CZECH TECHNICAL UNIVERSITY IN PRAGUE
Faculty of Nuclear Sciences and Physical Engineering

Machine learning for prediction of energy in condensed matter physics

Aplikace strojového učení k predikci energií ve fyzice pevných látek

Diploma Thesis

Author:	**Bc. Jiří Chmel**
Supervisor:	**doc. RNDr. Jan Vybíral, Ph.D.**
Academic year:	2021/2022

České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská

Katedra: matematiky Akademický rok: 2020/2021

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student:	Bc. Jiří Chmel
Studijní program:	Aplikace přírodních věd
Studijní obor:	Aplikované matematicko-stochastické metody
Název práce (česky):	Aplikace strojového učení k predikci energií ve fyzice pevných látek
Název práce (anglicky):	Machine learning for prediction of energy in condensed matter physics

Pokyny pro vypracování:

1) Student se seznámí s metodou používanou k získávání dat o chemických sloučeninách.
2) Student se seznámí s přístupy používanými k získání vektorů popisu materiálů (tzv. deskriptory) ve fyzice pevných látek a vybrané aplikuje.
3) S využitím metod strojového učení student prozkoumá vztah mezi vlastnostmi materiálu (vazebná energie, šířka zakázaného pásu) a jeho geometrií.
4) Získané algoritmy student aplikuje na dostupné datasety z Fritz-Haberova Institutu v Berlíně a výsledky porovná s dostupnou literaturou.

Doporučená literatura:

1) L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, Big data of materials science - Critical role of the descriptor. Phys. Rev. Lett. 114, 2015, 105503.

2) L. M. Ghiringhelli, J. Vybiral, E. Ahmetchik, R. Ouyang, S. V. Levchenko, C. Draxl, M. Scheffler, Learning physical descriptors for materials science by compressed sensing. New Journal of Physics 19, 2017, 023017.

3) C. Sutton, L. M. Ghiringhelli, T. Yamamoto, Y. Lysogorskiy, L. Blumenthal, T. Hammerschmidt, J. R. Golebiowski, X. Liu, A. Ziletti, M. Scheffler, Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition, Npj Comput. Mater. 5, 2019, 111.

4) C. M. Bishop, Pattern recognition and machine learning, Springer, 2006.

5) T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: Data mining, inference, and prediction. Springer, New York, 2009.

Jméno a pracoviště vedoucího diplomové práce:

doc. RNDr. Jan Vybíral, Ph.D.
Katedra matematiky FJFI, ČVUT v Praze, Trojanova 13, 120 00 Praha 2

Jméno a pracoviště konzultanta:

Doc. Ing. Václav Šmídl, PhD.
ÚTIA AV ČR, Pod vodárenskou věží 4, 180 00 Praha 8

Datum zadání diplomové práce: 31.10.2020

Datum odevzdání diplomové práce: 3.5.2021

Doba platnosti zadání je dva roky od data zadání.

Acknowledgment:
I would like to thank doc. RNDr. Jan Vybíral, Ph.D. for his guidance and patience during this uneasy academic year.

Author's declaration:
I declare that this Research Project is entirely my own work and I have listed all the used sources in the bibliography.

Prague, January 15, 2022 Jiří Chmel

Contents

Introduction

something

Chapter 1

Methodology

The following chapter captures the theoretical background of the machine learning used in this work starting with the simplest methods of ordinary least squares and gradually building towards neural networks. The notation and conventions which will be used throughout this work are defined as stated below.

Definition 1 (The $\hat{N}$ notation)**.** The set of natural numbers $\{1, 2, \ldots, N\}$ is denoted as $\hat{N}$.

Definition 2 (Vector and Matrix Notation)**.** Vectors of real numbers $\boldsymbol{y} \in \mathbb{R}^N$, $N \in \mathbb{N}$ are denoted by

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \Big(y_1, y_2, \ldots, y_N\Big)^T . \tag{1.1}$$

Matrices of real numbers $\boldsymbol{X} \in \mathbb{R}^{N \times M}$, $M, N \in \mathbb{N}$ are denoted by

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1M} \\ x_{21} & x_{22} & \ldots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & \ldots & \ldots & x_{NM} \end{pmatrix} = \Big(\boldsymbol{x}_{\bullet 1}, \boldsymbol{x}_{\bullet 2}, \ldots, \boldsymbol{x}_{\bullet M}\Big) = \Big(\boldsymbol{x}_{1\bullet}, \boldsymbol{x}_{2\bullet}, \ldots, \boldsymbol{x}_{N\bullet}\Big)^T , \tag{1.2}$$

where $\boldsymbol{x}_{\bullet j} \in \mathbb{R}^N$ for $j \in \hat{M}$ are columns of $\boldsymbol{X}$ and $\boldsymbol{x}_{i\bullet} \in \mathbb{R}^M$ for $i \in \hat{N}$ are rows of $\boldsymbol{X}$.

Definition 3 (ℓ_p norm)**.** Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)^T \in \mathbb{R}^M$.

1. If $p \in [1, +\infty)$, then the ℓ_p norm of a vector $\boldsymbol{x} \in \mathbb{R}^M$ is

$$\|\boldsymbol{x}\|_p = \Big(\sum_{j=1}^{M} |x_j|^p\Big)^{\frac{1}{p}}. \tag{1.3}$$

2. The ℓ_0 norm of a vector $\boldsymbol{x} \in \mathbb{R}^M$ is

$$\|\boldsymbol{x}\|_0 = \#\{j : x_j \neq 0\}, \tag{1.4}$$

which counts the number of non-zero components of $\boldsymbol{x}$.

Remark 1. For $0 < p < 1$ the convexity is broken (the triangle inequality doesn't hold). Such function is then called quasinorm.

1.1 Regression Methods

The methods used in this work and their underlying theory is outlined in this section.

