1}(B))$$

*for any two measurable sets $A \in \mathbb{R}^m$ and $B \in \mathbb{R}^n$.*

**Definition 12.** *Let $X : \Omega \mapsto \mathbb{R}^m$ and $Y : \Omega \mapsto \mathbb{R}^n$ be two random variables, then the* marginal distribution *of a random variable $X$ defined as follows:*

$$\mu_X(A) = \mu_{X,Y}(A, \mathbb{R}^n)$$

*and for the $Y$ it holds similarly.*

**Definition 13.** *The random variables $X : \Omega \mapsto \mathbb{R}^m$ and $Y : \Omega \mapsto \mathbb{R}^n$ are called* independent *if for all two measurable sets $A \in \mathbb{R}^m$ and $B \in \mathbb{R}^n$ it holds that:*

$$\mu_{X,Y}(A, B) = \mu_X(A)\mu_Y(B)$$

**Assumption 3.** *Let us suppose, that there exists $\mathcal{G} \subset \mathcal{F}$, which is a sub-$\sigma$-algebra. Then, let $X : \Omega \mapsto \mathbb{R}^n$. Now, the $\sigma(X)$ can be denoted as the smallest $\sigma$-algebra, which contains all sets $X^{-1}(A)$, where $A \in \mathcal{B}(\mathbb{R}^n)$. However, it does not mean that all possible outcomes are known, so $\sigma(X) \subset \mathcal{F}$.*

With the assumption 3 the notion of a *conditional expectation* can be given.

**Definition 14.** *A random variable $Y \in L^1(\Omega, \mathcal{G}, \mathbb{P}; \mathbb{R}^n)Y$ is called the* conditional expectation *of a random variable $X \in L^1(\Omega, \mathcal{G}, \mathbb{P}; \mathbb{R}^n)$ with respect to the sub-$\sigma$-algebra $\mathcal{G}$ if we have*

$$\int_G X(\omega) d\mathbb{P}(\omega) = \int_G Y(\omega) d\mathbb{P}(\omega)$$

*for all $G \in \mathcal{G}$. It is written as $\mathbb{E}(X|\mathcal{G}) = Y$.*

Using the definition 14, the *conditional probability* can be defined as follows.

**Definition 15.** *The conditional probability of an event $A \in \mathcal{F}$ given the sub-$\sigma$-algebra $\mathcal{G}$ is a conditional expectation of the form $\mathbb{E}(f(\mathbf{x}), \mathcal{G})$ where*

$$P(A|\mathcal{G}) = \mathbb{E}(\mathbb{1}_A(X)|\mathcal{G})$$

*and*

$$\mathbb{1}_A(x) = \begin{cases} 1 \text{ if } & x \in A \\ 0 \text{ if } & x \notin A \end{cases}$$

*is the indicator function of a subset $A$ of a set $X$.*

**Definition 16.** *A family of probability distributions $\mu(\cdot, \omega)_{\omega \in \Omega}$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is called a* regular conditional distribution *of $\mathbf{x}$ given $\mathcal{G} \subset \mathcal{F}$ if*

$$\mu(A, \mathbf{x}) = \mathbb{E}(\mathbb{1}_A(\mathbf{x})|\mathcal{G})$$

*almost surely for all $A \in \mathcal{B}(\mathbb{R}^n)$.*

**Definition 17.** *The* posterior measure $\mu_{post}(A, \mathbf{y})$ *from the regular conditional probability measure defined as*

$$\mu_{post}(A, Y(\omega)) = \mathbb{E}(\mathbb{1}_A(X)|\sigma(Y))(\omega),$$

*where $\sigma(Y)$ is the sub-$\sigma$-algebra generated by $Y$.*

## 3.2 Bayesian approach

An alternative method for solving inverse problems is called *Bayesian approach*, presented and discussed in this subsection. The content of the following subsection is based mostly on [10, 18].

Let us recall the noisy linear inverse problem model, defined by equation (4)

$$\mathbf{y} = A\mathbf{x} + \boldsymbol{\eta} \tag{10}$$

since the vector $\mathbf{y} \in \mathbb{R}^n$ is assumed to be a measurement or an observation of the unknown variable $\mathbf{x} \in \mathbb{R}^m$, the model's uncertainty is expressed by an additive noise term.

This leads to the idea of using the *Bayesian inference* for solving linear inverse problems. According to [4], it is based on the following principles:

1. All variables included in the model are modelled as random variables.
2. The randomness describes the degree of information concerning their realizations.
3. The degree of information concerning these values is coded in the probability distributions.
4. The solution of the inverse problem is the *posterior probability distribution*.

Let us discuss these principles in linear inverse problems setting.

Since all the variables, known and unknown, are assumed to be samples from corresponding probability distributions, in order to construct a solution of a linear inverse problem (4), there are must be explained the relations between these distributions.

When the linear inverse problem is considered in a real-valued multi-dimensional vector spaces, the probability distribution function can be replaced by its *PDF*, according to the theorem 1 and the definition 10.

The probability distribution of an unknown quantity $\mathbf{x} \in \mathbb{R}^m$, which is called a *prior* distribution and denoted as $p(\mathbf{x})$ contains all available information about the variable $\mathbf{x} \in \mathbb{R}^m$ and generally is unknown. The construction of the prior distribution (and its corresponding prior PDF $\pi(\mathbf{x})$) will be discussed later in the following parts of the thesis.

The distribution of the observed data $\mathbf{y} \in \mathbb{R}^n$, denoted as $p(\mathbf{y})$ with its probability density $\pi(\mathbf{y})$ in a setting of a linear inverse problem (4) also is unknown, but can be computed under certain conditions, which are also will be discussed later.

The relation between the unknown quantity and its measurement is expressed through a *likelihood* function over a variable $\mathbf{y} \in \mathbb{R}^n$, given $\mathbf{x} \in \mathbb{R}^m$. For a linear inverse problem (4) it can be constructed using the additive noise term probability distribution $q(\boldsymbol{\eta})$ or its Lebesgue probability density $\nu(\boldsymbol{\eta})$.

**Assumption 4.** *The additive noise term $\boldsymbol{\eta} \in \mathbb{R}^n$ is an independent (from a random variable $\mathbf{x} \in \mathbb{R}^m$) sample from a multivariate centered Gaussian distribution $q(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\eta}; 0, \Gamma)$ with a PDF $\nu(\boldsymbol{\eta})$, where $\Gamma \in \mathbb{R}^{n \times n}$ is a covariance matrix.*

**Lemma 1.** *Under assumption 4 the likelihood density function of $\mathbf{y} \in \mathbb{R}^n$, given $\mathbf{x} \in \mathbb{R}^m$, is equal to:*

$$\pi(\mathbf{y}|\mathbf{x}) = \nu(\mathbf{y} - A\mathbf{x})$$

*for a linear inverse problem* (4).

*Proof.* The model of a linear inverse problem (4) defines a conditional probability density

$$\pi(\mathbf{y}|\mathbf{x}) = \pi_{\mathbf{y}|\mathbf{x}}(\mathbf{y}) = \pi_{A\mathbf{x}+\boldsymbol{\eta}}(\mathbf{y}) = \pi_{\boldsymbol{\eta}|\mathbf{x}}(\mathbf{y} - A\mathbf{x}) = \pi_{\boldsymbol{\eta}}(\mathbf{y} - A\mathbf{x}) = \nu(\mathbf{y} - A\mathbf{x})$$

due to assumption 4. □

For the further construction of a Bayesian solution the Bayes' formula is presented in the next subsection. It connects the densities of the unknown variable $\mathbf{x} \in \mathbb{R}^m$ with the likelihood function and marginal density of measurement $\mathbf{y} \in \mathbb{R}^m$.

## 3.3 Bayes' formula

Previously, the Bayesian approach was discussed and the components of the Bayesian solution were introduced. This subsection is devoted to present a Bayes' formula for obtaining a solution for a linear inverse problem as a *posterior density*.

**Definition 18.** *A conditional probability distribution $p(\mathbf{x}|\mathbf{y})$ (or its PDF $\pi(\mathbf{x}|\mathbf{y})$) is called a* Bayesian solution *of a linear inverse problem* (4)*, since it is describes the distribution of the unknown $\mathbf{x} \in \mathbb{R}^m$ given the observed data $\mathbf{y} \in \mathbb{R}^n$.*

The solution 18 is given by a Bayes' formula.

Let us recall the Bayes's formula for events $A$ and $B$ and their probability density $\phi$

$$\phi(A|B) = \frac{\phi(B|A)\phi(A)}{\phi(B)} \tag{11}$$

this relation is used to connect conditional, joint and marginal probabilities.

**Theorem 2.** *If assumption 4 holds and*

$$Z = \int_{\mathbb{R}^n} \nu(\mathbf{y} - A\mathbf{x})\pi(\mathbf{x})d\mathbf{x} > 0$$

*then the random variable $\mathbf{x}|\mathbf{y}$ has Lebesgue density which is given by the Bayes' formula*

$$\pi_{post}(\mathbf{x}|\mathbf{y}) = \frac{\nu(\mathbf{y} - A\mathbf{x})\pi(\mathbf{x})}{\pi(\mathbf{y})}$$

*where $\pi(\mathbf{y}) = Z$ is a PDF of the observed data $\mathbf{y} \in \mathbb{R}^n$ and $Z$ plays a role of a* normalization constant.

