Optimization and Algorithms in Sparse Regression

Optimization and Algorithms in Sparse Regression

Screening Rules, Coordinate Descent, and Normalization

Johan Larsson

LUND
UNIVERSITY

Thesis for the degree of Doctor of Philosophy

Thesis Advisors
Jonas Wallin and Małgorzata Bogdan

Faculty Opponent
Professor Mário A. T. Figueiredo (Instituto Superior Técnico, Lisbon, Portugal)

To be presented, with the permission of the Lund University School of Economics and Business Administration of Lund University, for public criticism in the Clark Kent lecture hall (Kentsalen) at the Department of Statistics on Sunday, the 34th of December 2024 at 24:00.

Organization
LUND UNIVERSITY
Department of Statistics
Box 7080
SE–220 07 Lund
Sweden

Document name
DOCTORAL DISSERTATION

Date of disputation
2024-05-24

Sponsoring organization

Author(s)
Johan Larsson

Title and subtitle
Optimization and Algorithms in Sparse Regression: Screening Rules, Coordinate Descent, and Normalization

Abstract
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Key words
power, victory, awesomeness

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
978-91-8104-076-0 (print)
978-91-8104-077-7 (pdf)

Recipient's notes

Number of pages
72

Price

Security classification

Optimization and Algorithms in Sparse Regression

Screening Rules, Coordinate Descent, and Normalization

Johan Larsson

LUND
UNIVERSITY

Cover illustration front: The elastic net path for a data set of diabetes patients.

Lund University School of Economics and Management
The Department of Statistics
Box 743, SE-220 07
Lund, Sweden

ISBN: 978-91-8104-076-0 (print)
ISBN: 978-91-8104-077-7 (electronic)

Printed in Sweden by Media-Tryck, Lund University, Lund 2024

There's a point when you go with what you've got.
Or you don't go.

—Joan Didion

Contents

Acknowledgements iii

Abstract v

Popular Science Summary vii

List of Publications ix

Introduction 1
1 Background . . . 1
2 Regularization . . . 5
3 Optimization . . . 13
4 Screening Rules . . . 16
5 Normalization . . . 16
6 Summary of the Papers . . . 16

Papers 27
I The Strong Screening Rule for SLOPE . . . 29
II Look-Ahead Screening Rules for the Lasso . . . 43
III The Hessian Screening Rule . . . 45
IV Benchopt: Reproducible, Efficient and Collaborative Optimization Benchmarks . . . 47
V Coordinate Descent for SLOPE . . . 49
VI Regularization and Scaling in Sparse Regression . . . 51

Acknowledgements

Abstract

Popular Science Summary

List of Publications

This thesis is based on the following publications.

I Johan Larsson, Małgorzata Bogdan, and Jonas Wallin (Dec. 6–12, 2020). "The Strong Screening Rule for SLOPE". in: *Advances in Neural Information Processing Systems 33*. 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. Vol. 33. Virtual: Curran Associates, Inc., pp. 14592–14603. ISBN: 978-1-71382-954-6

II Johan Larsson (Sept. 6, 2021). "Look-Ahead Screening Rules for the Lasso". In: *22nd European Young Statisticians Meeting - Proceedings*. 22nd European Young Statisticians Meeting. Ed. by Andreas Makridis, Fotios S. Milienos, Panagiotis Papastamoulis, Christina Parpoula, and Athanasios Rakitzis. Athens, Greece: Panteion university of social and political sciences, pp. 61–65. ISBN: 978-960-7943-23-1

III Johan Larsson and Jonas Wallin (Nov. 28–Dec. 9, 2022). "The Hessian Screening Rule". In: *Advances in Neural Information Processing Systems 35*. 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Ed. by Sanmi Koyejo, Sidahmed Mohamed, Alekh Agarwal, Danielle Belgrave, Kyunghyun Cho, and Alice Oh. Vol. 35. New Orleans, USA: Curran Associates, Inc., pp. 15823–15835. ISBN: 978-1-71387-108-8

IV Thomas Moreau, Mathurin Massias, Alexandre Gramfort, Pierre Ablin, Pierre-Antoine Bannier, Benjamin Charlier, Mathieu Dagréou, Tom Dupré la Tour, Ghislain Durif, Cassio F. Dantas, Quentin Klopfenstein, Johan Larsson, En Lai, Tanguy Lefort, Benoit Malézieux, Badr Moufad, Binh T. Nguyen, Alain Rakotomamonjy, Zaccharie Ramzi, Joseph Salmon, and Samuel Vaiter (Nov. 28–Dec. 9, 2022). "Benchopt: Reproducible, Efficient and Collaborative Optimization Benchmarks". In: *Advances in Neural Information Processing Systems 35*.

36th Conference on Neural Information Processing Systems (NeurIPS 2022). Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. New Orleans, USA: Curran Associates, Inc., pp. 25404–25421. ISBN: 978-1-71387-108-8

v Johan Larsson, Quentin Klopfenstein, Mathurin Massias, and Jonas Wallin (Apr. 25–27, 2023). "Coordinate Descent for SLOPE". in: *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*. AISTATS 2023. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. Valencia, Spain: PMLR, pp. 4802–4821

vi PAPER ON THE WAY

Introduction

Everything should be made as simple as possible, but not simpler.
—Albert Einstein

1 Background

With modern advances in science and technology, statistical models and the data on which they are fit are becoming increasingly complex. Data sets are expanding in size, often both in terms of the number of variables (features) as well as the number of observations. In some fields, this growth in complexity has been paralleled with more effective methods with which to collect observations, as in, for instance, crowd science and social media data. But in other areas, collecting data still amounts to a costly endeavor. In bioinformatics, for example, ethical concerns and rising requirements on the quality of data have only served to *raise* the costs of data collection. And as a result, the data collected in these fields is becoming *wider*: the ratio between the number of variables (features) and the number of observations is increasing (Table 1).

Table 1: Tall and wide data. Each row is an observation, for instance the measurement on a person in a study, and each column (feature) represents all the measurements on a variable for all the observations.

(a) Tall data

x_1	x_2	x_3
0	0.32	1
1	1	-1
⋮	⋮	⋮

(b) Wide data

x_1	x_2	x_3	$\cdots$
0	0.32	1	$\cdots$
1	1	-1	$\cdots$

The growth in the number of observations is a luxury problem, since it, at least as far as the model is concerned, provides only benefit[1]. But an expansion in the number of features (wider data) is a more delicate issue. The problem is that if all the features that we have collected are important, but to varying degree, then we are out of luck as far as understanding our data goes. Instead we have to more or less hope that there is a *sparse* representation (Figure 1) of our data that, with some acceptable loss of information, allows us to understand the problem that we are studying.

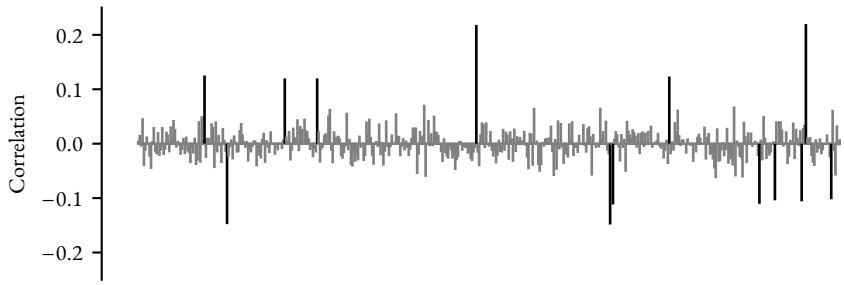


Figure 1: A relatively sparse signal. The plot shows standard Pearson correlations between the response vector y and each feature in the madelon data set (Guyon, Gunn, Ben-Hur, and Dror 2004). Correlations above 0.1 have been colored in black, the rest in gray.

We call this hope the *sparsity assumption*. And it can be motivated through the *bet-on-sparsity principle*: assume that the underlying model is sparse and use a sparse method to model it. If the assumption is correct, then our method has a chance of doing well. But if the assumption is incorrect, then our method will not work—but no other method would (Hastie, Robert Tibshirani, and Friedman 2009).

The success of neural networks and other complex models that model high-dimensional data well yet do not enforce sparsity does raise questions as to the validity of this principle. But in our setting, which, loosely speaking, is *explainable* methods for regression, it still bears relevance.

Technically speaking, we are interested in data sets that are made up of a $n \times p$

[1] The downsides are generally only related to computational issues, such as storing and processing this data.

matrix of features $\boldsymbol{X}$ and a response vector of length n, $\boldsymbol{y}$:

$$\boldsymbol{X} = \begin{bmatrix} 1.5 & 0.3 & \cdots & x_{1,p} \\ -0.9 & 0.1 & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} 0.2 \\ -0.9 \\ \vdots \\ y_n \end{bmatrix},$$

where we have inserted some arbitrary values for the sake of illustration. The data presented in Table 1 corresponds to $\boldsymbol{X}$ here.

In the simplest case, we assume that $\boldsymbol{y}$ is a linear combination of the features in $\boldsymbol{X}$ plus some noise, for instance measurement noise, which we write mathematically as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \beta_0 + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}$ is a vector of coefficients and β_0 the *intercept*. In this representation of the data, the coefficients $\boldsymbol{\beta}$ are the parameters that we are interested in estimating and represent the effect each feature has on the response vector $\boldsymbol{y}$.

Assuming this model is correct, a natural choice of model to fit this data with is linear regression, which is in fact exactly the model above provided that we, in addition, also assume that the noise $\boldsymbol{\epsilon}$ is normally distributed[2] with mean zero and constant variance.

Since in the presence of noise there generally exists no β that will fit the data perfectly, we must accept that the model is only an approximation. The natural follow-up question is then: what *is* a good approximation? To answer this question, we need to define some measure of error. The most common measure, at least as far as linear regression models go, is by far the sum of squared errors between the predicted response vector

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

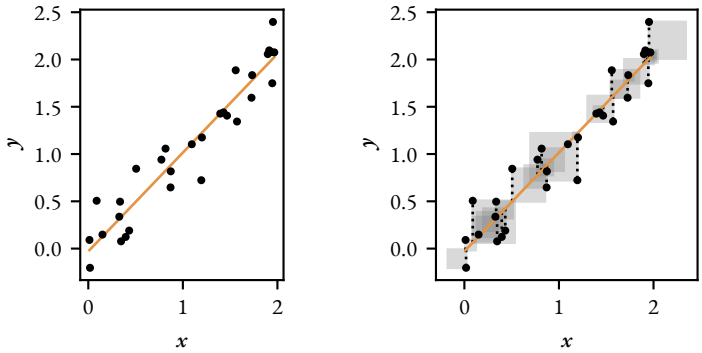
and the true response vector $\boldsymbol{y}$, that is

$$\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

The smaller this measure, the better the fit. Which means that we can pose the problem of finding the best model as the following optimization problem:

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2. \tag{1}$$

[2] Tecnically, this not in fact a core assumption of linear regression, but it is necessary for certain aspects of it.



(a) The slope of the orange line is β. The point where the line intersects the y-axis is the intercept β_0.

(b) The dotted lines are the residuals and the grey squares are the squared errors.

Figure 2: Simple ordinary least-squares linear regression for a one-feature problem.