1.1.1 Ordinary Least Squares

Let us assume we have a real setting with data points $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)$, where $\boldsymbol{x}_i \in \mathbb{R}^M$ for $i \in \hat{N}$ and $y_1, y_2, \ldots, y_N \in \mathbb{R}$. Generally, we want to describe the dependence of the output y_i on $\boldsymbol{x}_i$, in other words we want to model the relation $y_i = f(\boldsymbol{x}_i)$ for $i \in \hat{N}$. Here, we presume the model is linear in the coefficients: $f(\boldsymbol{x}_i) = \langle \boldsymbol{x}_i, \boldsymbol{b} \rangle$, where $\boldsymbol{b} = (b_1, b_2, \ldots, b_M)^T \in \mathbb{R}^M$ is the vector of the coefficients and $\langle \cdot, \cdot \rangle$ is the scalar product of two vectors. Our goal is to obtain the coefficients $\hat{\boldsymbol{b}} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_M)^T \in \mathbb{R}^M$. Generally, we want to add an absolute term called *bias* (or *intercept*) to our linear regression model. We elegantly do so by adding a column of ones to the matrix $\boldsymbol{X}$ which is then $N \times (M + 1)$ dimensional and the vector of coefficients $\hat{\boldsymbol{b}}$ is $(M + 1)$ dimensional. It will be assumed (unless explicitly said otherwise) that bias is included in the model.

The linear model of ordinary least squares (OLS) is the best known method of statistical learning. The linear regression model for a matrix of regressor $\boldsymbol{X} \in \mathbb{R}^{N \times M}$ has the form

$$y_i \approx \sum_{j=1}^{M} x_{ij} b_j = \langle \boldsymbol{x}_{i\bullet}, \boldsymbol{b} \rangle, \forall i \in \hat{N}. \tag{1.5}$$

We want to find the estimate of the vector of coefficients $\boldsymbol{b} = (b_1, \ldots, b_M)^T \in \mathbb{R}^M$. We perform the minimization of quadratic loss - the least squares

$$\hat{\boldsymbol{b}}_{OLS} = \underset{\boldsymbol{b} \in \mathbb{R}^M}{\operatorname{argmin}} J_{OLS}(\boldsymbol{b}) = \underset{\boldsymbol{b} \in \mathbb{R}^M}{\operatorname{argmin}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}||_2^2, \tag{1.6}$$

It is necessary that the datapoints are independently identically distributed (iid). The equation 1.6 does not tell us if the model we find is valid. In other words, we can find the ordinary least squares model for any dataset and the model quality must be evaluated afterwards. It is easily seen the problem of least squares is convex which means the solution can be found in a closed form

$$\begin{aligned} J_{OLS}(\boldsymbol{b}) &= ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}||_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \\ \frac{\partial J_{OLS}}{\partial \boldsymbol{b}} &= -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \end{aligned} \tag{1.7}$$

If $\boldsymbol{X}^T\boldsymbol{X}$ is regular, the unique solution can be recovered from $0 = \frac{\partial J_{OLS}}{\partial \boldsymbol{b}} \implies 0 = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$ which means that the coefficients $\hat{\boldsymbol{b}}$ are given as

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{1.8}$$

The predicted values at the training inputs are then

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{b}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}. \tag{1.9}$$

The matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is called the "hat" matrix because it puts the hat on vector $\boldsymbol{y}$ and it computes the orthogonal projection $\hat{\boldsymbol{y}}$ onto the hyperplane spanned by the columns of $\boldsymbol{X}$. Therefore the vector $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to the hyperplane.

If the $\boldsymbol{X}^T\boldsymbol{X}$ matrix is singular then there is not a unique solution. Methods which attempt to deal with such problems are subjects of the following sections.

1.1.2 Ridge Regression

Ridge regression solves the possible problem of matrix singularity from the previous section by adding a regularization term into the loss function:

$$\hat{\boldsymbol{b}}_{ridge} = \underset{\boldsymbol{b}\in\mathbb{R}^M}{\operatorname{argmin}}\, J_{ridge}(\boldsymbol{b}) = \underset{\boldsymbol{b}\in\mathbb{R}^M}{\operatorname{argmin}} \Big(||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}||_2^2 + \lambda||\boldsymbol{b}||_2^2\Big),\ \lambda > 0. \tag{1.10}$$

The concept of regularization is sometimes called weight-decay because we impose a penalty on the size of $\boldsymbol{b}$. The parameter λ controls the strength of the regularization. We get OLS coefficients for $\lambda \to 0^+$ and $\hat{\boldsymbol{b}}_{ridge} = 0$ for $\lambda \to +\infty$. The solution can be found using the very same procedure as in (1.7):

$$\begin{aligned} J_{ridge}(\boldsymbol{b}) &= ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}||_2^2 + \lambda||\boldsymbol{b}||_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) + \lambda\boldsymbol{b}^T\boldsymbol{b} \\ \frac{\partial J_{ridge}}{\partial \boldsymbol{b}} &= -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) + 2\lambda\boldsymbol{b}. \end{aligned} \tag{1.11}$$

Putting $0 = \frac{\partial J_{ridge}}{\partial \boldsymbol{b}} \implies 0 = -\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) + \lambda\boldsymbol{b} = -\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) + \lambda\boldsymbol{I}\boldsymbol{b}$, where $\boldsymbol{I}$ is the identity matrix. The solution can be easily extracted as

$$\hat{\boldsymbol{b}}_{ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{1.12}$$

It is important to standardize the columns of the matrix $\boldsymbol{X}$ before training to eliminate spurious behavior. The shape of the equation (1.12) shows the reason this procedure works: the regularization stabilizes the inverse of the matrix for some value of λ. The optimal value is usually chosen using cross validation.

1.1.3 Kernel Ridge Regression (KRR)

Kernel ridge regression (KRR) builds on top of ridge regression and allows modeling of nonlinear relationships. We put $\boldsymbol{x}_{i\bullet} = \boldsymbol{x}_i$ to make the notation less cumbersome in this section. The datapoints themselves are replaced with a feature vector $\boldsymbol{x}_i \to \phi(\boldsymbol{x}_i)$ where $\phi : \mathbb{R}^M \to \mathcal{F}$ is a nonlinear mapping to a higher dimensional feature space $\mathcal{F}$, $\dim(\mathcal{F}) \leq +\infty$. Now, we consider datapoints $(\phi(\boldsymbol{x}_1), y_i), \ldots, (\phi(\boldsymbol{x}_N), y_N)$ for the very same learning algorithm as in Section 1.1.2. In other words, we find ridge regression coefficients and create a linear model in feature space where datapoints $(\phi(\boldsymbol{x}_1), y_i), \ldots, (\phi(\boldsymbol{x}_N), y_N)$ are but we observe a nonlinear model in space where datapoints $(\boldsymbol{x}_1, y_i), \ldots, (\boldsymbol{x}_N, y_N)$ are.

