

# 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 <b>LUND UNIVERSITY</b> Department of Statistics Box 7080 SE-220 07 Lund Sweden	Document name <b>DOCTORAL DISSERTATION</b>	
	Date of disputation 2024-05-24	
Author(s) Johan Larsson	Sponsoring organization	
Title and subtitle Optimization and Algorithms in Sparse Regression: Screening Rules, Coordinate Descent, and Normalization		
<b>Abstract</b> Lorem ipsum dolor sit amet, consectetur 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, consectetur 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, consectetur 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, consectetur 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 62	Price
	Security classification	

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources the permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature \_\_\_\_\_

Date 1776-7-4



# 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.

© Johan Larsson 2024

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



Printed matter  
3041 0903

Media-Tryck is a Nordic Swan Ecolabel  
certified provider of printed material.  
Read more about our environmental  
work at [www.mediathyck.lu.se](http://www.mediathyck.lu.se)

**MADE IN SWEDEN**

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

—Joan Didion



# Contents

<b>Acknowledgements</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>Popular Science Summary</b>	<b>vii</b>
<b>List of Publications</b>	<b>ix</b>
<b>Introduction</b>	<b>I</b>
1    Background . . . . .	I
2    Regularization . . . . .	5
3    Solving the Problems . . . . .	8
4    Normalization . . . . .	9
5    Summary of the Papers . . . . .	9
<b>Papers</b>	<b>I7</b>
I    The Strong Screening Rule for SLOPE . . . . .	19
II   Look-Ahead Screening Rules for the Lasso . . . . .	33
III  The Hessian Screening Rule . . . . .	35
IV   BenchOpt: Reproducible, Efficient and Collaborative Optimization Benchmarks . . . . .	37
V    Coordinate Descent for SLOPE . . . . .	39
VI   Regularization and Scaling in Sparse Regression . . . . .	41



# 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, 2020–12). “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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. 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 Vaïter (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

All papers are reproduced with permission of their respective publishers.

# Introduction

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

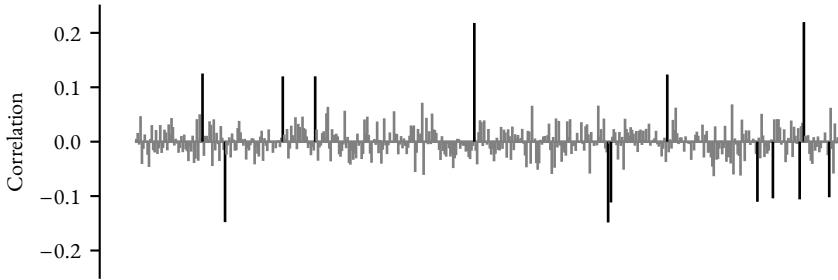
## I Background

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			(b) Wide data			
$x_1$	$x_2$	$x_3$	$x_1$	$x_2$	$x_3$	$\dots$
0	0.32	1	0	0.32	1	...
1	1	-1	1	1	-1	...
:	:	:				

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<sup>1</sup>. 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  $n \times p$

---

<sup>1</sup>The downsides are generally only related to computational issues, such as storing and processing this data.

matrix of features  $X$  and a response vector of length  $n$ ,  $y$ :

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

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

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

$$y = X\beta + \beta_0 + \epsilon,$$

where  $\beta$  is a vector of coefficients and  $\beta_0$  the *intercept*. In this representation of the data, the coefficients  $\beta$  are the parameters that we are interested in estimating and represent the effect each feature has on the response vector  $y$ .

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  $\epsilon$  is normally distributed<sup>2</sup> with mean zero and constant variance.

Since in the presence of noise there generally exists no  $\beta$  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

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

and the true response vector  $y$ , that is

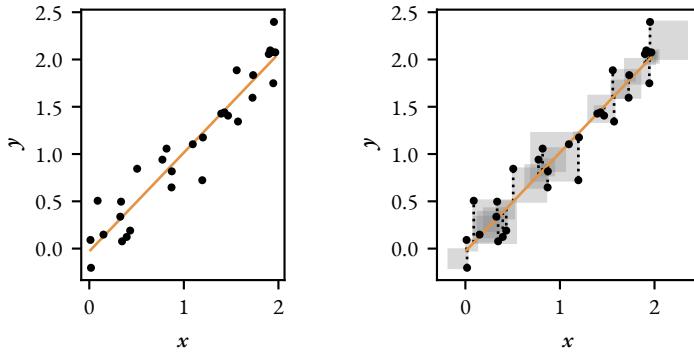
$$\|y - \hat{y}\|_2^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

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:

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

---

<sup>2</sup>Technically, this is 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  $\beta$ . The point where the line intersects the y-axis is the intercept  $\beta_0$ .