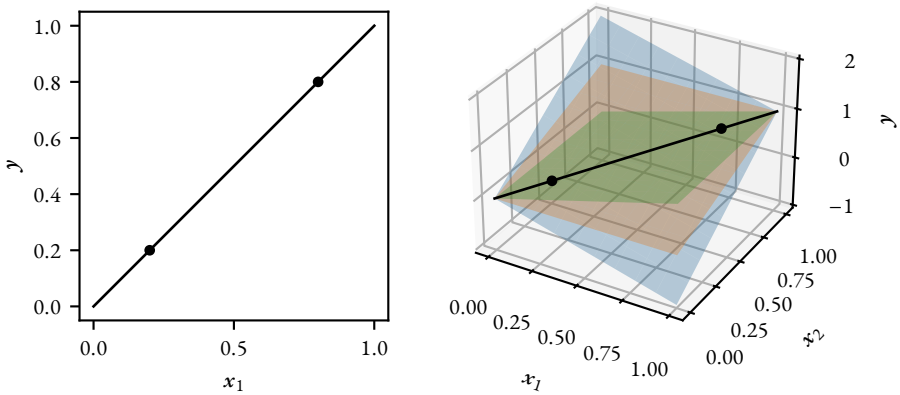
The factor of $1/2$ is included for convenience, for reasons that will become clear later on. This choice leads to the ordinary least-squares (OLS) regression model, which, for a simple case of a single feature, we have illustrated in Figure 2. There are many other ways to measure error, which all lead to different models, but in this thesis we will focus on the method of least squares.

The most common linear regression model is ordinary least-squares regression (OLS), in which we then the model that minimizes the sum of the squared residuals between the response vector $\boldsymbol{y}$ and the predictions ($\hat{\boldsymbol{y}}$) made by the model. For the simple case of a single feature, this model is illustrated in Figure 2.

If have many more observations than features ($n \gg p$), then this model might just do. But if we have many more features than observations ($p \gg n$), then we have a problem. The problem is that the model will be able to fit the data perfectly, but it will not generalize well to new data. This is because the model will be able to fit the noise in the data, and not the underlying signal. This is called *overfitting*. In fact, the coefficients β will not even be unique in this case.

The problem illustrated by Figure 3 is partly one of interpretation. If there are multiple sets of β that will work as well, what do we infer about our parameters? It is also a case of the curse of dimensionality, in the sense that, as we increase the dimensions of our feature space, our observations start to effectively grow further, and further apart, and occupy less and less of the available space.

In principle, the problem is one of over-parametrization. We have too many param-



(a) With one fature, the simple ordinarly least-squares regression line fits the data perfectly.

(b) With two features, multiple fits (planes) will fit the data perfectly.

Figure 3: A linear regression problem with two observations

eters for the amount of data that we have. This is, in principle, the same problem that one faces when fitting polynomial regresssion with an increasing number of degress. Eventually the model becomes too flexible and overfits (Figure 4).

These problems are the motivation for the use of *regularization* in regression, which we will turn to next.

2 Regularization

If the problem is that our model is over-parameterized, which we in the previous section saw is invariably the case when we employ linear regression in the $p \gg n$ scenario, an intuitive solution might be to restrict or altogheter remove some of the parameters. This is the idea behind *regularization*.

On the surface, this might seem like an awkward idea, since we are in some sense discarding information. But this is exactly what people in the $n \gg p$ have already done, but at a subjective level. The world is unquestionably high-dimensional and any situation in which we limit the number of features is only an artefact of how we have chosen to measure it. And if we are in the $n \gg p$ regime, it only means that we have already decided that some features are not important.

Regularization takes another stab at this problem and relies on data, rather than subjective knowledge, in order to choose which features it is that actually matter for

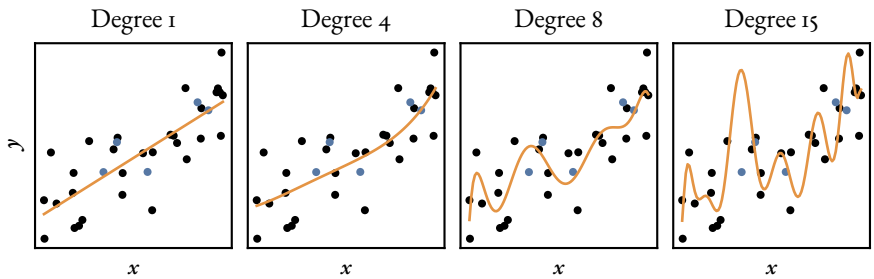


Figure 4: Polynomial regression of a one-feature regression problem. The data is generated from the simple linear model $y_i = 2x_i + \varepsilon$, where $\varepsilon \sim \text{Normal}(0, 0.5^2)$. The fits to the data become increasingly good as the degree of the polynomial increases, but when new data arrives (the blue points), we see that the model does not generalize well.

the model.

The simplest kind of regularization is called *best-subset selection*, in which we simply set a limit on how many features we allow in our model and fit all possible combinations of models until we find the one that best fits our data. Best subset selection can formally be posed as the following optimzation problem:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \\ &\text{subject to} && \|\boldsymbol{\beta}\|_0 \leq k, \end{aligned}$$

where k denotes the number of features that we allow in our model. The $\|\cdot\|_0$ norm[3] is the number of non-zero elements in a vector.

In other words, if $k = 2$ and $p = 3$, for instance, the following models would satisfy our constraints:

$$\boldsymbol{\beta} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

But the following model would not:

$$\boldsymbol{\beta} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

[3] Technically, it's not actually a real norm.

There are two problems with this method. The first is that the method involves no shrinkage, which James and Stein (1961) and Stein (1956) in pivotal work asserted was necessary for good performance.

The second is that it is computationally infeasible for large problems. The reason is that the problem is combinatorial and the number of possible models hence grows exponentially with the number of features. Interestingly, Bertsimas, King, and Mazumder (2016) has shown that the problem can actually be written as a mixed-integer optimization problem, which enables the use of modern optimization software. While this is a significant advancement, however, it is still the case that the problem is computationally infeasible for large problems (Hastie, Robert Tibshirani, and Ryan Tibshirani 2020).

2.1 The Lasso

The perhaps most obvious solution to this problem is to relax the constraint involved in the best-subset selection problem to something that makes the problem easier to solve. And in fact, the closest we can get to the best-subset selection problem without actually solving it is to use the ℓ_1 norm in place of the ℓ_0 norm. In other words, our problem now becomes

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \\ &\text{subject to} && \|\boldsymbol{\beta}\|_1 \leq t, \end{aligned} \tag{2}$$

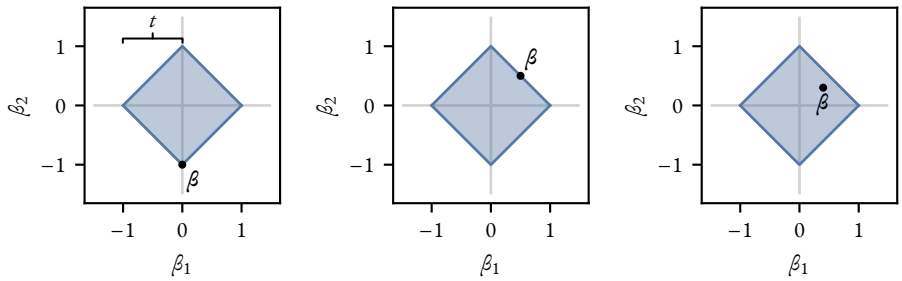
where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ and, as a consequence, we have replaced the integer-valued k with a real-valued (but positive) t. This problem is known as ℓ_1-regularized regression or, more commonly, the *lasso* (Robert Tibshirani 1996)[4].

The lasso was introduced to the statistics community by Robert Tibshirani (1996) but actually stems from much earlier research done in the field of signal processing by Santosa and Symes (1986). Donoho and Johnstone (1994, 1995) subsequently introduced the concept of the *basis pursuit* problem, which is closely related to the lasso, and developed much of the theoretical framework for the lasso.

We saw previously that the ℓ_0 constraint in best-subset selection puts a budget on the number of features allowed in the model. The ℓ_1 norm, in contrast, instead puts a budget on the *size* of the coefficients. This leads to both sparsity and shrinkage in the solution. In Figure 5, we have visualized how this constraint affects the solution of the least-squares objective.

There is now an extensive body of work on the lasso and it has spawned many different variants, such as the fused lasso (Robert Tibshirani, Saunders, Rosset, Zhu, and

[4] The lasso is sometimes written as an acronym (LASSO) for *least absolute shrinkage and selection operator*, but we will stick with the lower-case version here, which the authors themselves use in their recent work.



(a) A sparse solution, $\beta = [0, -1]^\intercal$.

(b) A dense solution, $\beta = [0.5, 0.5]^\intercal$.

(c) A solution in which the constraint is inactive, $\beta = [0.4, 0.3]^\intercal$. t is too small to have any effect on the solution.

Figure 5: The ℓ_1 norm ball in $\mathbb{R}^2$ with some possible solutions indicated by β. The ℓ_0 ball (for best-subset selection) and $k = 1$ would be lines of infinite length. along both of the axes.

Knight 2005), group lasso (Yuan and Lin 2005), adaptive lasso (Zou 2006), graphical lasso (Friedman, Hastie, and Robert Tibshirani 2008), and square-root lasso (Belloni, Chernozhukov, and Wang 2011). In this thesis, however, we will focus on the standard lasso.

Note, also, that the lasso is also not limited to regularized *linear* regression but can also be used for the entire family of generalized linear models, such as logistic, Poisson, multinomial, and multivariate regression, as well as survival models such as Cox regression. The use of the ℓ_1-norm penalty has also found its way into many other areas of statistics, signal processing, and machine learning, such as matrix factorization, clustering, and deep learning.

An interesting property of the lasso is that it is possible (and computationally feasible) to exactly solve the lasso problem for all possible values of $t \in [0, \infty)$. This is commonly referred to as the *lasso path* (Figure 6). It begins at $t = 0$, for which the constraint region is a point, which forces all of the coefficients to be exactly zero. As t increases, the constraint region grows, allowing the coefficients to enter the model. The reason for why it is possible to solve for the full path is that the solution vector β, as a function of t, is linear and continuous between the values of t for which features enter or leave the model.

A problem with the lasso, however, is that it does not deal with the case of correlated features as intuitiion (at least that of the author) would suggest. If two features are

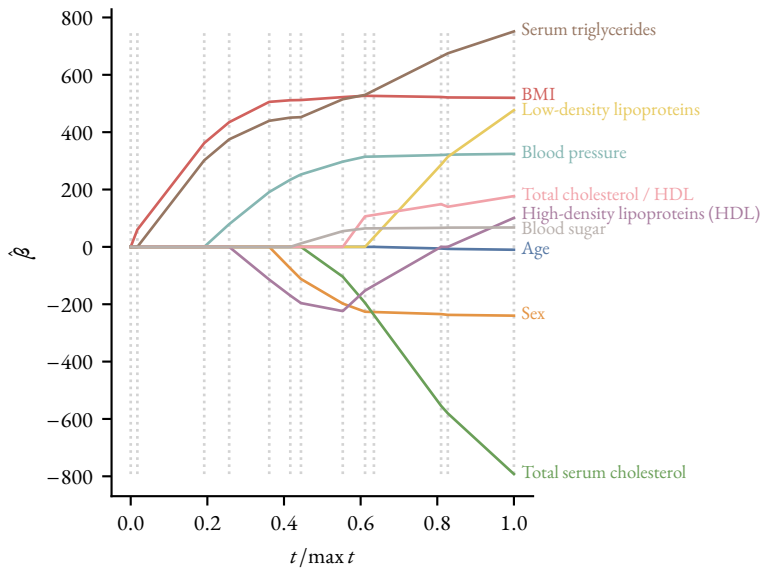


Figure 6: The lasso path for the diabetes data set (Efron, Hastie, Johnstone, and Robert Tibshirani 2004), which consists of $n = 442$ observations and $p = 10$ features. The path shows the coefficients as a function of the parameter t, which controls the size of the constraint region. The path is piecewise linear with kinks ocurring only when features enter or exit the model.

correlated highly "enough", the lasso will select one of them and set the other to zero. This is not a problem for the predicted response $\hat{\boldsymbol{y}}$, but it means that the estimated coefficients $\hat{\boldsymbol{\beta}}$ no longer provide a trustworthy estimate of variable (feature) importance. This effect is the result of the behavior of the ℓ_1 norm, which penalizes the size of the coefficients. If two features provide the same, or nearly the same, information about the response, then the optimization problem can attain a lower value by setting one of the coefficients to zero.