*Proof.* By lemma 1 the joint density $\pi(\mathbf{x}, \mathbf{y})$ is equal to

$$\pi(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y}|\mathbf{x})\pi(\mathbf{x}) = \nu(\mathbf{y} - A\mathbf{x})\pi(\mathbf{x}) \tag{12}$$

The marginal densities $\pi(\mathbf{y})$ and $\pi(\mathbf{x})$ by definition 12 are given by

$$\pi(\mathbf{y}) = \int\limits_{\mathbb{R}^m} \pi(\mathbf{x}, \mathbf{y}) d\mathbf{x} \tag{13}$$

$$\pi(\mathbf{x}) = \int\limits_{\mathbb{R}^n} \pi(\mathbf{x}, \mathbf{y}) d\mathbf{y} \tag{14}$$

since the marginal $\pi(\mathbf{y})$ is assumed to be positive the Bayes' formula (11) gives the following

$$\pi(\mathbf{x}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})}{\pi(\mathbf{y})} = \frac{\nu(\mathbf{y} - A\mathbf{x})\pi(\mathbf{x})}{Z} > 0 \tag{15}$$

for $\pi(\mathbf{x}) > 0$. Then, since $\pi(\mathbf{y}) = Z$, the following equality can be written as

$$Z = \int\limits_{\mathbb{R}^m} \pi(\mathbf{y}|\mathbf{x})\pi(\mathbf{x}) d\mathbf{x} = \int\limits_{\mathbb{R}^m} \nu(\mathbf{y} - A\mathbf{x})\pi(\mathbf{x}) d\mathbf{x} > 0 \tag{16}$$

giving the marginal density $\pi(\mathbf{y})$ and completing the proof. □

**Remark 2.** *The theorem 2 only holds if $\pi(\mathbf{y}) = Z > 0$, so it will be assumed that it is true for the rest of the work.*

Since the construction of a Bayesian solution (46) based on the unknown prior PDF $\pi(\mathbf{x})$ it cannot be computed without it. The standard approach is to choose some known prior density, such as Gaussian [19], in order to perform further calculations. Alternatively, the chosen prior distribution can provide some desired properties of the solution, for example, its smoothness. The process of solving a Bayesian inverse problem, using a chosen log-concave function as a prior is studied in [20].

In this subsection, the Bayes' formula for the construction of the Bayesian solution of a linear inverse problem (4) was introduced and proved. Since the problem (4) is considered as ill-posed, it begs the question whether the Bayesian approach allows to transform it into a well-posed problem.

## 3.4 Well-posedness of a Bayesian inverse problem

This subsection is devoted to prove that the Bayesian approach allows us to obtain a well-posed inverse problem. The main reference for this subsection's content is [21].

When recalling the definition of a well-posed problem 2 in application to the Baeysian solution $\pi_{post}(\mathbf{x}|\mathbf{y})$ it is understandable, that only the third condition of the well-posedness must be proved.

For the inverse problem (1) (and for the (4) also) the forward mapping $F$ is intractable in general case and for the problem solving it is approximated by some mapping $F_\theta$.

In order to prove that the Bayesian inverse problem (1) is well-posed, it is enough to prove that the small difference between the mappings $F$ and $F_\theta$ (*forward error*), leads to the small difference between the true posterior $\pi_{post}(\mathbf{x}|\mathbf{y})$ and the approximate posterior $\pi_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{y})$ (*modeling error*).

The basic fact has to be proved is given by

$$\|F - F_\theta\| = \mathcal{O}(\theta) \implies d(\pi_{post}, \pi_\theta) = \mathcal{O}(\theta) \tag{17}$$

for some unspecified metric $d(\cdot,\cdot)$ of PDFs and a small $\theta > 0$.

The first step in order to prove (17) is to denote the metric $d(\cdot,\cdot)$. In this subsection, the *Hellinger distance* will be used as that metric.

**Definition 19.** *The* Hellinger distance *between two PDFs $\pi \in L^1(\mathbb{R}^n)$ and $\pi' \in L^1(\mathbb{R}^n)$ is defined by*

$$d_H(\pi,\pi') = \left(\frac{1}{2}\int \left\|\sqrt{\pi(\mathbf{x})} - \sqrt{\pi'(\mathbf{x})}\right\|^2 d\mathbf{x}\right)^{\frac{1}{2}} = \frac{1}{\sqrt{2}}\|\sqrt{\pi} - \sqrt{\pi'}\|_{L^2}.$$

The definition 19 must be supplemented with the following lemma, which is justifying the normalization constant $\frac{1}{\sqrt{2}}$.

**Lemma 2.** *For any two probability densities $\pi \in L^1(\mathbb{R}^n)$ and $\pi' \in L^1(\mathbb{R}^n)$ it holds that*

$$0 \le d_H(\pi,\pi') \le 1$$

*Proof.* The proof for the left part of the inequality follows directly from the definition 19.

In order to prove that $d_H(\pi, \pi') \leq 1$ let us remind that $\int \pi(\mathbf{x})d\mathbf{x} = 1$ for any PDF and we have that

$$\begin{aligned} d_H(\pi, \pi') &= \left(\frac{1}{2}\int \left\|\sqrt{\pi(\mathbf{x})} - \sqrt{\pi'(\mathbf{x})}\right\|^2 d\mathbf{x}\right)^{\frac{1}{2}} \\ &= \left(\frac{1}{2}\int (\pi(\mathbf{x}) + \pi'(\mathbf{x}) - 2\sqrt{\pi(\mathbf{x})\pi'(\mathbf{x})}d\mathbf{x}\right)^{\frac{1}{2}} \\ &\leq \left(\frac{1}{2}\int (\pi(\mathbf{x}) + \pi'(\mathbf{x})d\mathbf{x}\right)^{\frac{1}{2}} \\ &= 1 \end{aligned}$$

which completes the proof. □

The following lemma is auxiliary to the further proof of the fact (17).

**Lemma 3.** *The function $f : \mathbb{R}^m \mapsto \mathbb{R}^n$ is such that*

$$\mathbb{E}^{\pi}\|f\|^2 + \mathbb{E}^{\pi'}\|f\|^2 =: f_2^2 < \infty \tag{18}$$

*then it holds that*

$$\left\|\mathbb{E}^{\pi} f - \mathbb{E}^{\pi'} f\right\| \leq 2 f_2 d_H(\pi, \pi')$$

*Proof.* Since $\mathbb{E}^{\pi} f = \int\limits_{\mathbb{R}^n} f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}$ it holds the following

$$\begin{aligned} \left\|\mathbb{E}^{\pi} f - \mathbb{E}^{\pi'} f\right\| &= \left\|\int\limits_{\mathbb{R}^n} f(\mathbf{x})\pi(\mathbf{x}) - f(\mathbf{x})\pi'(\mathbf{x})d\mathbf{x}\right\| \\ &= \left\|\int\limits_{\mathbb{R}^n} f(\mathbf{x})(\pi(\mathbf{x}) - \pi'(\mathbf{x}))d\mathbf{x}\right\| \\ &= \left\|\int\limits_{\mathbb{R}^n} f(\mathbf{x})(\sqrt{\pi(\mathbf{x})} - \sqrt{\pi'(\mathbf{x})})(\sqrt{\pi(\mathbf{x})} + \sqrt{\pi'(\mathbf{x})})d\mathbf{x}\right\| \end{aligned}$$

and it continues as follows

$$\begin{aligned}
&\leq \left(\frac{1}{2}\int\limits_{\mathbb{R}^n}\left\|\sqrt{\pi(\mathbf{x})}-\sqrt{\pi'(\mathbf{x})}\right\|^2 d\mathbf{x}\right)^{\frac{1}{2}}\\
&\cdot \left(2\int\limits_{\mathbb{R}^n}\|f(\mathbf{x})\|^2\left\|\sqrt{\pi(\mathbf{x})}+\sqrt{\pi'(\mathbf{x})}\right\| d\mathbf{x}\right)^{\frac{1}{2}}\\
&\leq d_H(\pi(\mathbf{x}),\pi'(\mathbf{x}))+\left(4\int\limits_{\mathbb{R}^n}\|f(\mathbf{x})\|^2(\pi(\mathbf{x})+\pi'(\mathbf{x}))d\mathbf{x}\right)^{\frac{1}{2}}\\
&= 2f_2 d_H(\pi,\pi')
\end{aligned}$$

the proof is complete. □

**Remark 3.** *Lemma 3 shows that if two PDFs are close in terms of the Hellinger distance, they are also close in the expectations, calculated w.r.t. to that distance assuming* (18)*.*

**Assumption 5.** *Let us assume that the likelihood functions, associated with $F(\mathbf{x})$ and $F_\theta(\mathbf{x})$, respectively, are given by the following equalities*

$$\begin{aligned}
f(\mathbf{x}) &= \nu(\mathbf{y}-F(\mathbf{x}))\\
f_\theta(\mathbf{x}) &= \nu(\mathbf{y}-F_\theta(\mathbf{x}))
\end{aligned}$$

*and the corresponding posterior PDFs are given by*

$$\begin{aligned}
\pi_{post}(\mathbf{x}) &= \frac{f(\mathbf{x})\pi(\mathbf{x})}{Z}\\
\pi_\theta(\mathbf{x}) &= \frac{f_\theta(\mathbf{x})\pi(\mathbf{x})}{Z_\theta}
\end{aligned}$$

*where $Z>0$ and $Z_\theta>0$ are the corresponding normalize constants.*

**Assumption 6.** *There exists: $\theta^+>0$ and $K_1,K_2>0$ such that for all $\theta\in(0,\theta^+)$ it holds the following*

1. $\left\|\sqrt{f(\mathbf{x})}-\sqrt{f_\theta(\mathbf{x})}\right\|\leq\theta\psi(\mathbf{x})$ *for function $\psi(\mathbf{x})$ such that $\mathbb{E}^\pi[\psi^2(\mathbf{x})]\leq K_1$*
2. $\sup\limits_{\mathbf{x}\in\mathbb{R}^n}\left[\|f(\mathbf{x})\|+\|f_\theta(\mathbf{x})\|\right]\leq K_2$

**Remark 4.** *Note, that assumption 6 only involves conditions on $f$ and its approximation $f_\theta$. While the content of this subsection emphasizes that this approximation may arise*

*due to the need of approximating the forward model $F$, another important scenario that the theory covers is an approximation due to perturbations of the data* $\mathbf{y}$. *Well-posedness results that guarantee the stability under data small perturbations of Bayesian data assimilation are posed in chapter 7 of [21].*

Under assumption 6 the following auxiliary lemma for characterizing the approximate normalization constant is proved.