We define the loss function of KRR

$$J_{KRR}(\boldsymbol{b}) = ||\boldsymbol{y} - \Phi\boldsymbol{b}||_2^2 + \lambda||\boldsymbol{b}||_2^2 = \sum_{i=1}^{N}\Big(y_i - \boldsymbol{b}^T\phi(\boldsymbol{x}_i)\Big)^2 + \lambda\boldsymbol{b}^T\boldsymbol{b}, \tag{1.13}$$

where $\lambda > 0$ and $\Phi = \begin{pmatrix} \phi^T(\boldsymbol{x}_1) \\ \vdots \\ \phi^T(\boldsymbol{x}_N) \end{pmatrix}$ is the mapping of matrix $\boldsymbol{X}$. Setting the gradient of J_{KRR} in (1.13) equal to zero gives

$$\boldsymbol{b} = -\frac{1}{\lambda}\sum_{i=1}^{N}\Big(y_i - \boldsymbol{b}^T\phi(\boldsymbol{x}_i)\Big)\phi(\boldsymbol{x}_i) = \sum_{i=1}^{N} a_i\phi(\boldsymbol{x}_i) = \Phi^T\boldsymbol{a}, \tag{1.14}$$

where we put

$$a_i = -\frac{1}{\lambda}\Big(y_i - \boldsymbol{b}^T\phi(\boldsymbol{x}_i)\Big) \text{ for } \forall i \in \hat{N}. \tag{1.15}$$

The result of (1.14) allows us to reformulate the loss function in terms of $\boldsymbol{a}$ instead of $\boldsymbol{b}$

$$J_{KRR}(\boldsymbol{a}) = \|\boldsymbol{y} - \Phi\Phi^T\boldsymbol{a}\|_2^2 + \lambda\|\Phi^T\boldsymbol{a}\|_2^2 = \|\boldsymbol{y} - \Phi\Phi^T\boldsymbol{a}\|_2^2 + \lambda\boldsymbol{a}^T\Phi\Phi^T\boldsymbol{a}. \tag{1.16}$$

Let us examine the result. We put $K = \Phi\Phi^T$. Therefore

$$K_{ij} = (\Phi\Phi^T)_{ij} = \langle\phi(\boldsymbol{x}_i),\ \phi(\boldsymbol{x}_j)\rangle = k(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{1.17}$$

where we introduce the kernel function $k : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$. The loss function then takes very elegant form

$$J_{KRR}(\boldsymbol{a}) = \|\boldsymbol{y} - K\boldsymbol{a}\|_2^2 + \lambda\boldsymbol{a}^T K\boldsymbol{a}. \tag{1.18}$$

This is the final form of the loss function for KRR with kernel K. Setting the gradient of J_{KRR} with respect to $\boldsymbol{a}$ to zero gives the final solution

$$\boldsymbol{a} = (K + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}, \tag{1.19}$$

and the original coefficients

$$\boldsymbol{b} = \Phi^T(K + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}. \tag{1.20}$$

The prediction y_{pred} for new datapoint $\boldsymbol{x}$ can be expressed elegantly as

$$y_{pred} = \boldsymbol{b}^T\phi(\boldsymbol{x}) = \boldsymbol{a}^T\Phi\phi(\boldsymbol{x}) = \boldsymbol{y}^T(K + \lambda\boldsymbol{I})^{-1}\Phi\phi(\boldsymbol{x}) = \boldsymbol{y}^T(K + \lambda\boldsymbol{I})^{-1}\kappa(\boldsymbol{x}). \tag{1.21}$$

where $\kappa(\boldsymbol{x}) = \begin{pmatrix}\phi^T(\boldsymbol{x}_1)\\ \vdots \\ \phi^T(\boldsymbol{x}_N)\end{pmatrix}\phi(\boldsymbol{x}) = \begin{pmatrix}\langle\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x})\rangle\\ \vdots \\ \langle\phi(\boldsymbol{x}_N), \phi(\boldsymbol{x})\rangle\end{pmatrix} = \begin{pmatrix}k(\boldsymbol{x}_1, \boldsymbol{x})\\ \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x})\end{pmatrix}$. We shall see that we can avoid working with the mapping $\phi(\boldsymbol{x})$ which can be even infinitely dimensional. The object of our interest is the kernel function k and we will show it is all we need in the following section.

So far, we dealt with the kernel K and its kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ without specifying any needed properties of these mathematical objects. The following theorem explains why we can avoid working with the cumbersome mapping ϕ and justifies our previous steps.

Theorem 1 (Mercer)**.** *To guarantee that the symmetric continuous function $k(\boldsymbol{x}, \boldsymbol{y}) : C \times C \to \mathbb{R}$ on compact set $C \subset \mathbb{R}^N$ has an expansion*

$$k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{N_{\mathcal{F}}} \lambda_j\psi_j(\boldsymbol{x})\psi_j(\boldsymbol{y}) = \langle\phi(\boldsymbol{x}), \phi(\boldsymbol{y})\rangle \tag{1.22}$$

with $\lambda_j > 0$ and $\phi : \mathbb{R}^N \to \mathcal{F}$ with $\dim(\mathcal{F}) = N_{\mathcal{F}} \leq +\infty$, it is necessary and sufficient that the function k is a kernel of a positive integral operator on $L_2(C)$:

$$\forall f \in L_2(C) : \int_C\int_C k(\boldsymbol{x}, \boldsymbol{y})f(\boldsymbol{x})f(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} \geq 0. \tag{1.23}$$

Proof. Can be found in [?]. □

It is easy to see a possible realization of the mapping can have the form $\phi(\boldsymbol{x}) = \left(\sqrt{\lambda_1}\psi_1(\boldsymbol{x}),\ \sqrt{\lambda_2}\psi_2(\boldsymbol{x}), \dots\right)$.