This is a problem that the *elastic net*—the topic of the next section—is designed to overcome.

2.2 The Elastic Net

The elastic net is a combination of the lasso and ridge regression[5], which can be written as the following optimization problem:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \\ &\text{subject to} && \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2 \leq t_1, \end{aligned}$$

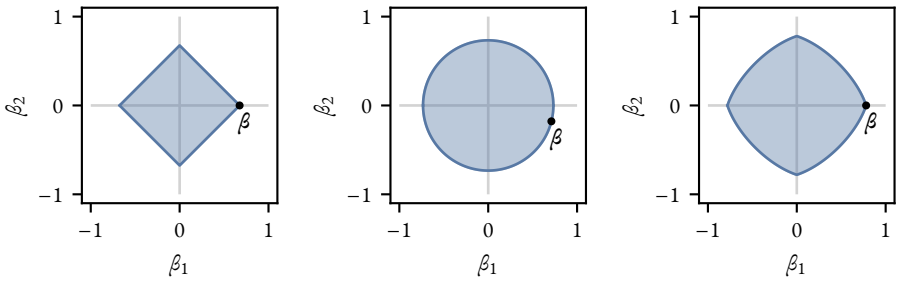
The difference compared to the lasso is that we transformed our constraint into a linear combination of the ℓ_1 and ℓ_2 norms. Setting $\alpha = 1$, the problem would once again become the lasso. With $\alpha = 0$, we would instead have ridge regression. Any value $\alpha \in (0, 1)$ yields a combination of the two (Figure 7).

The elastic net was first proposed by Zou and Hastie (2005). In addition to dealing with the problems encountered in using the lasso for highly correlated features, the elastic net also yields improved predictive performance in many situations. The latter fact is perhaps not so surprising given that it is a combination of methods that essentially assume different structure in the data. The lasso undoubtably works best when the true signal is sparse, while ridge regression handles the situation where there are weak signals better. It is then quite natural that there should for many data sets exist a α that is smaller than one but larger than zero that yields the best performance.

2.3 SLOPE

Another way of dealing with the problem of correlated features is to use *Sorted L-One Penalized Estimation* (SLOPE) (Bogdan, Berg, Sabatti, Su, and Candès 2015; Bogdan, Berg, Su, and Candès 2013; Zeng and Figueiredo 2014). SLOPE is a generalization of both the lasso and *the octagonal shrinkage and clustering algorithm for regression*

[5] Ridge regression is also known as Tikhonov regression.



(a) When $\alpha = 1$, the constraint is the lasso (ℓ_1-norm) ball. And in this case the solution is sparse.

(b) When $\alpha = 0$, the constraint is the ridge (ℓ_2-norm) ball. Here, the solution is not sparse.

(c) When $\alpha \in (0, 1)$, the constraint region is a combination of the ℓ_1 and ℓ_2 balls. In this case we have $\alpha = 1/2$. Once again, the solution is sparse.

Figure 7: The constraint regions for the elastic net for different values of α

(OSCAR) (Bondell and Reich 2008). It is represented by the following optimization objective:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \\ &\text{subject to} && \sum_{j=1}^{p} w_i |\beta|_{(i)} \leq t, \end{aligned}$$

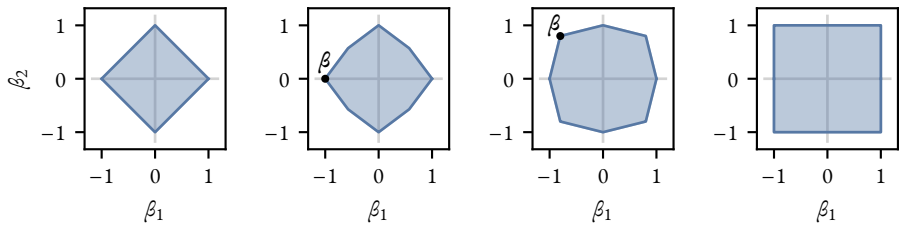
where $\boldsymbol{w}$ is a non-increasing and non-negative sequence of penalization weights and where we define the subscript operator (i) such that

$$|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(p)}.$$

The left-hand side of the constraint in SLOPE is, perhaps somewhat surprisingly, actually a norm: the *sorted ℓ_1 norm*.

The perhaps most salient feature of SLOPE is that it clusters coefficients (Figueiredo and Nowak 2014; Schneider and P. Tardivel 2022). Please refer to Figure 8 for an example of this. This makes it perfectly fit to handle the case when features are highly correlated, which he highlighted as an issue of the lasso before. If the lasso set one of the coefficients to zero, SLOPE will instead set them to exactly the same value (in absolue terms). This is a property that is not shared by the elastic net, which handles correlation (although not quite as delicately), but does not cluster coefficients.

But as we mentioned previously, SLOPE is actually a generalization of lasso and thus contains it as a special case (Figure 8a), which is attained by setting all of the elements of the penalization weight vector $\boldsymbol{w}$ to the same value. On the opposite end, setting only the first element to a non-zero value and the remaining ones to zero yields the infinity norm (Figure 8d). The latter is not of particular interest in practice.



(a) When $\boldsymbol{w} = [1, 1]^\intercal$, the constraint region is the lasso ball, $\|\boldsymbol{\beta}\|_1$.

(b) Here we have $\boldsymbol{w} = [1, 3/4]^\intercal$, but no clustering occurs.

(c) With $\boldsymbol{w} = [1, 1/4]^\intercal$, the solution is now clustered.

(d) When $\boldsymbol{w} = [1, 0]^\intercal$, the constraint is the infinity norm, $\|\boldsymbol{\beta}\|_\infty$.

Figure 8: SLOPE balls (the sorted ℓ_1 norm) for various choices of the penalization weight vector $\boldsymbol{w}$. It is the kinks at the boundaries, ocurring when $|\beta_1| = |\beta_2|$, that induce clustering. The larger the difference between adjacent values in $\boldsymbol{w}$, the stronger the clustering effect becomes.

SLOPE also has other appealing properties, such as the ability to (under certain assumptions on the design) control of false discovery rate[6] (Bogdan, Berg, Sabatti, Su, and Candès 2015) and recover of sparsity and ordering patterns in the solution (Bogdan, Dupuis, Graczyk, Kołodziejek, Skalski, P. Tardivel, and Wilczyński 2022). Another key feature is that the problem is also convex, which has implications that we will delve into later. And it puts SLOPE apart from other more complicated models such as minimax concave penalty (MCP) (Zhang 2010) and smoothly clipped absolute deviation (SCAD) (Fan and Li 2001).

We have so far only discussed the models that we are interested in fitting, but have said nothing about *how* we fit them. We will turn to this issue in the next section.

[6] In terms of the number of coefficients correctly identified as non-zero over the number of total discoveries (selected features).

3 Optimization

In the previous section we introduced the bulk of the statistical models this thesis will revolve around. They certainly have many interesting theoretical properties, which we have only touched upon briefly, but it is actually not the statistical theory of these problems that we will concern ourselves with in this thesis. Instead, we will be interested in the *numerical* aspects of these problems. That is: how do we actually solve them? And, moreover, how do we do this as efficiently as possible?

We have already introduced many optimization problems and have more or less assumed that we can solve them. This assumption is by no means wrong: methods for fitting the lasso, elastic net, and SLOPE are readily available for free in many programming languages and can be installed via a few lines of code. For instance, to fit the full lasso path to the diabetes data that we ecountered previously (see Figure 6 for the result), we only need to call the following R code.

```
library(lars)

data(diabetes)
fit <- lars(diabetes$x, diabetes$y, type = "lasso")
```

Behind the scenes, however, the method invoked through this command actually involve a complicated optimization algorithm into which considerable effort has been put in order to ensure that what you get in `fit` is reliable—and that you get it *fast*.

3.1 Direct Methods

The first optimization problem that we encountered in this text was ordinary least-squares regression, which we formally defined in Problem (1). Naively speaking, the solution to this problem is actually relatively straightforward. Letting

$$f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$$

be the objective function that we want to minimize, we simply set the gradient of it to zero:

$$\nabla f(\beta) = X^\intercal(X\beta - y) = 0 \implies$$
$$X^\intercal X\beta = X^\intercal y.$$

This system[7] is called the *normal equations*. Solving the system in β yields the ordinary least squares estimate, which, for a one-dimensional problem is equivalent to locating

[7] We have ignored the intercept β_0 here for simplicity, but it could be incorporated easily by prepending a vector of ones to X.

the "bottom" of the function $f(\beta)$ in Figure 9a. It might be tempting to want to simply invert $\boldsymbol{X}^\intercal \boldsymbol{X}$ here and premultiply by both sides to yield an explicit solution of the form

$$\beta = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}$$

but this is typically a bad idea since the inverse need not exist or be numerically stable. A better option is to use a method such as the QR decomposition and solve the resulting system through forward or backwards elimination, which is both more stable and more efficient, and all modern software use some variation of this approach.

Regardless, however, OLS can be solved directly and with accuracy at machine precision. This property is shared by ridge regression in which we can attain a solution simply by adding a diagonal matrix[8] to $\boldsymbol{X}^\intercal \boldsymbol{X}$ and solving as before. The key reason for why this is the case is that OLS is a differentiable and quadratic problem, which means that it is *convex* and hence has a global solution (Figure 9a), unlike, for instance, the problem in Figure 9b, which is non-convex (actually a third-degree polynomial) and hence has a local minimum.

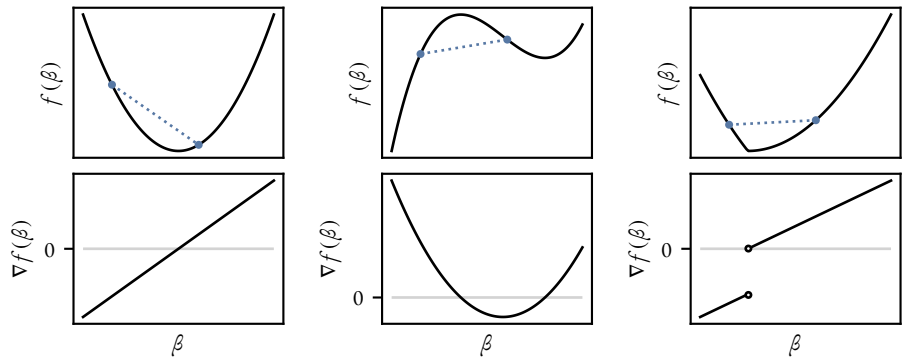
All the problems that we have covered so far: ordinary least-squares regression, the lasso, the elastic net, and SLOPE are all convex problems, which is the class of problems this thesis focuses on. Being convex, however, does not necessarily mean that the problem is easy to solve. The lasso (Problem (2)), for instance, is a convex problem, but the involvement of the inequality constraint means that we cannot solve it directly, at least not for any given t.