(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.

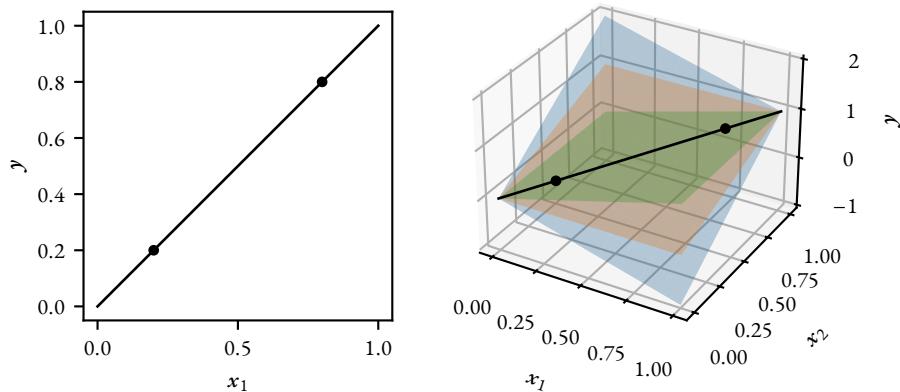
The factor of  $1/2$  is included for convenience, for reasons that will become clear later on. This choice leads to the ordinary least-squares (OLS) regression model, which, for a simple case of a single feature, we have illustrated in Figure 2. There are many other ways to measure error, which all 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  $y$  and the predictions ( $\hat{y}$ ) made by the model. For the simple case of a single feature, this model is illustrated in Figure 2.

If we 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  $\beta$  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  $\beta$  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 feature, the simple ordinary 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 regression with an increasing number of degrees. Eventually the model becomes too flexible and overfits (Figure 4).

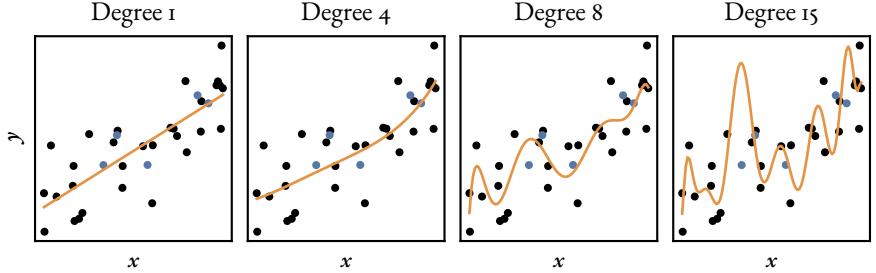
These problems are the motivation for the use of *regularization* in regression, which we will turn to next.

## 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 altogether 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 optimization problem:

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

where  $k$  denotes the number of features that we allow in our model. The  $\|\cdot\|_0$  norm<sup>3</sup> 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:

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

But the following model would not:

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

---

<sup>3</sup>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  $\ell_1$  norm in place of the  $\ell_0$  norm. In other words, our problem now becomes

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

where  $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$  and, as a consequence, we have replaced the integer-valued  $k$  with a real-valued (but positive)  $t$ . This problem is known as  $\ell_1$ -regularized regression or, more commonly, the *lasso* (Robert Tibshirani 1996)<sup>4</sup>.