The significance of this result is that we do not need to know the shape ϕ. This is often called the kernel trick. The dimensionality of ϕ is infinite for Gaussian kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma\|\boldsymbol{x} - \boldsymbol{y}\|_2^2)$ which can be seen from the following decomposition:

$$\begin{aligned}\langle\phi(\boldsymbol{x}), \phi(\boldsymbol{y})\rangle = k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma\|\boldsymbol{x} - \boldsymbol{y}\|_2^2) = \exp(-\gamma\|\boldsymbol{x}\|_2^2 + 2\gamma\langle\boldsymbol{x}, \boldsymbol{y}\rangle - \gamma\|\boldsymbol{y}\|_2^2) = \\ = \exp(-\gamma\|\boldsymbol{x}\|_2^2)\exp(2\gamma\langle\boldsymbol{x}, \boldsymbol{y}\rangle)\exp(-\gamma\|\boldsymbol{y}\|_2^2)).\end{aligned} \tag{1.24}$$

Taking the middle term and using the fact that Taylor expansion of e^x exists for all $x \in \mathbb{R}$:

$$\begin{aligned}\exp(2\gamma\langle \boldsymbol{x}, \boldsymbol{y}\rangle) &= 1 + 2\gamma\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \frac{(2\gamma)^2\langle \boldsymbol{x}, \boldsymbol{y}\rangle^2}{2} + \frac{(2\gamma)^3\langle \boldsymbol{x}, \boldsymbol{y}\rangle^3}{6} + \cdots = \\ &= 1 + 2\gamma(x_1 y_1 + x_2 y_2 + \cdots + x_N y_N) + \cdots = \\ &= \Big\langle (1, \sqrt{2\gamma}x_1, \sqrt{2\gamma}x_2, \dots, \sqrt{2\gamma}x_N, \dots)^T, (1, \sqrt{2\gamma}y_1, \sqrt{2\gamma}y_2, \dots, \sqrt{2\gamma}y_N, \dots)^T \Big\rangle,\end{aligned} \tag{1.25}$$

where we do not list higher order expansion terms for visibility. The mapping function can be expressed as

$$\phi(\boldsymbol{x}) = \exp(-\gamma\|\boldsymbol{x}\|_2^2)\Big(1, \sqrt{2\gamma}x_1, \sqrt{2\gamma}x_2, \dots, \sqrt{2\gamma}x_N, \dots\Big)^T, \tag{1.26}$$

and its dimension is infinite because of the Taylor expansion we used.

New kernels can be constructed from already developed kernels. We list a few of the techniques in Table 1.1.

Construction Technique
$k(\boldsymbol{x}, \boldsymbol{y}) = ck_1(\boldsymbol{x}, \boldsymbol{y})$
$k(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x})k_1(\boldsymbol{x}, \boldsymbol{y})f(\boldsymbol{y})$
$k(\boldsymbol{x}, \boldsymbol{y}) = q(k_1(\boldsymbol{x}, \boldsymbol{y}))$
$k(\boldsymbol{x}, \boldsymbol{y}) = \exp(k_1(\boldsymbol{x}, \boldsymbol{y}))$
$k(\boldsymbol{x}, \boldsymbol{y}) = k_1(\boldsymbol{x}, \boldsymbol{y}) + k_2(\boldsymbol{x}, \boldsymbol{y})$
$k(\boldsymbol{x}, \boldsymbol{y}) = k_1(\boldsymbol{x}, \boldsymbol{y})k_2(\boldsymbol{x}, \boldsymbol{y})$
$k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{y}$

Table 1.1: The $k_1(\boldsymbol{x}, \boldsymbol{y})$ and $k_2(\boldsymbol{x}, \boldsymbol{y})$ are valid kernels, constant $c > 0$, f is a function defined on $\mathbb{R}^N$, q is a polynomial with nonnegative coefficients and $\boldsymbol{A}$ is a symmetric positive semidefinite matrix

We can construct the Gaussian kernel from the linear kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T\boldsymbol{y}$ which is a trivial identity. We use the second and the fourth technique in Table 1.1:

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma\|\boldsymbol{x}\|_2^2 + 2\gamma\langle \boldsymbol{x}, \boldsymbol{y}\rangle\gamma + \|\boldsymbol{y}\|_2^2) = \exp(-\gamma\boldsymbol{x}^T\boldsymbol{x})\exp(2\gamma\boldsymbol{x}^T\boldsymbol{y})\exp(-\gamma\boldsymbol{y}^T\boldsymbol{y}). \tag{1.27}$$

Some commonly used kernels are listed in Table 1.2.

	Kernels
Gaussian	$\exp(-\gamma\|\boldsymbol{x} - \boldsymbol{y}\|_2^2)$
Laplacian	$\exp(-\gamma\|\boldsymbol{x} - \boldsymbol{y}\|_1)$
Sigmodial,	$\tanh(\kappa(\boldsymbol{x} \cdot \boldsymbol{y}) + \theta)$
Polynomial	$(\boldsymbol{x} \cdot \boldsymbol{y} + \theta)^d$

Table 1.2: Commonly used kernels. $\gamma > 0$, $\kappa \in \mathbb{R}$, $\theta \in \mathbb{R}$, $d \in \mathbb{N}$.

We are concerned with the Gaussian and Laplacian kernels because of their form. These two kernels have the property $k(\boldsymbol{x}, \boldsymbol{y}) = k(\|\boldsymbol{x} - \boldsymbol{y}\|_p)$ where $p \geq 1$ and are called radial basis functions. This property will play an important role in the carried out experiments.

Kernel ridge regression with Gaussian or Laplacian kernel have two parameters λ and γ which have to be optimized outside of the training procedure. Such numbers are called hyperparameters and the following section briefly outlines the methods used for their optimization as well as other important related practices.