**Lemma 4.** *Assuming 6 there exists $\theta^* > 0$ and $c_1, c_2 \in \mathbb{R}^+$ such that the following holds*

$$|Z - Z_\theta| \leq c_1\theta$$
$$Z > c_2, Z_\theta > c_2$$

*for $\theta \in (0, \theta^*)$.*

*Proof.* According to the theorem 2 and the assumption 5 the normalization constants $Z, Z_\theta$ can be evaluated as

$$Z = \int_{\mathbb{R}^n} f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}$$
$$Z_\theta = \int_{\mathbb{R}^n} f_\theta(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}$$

then it holds the following

$$\begin{aligned}
|Z - Z_\theta| &= \int_{\mathbb{R}^n} (f(\mathbf{x}) - f_\theta(\mathbf{x}))\pi(\mathbf{x})d\mathbf{x} \\
&\leq \left(\int_{\mathbb{R}^n} \left\|\sqrt{f(\mathbf{x})} - \sqrt{f_\theta(\mathbf{x})}\right\| \pi(\mathbf{x})d\mathbf{x}\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^n} \left\|\sqrt{f(\mathbf{x})} + \sqrt{f_\theta(\mathbf{x})}\right\| \pi(\mathbf{x})d\mathbf{x}\right)^{\frac{1}{2}} \\
&\leq \left(\int_{\mathbb{R}^n} \theta^2\psi^2(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^n} K_2\pi(\mathbf{x})d\mathbf{x}\right)^{\frac{1}{2}} \\
&\leq \sqrt{K_1K_2}\theta
\end{aligned}$$

where $\theta \in (0, \theta^+)$. And if there is $\theta \leq \theta^* = \min\left[\frac{Z}{2\sqrt{K_1K_2}}, \theta^+\right]$ then takes place the following inequality

$$Z_\theta \geq Z - |Z - Z_\theta| \geq \frac{1}{2}Z$$

and $c_1, c_2$ are taken as $c_1 = \sqrt{K_1K_2}$, $c_2 = \frac{1}{2}Z$. □

After preliminary studies, the theorem of a well-posedness of a posterior density is presented.

**Theorem 3.** *Under the assumption 6 the following holds*

$$d_H(\pi_{post}, \pi_\theta) \leq c\theta$$

*where* $\theta \in (0, \theta^*)$ *for some* $\theta^* > 0$ *and* $c \in \mathbb{R}^+$*, which is independent of* $\theta$*.*

*Proof.* Let us separate the Hellinger distance for two terms: the first will be the distance between the true and approximate normalization constants, and the second will be the distance between the true and approximate likelihoods, given by

$$\begin{aligned} d_H(\pi_{post}, \pi_\theta) &= \frac{1}{\sqrt{2}} \left\| \sqrt{\pi_{post}} - \sqrt{\pi_\theta} \right\|_{L^2} \\ &= \frac{1}{\sqrt{2}} \left\| \sqrt{\frac{f\nu}{Z}} - \sqrt{\frac{f\nu}{Z_\theta}} + \sqrt{\frac{f\nu}{Z_\theta}} - \sqrt{\frac{f_\theta\nu}{Z_\theta}} \right\|_{L^2} \\ &\leq \frac{1}{\sqrt{2}} \left\| \sqrt{\frac{f\nu}{Z}} - \sqrt{\frac{f\nu}{Z_\theta}} \right\|_{L^2} + \frac{1}{\sqrt{2}} \left\| \sqrt{\frac{f\nu}{Z_\theta}} - \sqrt{\frac{f_\theta\nu}{Z_\theta}} \right\|_{L^2} \end{aligned}$$

then, under the auxiliary lemma 4 for some $\theta \in (0, \theta^*)$ the following inequalities holds

$$\begin{aligned} \left\| \sqrt{\frac{f\nu}{Z}} - \sqrt{\frac{f\nu}{Z_\theta}} \right\|_{L_2} &= \left\| \frac{1}{\sqrt{Z}} - \frac{1}{\sqrt{Z_\theta}} \right\| \left( \int_{\mathbb{R}^n} f(\mathbf{x})\nu(\mathbf{x}) d\mathbf{x} \right)^{\frac{1}{2}} \\ &= \frac{\|Z - Z_\theta\|}{\left(\sqrt{Z} + \sqrt{Z_\theta}\right)\sqrt{Z_\theta}} \\ &\leq \frac{c_1}{2c_2}\theta \end{aligned} \tag{19}$$

and

$$\frac{1}{\sqrt{2}} \left\| \sqrt{\frac{f(\mathbf{x})\nu}{Z_\theta}} - \sqrt{\frac{f_\theta(\mathbf{x})\nu}{Z_\theta}} \right\|_{L_2} = \frac{1}{Z_\theta} \left( \int_{\mathbb{R}^n} \left\| \sqrt{f(\mathbf{x})} - \sqrt{f_\theta(\mathbf{x})} \right\|^2 \nu(\mathbf{x}) d\mathbf{x} \right)^{\frac{1}{2}} \leq \sqrt{\frac{K_1}{c_2}}\theta \tag{20}$$

combining (19) and (20) the estimation for Hellinger distance is evaluated as

$$d_H(\pi_{post}, \pi_\theta) \leq \frac{1}{\sqrt{2}} \frac{c_1}{2c_2}\theta + \frac{1}{\sqrt{2}} \sqrt{\frac{K_1}{c_2}}\theta = c\theta$$

since $c = \frac{1}{\sqrt{2}} \frac{c_1}{2c_2}\theta + \frac{1}{\sqrt{2}} \sqrt{\frac{K_1}{c_2}}$, is is independent of $\theta$. □

The Bayesian inverse problem is proved to be well-posed, however, the form of its solution as a probability density cannot be used without its point estimation, which allows obtaining the solution in a classical vector form from a Bayesian one. In the next subsection, the two most useful methods for single-point estimation are described.

## 3.5 Point estimation of the posterior distribution

The Bayesian solution for inverse problems presented in the previous chapter in the form of the posterior density $\pi_{post}(\mathbf{x}|\mathbf{y})$ also could be used for producing a single estimate for obtaining a solution $\mathbf{x}^* \in \mathbb{R}^m$ as a vector, in a deterministic setting. For that aim, this chapter will be presented two estimators: a *maximum a posteriori (MAP)* estimator and a *conditional mean (CM)* estimator.

### 3.5.1 Maximum a posteriori estimator

The following part of the subsection 3.5 introduces a single-point MAP estimator for obtaining a vector form solution $\mathbf{x}^* \in \mathbb{R}^m$.

**Definition 20.** *The vector*

$$\mathbf{x}_{MAP} = \arg\max_{\mathbf{x}\in\mathbb{R}^m} \pi_{post}(\mathbf{x}|\mathbf{y})$$

*is called a* maximum a posteriori estimator *of a posterior PDF* $\pi_{post}(\mathbf{x}|\mathbf{y})$.

The definition 20 of a MAP estimator shows that its computation is an *optimization problem*. In order to solve it more efficiently, instead of a maximization of a posterior density, the minimization of a negative *log-posterior density* $L_p(\mathbf{x}) = -\ln \pi_{post}(\mathbf{x}|\mathbf{y})$ usually takes place and is given by

$$\mathbf{x}_{MAP} = \arg\min_{\mathbf{x}\in\mathbb{R}^m} L_p(\mathbf{x}) \tag{21}$$

and the MAP estimator $\mathbf{x}_{MAP}$ of $\mathbf{x}^* \in \mathbb{R}^n$ given $\mathbf{y} \in \mathbb{R}^m$ is analogous to the mode.

### 3.5.2 Conditional mean estimator

The following part of the subsection 3.5 introduces a single-point CM estimator for obtaining a vector form solution $\mathbf{x}^* \in \mathbb{R}^m$.

**Definition 21.** *The vector*

$$\mathbf{x}_{CM} = \mathbb{E}\left[\pi_{post}(\mathbf{x}|\mathbf{y})\right] = \int\limits_{\mathbb{R}^m} \mathbf{x}\pi_{post}(\mathbf{x}|\mathbf{y})d\mathbf{x}$$

*is called a* conditional mean *of a posterior PDF* $\pi_{post}(\mathbf{x}|\mathbf{y})$.

The definition 21 describes an *integration problem*, which is most likely high-dimensional. In order to solve it the *Monte-Carlo* method of integration [22] is performed.

# 4 VARIATIONAL AUTOENCODERS

In this chapter the notion of the *variational autoencoder (VAE)* is considered as a tool for solving the data compression problem. The structure of a VAE is explained based on the structure of the *autoencoder*, which represents the working process of an *encoder-decoder system*.

In the following sections, the concepts, mentioned above, is considered separately and in the complex. After the theoretical discussion, the VAE optimization objective, *evidence lower bound (ELBO)*, is discussed and an algorithm of its estimation and maximization is presented.

## 4.1 Encoders and autoencoders

This subsection contains the discussion about the compression of data by the encoder-decoder system, based on the *dimensionality reduction* process.

### 4.1.1 Encoder-decoder system

The following part of the subsection 4.1 is devoted to the discussion about the *dimensionality reduction* process, which lies under the *encoder-decoder system* working.