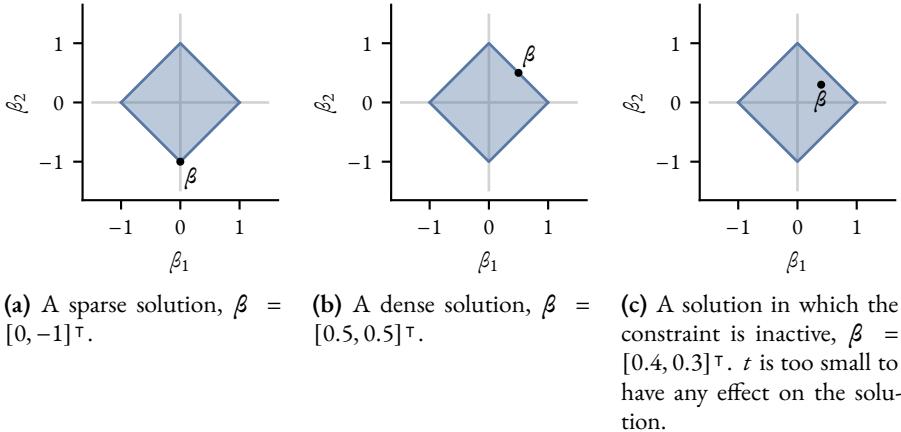
The lasso was introduced to the statistics community by Robert Tibshirani (1996) but actually stems from much earlier research done in the field of signal processing by Santosa and Symes (1986). Donoho and Johnstone (1994, 1995) subsequently introduced the concept of the *basis pursuit* problem, which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

We saw previously that the  $\ell_0$  constraint in best-subset selection puts a budget on the number of features allowed in the model. The  $\ell_1$  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,

---

<sup>4</sup>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.



**Figure 5:** The  $\ell_1$  norm ball in  $\mathbb{R}^2$  with some possible solutions indicated by  $\beta$ . The  $\ell_0$  ball (for best-subset selection) and  $k = 1$  would be lines of infinite length along both of the axes.

and Knight 2005), group lasso (Yuan and Lin 2005), 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  $\ell_1$ -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.

## 2.2 Elastic Net

## 2.3 SLOPE

# 3 Solving the Problems

## 3.1 Optimization

Originally, Robert Tibshirani (1996) proposed to solve the lasso via an iterative method that used a modification of an algorithm from Lawson and Hanson (1995), but it scales badly and is inapplicable when  $p > n$ .

### 3.2 Screening Rules

## 4 Normalization

## 5 Summary of the Papers

### 5.1 Paper I

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.

### 5.2 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.

### 5.3 Paper III

In this paper, we address the challenge of predictor screening rules in  $\ell_1$ -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  $\ell_1$ -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  $\ell_1$ -regularized regression problems.

### 5.4 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:  $\ell_2$ -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.

### 5.5 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.

## 5.6 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, A., V. Chernozhukov, and L. 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](https://doi.org/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](https://doi.org/10.1214/15-AOS1388).
- 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](https://doi.org/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](https://doi.org/10.2307/2291512). JSTOR: [2291512](#).
- 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](https://doi.org/10.1093/biostatistics/kxm045).
- 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 L. K. Saul, Y. Weiss, and L. 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](https://doi.org/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 I: 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 Papastamoulis, 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, 2020–12). "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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. 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](https://doi.org/10.1137/1.9781611971217).
- 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.
- 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](https://doi.org/10.1137/0907087).
- 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 (Methodological)* 58.1, pp. 267–288. ISSN: 0035-9246. DOI: [10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/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 (Statistical Methodology)* 67.1, pp. 91–108. ISSN: 1467-9868. DOI: [10.1111/j.1467-9868.2005.00490.x](https://doi.org/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 (Statistical Methodology)* 68.1, pp. 49–67. ISSN: 1467-9868. DOI: [10.1111/j.1467-9868.2005.00532.x](https://doi.org/10.1111/j.1467-9868.2005.00532.x).