1.1.4 The Least Absolute Shrinkage and Selection Operator (LASSO)

The LASSO emerged as a technique to obtain low-dimensional solutions to regression problems and interestingly enough, long before the underlying theory was developed and understood thoroughly. Since its establishment as a useful method, LASSO made its way into the portfolio of virtually every machine learning engineer. However, the proper use of the method with all the constrains fulfilled is not always done as it should be. This section establishes the theory needed to define LASSO with careful attention towards the use case in this work.

Definition 4 (k-sparse vectors)**.** Let $k \in \mathbb{N}$ such that $k < M$. A vector $\boldsymbol{x} \in \mathbb{R}^M$ is called k-sparse if $||x||_0 \leqslant k$. The set of all k-sparse vectors is

$$\mathbb{R}_k^M = \{\boldsymbol{x} \in \mathbb{R}^M : ||x||_0 \leqslant k\} \tag{1.28}$$

Remark 2. It is easy to see that for every, $\boldsymbol{x} \in \mathbb{R}^M$ there is a permutation π: $\hat{M} \mapsto \hat{M}$ such that

$$|x_{\pi(1)}| \geqslant |x_{\pi(2)}| \geqslant \cdots \geqslant |x_{\pi(M)}| \geqslant 0. \tag{1.29}$$

The vector $\boldsymbol{x}^* \in \mathbb{R}^M$ with components $x_j^* = |x_{\pi(j)}|$ for $j \in \hat{M}$ is called nonincreasing rearrangement of $\boldsymbol{x}$.

Definition 5 (The Best k-term Approximation)**.** Let $k \leqslant M$ and ℓ_p be a norm, $p > 1$. The best k-term approximation $\sigma_k(\boldsymbol{x})_p$ of $\boldsymbol{x} \in \mathbb{R}^M$ is

$$\sigma_k(\boldsymbol{x})_p = \inf_{\tilde{\boldsymbol{x}} \in \mathbb{R}_k^M} ||\boldsymbol{x} - \tilde{\boldsymbol{x}}||_p = \Big(\sum_{j=k+1}^{M} |x_j^*|^p \Big)^{\frac{1}{p}}. \tag{1.30}$$

ℓ_0 Minimization and Basis Pursuit

Definition 6 (ℓ_0 Minimization)**.** Let $\boldsymbol{x} \in \mathbb{R}^M$, $\boldsymbol{A} \in \mathbb{R}^{N \times M}$ be known and $\boldsymbol{y} \in \mathbb{R}^N$ be known. The ℓ_0 minimization problem is defined as

$$\min_{\boldsymbol{x} \in \mathbb{R}^M} ||\boldsymbol{x}||_0 \; subject \; to \; \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{1.31}$$

Remark 3. It will be shown that ℓ_0 minimization is numerically a very expensive optimization problem. For this purpose, we introduce the classes of complexity:

- P class - all decision problems which can be solved in polynomial time.
- NP class - a candidate for solution can be tested in polynomial time.
- NP-hard class - decision problems for which all their solving algorithms can be transformed in polynomial time into a solving algorithm of any other NP problem.
- NP-complete class - those decision problems which are NP-hard and NP.

Here, we will present without a proof a problem from complexity theory called Three Cover Problem which is NP-complete.

Three Cover Problem

Let $N \in \mathbb{N}$ be divisible by 3 and $M \in \mathbb{N}$. We define a system $\{T_j : j \in \hat{M}\}$ of subsets of $\hat{N}$ and $\#T_j = 3$ for $\forall j \in \hat{M}$. **Decision problem:** Establish the existence of a subsystem $\{T_j : j \in J\}$ for which holds:

1. $\bigcup_{j\in J} T_j = \hat{N}$,

2. $T_i \bigcap T_j = \emptyset$ for $i, j \in J, i \neq j$.

Theorem 2. *The ℓ_0 minimization problem is NP-hard.*

Proof. The problem (1.31) will be reformulated as the Three Cover Problem. Using the notation in the definiton of the Three Cover problem we construct a matrix $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ which columns $\boldsymbol{a}_j$ are the characteristic function of the given T_j. Therefore the components of $\boldsymbol{a}_j$ are defined as:

$$a_{ij} := \begin{cases} 1 & \text{if } i \in T_j \\ 0 & \text{if } i \notin T_j. \end{cases}$$

The vector and matrix multiplication gives

$$\boldsymbol{A}\boldsymbol{x} = \sum_{j=1}^{M} x_j \boldsymbol{a}_j.$$

It is easy to see from the construction itself that the matrix $\boldsymbol{A}$ can be constructed in polynomial time. Let's presume $\boldsymbol{x}$ is the solution of the ℓ_0 minimization problem with $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} = (1, \ldots, 1)^T$. The vector and matrix multiplication causes the amount of nonzero components of $\boldsymbol{x}$ to be at most three times bigger:

$$N = \|\boldsymbol{y}\|_0 = \|\boldsymbol{A}\boldsymbol{x}\|_0 \leqslant 3\|\boldsymbol{x}\|_0 \Leftrightarrow \|\boldsymbol{x}\|_0 \geqslant N/3.$$

We will show: The Exact Cover problem has a solution if and only if $\|\boldsymbol{x}\|_0 = N/3$.
$\Rightarrow$: Then $\exists J \subset \hat{M}$ and the amount of columns needed is precisely $N/3$, that is $|J| = N/3$ and

$$(1, \ldots, 1)^T = \sum_{j\in J} \boldsymbol{a}_j = \sum_{j=1}^{M} \boldsymbol{x}_j \boldsymbol{a}_j.$$

Now, it is easy to see that $\boldsymbol{x}$ has nonzero components which are ones and only for indices in J which gives $\|x\|_0 = |J| = N/3$.
$\Leftarrow$: Assuming $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ with $\|x\|_0 = N/3$ then we choose a subsystem $\{T_j : j \in supp(\boldsymbol{x})\}$. □

With ℓ_0 minimization being too difficult to solve for any $\boldsymbol{A}$ and $\boldsymbol{y}$, we are force to find a feasible compromise. We demand the problem to be convex and also promote sparsity. Convexity will be ensured if we choose to use ℓ_p norm where $p \geqslant 1$. Sparsity will be possible for $p \leqslant 1$. Therefore we are left with no other choice than $p = 1$ and explore whether such optimization problem can work for our purposes. Turns out it can for certain matrices.