**Definition 22.** *The* dimensionality reduction process *of vectors in $\mathbb{R}^m$ into a space $\mathbb{R}^k$ of a smaller dimension is defined by the following two steps:*

1) *The* encoding mapping $\mu_E : \mathbb{R}^m \mapsto \mathbb{R}^k$ *transforms the vector $\mathbf{x} \in \mathbb{R}^m$ into a vector $\mathbf{z} \in \mathbb{R}^k$, $k < m$.*

2) *The* decoding mapping $\mu_D : \mathbb{R}^k \mapsto \mathbb{R}^m$ *transforms the vector $\mathbf{z} = \mu_E(\mathbf{x})$ into a vector*
$$\mathbf{x}' = \mu_D(\mathbf{z}) = \mu_D(\mu_E(\mathbf{x})) \in \mathbb{R}^m.$$

*The* data loss *(or reconstruction error) is determined as a function:*

$$f_{loss}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^2_{\mathbb{R}^m}.$$

**Remark 5.** *In what follows,* $\boldsymbol{Z}$ *stands for a latent space. previously, the space* $\mathbb{R}^k = \boldsymbol{Z}$ *was given as an example of a latent space with* $\dim \boldsymbol{Z} = k$*. The variables* $\mathbf{z} \in \boldsymbol{Z}$ *are called* latent *variables, respectively.*

The definition 22 is illustrated by figure 1.

The encoding process produces a data loss, when the data from the initial space $\mathbb{R}^m$ loses some of the features and they cannot be recovered later. Due to the data loss, $\mathbf{x}' \neq \mathbf{x}$.

The definition 22 describes an idea of the *encoder-decoder system* working process. This system's main aim is to find the pair of mappings $\mu_E$ and $\mu_D$, which gives a minimum to a data loss function

$$(\mu_E^*, \mu_D^*) = \underset{(\mu_E, \mu_D)}{\arg\min} \sum_{i=1}^{N} f_{loss}(\mathbf{x}_n, \mu_D(\mu_E(\mathbf{x}_i))) \tag{22}$$

where variables $\mathbf{x}_i$ are united in a *dataset* $\boldsymbol{D} = \{\mathbf{x}_i\}_{i=1}^N$.

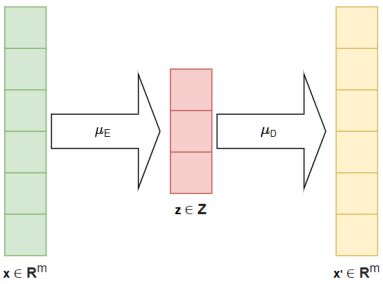


**Figure 1.** The encoder-decoder system's scheme of work. The vector of the initial data $\mathbf{x} \in \mathbb{R}^m$ is coded into a vector $\mathbf{z}$ in a latent space $\boldsymbol{Z}$ and then decoded back to the initial space. Due to the loss of some features, the output vector $\mathbf{x}' \in \mathbb{R}^m$ is different from the input.

The optimization problem, formulated by equation (22) can be solved, using the artificial neural network (ANN). For a deeper understanding of the structure of an ANN, as well as for an introduction to the basic concepts and terms associated with it, see [23, 24]. Next part of the subsection 4.1 contains information about a particular type of an ANN, called an *autoencoder*.

### 4.1.2 Autoencoders

The optimization task for the dimensionality reduction process, which arose as the equation (22) can be solved using ANNs for *learning* the best pair of the mappings $\mu_E$, $\mu_D$.

For that purpose, the special type of a neural network is defined below, under a following assumption.

Below, The encoder and decoder mappings are parameterized in such a way

$$\mu^E = \mu^E_{\boldsymbol{\theta}} \quad \text{and} \quad \mu^D = \mu^D_{\boldsymbol{\phi}}$$

and parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ are vectors in some multi-dimensional real-valued vector space.

**Definition 23.** *The* autoencoder *is a feed-forward ANN, which performed the unsupervised learning of the mappings $\mu_E$ and $\mu_D$ in order to perform the dimensionality reduction process and give a minimum to a data loss function, formulated by equation* (22). *The autoencoder is defined by its working process scheme:*

1. *The input vector $\mathbf{x} \in \boldsymbol{D}$ is compressed into a vector $\mathbf{z} \in \boldsymbol{Z}$ by the encoder mapping $\mu_E$.*
2. *The latent variable $\mathbf{z} \in \boldsymbol{Z}$ is transformed into the output vector $\mathbf{x}' \in \mathbb{R}^m$ by the decoder mapping $\mu_D$.*
3. *The data loss function value is optimized over the average vector $\mathbf{x} \in \boldsymbol{D}$ and the output vector $\mathbf{x}' \in \mathbb{R}^m$.*
4. *The result of the optimization is backpropagated [25] through an ANN.*
5. *The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are updated after the backpropagation.*

The assumption 4.1.2 allows us to look at the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ as the weights and biases of an ANN.

The process, defined by 23 is performed for all vectors from a *training dataset* $\boldsymbol{D}$ and after that the optimized parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ of the encoder and decoder mappings are obtained.

The illustration of the autoencoder architecture is presented on figure 2.

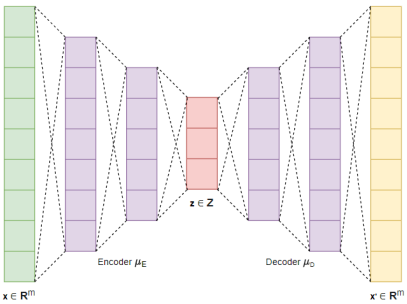


**Figure 2.** The autoencoder architecture. The initial vector's $\mathbf{x} \in \mathbb{R}^m$ dimensionality reduces consequently by the encoding layers with parameter vector $\boldsymbol{\phi}$. Then, the coded representation of an initial vector, $\mathbf{z} \in \boldsymbol{Z}$ is consequently decoded by the decoding layers with parameter vector $\boldsymbol{\theta}$ and transforms into a vector $\mathbf{x}' \in \mathbb{R}^n$.

For the studying of construction and training of a *linear autoencoder*, see [26]. The next part of the thesis uses mostly the architecture of the autoencoder and its scheme of work, described by definition 23.

## 4.2 Variational autoencoders preliminaries

This subsection contains a discussion about the modeling approach and structure of a network, which are used in order to construct and determine a *variational autoencoder*.

### 4.2.1 Probabilistic models

The previous subsection introduced the concept of an autoencoder, as an ANN and had a description of its scheme of work as a definition 23. In subsection 4.1 the data $\mathbf{x} \in \boldsymbol{D}$ was presented as a vector in multi-dimensional real-valued space. That type of representation is usually called a *deterministic model* since it excludes any randomness in the data and guarantees the same output vector for a given input. This subsection introduces the concept of *probabilistic model* and its application to autoencoders.

**Assumption 7.** *The input data in form of a vector* $\mathbf{x} \in \boldsymbol{D}$ *is sampled from the unknown underlying probability distribution* $p_{true}(\mathbf{x})$ *and the dataset* $\boldsymbol{D}$ *is consisted of* independent identically distributed (i.i.d) *samples* $\mathbf{x}_i$ *from the probability distribution* $p_{true}(\mathbf{x})$.

The assumption 7 allows us to build more flexible and realistic model for the training process.

Since the true probability distribution is unknown, the parameterized model $p_{\boldsymbol{\theta}}(\mathbf{x})$ is chosen to approximate the true distribution $p_{true}(\mathbf{x})$ unknown parameters

$$p_{true}(\mathbf{x}) \approx p_{\boldsymbol{\theta}}(\mathbf{x})$$

In this case, the objective of the autoencoder working process is to find the best parameter vector $\boldsymbol{\theta}$ in order to approximate the unknown distribution $p_{true}(\mathbf{x})$ with the parameterized distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$.

*Conditional probabilistic models* where the unknown probability distribution $p_{true}(\mathbf{y}|\mathbf{x})$ over unknown $\mathbf{y}$, given the measured variable $\mathbf{x}$ can be treated in the same manner with the difference that data, on what model is conditioned, does not optimize over their value and is considered as input for the model.

The VAE is based on the probabilistic model of data, which allows to add more flexibility to the output result. The prior probability distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ in this framework is available through the dataset of the observed samples $\boldsymbol{D}$ which are realizations of this distribution, so it can be expressed as

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{D}) \tag{23}$$

where $\boldsymbol{D} = \{\mathbf{x}_i\}_{i=1}^{N}$ are the observed training samples.

### 4.2.2 Bayesian networks

As it was determined in the previous part of this subsection, the VAE basis is a *probabilistic model* of the data. This model needs to be organized in a network in order to perform the process of learning. In this part of the thesis the structure of this network is presented.

**Assumption 8.** *Let us suppose that all initial data, which is some number $N$ of variables from the initial space $\mathbb{R}^m$, sampled by $\mathbf{x} \sim p_{true}(\mathbf{x})$ (or $\mathbf{x} \sim p_{true}(\mathbf{y}|\mathbf{x})$ for conditional models) independently, is united is the following dataset: $\boldsymbol{D} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$.*

Since all samples are independent identically distributed (i.i.d) and united in one dataset $\boldsymbol{D}$, according to the assumption 8, they can be organized into a network $\boldsymbol{N}$, which is called a *Bayesian network*.

For giving a definition of the Bayesian network, let us recall the definition of a directed acyclic graph (DAG), based on the following auxiliary definitions from [27].

**Definition 24.** *A directed graph is a pair of sets $G = (V, A)$, where set $V = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ is a set of vertices and set $A$ is a set of ordered pairs (called arcs) of vertices of $V$.*

**Definition 25.** *The* directed acyclic graph (DAG) *is a directed graph with no directed cycles.*

An example of a DAG based on the definition 25 is presented on figure 3.