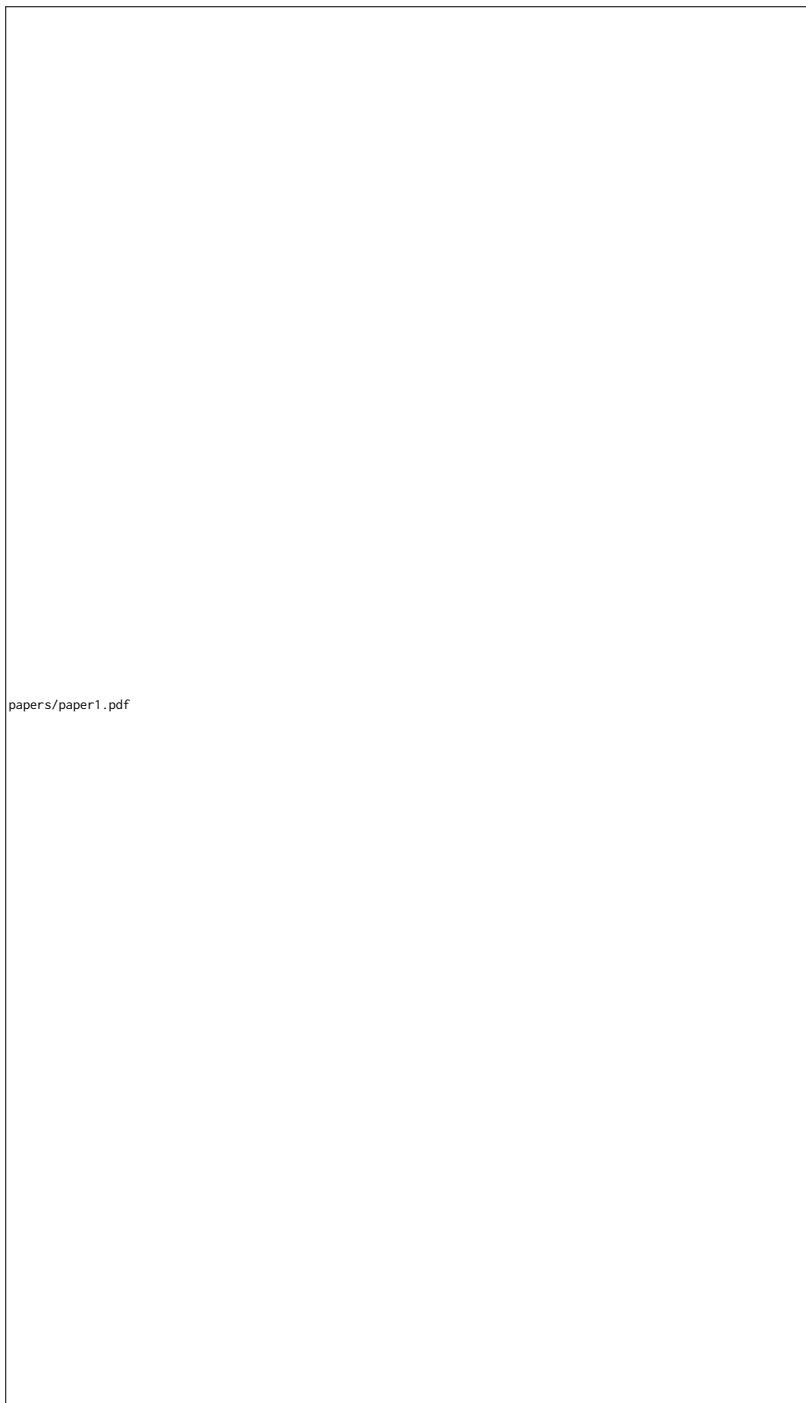


# Papers