Somewhat remarkably, however, there actually exist methods that *can* solve the full lasso path directly, which means they can also indirectly solve the lasso for a single t. This class of methods are called *homotopy algorithms* since they solve the problem for all values they are parameterized by (in this case t). The first homotopy method for the lasso was introduced by Osborne, Presnell, and Turlach (2000) but it is the LARS algorithm (Efron, Hastie, Johnstone, and Robert Tibshirani 2004), that we already saw in action at the beginning of the section, which popularized the method.

In essence, homotopy methods for lasso are based on the idea that the lasso can be solved directly if we know the support of the solution (the identity of the non-zero coefficients). Based on this idea, we start with the empty support at $t = 0$. From this point, it is possible to say which feature(s) will become active first and then solve the problem (directly) for this support set. We can then proceed to the next support set, rinse, and repeat. We give a rough, but slightly more formalized, description of the method in **??** 1.

At the time that these methods were introduced, they offered a remarkable boost in efficiency compared to the original algorithm used by Robert Tibshirani (1996), which

[8] This procedure referes to the *unconstrained* form of ridge regression, which have not yet—but will soon—introduce.



(a) A convex and smooth function. The minimum ocurs when $\nabla f(\beta) = 0$.

(b) A non-convex but smooth function. The derivative offers no information about the global minimum (which is $-\infty$).

(c) A convex but non-smooth function. There is a global minimum, but the derivative does not exist at this point.

Figure 9: Three different kinds of functions. We show the objective value $f(\beta)$ and the gradient $\nabla f(\beta)$ for each. In each case, our objective is to find the minimum of the function. Only the first and last are convex (the complete line segment between two points on the function lies above the function) and have a global minimum.

Algorithm 1: A rough outline of the homotopy method for the lasso path. The steps in lines 3 and 4 represent the critical aspect of the algorithm, and are omitted here for brevity, but they are not particularly demanding computationally. The costs come from the number of iterations that are needed to solve the entire path.

Input: $\beta^{(0)} \leftarrow \mathbf{0}$, $t \leftarrow 0$, $\mathcal{A} \leftarrow \emptyset$, $i \leftarrow 0$

1 **repeat**
2 $\quad i \leftarrow i + 1$;
3 $\quad t \leftarrow$ next value for which the support changes;
4 $\quad \mathcal{A} \leftarrow$ support at t;
5 $\quad \beta^{(i)}_{\mathcal{A}} \leftarrow \arg\min_{\beta \in \mathbb{R}^{|\mathcal{A}|}} f(\beta)$;
6 $\quad \beta^{(i)}_{\mathcal{A}^C} \leftarrow \mathbf{0}$;
7 **until** $|\mathcal{A}| = p$;

consisted of an iterative method based on an algorithm by Lawson and Hanson (1995). This method scaled badly with p and was altogether inapplicable when $p > n$.

It is a well-known fact that the elastic net can be recast as a lasso problem, which means that the homotopy methods for the lasso can be used for the elastic net as well. There also exists homotopy methods for SLOPE (Dupuis and P. J. C. Tardivel 2023; Nomura 2020), since SLOPE shares the piecewise-linear property of the lasso path (although the SLOPE path is typically more complicated, with even more kinks).

Even if these homotopy methods provided a much-wanted upgrade compared to the original method for the high-dimensional regime, it is nevertheless this domain that they ultimately struggle to deal with. The root of this problem is that there are at least $\min(n, p)$ changes in support along the full lasso path—and in the worst case as many as $(3^p + 1)/2$ such changes (Mairal and Yu 2012). The algorithm has to solve an equivalent number of OLS problems, albeit at a complexity much reduced from that of solving the full problem, which, in the end, means that the method has found itself outperformed by iterative optimization methods (Friedman, Hastie, and Robert Tibshirani 2010), which we will introduce in the next section, starting with *proximal gradient descent*.

3.2 Proximal Gradient Descent

but it is also non-smooth, which means that the gradient does not exist everywhere. The problem in Figure 9c is actually corresponds to a one-dimensional lasso problem where the optimum is achieved at a point where the derivative does not exist.

4 Screening Rules

5 Normalization

6 Summary of the Papers

6.1 Paper 1

In this paper, we address the challenge of extracting relevant features from data sets where the number of observations, n, is significantly smaller than the number of predictors, p. We focus on the Sorted L-One Penalized Estimation (SLOPE)—a generalization of the lasso—as a promising method in this context. However, current numerical procedures for SLOPE lack the efficiency that lasso tools possess, especially when estimating a complete regularization path. A key component of lasso's efficiency is predictor screening rules, which allow predictors to be discarded before model estimation. This

paper is the first to establish such a rule for SLOPE. We develop a SLOPE screening rule by examining its subdifferential and demonstrate that this rule is a generalization of the strong rule for the lasso. Although our rule is heuristic and may occasionally discard predictors erroneously, we show that such instances are rare and can be easily safeguarded against by a simple check of the optimality conditions. Our numerical experiments reveal that the rule performs well in practice, leading to significant improvements for data in the $p \gg n$ domain, and incurs no additional computational overhead when $n > p$. This paper, therefore, presents a significant advancement in the efficiency of SLOPE, particularly in high-dimensional settings.

6.2 Paper II

In this paper, we focus on the lasso, a widely used method for inducing shrinkage and sparsity in the solution vector of regression problems, especially when the number of predictors outweighs the number of observations. Solving the lasso in such high-dimensional settings can be computationally challenging. However, this challenge can be mitigated through the use of screening rules that discard predictors before fitting the model, resulting in a reduced problem. We introduce a new screening strategy, termed look-ahead screening. This method employs safe screening rules to identify a range of penalty values for which a specific predictor cannot enter the model, thereby screening predictors along the remaining path. Our experiments demonstrate that these look-ahead screening rules outperform the active warm-start version of the Gap Safe rules, marking a significant advancement in the efficiency of solving high-dimensional lasso problems.

6.3 Paper III

In this paper, we address the challenge of predictor screening rules in l1-regularized regression problems, such as the lasso. These rules, which eliminate predictors from the design matrix before fitting a model, have significantly improved the speed of solving such problems. However, current state-of-the-art screening rules struggle with highly-correlated predictors, often becoming overly conservative. To tackle this issue, we introduce a new screening rule: the Hessian Screening Rule. This rule leverages second-order information from the model to provide more accurate screening and higher-quality warm starts. Our proposed rule outperforms all other alternatives we studied on datasets with high correlation for both l1-regularized least-squares (the lasso) and logistic regression. It also delivers the best performance overall on the real datasets we examined. This paper, therefore, presents a significant advancement in dealing with highly-correlated predictors in l1-regularized regression problems.

6.4 Paper IV

In this paper, we tackle the challenges posed by the rapid development of machine learning research, particularly in the area of numerical validation. Researchers often face a multitude of methods to compare, lack of transparency and consensus on best practices, and the tedious task of re-implementing work. This often results in partial validation, which can lead to incorrect conclusions and hinder research progress. To address these issues, we introduce Benchopt, a collaborative framework designed to automate, reproduce, and publish optimization benchmarks in machine learning across different programming languages and hardware architectures. Benchopt simplifies the benchmarking process by providing a ready-to-use tool for running, sharing, and extending experiments. We demonstrate its wide applicability through benchmarks on three standard learning tasks: ℓ_2-regularized logistic regression, Lasso, and ResNet18 training for image classification. These benchmarks reveal key practical findings that provide a more nuanced view of the state-of-the-art for these problems, emphasizing that the details matter in practical evaluation. We believe that Benchopt will encourage collaborative work in the community and improve the reproducibility of research findings.

6.5 Paper V

In this paper we delve into the Sorted L-One Penalized Estimation (SLOPE), an extension of the renowned lasso regression method. Despite the promising statistical properties of SLOPE, its adoption has been limited due to the inefficiency of existing algorithms in high-dimensional contexts. To overcome this challenge, we introduce a novel, faster algorithm that solves the SLOPE optimization problem.

Our algorithm merges the techniques of proximal gradient descent and proximal coordinate descent, significantly enhancing the efficiency of the SLOPE method. We also shed new light on the directional derivative of the SLOPE penalty and its associated SLOPE thresholding operator, and provide assurances of convergence for our proposed solver. Through comprehensive benchmarks on both simulated and real data, we demonstrate that our method outperforms a host of competing algorithms. This paper is a significant contribution as it broadens the applicability of the SLOPE method in high-dimensional settings, potentially paving the way for its wider use in the field.

6.6 Paper VI

In this paper, we explore the sensitivity of regularized methods, such as the lasso and ridge regression, to the scales of the features in the data. It's standard practice to normalize features to ensure they share the same scale. While standardization is common

for continuous data, binary data, particularly when high-dimensional and sparse, is often not scaled at all. We demonstrate that this choice can significantly impact the estimated model when the binary features are imbalanced, and that these effects also depend on the type of regularization used. Specifically, we show that the size of a feature's corresponding coefficient in the lasso is directly related to its class imbalance, and this effect depends on the normalization used. We propose potential solutions to this issue and discuss the case when data is mixed, containing both continuous and binary features. This paper, therefore, provides valuable insights into the impact of feature scaling on regularized methods and offers practical solutions for handling mixed data.

Bibliography

Belloni, Alexandre, Victor Chernozhukov, and Lie Wang (Dec. 2011). "Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming". In: *Biometrika* 98.4, pp. 791–806. ISSN: 0006-3444. DOI: 10.1093/biomet/asr043.

Bertsimas, Dimitris, Angela King, and Rahul Mazumder (Apr. 1, 2016). "Best Subset Selection via a Modern Optimization Lens". In: *The Annals of Statistics* 44.2, pp. 813–852. ISSN: 0090-5364. DOI: 10.1214/15-AOS1388.

Bogdan, Małgorzata, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès (Sept. 2015). "SLOPE – Adaptive Variable Selection via Convex Optimization". In: *The annals of applied statistics* 9.3, pp. 1103–1140. ISSN: 1932-6157. DOI: 10.1214/15-AOAS842. pmid: 26709357.

Bogdan, Małgorzata, Ewout van den Berg, Weijie Su, and Emmanuel J. Candès (Oct. 29, 2013). *Statistical Estimation and Testing via the Sorted L1 Norm*. DOI: 10.48550/arXiv.1310.1969. arXiv: 1310.1969 [math, stat]. URL: http://arxiv.org/abs/1310.1969 (visited on 04/16/2020). preprint.

Bogdan, Małgorzata, Xavier Dupuis, Piotr Graczyk, Bartosz Kołodziejek, Tomasz Skalski, Patrick Tardivel, and Maciej Wilczyński (Mar. 22, 2022). *Pattern Recovery by SLOPE*. DOI: 10.48550/arXiv.2203.12086. arXiv: 2203.12086 [math, stat]. URL: http://arxiv.org/abs/2203.12086 (visited on 06/03/2022). preprint.

Bondell, Howard D. and Brian J. Reich (Mar. 2008). "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR". In: *Biometrics* 64.1, pp. 115–123. ISSN: 0006-341X. DOI: 10.1111/j.1541-0420.2007.00843.x. JSTOR: 25502027.

Donoho, David L. and Iain M. Johnstone (Aug. 1994). "Ideal Spatial Adaptation by Wavelet Shrinkage". In: *Biometrika* 81.3, pp. 425–455. ISSN: 0006-3444. DOI: 10.2307/2337118. JSTOR: 2337118.

— (1995). "Adapting to Unknown Smoothness via Wavelet Shrinkage". In: *Journal of the American Statistical Association* 90.432, pp. 1200–1224. ISSN: 0162-1459. DOI: 10.2307/2291512. JSTOR: 2291512.