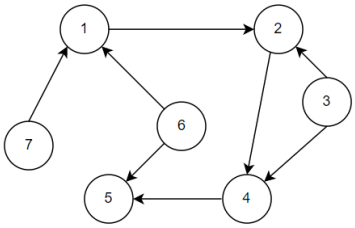


**Figure 3.** The example of a DAG.

**Definition 26.** Bayesian network *is a directed probabilistic model, which means that all the variables $\mathbf{x}_i$ from the dataset $\boldsymbol{D}$ are organized into a* DAG.

In the Bayesian network, each vertice of a DAG represents one variable $\mathbf{x}_i \in \boldsymbol{D}$ and edges represent conditional dependencies: if two nodes do not have a path from one to another they are *conditionally independent*.

Under the assumption 8, for the Bayesian network, the log-probability assigned to the data by the unconditional model is given by

$$\ln p_{\boldsymbol{\theta}}(\boldsymbol{D}) = \sum_{\mathbf{x} \in \boldsymbol{D}} \ln p_{\boldsymbol{\theta}}(\mathbf{x}) \tag{24}$$

this equation can be reformulated for the conditional model as follows

$$\ln p_{\boldsymbol{\theta}}(\boldsymbol{D}) = \sum_{\mathbf{x} \in \boldsymbol{D}} \ln p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) \tag{25}$$

For the further work the unconditional case will be taken into account, but the same techniques are applicable for conditional models too.

**Definition 27.** *The vertice $\mathbf{x}_i$ of the DAG is a* parent variable *for the vertice $\mathbf{x}_j$ if there is an arrow from $\mathbf{x}_i$ to $\mathbf{x}_j$. In that case, $\mathbf{x}_j$ is called a* child variable.

Let us consider a set of all parent variables in sense of definition 27 in the DAG for the vertice $\mathbf{x}_i$ as a mapping $P : \boldsymbol{D} \mapsto 2^{\boldsymbol{D}}$ with the property $\mathbf{x}_i \notin P(\mathbf{x}_i)$ for all $i$.

Based on previous assumptions and definitions we can consider the joint distribution of the dataset $\boldsymbol{D}$ as a following product of the prior and conditional distributions

$$p_\theta(\mathbf{x}_1, ..., \mathbf{x}_n) = \prod_{i=1}^{n} p_\theta(\mathbf{x}_i | P(\mathbf{x}_i)) \tag{26}$$

with the convention that $p_{\boldsymbol{\theta}}(\mathbf{x}_i|P(\mathbf{x}_i)) = p_{\boldsymbol{\theta}}(\mathbf{x}_i)$ if $P(\mathbf{x}_i) = \emptyset$, i.e., $\mathbf{x}_i$ has no parent nodes.

When the initial data from the dataset $\boldsymbol{D}$ is organized in a Bayesian network, the goal is to parametrize the conditional distribution $p_\theta(\mathbf{x}_i|P(\mathbf{x}_i))$ using the VAE. In that case the parents variables of a variable $\mathbf{x} \in \boldsymbol{D}$ are taken as input for the neural network and the parameters $\boldsymbol{\nu}$ over that variable are produced by

$$\boldsymbol{\nu} = \boldsymbol{N}(P(\mathbf{x})) \tag{27}$$

$$p_\theta(\mathbf{x}|P(\mathbf{x})) = p_\theta(\mathbf{x}|\boldsymbol{\nu}) \tag{28}$$

Here the function $\boldsymbol{N}$ is representing the neural network.

### 4.2.3 Deep latent variable models

In the previous part of this subsection, the initial data was organized into a dataset $\boldsymbol{D}$ in order to build the Bayesian network. The next step is to extend the set of variables used for the learning process by transforming the probabilistic model into a *deep latent variable model (DLVM)*.

Previously, only the observed data sampled from the unknown distribution $p_{true}(\mathbf{x})$, organized into a set $\boldsymbol{D} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, was taken into account. This approach for modeling is so-called *fully-observed modeling*.

In order to extend the model, the latent variables $\mathbf{z} \in \boldsymbol{Z}$ can be also considered as a part of the data model. The new *model with latent variables* allows us to express the approximate distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ as a marginalization over a latent variable with the following equality

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} \tag{29}$$

where the latent variables are added to the model and forms a joint distribution $p_\theta(\mathbf{x}, \mathbf{z})$ over the both observed and latent variables.

Using the structure of the model with latent variables, the DLVM can be denoted.

**Definition 28.** *The model with latent variables, described by the distribution $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$ is called a* deep latent variable model (DLVM)*, if the vector $\boldsymbol{\theta}$ of the approximate distribution parameters is represented by weights of an ANN.*

**Example 2.** *The simplest case of the DLVM can be described as a factorization*

$$p_\theta(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$$

*where the distribution $p(\mathbf{z})$ is called a* prior distribution*, since it does not depend on any observations, the joint distribution $p_\theta(\mathbf{x}, \mathbf{z})$ is considered as a* generative model*, and the conditional distribution $p_\theta(\mathbf{x}|\mathbf{z})$ is a* posterior distribution*.*

The equation (29) defines the marginal likelihood over a variable $\mathbf{x} \in \mathbb{R}^m$ as an integral that cannot be evaluated numerically or sufficiently estimated, so that function is, in general, intractable. This leads to the intractability of the posterior distribution because these entities are connected by a basic identity

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})} \tag{30}$$

by equation (30) the most difficult part of the DLVM learning is shown, since in order to obtain a tractable posterior, we have to evaluate a tractable likelihood, and vice versa.

The DLVM, as a directed probabilistic model with latent variables (and a representation as in the example 2), which observed data is organized as a Bayesian network is a basis for defining a variational autoencoder.

## 4.3 Definition of a variational autoencoder

In subsection 5.2 the concepts of a DLVM and a Bayesian network were introduces as important parts for the further work with a VAE. The following subsection relies mostly on the article [28].

Let us to sum up contents of the subsection 4.2 by several assumptions.

**Assumption 9.** *The set of observed data $\boldsymbol{D}$ consists of i.i.d samples $\mathbf{x}_i$ from the unknown distribution $p_{true}(\mathbf{x})$. This distribution has an approximation as a parameterized distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$*

$$p_{true}(\mathbf{x}) \approx p_{\boldsymbol{\theta}}(\mathbf{x})$$

*then, the approximate distribution can be parameterized by an ANN training.*

**Assumption 10.** *The (prior) distribution of latent variables $\mathbf{z} \in \boldsymbol{Z}$ is chosen to be a tractable Gaussian distribution with zero mean and a covariance matrix $\Sigma$*

$$\mathbf{z} \sim p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, \Sigma)$$

*where $\Sigma = \sigma^2 \mathbb{I}^{k \times k}$.*

Recalling the example 2 we again can see the intractability of the DLVM, described by a joint distribution $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$ and the identity (30)

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})} \tag{31}$$

the conditional distribution $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ is a *posterior distribution* and the goal of the parameter estimation of a DLVM.

However, under the assumptions 9 and 10 we can formulate the way of the approximation of the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ by the other model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, which is interpreted as a modeling objective of a VAE.

**Definition 29.** *The DLVM $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ is called a* parametric inference model *if it approximates the intractable posterior distribution.*

**Remark 6.** *The parameter vector $\boldsymbol{\phi}$ is obtained through the learning process of an ANN.*

In case, when $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \approx p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$, these two conditional distributions can be considered as targets for the learning process of two neural networks. The same approach arose in the definition 23 of an autoencoder, but instead of searching for the best parameters for mappings, now the parameters of the posterior and inference model are taken as weights for ANNs.

**Definition 30.** *The* variational autoencoder *is a pair of two ANNs, such as encoder and decoder, which are parametrizing the following DLVMs and optimize the distance between them:*

1. *The parametric inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, or a* stochastic encoder*, which parameters $\boldsymbol{\phi}$ are learned by an ANN $\mu_E(\boldsymbol{\phi}) : \mathbb{R}^m \mapsto \mathbb{R}^k$.*

2. *The posterior model $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ or a* stochastic decoder *which parameters $\boldsymbol{\theta}$ are learned by an ANN $\mu_D(\boldsymbol{\theta}) : \mathbb{R}^k \mapsto \mathbb{R}^m$.*

*After the forward propagation of the sampled data $\mathbf{x} \in \boldsymbol{D}$ through the networks $\mu_E(\boldsymbol{\phi})$ and $\mu_D(\boldsymbol{\theta})$, the distance between models is calculated and backpropagated in order to perform a learning process.*

The definition 30 is concentrated on the probabilistic models, which are the objectives of study for this thesis. The basic structure of the VAE in terms of ANNs remains the same, as for the simple autoencoder and the dimensionality reduction process is also a basic process, performed by the variational autoencoder.

According to the content of this subsection, the next task which arises for study of a VAE is a minimization of the distance between the true posterior and the inference model.

## 4.4 Evidence lower bound

As it was declared in the previous subsection, the aim of the VAE's work is to optimize the posterior distribution $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ by its parametric inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, which leads to the task of the minimization of the distance between these distributions. For this purpose the concept of the ELBO is introduced.

The distance between two probability distributions can be calculated as Kullback-Liebler (KL) divergence [29].