Definition 7 (Basis Pursuit)**.** Let $\boldsymbol{x} \in \mathbb{R}^M$, $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ be known and $\boldsymbol{y} \in \mathbb{R}^N$ be known. The ℓ_1 minimization problem called Basis Pursuit is defined as

$$\min_{x\in\mathbb{R}^M} \|\boldsymbol{x}\|_1 \; subject\ to\ \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{1.32}$$

Null Space Property

Remark 4. Before we define Null Space Property, we will introduce useful notation which will be used onward. For $T \subset \hat{M}$ we denote by $T^C = \hat{M}\backslash T$ the complement of T in $\hat{M}$. For $\boldsymbol{v} \in \mathbb{R}^M$, we denote $\boldsymbol{v}_T$ the vector in $\mathbb{R}^{\#T}$ which contains the coordinates of $\boldsymbol{v}$ indexed by T or the vector in $\mathbb{R}^M$ which equals $\boldsymbol{v}$ on T and has zero components on T^C.

Definition 8 (Null Space Property)**.** Let $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ and $k \in \hat{M}$. Then $\boldsymbol{A}$ has the Null Space Property (NSP) of order k if

$$\|\boldsymbol{v}_T\|_1 < \|\boldsymbol{v}_{T^C}\|_1 : \forall \boldsymbol{v} \in ker\boldsymbol{A}\backslash\{\boldsymbol{0}\} \; and \; \forall T \subset \hat{M} \; with \; |T| \leqslant k. \tag{1.33}$$

Remark 5. The Null Space Property of a matrix says that the components of vectors of the kernel are 'not so different from each other' or not supported solely on a few components. It is straight forward to see that the inequality in (1.33) can be equivalently expressed as $\|\boldsymbol{v}\|_1 < 2\|\boldsymbol{v}_{T^C}\|_1$ or $2\|\boldsymbol{v}_T\|_1 < \|\boldsymbol{v}_T\|_1$. The following theorem shows the relation between k-sparse solutions of (1.32) and NSP.

Theorem 3. *Let $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ and $k \in \hat{M}$. Then,*

every k-sparse vector $\boldsymbol{x} \in \mathbb{R}^M$ is the unique solution of (1.32) $\Leftrightarrow$ $\boldsymbol{A}$ has the NSP of order k.

Proof.
$\Rightarrow$: Let $\boldsymbol{v} \in ker\boldsymbol{A}\backslash\{\boldsymbol{0}\}$, $T \subset \hat{M}$, $|T| \leqslant k$ arbitrary. Then from the presumption is $\boldsymbol{v}_T$ the unique solution of (1.32). Also,

$$\boldsymbol{0} = \boldsymbol{A}\boldsymbol{v} = \boldsymbol{A}(\boldsymbol{v}_T + \boldsymbol{v}_{T^C}) \Leftrightarrow \boldsymbol{A}(-\boldsymbol{v}_{T^C}) = \boldsymbol{A}(\boldsymbol{v}_T). \tag{1.34}$$

Since the solution is unique and $-\boldsymbol{v}_{T^C} \neq \boldsymbol{v}_T$ it must hold $\|\boldsymbol{v}_T\|_1 < \|\boldsymbol{v}_{T^C}\|_1$ which means $\boldsymbol{A}$ has NSP of order k.
$\Leftarrow$: Let $\boldsymbol{x} \in \mathbb{R}^M$ be a k-sparse vector with supp($\boldsymbol{x}$) = T. We have to show that this vector is the unique solution of (1.32). That means, that if $\boldsymbol{z} \in \mathbb{R}^M$ is also a solution of (1.32) then $\|\boldsymbol{x}\|_1 < \|\boldsymbol{z}\|_1$ for every such $\boldsymbol{z}$. Using the fact that both $\boldsymbol{x}, \boldsymbol{z}$ are solutions $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{A}\boldsymbol{z}$, we get $(\boldsymbol{x} - \boldsymbol{z}) \in ker\boldsymbol{A}\backslash\{\boldsymbol{0}\}$. The implication then concludes from the inequality

$$\|\boldsymbol{x}\|_1 \leqslant \|\boldsymbol{x} - \boldsymbol{z}_T\|_1 + \|\boldsymbol{z}_T\|_1 = \|(\boldsymbol{x} - \boldsymbol{z})_T\|_1 + \|\boldsymbol{z}_T\| < \|(\boldsymbol{x} - \boldsymbol{z})_{T^C}\|_1 + \|\boldsymbol{z}_T\|_1 = \|\boldsymbol{z}_{T^C}\|_1 + \|\boldsymbol{z}_T\|_1 = \|\boldsymbol{z}\|_1,$$

where we used (in order) the triangle inequality, the k-sparsity of $\boldsymbol{x}$, the NSP of $\boldsymbol{A}$, the k-sparsity of $\boldsymbol{x}$ and then the additivy of ℓ_1 norm. □

Remark 6. The theorem above implies solutions of problem (1.31) and (1.32) can overlap - if $\hat{\boldsymbol{x}}$ is also a solution of (1.31) and $\boldsymbol{x}$ is a k-sparse solution of (1.32) with $\boldsymbol{A}$ with NSP of order k then $\|\hat{\boldsymbol{x}}\|_0 \leqslant \|\boldsymbol{x}\|_0 \leqslant k$ and Theorem 2 says $\hat{\boldsymbol{x}}$ is a solution of (1.32) and $\hat{\boldsymbol{x}} = \boldsymbol{x}$. In other words, there is a class of matrices for which the problem of ℓ_0 minimization can be solved in polynomial time and that is done using the Basis Pursuit problem since the solutions coincide.