Dupuis, Xavier and Patrick J C Tardivel (Oct. 28, 2023). *The Solution Path of SLOPE*. HAL: hal-04100441v2. URL: https://hal.science/hal-04100441v2 (visited on 01/17/2024). preprint.

Efron, Bradley, Trevor Hastie, Iain M. Johnstone, and Robert Tibshirani (Apr. 2004). "Least Angle Regression". In: *Annals of Statistics* 32.2, pp. 407–499. ISSN: 0090-5364. DOI: 10.1214/009053604000000067.

Fan, Jianqing and Runze Li (Dec. 1, 2001). "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties". In: *Journal of the American Statistical Association* 96.456, pp. 1348–1360. ISSN: 0162-1459. DOI: 10/fd7bfs.

Figueiredo, Mário A. T. and Robert D. Nowak (Sept. 13, 2014). "Sparse Estimation with Strongly Correlated Variables Using Ordered Weighted L1 Regularization". DOI: 10.48550/arXiv.1409.4005. arXiv: 1409.4005 [stat].

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (July 2008). "Sparse Inverse Covariance Estimation with the Graphical Lasso". In: *Biostatistics* 9.3, pp. 432–441. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxm045.

— (Jan. 2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1–22. DOI: 10.18637/jss.v033.i01.

Guyon, Isabelle, Steve Gunn, Asa Ben-Hur, and Gideon Dror (Dec. 13–18, 2004). "Result Analysis of the NIPS 2003 Feature Selection Challenge". In: *Advances in Neural Information Processing Systems 17*. Neural Information Processing Systems 2004. Ed. by Lawrence K. Saul, Yair Weiss, and Léon Bottou. Vancouver, BC, Canada: MIT Press, pp. 545–552. ISBN: 978-0-262-19534-8.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag. ISBN: 978-0-387-84857-0.

Hastie, Trevor, Robert Tibshirani, and Ryan Tibshirani (Nov. 2020). "Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons". In: *Statistical Science* 35.4, pp. 579–592. ISSN: 0883-4237. DOI: 10.1214/19-STS733.

James, Willard and Charles Stein (1961). "Estimation with Quadratic Loss". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Vol. 4.1. Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, USA: University of California Press, pp. 361–380.

Larsson, Johan (Sept. 6, 2021). "Look-Ahead Screening Rules for the Lasso". In: *22nd European Young Statisticians Meeting - Proceedings*. 22nd European Young Statisticians Meeting. Ed. by Andreas Makridis, Fotios S. Milienos, Panagiotis Papasta-

moulis, Christina Parpoula, and Athanasios Rakitzis. Athens, Greece: Panteion university of social and political sciences, pp. 61–65. ISBN: 978-960-7943-23-1.

Larsson, Johan, Małgorzata Bogdan, and Jonas Wallin (Dec. 6–12, 2020). "The Strong Screening Rule for SLOPE". In: *Advances in Neural Information Processing Systems 33*. 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. Vol. 33. Virtual: Curran Associates, Inc., pp. 14592–14603. ISBN: 978-1-71382-954-6.

Larsson, Johan, Quentin Klopfenstein, Mathurin Massias, and Jonas Wallin (Apr. 25–27, 2023). "Coordinate Descent for SLOPE". In: *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*. AISTATS 2023. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. Valencia, Spain: PMLR, pp. 4802–4821.

Larsson, Johan and Jonas Wallin (Nov. 28–Dec. 9, 2022). "The Hessian Screening Rule". In: *Advances in Neural Information Processing Systems 35*. 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Ed. by Sanmi Koyejo, Sidahmed Mohamed, Alekh Agarwal, Danielle Belgrave, Kyunghyun Cho, and Alice Oh. Vol. 35. New Orleans, USA: Curran Associates, Inc., pp. 15823–15835. ISBN: 978-1-71387-108-8.

Lawson, Charles L. and Richard J. Hanson (1995). *Solving Least Squares Problems*. 2nd ed. Classics in Applied Mathematics. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics. 351 pp. ISBN: 978-0-89871-356-5. DOI: 10.1137/1.9781611971217.

Mairal, Julien and Bin Yu (June 2012). "Complexity Analysis of the Lasso Regularization Path". In: *Proceedings of the 29th International Conference on Machine Learning*. International Conference on Machine Learning 2012. Edinburgh, United Kingdom, pp. 1835–1842.

Moreau, Thomas, Mathurin Massias, Alexandre Gramfort, Pierre Ablin, Pierre-Antoine Bannier, Benjamin Charlier, Mathieu Dagréou, Tom Dupré la Tour, Ghislain Durif, Cassio F. Dantas, Quentin Klopfenstein, Johan Larsson, En Lai, Tanguy Lefort, Benoit Malézieux, Badr Moufad, Binh T. Nguyen, Alain Rakotomamonjy, Zaccharie Ramzi, Joseph Salmon, and Samuel Vaiter (Nov. 28–Dec. 9, 2022). "Benchopt: Reproducible, Efficient and Collaborative Optimization Benchmarks". In: *Advances in Neural Information Processing Systems 35*. 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. New Orleans, USA: Curran Associates, Inc., pp. 25404–25421. ISBN: 978-1-71387-108-8.

Nomura, Shunichi (Oct. 29, 2020). *An Exact Solution Path Algorithm for SLOPE and Quasi-Spherical OSCAR*. DOI: 10.48550/arXiv.2010.15511. arXiv: 2010.15511. URL: http://arxiv.org/abs/2010.15511 (visited on 05/27/2021). preprint.

Osborne, Michael R., Brett Presnell, and Berwin A. Turlach (July 1, 2000). "A New Approach to Variable Selection in Least Squares Problems". In: *IMA Journal of Numerical Analysis* 20.3, pp. 389–403. ISSN: 1464-3642. DOI: 10.1093/imanum/20.3.389.

Santosa, Fadil and William W. Symes (Oct. 1986). "Linear Inversion of Band-Limited Reflection Seismograms". In: *SIAM Journal on Scientific and Statistical Computing* 7.4, pp. 1307–1330. ISSN: 0196-5204. DOI: 10.1137/0907087.

Schneider, Ulrike and Patrick Tardivel (Oct. 1, 2022). "The Geometry of Uniqueness, Sparsity and Clustering in Penalized Estimation". In: *The Journal of Machine Learning Research* 23.331, pp. 1–36. ISSN: 1532-4435.

Stein, Charles (1956). "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Ed. by Jerzy Neyman. Vol. 3.1. Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, USA: University of California Press.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society: Series B* 58.1, pp. 267–288. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1996.tb02080.x. JSTOR: 2346178.

Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight (Feb. 2005). "Sparsity and Smoothness via the Fused Lasso". In: *Journal of the Royal Statistical Society: Series B* 67.1, pp. 91–108. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2005.00490.x.

Yuan, Ming and Yi Lin (Dec. 21, 2005). "Model Selection and Estimation in Regression with Grouped Variables". In: *Journal of the Royal Statistical Society: Series B* 68.1, pp. 49–67. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2005.00532.x.

Zeng, Xiangrong and Mário A. T. Figueiredo (Oct. 2014). "Decreasing Weighted Sorted L1 Regularization". In: *IEEE Signal Processing Letters* 21.10, pp. 1240–1244. ISSN: 1070-9908, 1558-2361. DOI: 10.1109/LSP.2014.2331977.

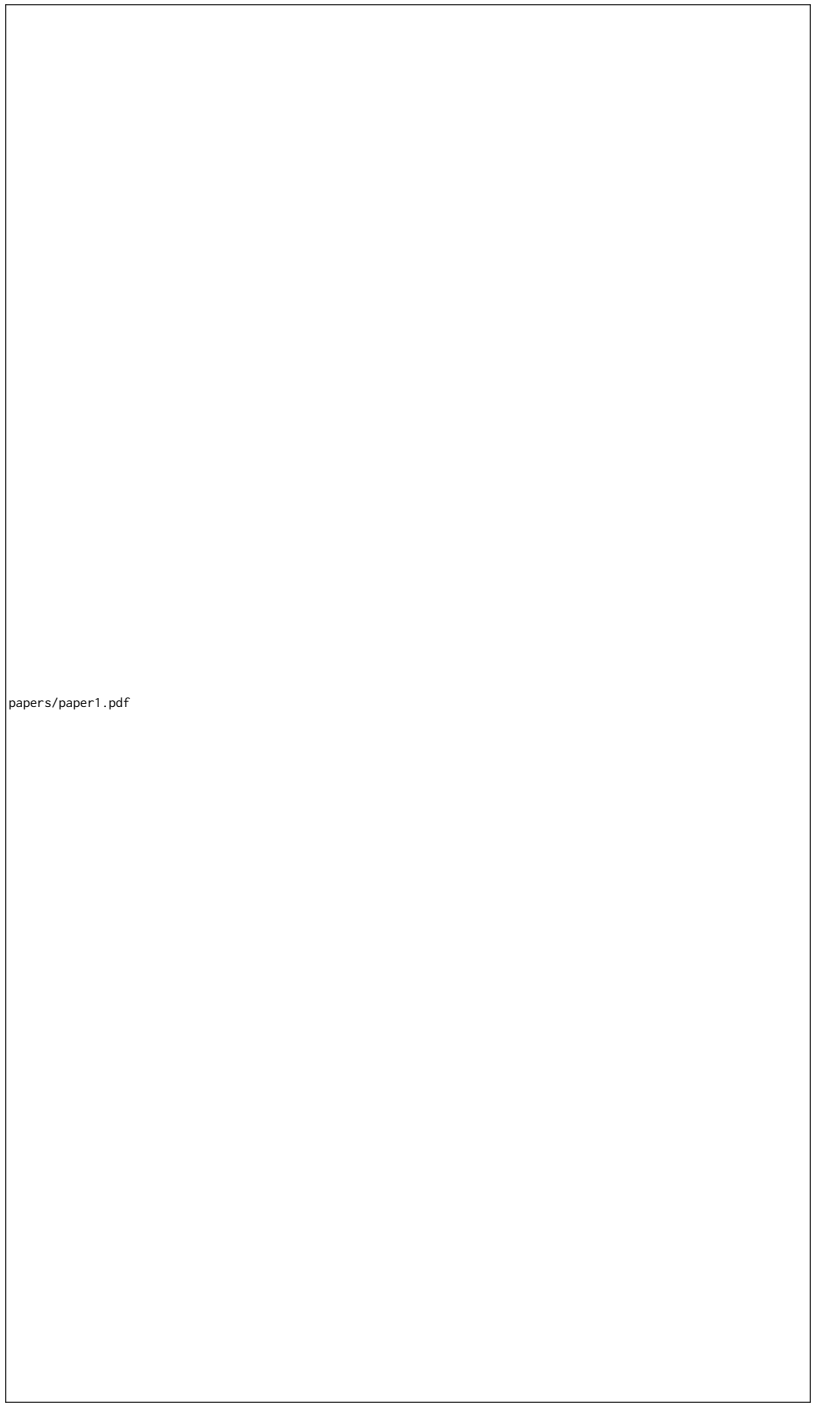
Zhang, Cun-Hui (Apr. 2010). "Nearly Unbiased Variable Selection under Minimax Concave Penalty". In: *The Annals of Statistics* 38.2, pp. 894–942. ISSN: 0090-5364. DOI: 10/bp22zz.