**Definition 31.** *The* KL-divergence *between distributions $p(x)$ and $q(x)$ over a continuous random variable $x \in \mathbb{R}$ is written as a following function*

$$D_{KL}(p(x)\|q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx = \mathbb{E}_p \left[ \ln \frac{p(x)}{q(x)} \right]$$

**Lemma 5.** *By definition 31 $D_{KL}(p(x)\|q(x)) = 0$ if and only if $p(x) = q(x)$.*

*Proof.* The proof is presented in [29]. □

**Definition 32.** *The function* $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}} : \mathbb{R}^n \to \mathbb{R}$ *defined by*

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(x) = \mathbb{E}_{q_{\boldsymbol{\phi}}(z|x)} \left[ \ln p_{\boldsymbol{\theta}}(z, x) - \ln q_{\boldsymbol{\phi}}(z|x) \right]$$

*is called the* evidence lower bound (ELBO)*, where* $p(x)$ *and* $q(x)$ *are distributions over a continuous random variable* $x \in \mathbb{R}$*,* $z \in \boldsymbol{Z}$ *is a latent variable, and* $\boldsymbol{\phi}$ *and* $\boldsymbol{\theta}$ *are the parameters of these distributions* $q_{\boldsymbol{\phi}}(x)$ *and* $p_{\boldsymbol{\theta}}(x)$*, respectively.*

The definitions 31 and 32 allows us to formulate the further optimization task in VAEs framework as a theorem.

In order to minimize the KL-divergence between the parametric inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and the posterior distribution $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$, the ELBO of these two entities has to be maximized, since the following theorem for the log-likelihood $\ln p_{\boldsymbol{\theta}}(\mathbf{x})$ holds.

**Theorem 4** (ELBO as an optimization objective)**.** *It holds that*

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \ln p_{\boldsymbol{\theta}}(\mathbf{x}) - D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) \leq \ln p_{\boldsymbol{\theta}}(\mathbf{x})$$

*for each* $\mathbf{x} \in \mathbb{R}^n$*.*

*Proof.* The KL-divergence between $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ by definition 31 is evaluated as follows

$$D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[ \ln \frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})} \right]. \tag{32}$$

In the equation 32 the posterior distribution $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ is unknown, but it can be replaced with the fraction of the likelihood and the marginal distribution according to the Bayes' formula

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})}. \tag{33}$$

By substituting equality (33) into (32) we obtain

$$\begin{aligned} D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[ \ln \frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})} \right] \\ &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[ \ln \frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{x})} \right] \\ &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[ \ln \frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{x})} \right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[ \ln p_{\boldsymbol{\theta}}(\mathbf{x}) \right] \end{aligned} \tag{34}$$

where $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[ \ln p_{\boldsymbol{\theta}}(\mathbf{x}) \right] = \ln p_{\boldsymbol{\theta}}(\mathbf{x})$.

Then equality (34) can be rearranged as follows

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\ln p_{\boldsymbol{\theta}}(\mathbf{x})\right] = D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\right]. \tag{35}$$

Now three entities are presented in the equation 35. The first contains the marginal log-likelihood $\ln p_{\boldsymbol{\theta}}(\mathbf{x})$. The second is a KL-divergence between the inference model and the posterior distribution, which is non-negative, according to the lemma 5.

The third entity is by definition 32 the ELBO, and it is given by

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\ln p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right]$$

The ELBO is a lower bound on the log-likelihood marginal distribution and an optimization objective of VAEs

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \ln p_{\boldsymbol{\theta}}(\mathbf{x}) - D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) \leq \ln p_{\boldsymbol{\theta}}(\mathbf{x}) \tag{36}$$

the inequality 36 holds and, since the KL-divergence is non-negative, the ELBO represents the objective of the maximization for VAE. □

In order to approximate the intractable posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ by its parametric inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, the minimization of the ELBO $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})$ is performed w.r.t the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ according to the theorem 4. In that case the maximization of the ELBO will solve two problems in one:

1. The maximization of the marginal likelihood $p_{\boldsymbol{\theta}}(\mathbf{x})$, which means, that the generative model $p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})$ becomes better.

2. The minimization of the KL-divergence between the parametric inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and the true posterior distribution $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$, which makes the inference model more accurate.

By definition 32, ELBO is a complicated structured function with two parameters that need to be optimized. It leads to the necessity of a so-called *joint optimization* w.r.t. both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. So the next goal is to implement an optimization method to perform that process.

## 4.5 Stochastic gradient-based optimization

In the previous subsection, the ELBO function was taken into account as an optimization objective for the VAE. Since its construction it requires a joint optimization, the stochastic gradient-based (SGB) descent is used for solving the optimization problem.

For the VAE the objective of the optimization is a sum (or average) value of the ELBO function per all single datapoints $\mathbf{x} \in \boldsymbol{D}$

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{D}) = \sum_{\mathbf{x}\in\boldsymbol{D}} \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \sum_{\mathbf{x}\in\boldsymbol{D}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[\ln p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right] \tag{37}$$

however, when the size of a dataset $\boldsymbol{D}$ is large, the equality (37) becomes intractable. Instead of calculating ELBOs per all datapoints, the *minibatch* optimization can be implemented: on each step of an optimization process values of the ELBO function will be evaluated for a randomly chosen subset $\boldsymbol{M}$ of the dataset $\boldsymbol{D}$ ($\dim \boldsymbol{M} < \dim \boldsymbol{D}$). The optimization objective in that case is

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{M}) = \sum_{\mathbf{x}\in\boldsymbol{M}} \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \sum_{\mathbf{x}\in\boldsymbol{M}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[\ln p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right] \tag{38}$$

since the equation (38) consist of a sum of a single datapoint values of the ELBO function, for the further work the following simplified objective function is used

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[\ln p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right] \tag{39}$$

by the evaluation of (39) and its gradient $\nabla\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})$ w.r.t. $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ the SGB optimization is performed.

The SGB optimization process can be considered as a stochastic approximation of gradient descent optimization since it replaces the actual gradient (calculated from the entire dataset) with an estimate thereof (computed for a randomly selected minibatch from a dataset). For a deeper understanding of theoretical aspects of a stochastic approximation and the proof of its convergence, see [30,31]. For more information about the algorithm of a gradient descent optimization for learning an ANN, see [32]. This subsection is devoted to obtaining an estimator for the objective function $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})$.

**Theorem 5.** *The estimator of the objective function $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})$ can be obtained for a single datapoint $\mathbf{x} \in \mathcal{D}$ in a following form*

$$\hat{\mathcal{L}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \ln p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \tag{40}$$

*where the variable* $\mathbf{z}$ *is expressed as a transformation* $\mathbf{z} = g(\boldsymbol{\phi}, \mathbf{x}, \boldsymbol{\epsilon})$ *and the auxiliary variable* $\boldsymbol{\epsilon}$ *is a single sample from a noise distribution* $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$ *under the following mild assumptions:*

- *the latent variables* $\mathbf{z} \in \boldsymbol{Z}$ *are continuous random variables;*
- *the inference model* $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ *and the generative model* $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$ *are differentiable functions w.r.t. its parameters* $\boldsymbol{\phi}$ *and* $\boldsymbol{\theta}$*, respectively.*

*Proof.* The ELBO function of a single datapoint $\mathbf{x} \in \mathbb{R}^m$ can be written as in equation (39)

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\ln p_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right]$$

Then, its gradient w.r.t. the generative model parameter $\boldsymbol{\theta}$ is

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) &= \nabla_{\boldsymbol{\theta}}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\ln p_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\nabla_{\boldsymbol{\theta}}\left(\ln p_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right)\right] \\
&\approx \nabla_{\boldsymbol{\theta}}\left(\ln p_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right) \\
&= \nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{x}).
\end{aligned}$$

The evaluation of the gradient of the objective function $f(\mathbf{x})$ cannot be performed w.r.t. the inference parameter vector $\boldsymbol{\phi}$, because this gradient cannot be backpropagated through the random variable $\mathbf{z}$.

In order to perform the differentiation, the *reparametrization trick* (see section 2.4 in [33]) must be performed. The latent variable $\mathbf{z} \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ is presented as a following invertible and differentiable transformation of some other random variable $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$ with given $\mathbf{x}$ such that

$$\mathbf{z} = g(\boldsymbol{\phi}, \mathbf{x}, \boldsymbol{\epsilon}), \tag{41}$$

where the distribution $p(\boldsymbol{\epsilon})$ is independent of the variables $\mathbf{x}$ or $\boldsymbol{\phi}$.

The transformation (41) gives a deterministic representation to the random variable $\mathbf{z}$.

Therefore, the expectation can also be rewritten using equality (41) as

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\mathbf{x}] = \int q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})h(\mathbf{z})d\mathbf{z} = \int p(\boldsymbol{\epsilon})h(g(\boldsymbol{\phi}, \mathbf{x}, \boldsymbol{\epsilon}))d\boldsymbol{\epsilon} \tag{42}$$

there the function $h(\mathbf{z})$ is a probability density function for the latent variable $\mathbf{z}$.

Using the transformation (41) the ELBO function can be rewritten as follows

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\ln p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right] = \mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\ln p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right] \quad (43)$$

and the gradient of ELBO can also be reformulated using the equation (41) and the representation of the ELBO function, given by equation (43)

$$\begin{aligned}
\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}}\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) &= \nabla_{\boldsymbol{\theta},\boldsymbol{\phi}}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\ln p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right] \\
&= \nabla_{\boldsymbol{\theta},\boldsymbol{\phi}}\mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\ln p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right] \\
&= \mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}}(\ln p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}))\right] \\
&\approx \nabla_{\boldsymbol{\theta},\boldsymbol{\phi}}(\ln p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})) \qquad (44)
\end{aligned}$$

the result, obtained by the evaluation of equation 44 is a gradient of an estimator of the ELBO function. The estimator is evaluated as

$$\hat{\mathcal{L}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \ln p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \quad (45)$$

and this concludes the proof. □

The previous theorem allows us to formulate the algorithm 1 of an ELBO stochastic optimization.

The performing algorithm 1 requires computation of the $\ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, which depends on the chosen reparametrization function given as (41).