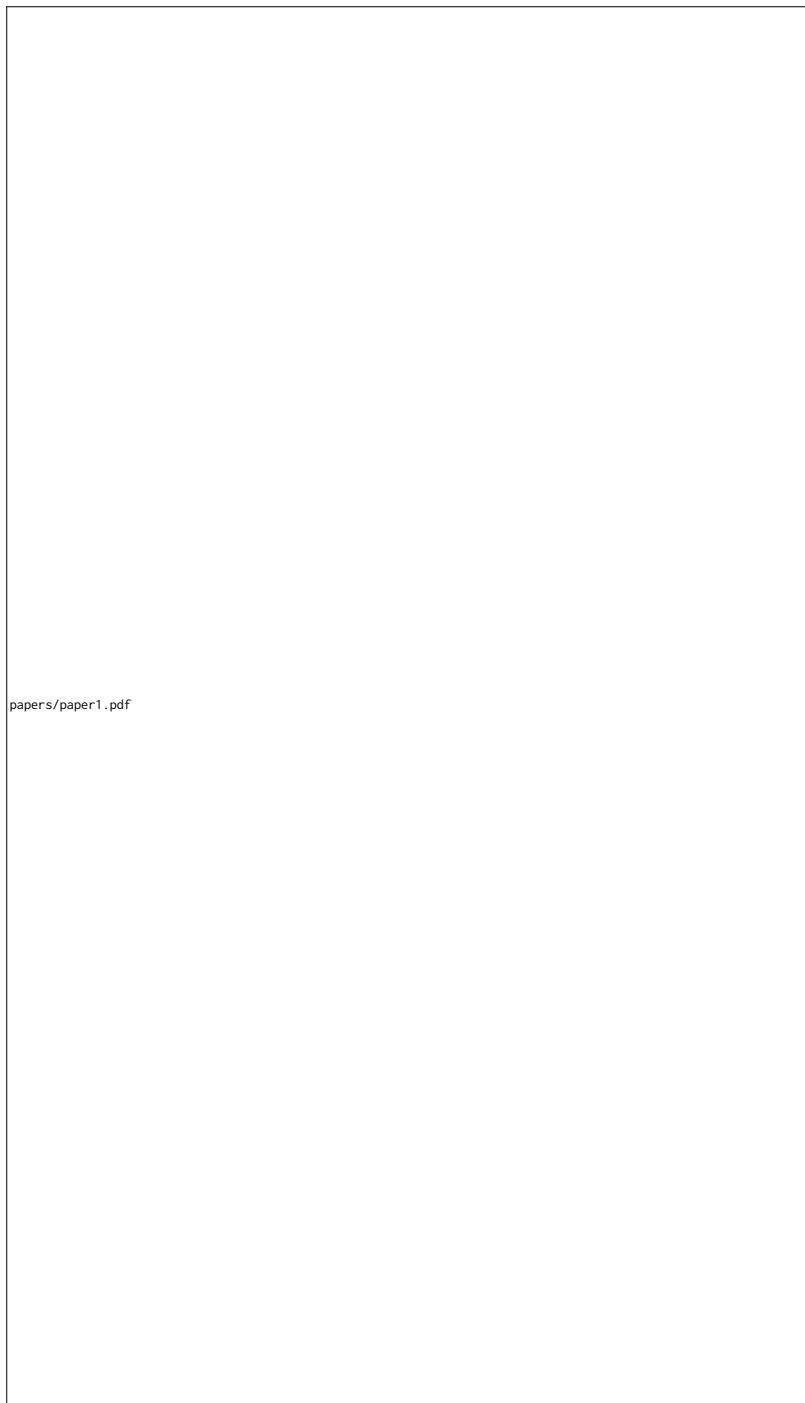


I

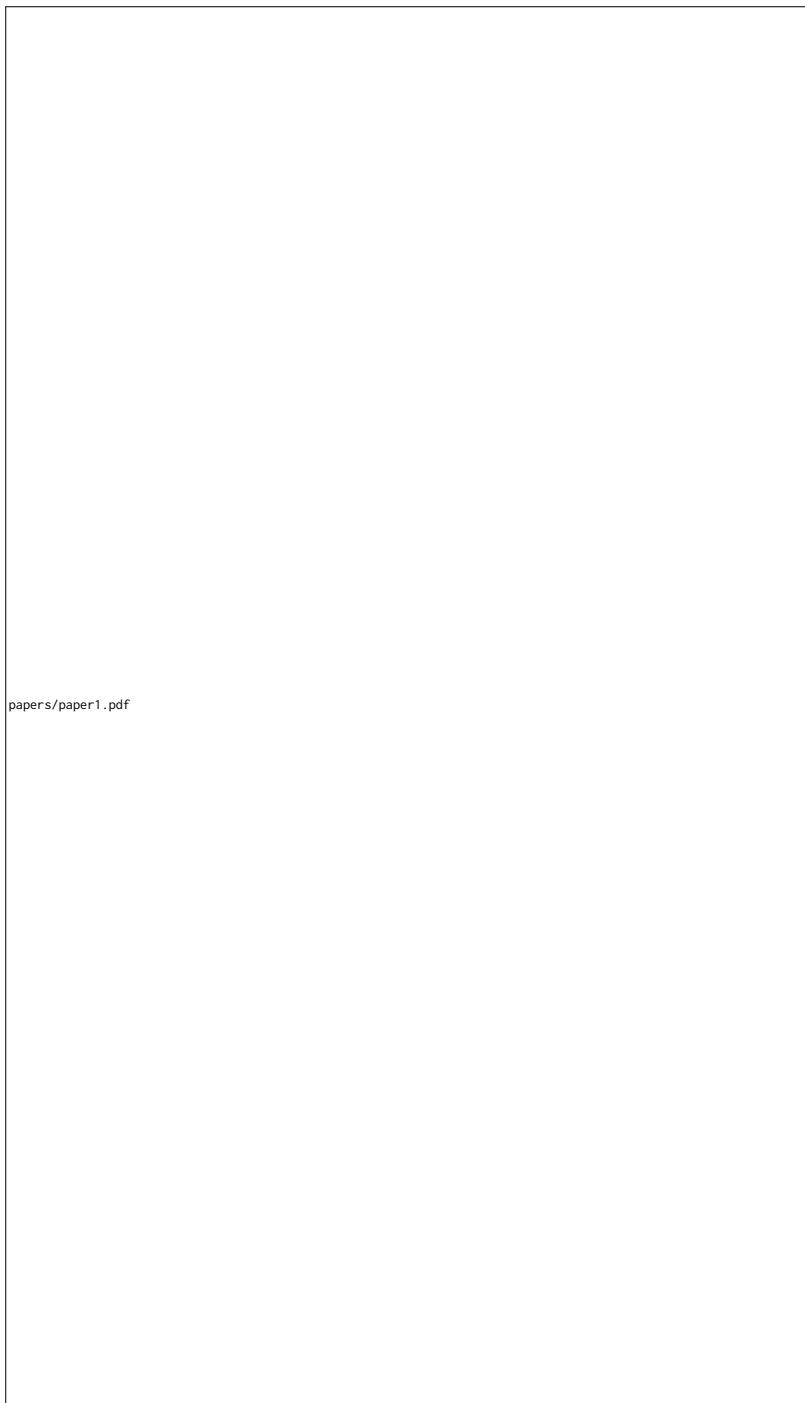