Remark 7. It is easy to see from the definition of NSP that $\hat{\boldsymbol{A}} = \boldsymbol{M}\boldsymbol{A}$ where $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ has NSP of order k and $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ is full rank then $\hat{\boldsymbol{A}}$ also has NSP of order k.

Restricted Isometry Property

The Null Space Property is rather impractical because finding matrices which satisfy the condition is difficult. Therefore we define a stronger property of $\boldsymbol{A}$ which implies NSP.

Definition 9 (Restricted Isometry Property)**.** Let $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ and $k \in \hat{M}$. The restricted isometry constant $\delta_k = \delta_k(\boldsymbol{A})$ of $\boldsymbol{A}$ of order k is the smallest $\delta \geqslant 0$ such that

$$(1 - \delta)\|\boldsymbol{x}\|_2^2 \leqslant \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \leqslant (1 + \delta)\|\boldsymbol{x}\|_2^2 \; \forall \boldsymbol{x} \in \mathbb{R}_k^M. \tag{1.35}$$

We say $\boldsymbol{A}$ satisfies the Restricted Isometry Property (RIP) of order k with the constant δ_k if $\delta_k < 1$.

Remark 8. The condition (1.35) means that $\boldsymbol{A}$ is almost isometrical on the set of k-sparse vectors. Trivial observation $\delta_1(\boldsymbol{A}) \leqslant \delta_2(\boldsymbol{A}) \leqslant \cdots \leqslant \delta_k(\boldsymbol{A})$. The following theorem says that RIP implies NSP.

Theorem 4 (RIP ⇒ NSP)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ and $k \in \mathbb{N}$ such that $k \leqslant M/2$. Then*

$$\delta_{2k}(\boldsymbol{A}) < 1/3 \Rightarrow \boldsymbol{A} \textit{ has NSP of order } k.$$

Proof. Let $\boldsymbol{v} \in ker\boldsymbol{A}$ and $T \subset \hat{M}$ with $|T| \leqslant k$. We will prove the inequality

$$\|\boldsymbol{v}_T\|_2 \leqslant \frac{\delta_{2k}}{1-\delta_k} \cdot \frac{\|\boldsymbol{v}\|_1}{\sqrt{k}}, \tag{1.36}$$

because then under the assumption $\delta_k \leqslant \delta_{2k} < 1/3$, we get $\|\boldsymbol{v}_T\|_1 \leqslant \sqrt{k}\|\boldsymbol{v}_T\|_2 < \|\boldsymbol{v}\|_1/2$ where the Hölder's inequality gives the first inenquality and (1.36) gives the sharp inequality which combined with the note in Remark 6 gives NSP of order k. First, we will prove a small useful statement:

$$\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^M_k \textit{ such that } supp(\boldsymbol{x}) \bigcap supp(\boldsymbol{z}) = \emptyset \textit{ and } \boldsymbol{A} \textit{ has } NSP \textit{ of order } 2k \Rightarrow |\langle \boldsymbol{Ax}, \boldsymbol{Az}\rangle| \leqslant \delta_{2k}\|\boldsymbol{x}\|_2\|\boldsymbol{z}\|_2. \tag{1.37}$$

Proof of the statement. It is easy to consider the validity of the following implication

$$\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^M_k, \|\boldsymbol{x}\|_2 = \|\boldsymbol{z}\|_2 = 1 \text{ such that } supp(\boldsymbol{x}) \bigcap supp(\boldsymbol{z}) = \emptyset \Rightarrow \boldsymbol{x} \pm \boldsymbol{z} \in \mathbb{R}^M_{2k} \text{ and } \|\boldsymbol{x} \pm \boldsymbol{z}\|_2^2 = 2.$$

Taking the RIP of $\boldsymbol{A}$ for $\boldsymbol{x} \pm \boldsymbol{z}$

$$2(1-\delta_{2k}) \leqslant \|\boldsymbol{A}(\boldsymbol{x} \pm \boldsymbol{z})\|_2^2 \leqslant 2(1+\delta_{2k}),$$

and combining it with the polarization identity gives

$$|\langle \boldsymbol{Ax}, \boldsymbol{Az}\rangle| = \frac{1}{4}\left|\|\boldsymbol{A}(\boldsymbol{x}+\boldsymbol{z})\|_2^2 - \|\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{z})\|_2^2\right| \leqslant \frac{1}{4}\left|2(1+\delta_{2k}) - 2(1-\delta_{2k})\right| \leqslant \delta_{2k}.$$

Finally, we plug in $\tilde{x} = x/\|x\|_2$ and $\tilde{z} = z/\|z\|_2$ and get the statement $|\langle \boldsymbol{Ax}, \boldsymbol{Az}\rangle| \leqslant \delta_{2k}\|\boldsymbol{x}\|_2\|\boldsymbol{z}\|_2$.

Let $\boldsymbol{v} \in ker\boldsymbol{A}$ and let us consider a nonincreasing rearrangement $\boldsymbol{v}^*$ of $\boldsymbol{v}$. Then we slice the components of $\boldsymbol{v}^*$ into sets of size k where the last set can be smaller:

$$T_0 = \{1, \ldots, k\}, T_1 = \{k+1, \ldots, 2k\}, T_2 = \{2k+1, \ldots, 3k\}, \text{etc.}$$

Then

$$\boldsymbol{Av}_{T_0} = \boldsymbol{A}(-\boldsymbol{v}_{T_1} - \boldsymbol{v}_{T_2} - \ldots). \tag{1.38}$$