**Remark 7.** *When computing the function* $\ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ *of a given data sample* $\mathbf{x} \in \boldsymbol{D}$ *and a given latent variable* $\mathbf{z} \in \boldsymbol{Z}$ *the following relation can be used*

$$\ln q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = \ln p(\boldsymbol{\epsilon}) - \ln d_{\boldsymbol{\phi}}(\mathbf{x},\boldsymbol{\epsilon})$$

*since, according to the theorem 5 the transformation* $g$ *is an invertible function. Here the function* $d_{\boldsymbol{\phi}}(\mathbf{x},\boldsymbol{\epsilon})$ *is an absolute value of a determinant of a Jacobian matrix* $\frac{\delta \mathbf{z}}{\delta \boldsymbol{\epsilon}}$.

The remark 7 shows that the choice of a transformation $g$ for performing a reparametrization trick is important in order to obtain a flexible and easy-computed inference model.

**Algorithm 1** The stochastic gradient-based optimization of an ELBO.

Input:
$\boldsymbol{D}$: the dataset of training samples
$q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$: the parametric inference model
$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$: the generative model

Output: $\boldsymbol{\phi}, \boldsymbol{\theta}$: learned parameters of the corresponding models.

1. initialize parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$;
2. **while** SGB-descent not converged **do**
   (a) sample random minibatch of data: $\boldsymbol{M} \sim \boldsymbol{D}$;
   (b) sample random noise $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$ for every datapoint in minibatch $\boldsymbol{M}$;
   (c) compute the estimator of the ELBO function and its gradients w.r.t. the parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$ , as it shown by theorem 5;
   (d) update the parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$.
3. **end**

# 5 INVERSE PROBLEM SOLUTION USING VAE

The following chapter is devoted to combining the theoretical results, obtained in chapters 3 and 4 and to presenting a scheme of VAE work when solving a linear inverse problem. The methodology presented in this chapter was inspired by [28, 33]. In the first subsection, the linear inverse problem with a Bayesian solution is recalled and the formulation of its obtaining is presented in a two-step form. The next two subsections reveal the details of the VAE work to perform these steps.

## 5.1 Inverse problem statement in a VAE perspective

Let us first recall the linear inverse problem statement given by equation (4)

$$\mathbf{y} = A\mathbf{x} + \boldsymbol{\eta}$$

where $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$, $\boldsymbol{\eta} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times m}$.

Its Bayesian solution in a form of a posterior density as it was presented in theorem 2 is given by

$$\begin{aligned} \pi_{post}(\mathbf{x}|\mathbf{y}) &= \frac{\pi(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})}{\pi(\mathbf{y})} \\ &= \frac{\nu(\mathbf{y} - A\mathbf{x})\pi(\mathbf{x})}{Z} \end{aligned} \tag{46}$$

where the normalization constant $Z = \int\limits_{\mathbb{R}^m} \nu(\mathbf{y} - A\mathbf{x})\pi(\mathbf{x})d\mathbf{x}$ assumed to be strictly positive.

The solution in form (46) has two main difficulties when computing it:

1. The dimensions of the spaces $\mathbb{R}^m$, $\mathbb{R}^n$ are usually large in real problems and it leads to computationally intensive calculations.
2. The prior density $\pi(\mathbf{x})$ should express all available expert information and data.

Both of these obstacles can be resolved using a variational autoencoder. Since it is based on a dimensionality reduction principle (see definition 22) it can be used for learning a best-fitting prior density coded representation in a latent space $\boldsymbol{Z}$ of a lower dimension.

The decoder, which is also trained by a VAE then will decode the sample from a posterior density over a latent variable with given observed data $\mathbf{y} \in \mathbb{R}^n$ into a vector form solution $\mathbf{x}^* \in \mathbb{R}^m$.

Summarizing it up, there arises 2 tasks in order to solve (4) by a variational autoencoder:

1. Construction of a prior density $\pi(\mathbf{x})$ by learning of a variational autoencoder.
2. Construction of a posterior density for the obtained prior and known observed data $\mathbf{y} \in \mathbb{R}^n$.

These tasks are resolved in complex using a VAE framework and discussed in the following subsections.

## 5.2 Construction of a prior density

Previously there appeared a task of seeking an accurate prior density, which must describe all existed knowledge and beliefs about the model of a particular problem. This subsection contains the idea of obtaining a so-called *data-driven* prior density, which uses available empirical examples of data in order to model its distribution, and, therefore, its PDF.

**Assumption 11.** *Let it be a dataset consisted of i.i.d samples*

$$\boldsymbol{D} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\} = \{\mathbf{x}_i\}_{i=1}^N$$

*from an unknown probability distribution $p(\mathbf{x})$ with a PDF $\pi(\mathbf{x})$.*

With the assumption 11 the information about the prior probability distribution $p(\mathbf{x})$ and its density $\pi(\mathbf{x})$ are consisted in a set $\boldsymbol{D}$. Therefore, using a technique from subsubsection 4.2.1, the prior density can be expressed as $\pi(\mathbf{x}) = \pi(\mathbf{x}|\boldsymbol{D})$. However, this form of a posterior probability density requires performing non-trivial calculations in a high-dimensional space $\mathbb{R}^m$. Moreover, in general case it also might be not amenable to Bayesian computation.

The VAE framework in subsubsection 4.2.3 introduced an alternative technique for prior density evaluation through its marginalization over a latent variable $\mathbf{z} \in \boldsymbol{Z}$, which is

sampled from a relatively simple distribution $p(\mathbf{z})$, such as Gaussian, as it was presented in the assumption 10. Then, the prior density can be evaluated as in (29)

$$\pi(\mathbf{x}) = \int_{\boldsymbol{Z}} \pi(\mathbf{x}, \mathbf{z}|\{\mathbf{x}_i\}_{i=1}^{N})d\mathbf{z} = \int_{\boldsymbol{Z}} \pi(\mathbf{x}, \mathbf{z}|\boldsymbol{D})d\mathbf{z} \tag{47}$$

where dimension of a latent space $\boldsymbol{Z}$ is reduced in comparison to the original dimension of an initial space $\mathbb{R}^m$.

The generative model $\pi(\mathbf{x}, \mathbf{z}|\boldsymbol{D})$ is also evaluated as

$$\pi(\mathbf{x}, \mathbf{z}|\boldsymbol{D}) = \pi(\mathbf{z})\pi(\mathbf{x}|\mathbf{z}, \boldsymbol{D}) \tag{48}$$

where the density $\pi(\mathbf{x}|\mathbf{z}, \boldsymbol{D})$ is an training objective for VAE since it is approximate by a parameterized stochastic decoder $\pi_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$.

As it was determined in subsection 4.3, in order to perform a training process there need to be implemented a stochastic encoder in a form of a conditional probability density $\rho_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, which approximates the VAE true posterior density $\pi(\mathbf{z}|\mathbf{x})$.

Next actions, performing in the optimization of a VAE, were discussed in the subsections 4.4 and 4.5 and remains the same with described neural networks structure.

After performing a learning process, the VAE gives the evaluated approximate prior density $\pi(\mathbf{x})$ in a form of a generative model, where the samples are created from a known Gaussian distribution of a latent variable $\mathbf{z}$ and parameters $\boldsymbol{\phi}$,$\boldsymbol{\theta}$ of an encoder and decoder.

However, that training process does not provide the posterior $\pi_{post}(\mathbf{x}|\mathbf{y})$ since the observed data $\mathbf{y} \in \mathbb{R}^n$ was not involved in it. This leads to the second part of solving the inverse problem: evaluating a posterior density $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{D})$ based on the obtained prior and a decoding mapping $\mu_{\boldsymbol{\theta}}$.

## 5.3 Construction of a posterior density over a latent variable

Contents of this subsection describe the process of obtaining the Bayesian solution of a problem (4) as a posterior density over a latent variable $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{D})$ based on the training set $\boldsymbol{D}$ and the obtaining of a solution in a vector form $\mathbf{x}^* \in \mathbb{R}^m$ as a decoded sample from this probability density function.

Aiming to obtain a density $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{D})$ up to its proportion in a latent space $\boldsymbol{Z}$ of a lower dimensionality, the decoder is set to be a decoding mapping, i.e. $\mathbf{x} = \mu_{\boldsymbol{\theta}}(\mathbf{z})$.

This replacement is based on the fact that the mapping $\mu_{\boldsymbol{\theta}} : \boldsymbol{Z} \mapsto \mathbb{R}^m$ is learned from a set of training samples $\boldsymbol{D}$ on a previous step. Taking it, the posterior density over a latent variable $\mathbf{z} \in \boldsymbol{Z}$ can be computed up to its proportionality

$$\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{D}) \propto \pi(\mathbf{y}|\mathbf{x} = \mu_{\boldsymbol{\theta}}(\mathbf{z}))\pi(\mathbf{z}) \tag{49}$$

where the likelihood $\pi(\mathbf{y}|\mathbf{x} = \mu_{\boldsymbol{\theta}}(\mathbf{z})) = \nu(\mathbf{y} - A\mu_{\boldsymbol{\theta}}(\mathbf{z}))$.

**Remark 8.** *The well-posedness of the posterior density $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{D})$ w.r.t. the Hellinger distance can be proved in a manner akin on the theorem 3 under conditions, formulated in [28].*

The posterior defined by (49) expresses the relation between latent variables sampled from the known distribution and given observed data $\mathbf{y} \in \mathbb{R}^n$. The samples, generated by the associated probability distribution $p(\mathbf{z}|\mathbf{y}, \boldsymbol{D})$ with a density $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{D})$ are in the latent space. Nevertheless, since after the learning of a VAE the decoding mapping $\mu_{\boldsymbol{\theta}}$ is obtained, these samples decoded by it represent a solution for the problem (4).