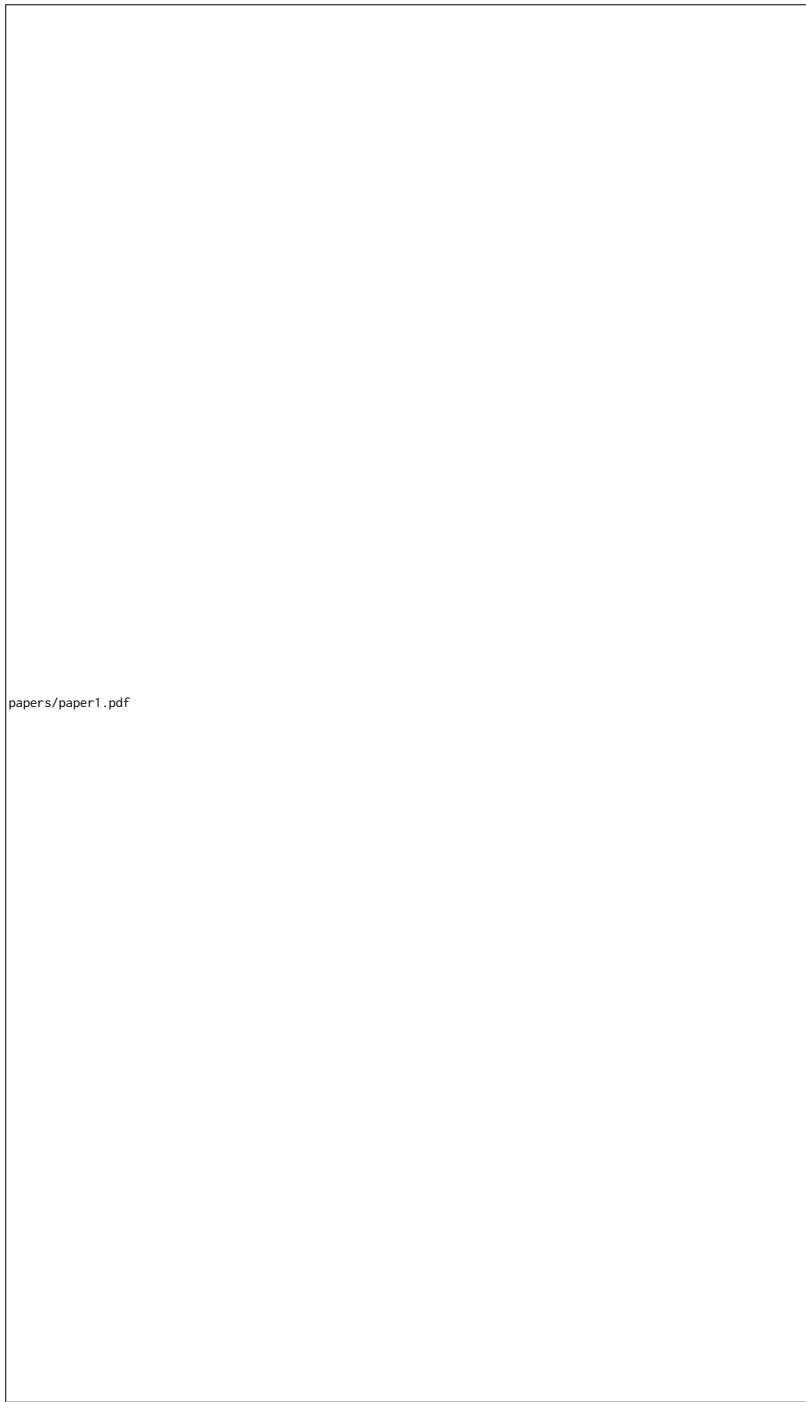
papers/paper1.pdf