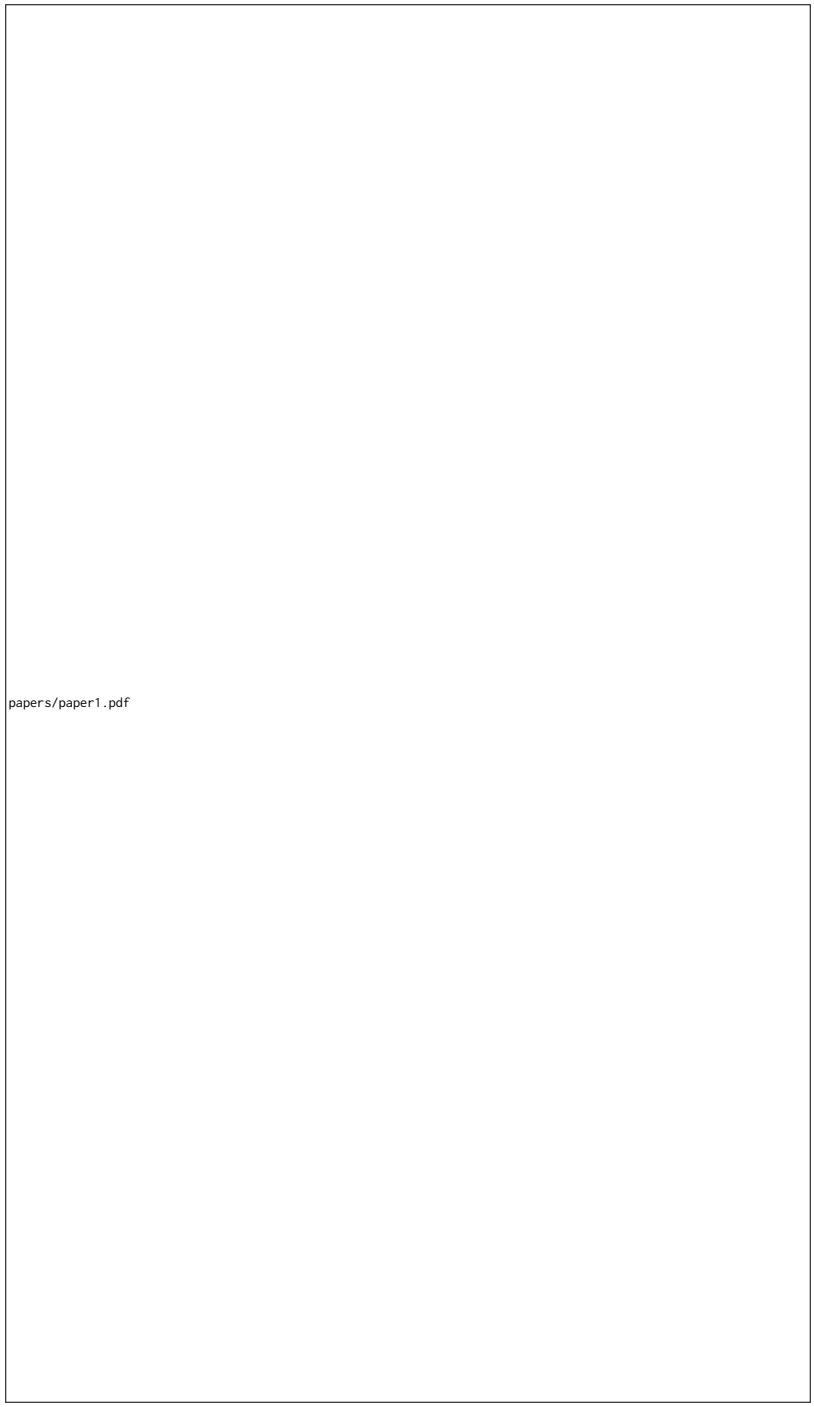
Zou, Hui (Dec. 1, 2006). "The Adaptive Lasso and Its Oracle Properties". In: *Journal of the American Statistical Association* 101.476, pp. 1418–1429. ISSN: 0162-1459. DOI: 10.1198/016214506000000735.

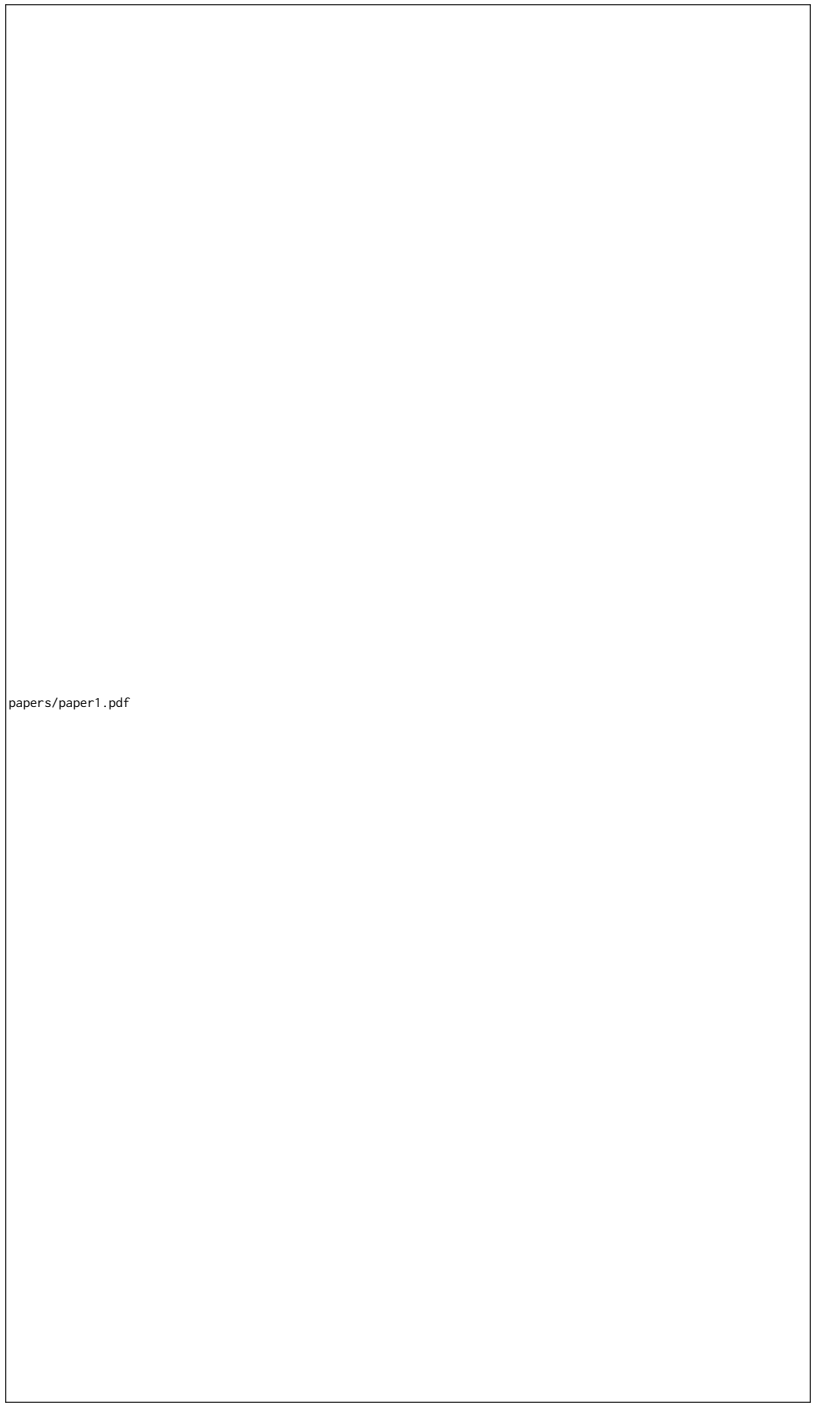
Zou, Hui and Trevor Hastie (2005). "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.2, pp. 301–320. ISSN: 1369-7412.