The solution in a vector form $\mathbf{x}^* \in \mathbb{R}^m$ in that case can be obtained as a sample from the posterior distribution, associated with the defined by (49) posterior density in such a way:

1. The latent variable $\mathbf{z}^* \in \boldsymbol{Z}$ is sampled from a posterior distribution $p(\mathbf{z}|\mathbf{y}, \boldsymbol{D})$.
2. The obtained variable $\mathbf{z}^*$ sampled w.r.t. the observed data $\mathbf{y} \in \mathbb{R}^n$ decodes by a decoding mapping $\mu_{\boldsymbol{\theta}} : \boldsymbol{Z} \mapsto \mathbb{R}^m$ aiming to obtain a vector $\mathbf{x}^* = \mu_{\boldsymbol{\theta}}(\mathbf{z}^*)$.

The two-step scheme of a linear inverse problem solving 5.1 with a variational autoencoder was discussed and can be formulated as the algorithm in the next subsection.

## 5.4 Algorithm of solving a linear inverse problem with a VAE

The subsection is summarizing the content of previous parts 5.1, 5.2 and 5.3 as a following algorithm.

**Algorithm 2** The solving process of a linear inverse problem using a VAE

Input:
$\boldsymbol{D} = \{\mathbf{x}_i\}_{i=1}^N$: the dataset of training samples
$\mathbf{y} \in \mathbb{R}^n$: the given observation
$A \in \mathbb{R}^{n \times m}$: matrix of a forward operator
$\boldsymbol{\eta} \in \mathbb{R}^n$: noise term, sampled independently from a Gaussian distribution $\mathcal{N}(0, \Gamma)$

Output: $\mathbf{x}^* \in \mathbb{R}^m$: the vector sample from a posterior distribution with the density $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{D})$

1. initialize a latent variable PDF $\pi(\mathbf{z}) = \mathcal{N}(0, \sigma^2 \mathbb{I}^{k \times k})$;
2. initialize parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$;
3. set a stochastic encoder as $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = q(\mu_{\boldsymbol{\phi}}(\mathbf{x}))$, where $\mu_{\boldsymbol{\phi}}(\mathbf{x}) : \mathbb{R}^m \mapsto \boldsymbol{Z}$ is an ANN, specifying parameters for the encoder;
4. set a stochastic decoder as $\pi_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \pi(\mu_{\boldsymbol{\theta}}(\mathbf{z}))$, where $\mu_{\boldsymbol{\theta}}(\mathbf{z}) : \boldsymbol{Z} \mapsto \mathbb{R}^m$ is an ANN, specifying parameters for the decoder;
    - using the decoder, set a generative model $\pi(\mathbf{x}, \mathbf{z})$;
5. perform a VAE learning process, according to the algorithm 1 and obtain parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$;
6. reset a stochastic decoder $\pi_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ to a mapping $\mu_{\boldsymbol{\theta}}(\mathbf{z}) : \boldsymbol{Z} \mapsto \mathbb{R}^m$;
7. compute the posterior density up to its proportionality $\pi(\mathbf{z}|\mathbf{y}) \propto \mathcal{N}(\mathbf{y} - A\mu_{\boldsymbol{\theta}}(\mathbf{z}); 0, \Gamma)\mathcal{N}(\mathbf{z}; 0, \sigma^2 \mathbb{I}^{k \times k})$;
8. sample a latent variable $\mathbf{z}^* \sim \pi(\mathbf{z}|\mathbf{y})$;
9. decode the latent variable into a variable in the initial space $\mathbf{x}^* = \mu_{\boldsymbol{\theta}}(\mathbf{z}^*)$.
10. **end**

Step 8 can be performed, using, for example, a Monte Carlo Markov Chain algorithm [28, 34].

# 6 DISCUSSION

## 6.1 Current study

In process of a current study the objectives, defined in subsection 1.2 were discussed and achieved. Chapter 2 described the notion of an inverse problem and its ill-posedness with a concept of a deterministic solution in a vector form. The alternative Bayesian approach to linear inverse problems was studied in chapter 3 including the construction of a Bayesian solution as a conditional probability density over an unknown variable, given the observed data. Chapter 4 was devoted to studying the complex notion of a variational autoencoder with corresponding proof of the fact, that ELBO can be used as its optimization objective.

The main result of the study, proposed in chapter 5 incorporates the algorithm of solving a linear inverse problem using a variational autoencoder as a tool for learning a prior distribution, based on a training dataset, which contains all existed knowledge about the probability density of an unknown. Although this result is strictly theoretical, the numerical tests, provided in [28] show that for the particular imaging linear inverse problems this method shows sufficiently good results.

## 6.2 Future work

The continuity of this study might be processed in several directions. First, since this thesis was limited to studying linear problems only, it can be extended to non-linear inverse problems and study the possibility to apply the VAE to obtain a well-posed posterior distribution, as a solution to a non-linear inverse problem. Second, since the result of the research is theoretical, there could be provided additional numerical tests for solving classical deconvolution problems aiming to compare the methodology, presented in the thesis with results, obtained via regularization methods. Thirdly, since a process of sampling from a posterior distribution and obtaining point estimations were left out of the scope of this study, the discussion about choosing the most sufficient algorithms for that purpose also remains open.

# 7 CONCLUSION

This thesis was aimed at obtaining a methodology for solving a linear inverse Bayesian problem using a variational autoencoder. This technique was developed and presented in the form of an algorithm. It is based on the results of the study of inverse problems in their classical and Bayesian formulation. The Bayesian approach was then embedded in the structure of the VAE, which was also studied in the thesis with its basic principle, structure, and optimization objective. The combination of these studies led to the opportunity to construct a data-driven prior density in a latent space of a lower dimensionality using a VAE.

# REFERENCES


[1] M. Bertero and P. Boccacci. *Introduction to Inverse Problems in Imaging*. CRC Press, 1st edition, 1998.

[2] Manya V. Afonso, José M. Bioucas-Dias, and Mário A. T. Figueiredo. Fast image recovery using variable splitting and constrained optimization. *IEEE Transactions on Image Processing*, 19(9):2345–2356, 2010.

[3] G. Freiling and V.A. Yurko. *Inverse Sturm-Liouville Problems and Their Applications*. Nova Science Publishers, 1st edition, 2001.

[4] Jari P. Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*. Springer Science+Business Media, Inc., 1st edition, 2005.

[5] Alexander G. Ramm. *Inverse Problems*. Springer New York, NY, 1st edition, 2005.

[6] Sheldon J. Axler. *Linear Algebra Done Right*. Springer Science+Business Media, Inc., 3rd edition, 2015.

[7] Christian Clason. Regularization of inverse problems. https://arxiv.org/abs/2001.00617, 2020.

[8] Jun Lu. Matrix decomposition and applications. https://arxiv.org/abs/2201.00145, 2022.

[9] Roger Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.

[10] Hanne Kekkonen. Bayesian inverse problems, 2019. [Online; accessed March, 24, 2022].

[11] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Springer Dordrecht, 1st edition, 1996.

[12] J. Hadamard. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. Dover Publications, INC., 2nd edition, 2014.

[13] Per Christian Hansen. The truncatedSVD as a method for regularization. *BIT Numerical Mathematics*, 27(4):534–553, 1987.

[14] Sergei Pereverzyev and Eberhard Schock. Morozov's discrepancy principle for tikhonov regularization of severly ill-posed problems in finite-dimensional subspaces. *Numerical Functional Analysis and Optimization*, 21(7), 2000.

[15] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Resource Center, 2nd edition, 2001.

[16] Tapio Helin. Bayesian inversion: Theoretical perspective, 2016. [Online; accessed March, 14, 2022].

[17] Patrick Billingsley. *Probability and measure*. New York: John Wiley & Sons, 3rd edition, 1995.

[18] Masoumeh Dashti and Andrew M. Stuart. The bayesian approach to inverse problems. https://arxiv.org/abs/1302.6989, 2013.

[19] B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten. Bayesian inverse problems with gaussian priors. https://arxiv.org/abs/1103.2692, 2011.

[20] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient bayesian computation by proximal markov chain monte carlo: when langevin meets moreau. https://arxiv.org/abs/1612.07471, 2016.

[21] Daniel Sanz-Alonso, Andrew M. Stuart, and Armeen Taeb. Inverse problems and data assimilation. https://arxiv.org/abs/1810.06191, 2018.

[22] Adrian Barbu and Song-Chun Zhu. *Monte Carlo Methods*. Springer Singapore, 1st edition, 2020.

[23] Kevin Gurney. *An Introduction to Neural Networks*. Taylor & Francis e-Library, 1st edition, 1997.

[24] Simon S. Haykin. *Neural networks and learning machines*. Pearson Education, 3rd edition, 2009.

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 1st edition, 2016.

[26] Elad Plaut. From principal subspaces to principal components with linear autoencoders. https://arxiv.org/abs/1804.10253, 2018.

[27] Jørgen Bang-Jensen and Gregory Z. Gutin. *Digraphs*. Springer London, 2nd edition, 2009.

[28] Matthew Holden, Marcelo Pereyra, and Konstantinos C. Zygalakis. Bayesian imaging with data-driven priors encoded by neural networks: Theory, methods, and algorithms. https://arxiv.org/abs/2103.10182, 2021.

[29] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[30] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[31] Julius R. Blum. Approximation Methods which Converge with Probability one. *The Annals of Mathematical Statistics*, 25(2):382 – 386, 1954.

[32] Jiawei Zhang. Gradient descent based optimization algorithms for deep learning models training. http://arxiv.org/abs/1903.03614, 2019.

[33] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. https://arxiv.org/abs/1312.6114, 2013.

[34] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: Modifying old algorithms to make them faster. *Statistical Science*, 28(3):424 – 446, 2013.