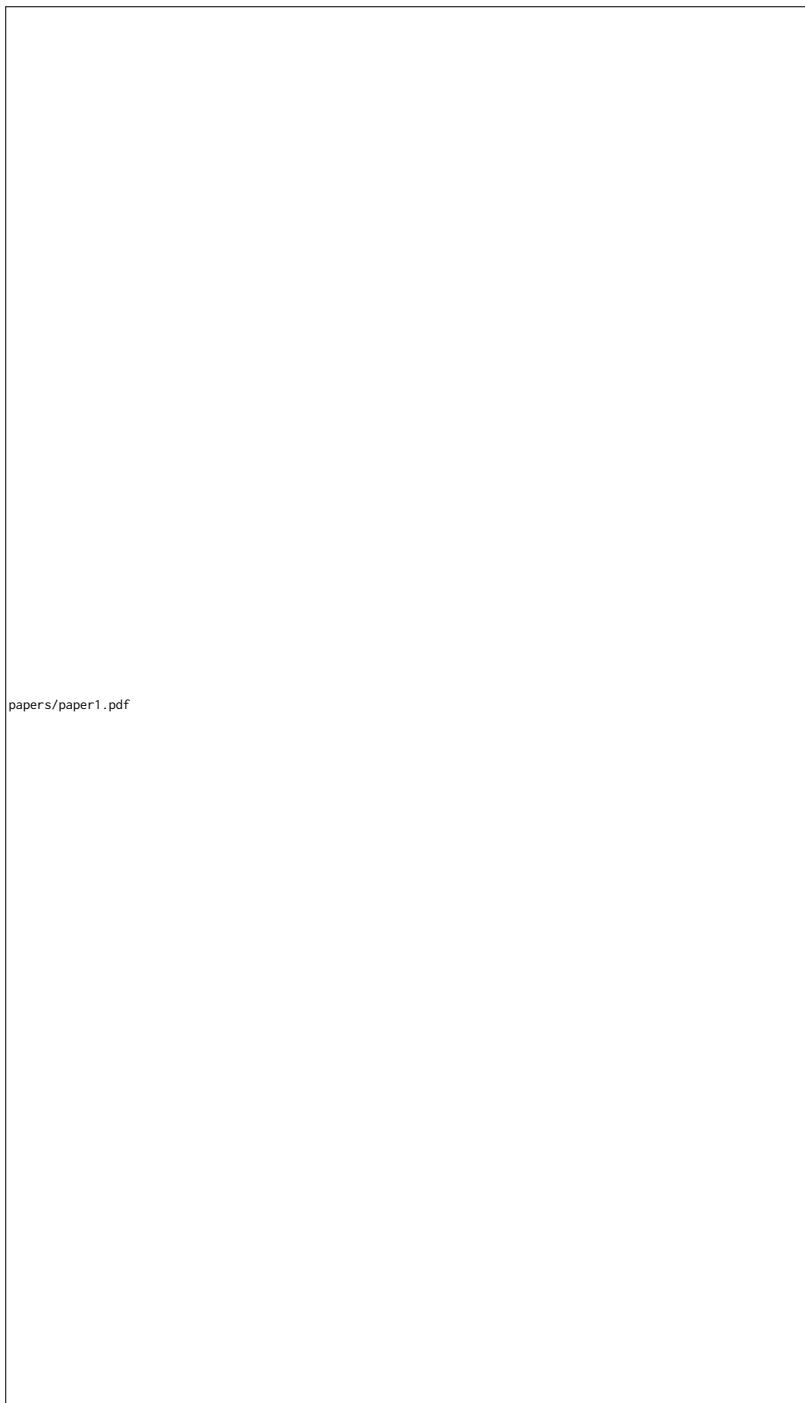
papers/paper1.pdf