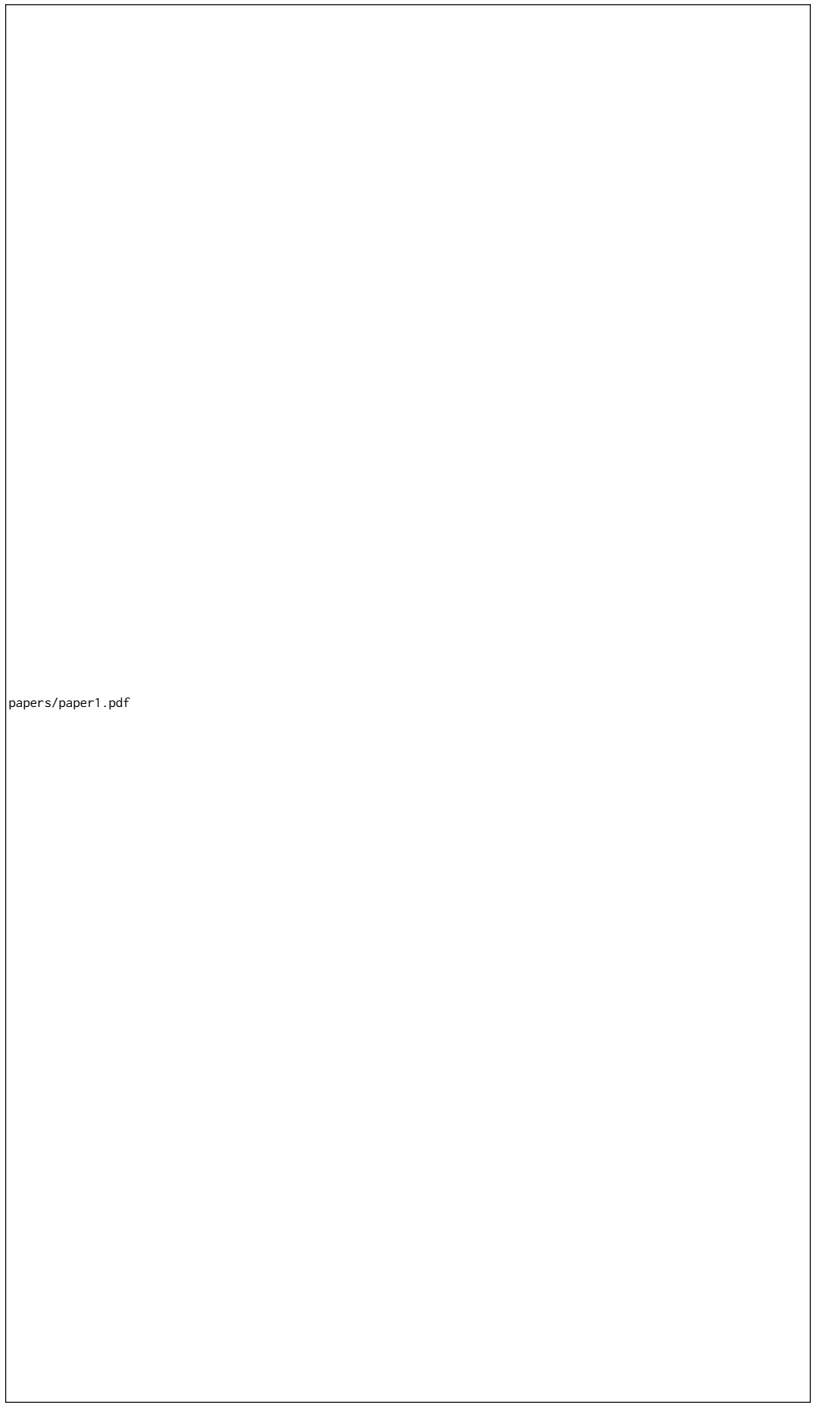
Papers

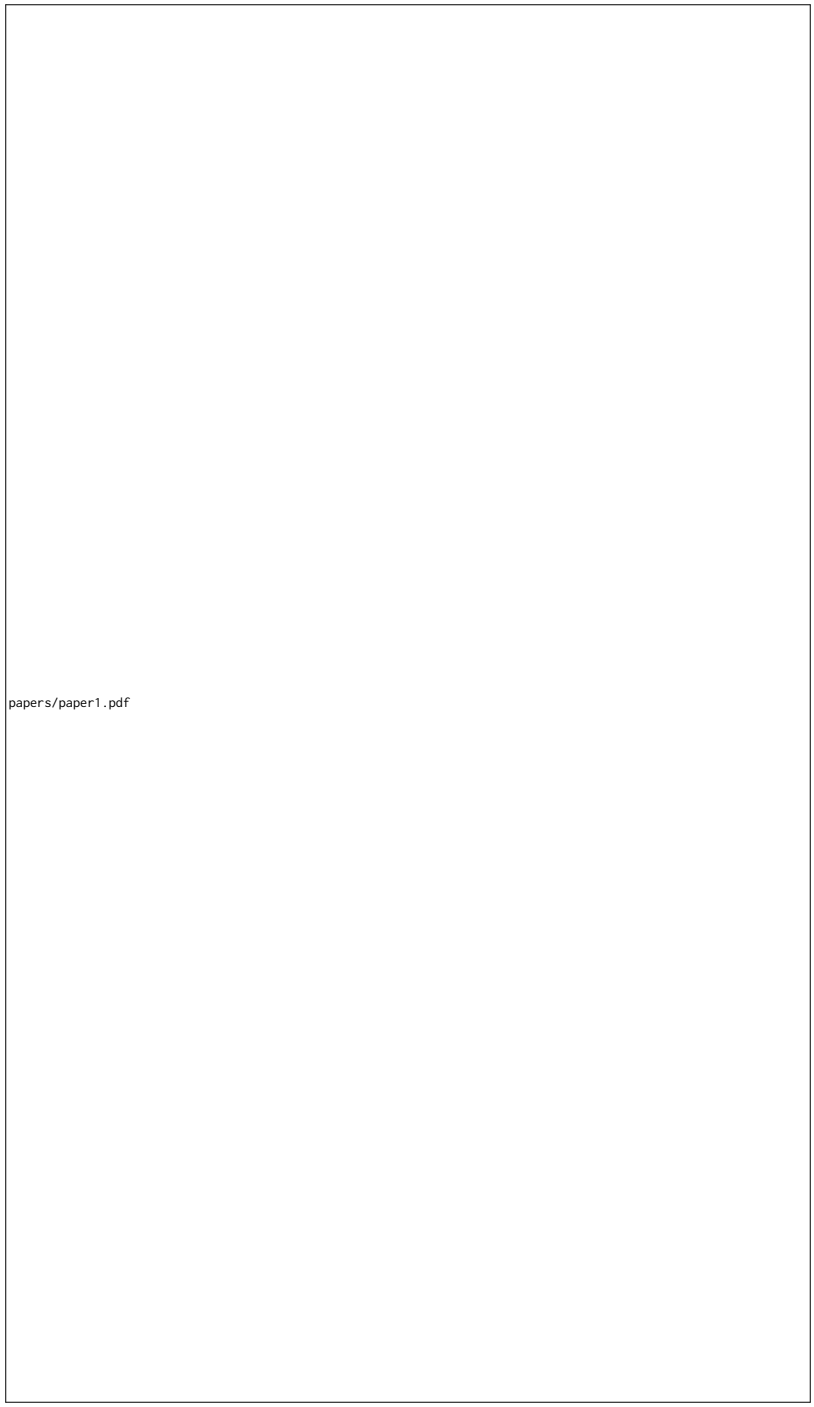
I

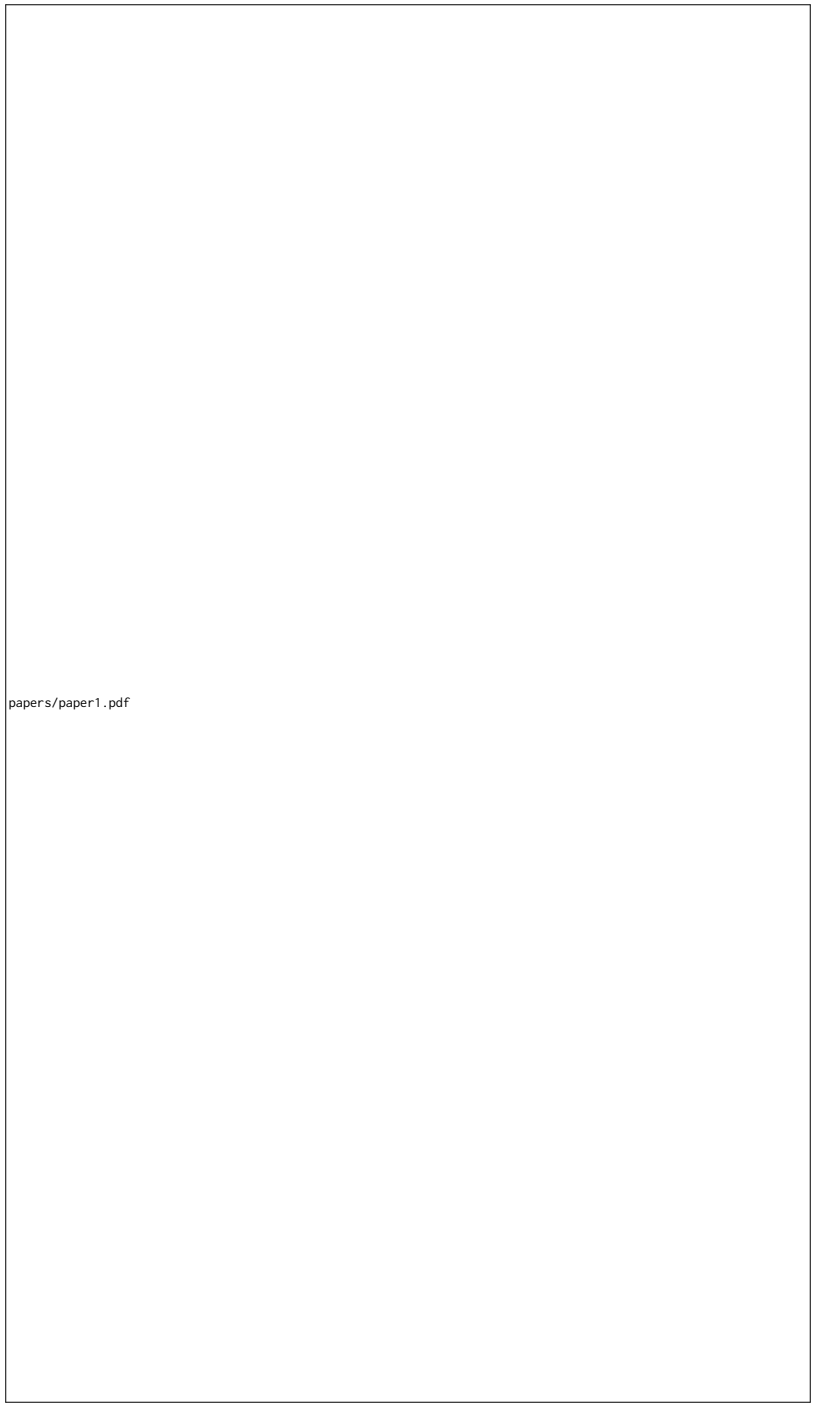

papers/paper1.pdf


papers/paper1.pdf

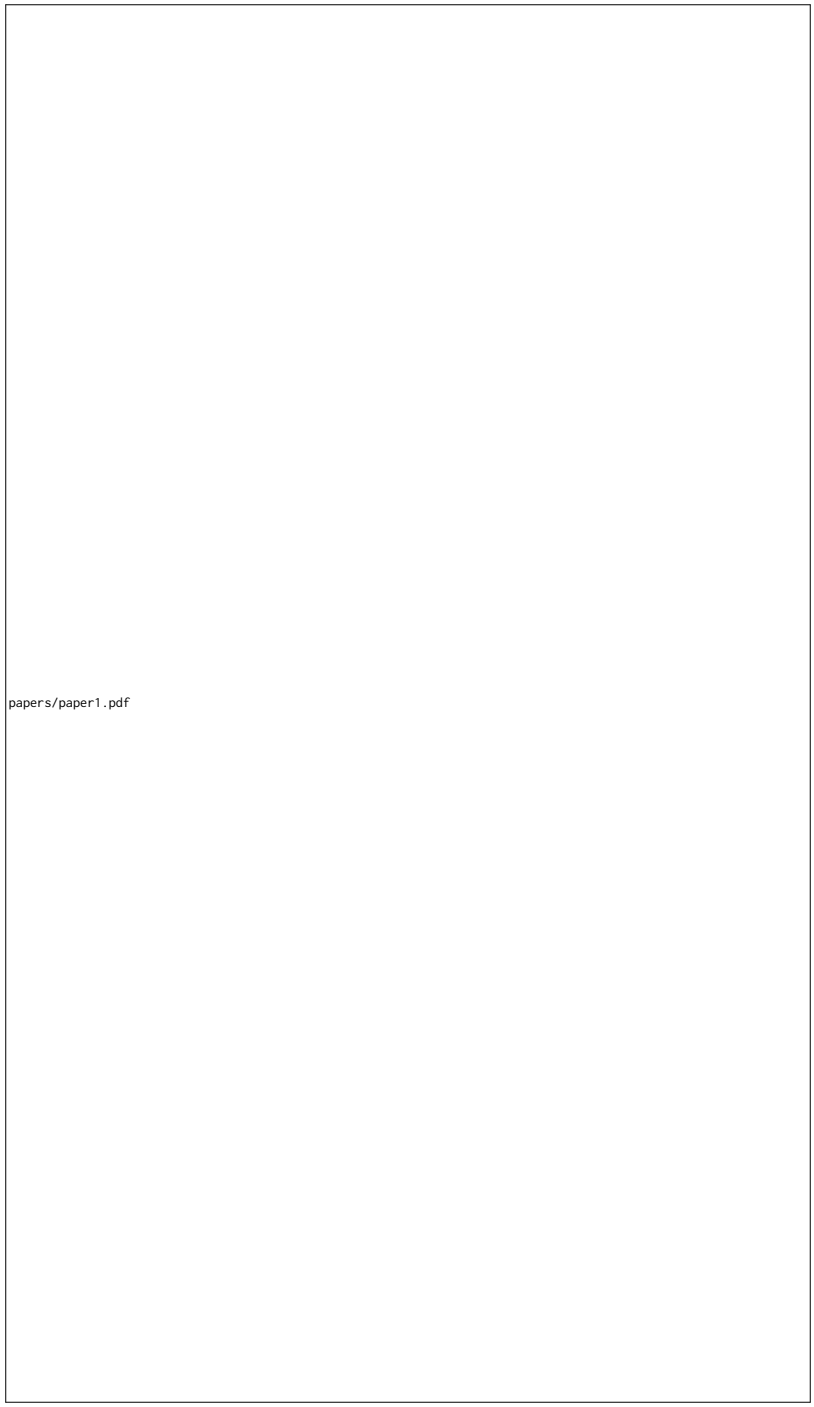

papers/paper1.pdf

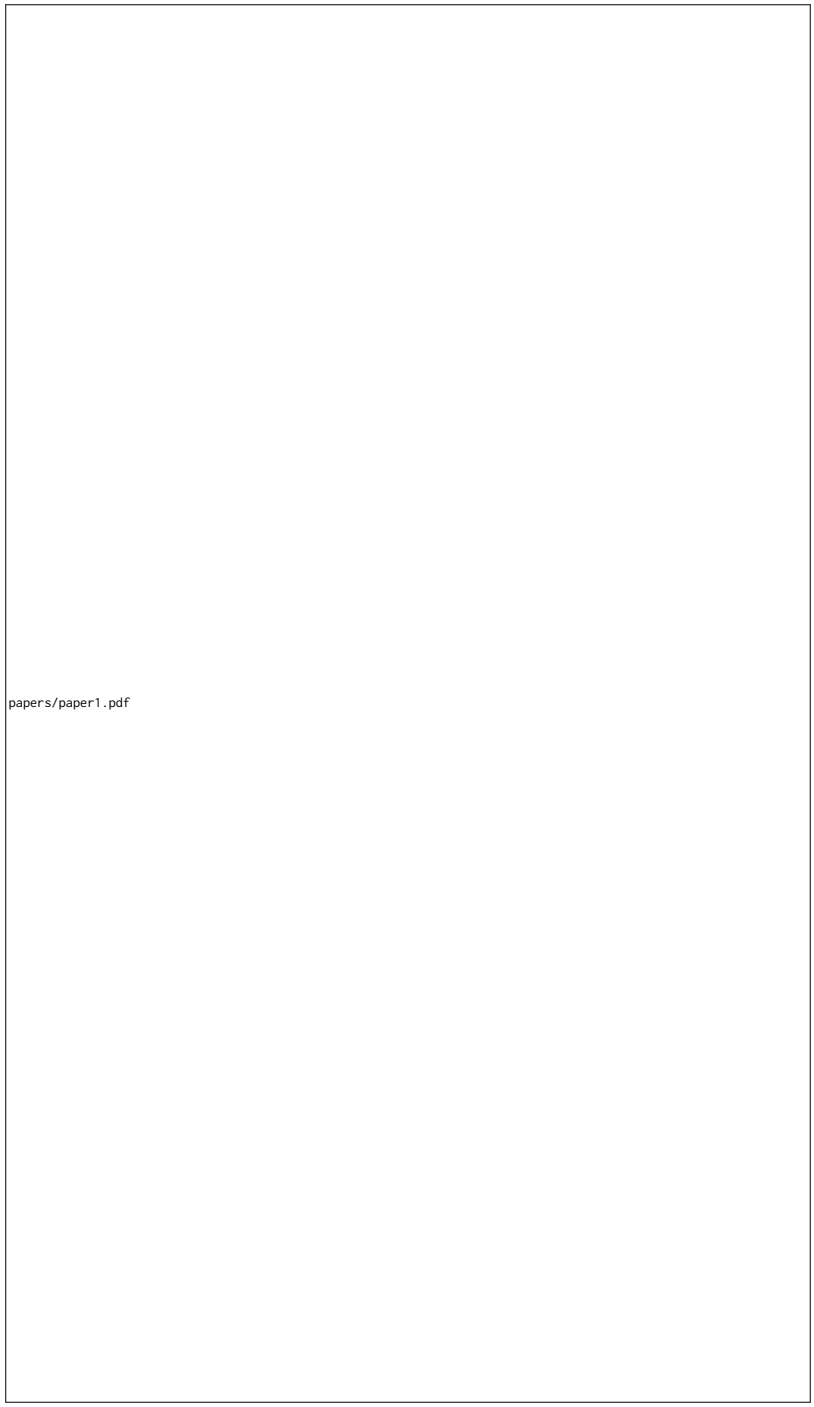