We construct an estimation

$$\|\boldsymbol{v}_{T_0}\|_2^2 \leqslant \frac{\|\boldsymbol{v}_{T_0}\|_2^2}{1-\delta_k} = \frac{1}{1-\delta_k}\langle \boldsymbol{Av}_{T_0}, \boldsymbol{A}(-\boldsymbol{v}_{T_1}) + \boldsymbol{A}(-\boldsymbol{v}_{T_2}) + \ldots\rangle = \frac{1}{1-\delta_k}\sum_{j\geqslant 1}\langle \boldsymbol{Av}_{T_0}, \boldsymbol{A}(-\boldsymbol{v}_{T_j})\rangle \leqslant$$

$$\leqslant \frac{1}{1-\delta_k}\sum_{j\geqslant 1}\langle \boldsymbol{Av}_{T_0}, \boldsymbol{A}(-\boldsymbol{v}_{T_j})\rangle \leqslant \frac{1}{1-\delta_k}\delta_{2k}\sum_{j\geqslant 1}\|\boldsymbol{v}_{T_0}\|_2\|\boldsymbol{v}_{T_j}\|_2,$$

where we applied the definition of ℓ_2 norm through scalar product together with (1.38) in the first equality and the proved statement (1.37) in the last inequality. Dividing the inequality by $\|\boldsymbol{v}_{T_0}\|_2 \neq 0$ finally gives

$$\|\boldsymbol{v}_{T_0}\|_2 \leqslant \frac{\delta_{2k}}{1-\delta_k}\sum_{j\geqslant 1}\|\boldsymbol{v}_{T_j}\|_2. \tag{1.39}$$

The proof is finished through the following chain of inequalities

$$\sum_{j\geqslant 1}\|\boldsymbol{v}_{T_j}\|_2 = \sum_{j\geqslant 1}\Big(\sum_{l\in T_j}|\boldsymbol{v}_l|^2\Big)^{1/2} \leqslant \sum_{j\geqslant 1}\Big(k\max_{l\in T_j}|\boldsymbol{v}_l|^2\Big)^{1/2} = \sum_{j\geqslant 1}\sqrt{k}\max_{l\in T_j}|\boldsymbol{v}_l| \leqslant$$

$$\leqslant \sum_{j\geqslant 1}\sqrt{k}\min_{l\in T_{j-1}}|\boldsymbol{v}_l| \leqslant \sum_{j\geqslant 1}\sqrt{k}\Big(\sum_{l\in T_{j-1}}\frac{1}{k}|\boldsymbol{v}_l|\Big) = \sum_{j\geqslant 1}\frac{\|\boldsymbol{v}_{T_{j-1}}\|_1}{\sqrt{k}} = \frac{\|\boldsymbol{v}\|_1}{\sqrt{k}}.$$

Plugging the above result into (1.39) gives the inequality (1.36) since $T = T_0$. □

Corollary 1. Let $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ and $k \in \mathbb{N}$ such that $k \leqslant M/2$. Then,

$$\delta_{2k} < 1/3 \Rightarrow \text{every } k\text{-sparse vector } \boldsymbol{x} \text{ is the unique solution of (1.32).}$$

Proof. Combining the Theorem 2 and 3 immediately gives the statement. □

Remark 9. In a very nonrigorous reductionist manner, we can symbolically note

$$\text{RIP} \Rightarrow \text{NSP} \Rightarrow \ell_1 \text{ solution} \Rightarrow \ell_0 \text{ solution.}$$

Stability and Robustness

So far, we assumed $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ but that is not the case for a real setting. The input will always be influenced by errors $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}$. We will also want to recover vectors or their approximations which are not exactly sparse.

Definition 10 (Modified Basis Pursuit)**.** Let $\boldsymbol{x} \in \mathbb{R}^M$, $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ be known and $\boldsymbol{y} \in \mathbb{R}^N$ be known. Let $\eta \geqslant 0$. Then we define

$$\min_{\boldsymbol{x}\in\mathbb{R}^M}\|\boldsymbol{x}\|_1 \textit{ subject to } \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2 \leqslant \eta. \tag{1.40}$$

Theorem 5. *Let $\delta_{2k} < \sqrt{2} - 1$ and $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2 \leqslant \eta$. Then the solution $\hat{x}$ of (1.40) satisfies*

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \leqslant \frac{C\sigma_k(\boldsymbol{x})_1}{\sqrt{k}} + D\eta, \tag{1.41}$$

where $C, D > 0$ are two universal constants.

Proof. We will once again make use of the statement (1.37) we proved during the proof of Theorem 3.

1.2 Data Transformation Methods

In both traditional statistical inference (LASSO, principal component analysis, etc.) and machine learning, it is either advantages or even required to perform some kind of transformation of the data. The most common purpose is to improve the performance of the model. The given transformation can have a physical meaning or interpretability, the motivation to perform such transformation can even be initiated by the context of the underlying problem.

1.2.1 Feature Standardization

The purpose of data standardization is to remove the difference of scale between features of the data. The standardization used in this work is fairly common and has the following form

$$\boldsymbol{x}'_{\bullet i} = \frac{\boldsymbol{x}_{\bullet i} - \bar{\boldsymbol{x}}_{\bullet i}}{\sigma_i}, \tag{1.42}$$

where $\bar{\boldsymbol{x}}_{\bullet i}$ is the mean of the column $\boldsymbol{x}_{\bullet i}$ of the matrix of regressors and σ_i is the standard deviation of said column. A special case of standardization is mean-centering which we get when σ_i is set to one for all columns. The idea of standardization comes from the assumption that the data was sampled from standard normal distribution with zero mean and unit variance.

1.2.2 Feature Normalization

Feature normalization performs scaled of feature to interval (0,1). The transformation is given by

$$\boldsymbol{x}''_{\bullet i} = \frac{\boldsymbol{x}_{\bullet i} - \min_j(\boldsymbol{x}_{\bullet j})}{\max_j(\boldsymbol{x}_{\bullet j}) - \min_j(\boldsymbol{x}_{\bullet j})}, \tag{1.43}$$

, where we define minimum of jth feature as $\min_j(\boldsymbol{x}_{\bullet j})$ and maximum of jth as $\max_j(\boldsymbol{x}_{\bullet j})$. Feature normalization can improve the performance of the model when the features are on different scaled by orders of magnitude. This usually happens when dealing with physical problems where variables have different units and therefore the method used is more sensitive to some features than it should be.

1.3 Model Validation Methods

Chapter 2

Feature Engineering

2.1 Density Functional Theory Data

2.2 Material Descriptors

Chapter 3

Classification Problem of Binary Compounds Experiment

something

Chapter 4

Transparent Conductiong Oxides Experiment

something

Chapter 5

Materials Project Experiment

something

Conclusion

something