papers/paper1.pdf


papers/paper1.pdf

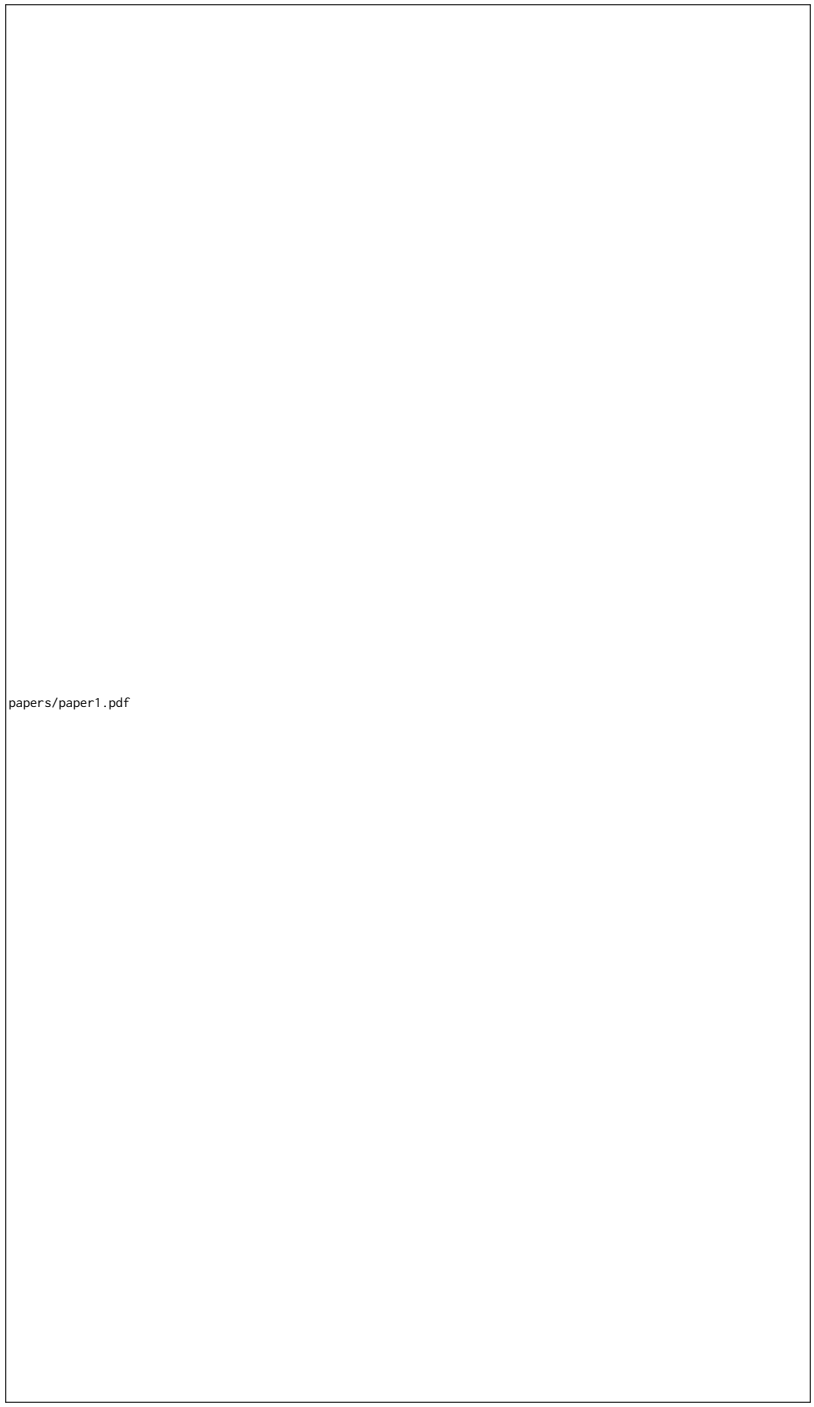

papers/paper1.pdf

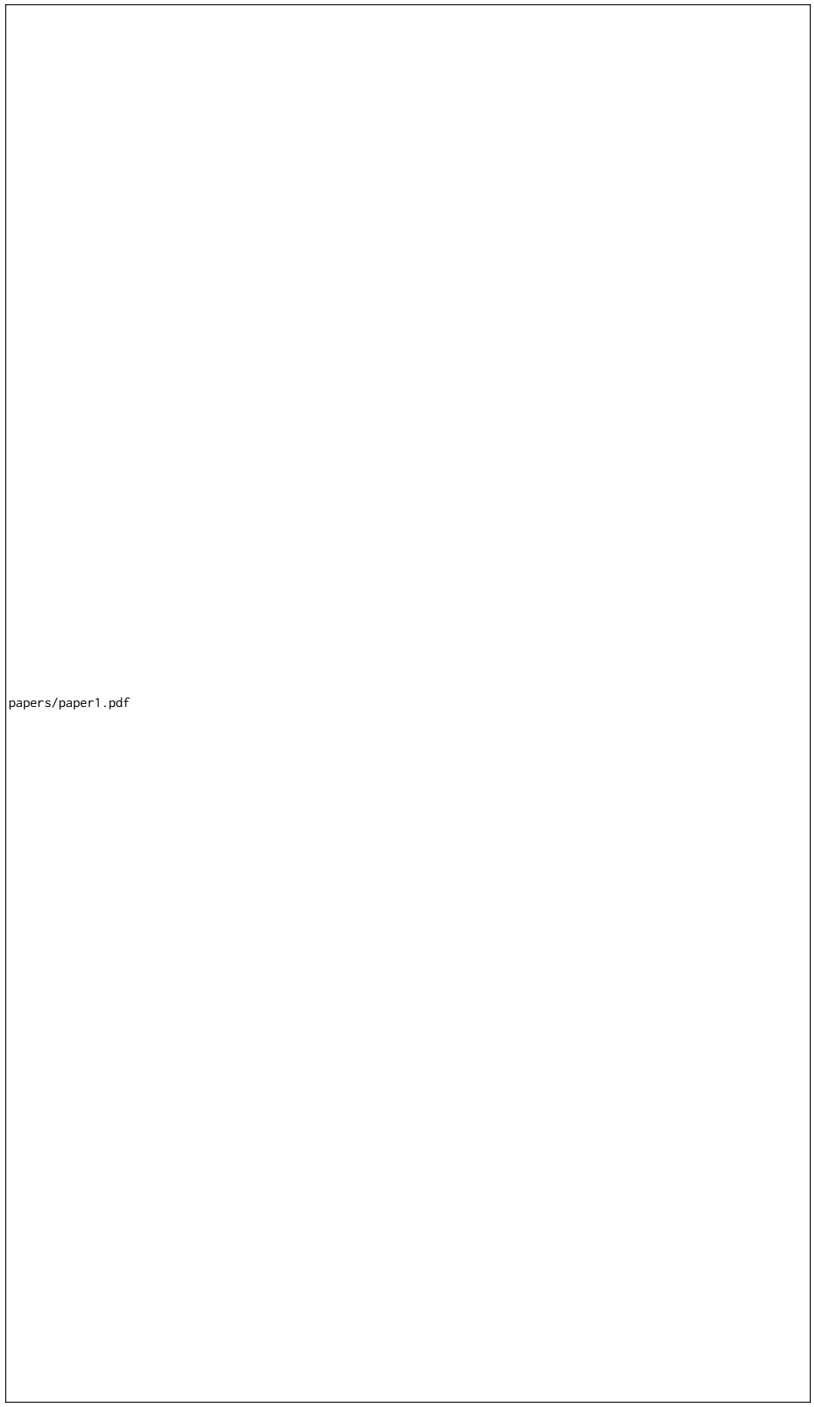
papers/paper1.pdf


papers/paper1.pdf


papers/paper1.pdf

papers/paper1.pdf


papers/paper1.pdf


papers/paper1.pdf

II

III

IV

V

VI