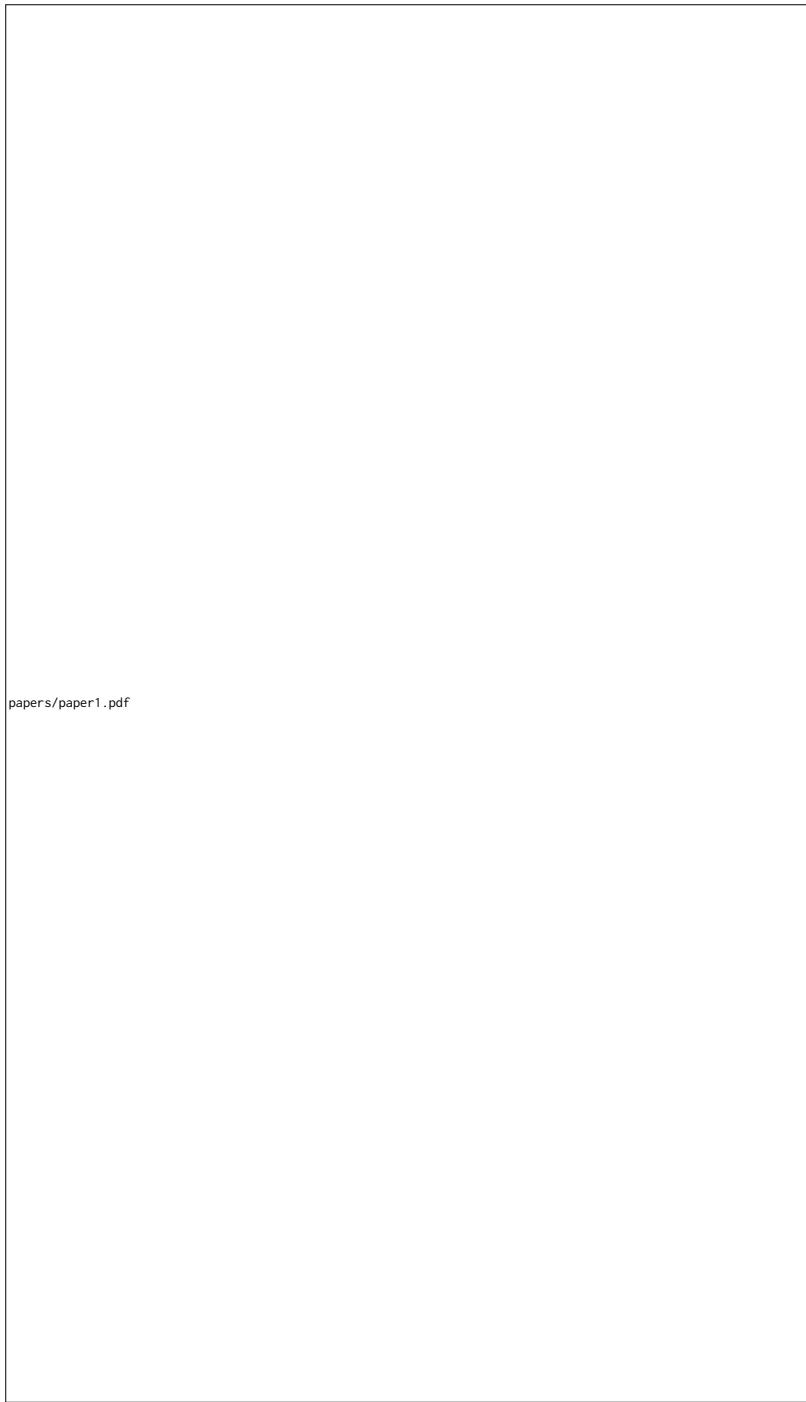
papers/paper1.pdf



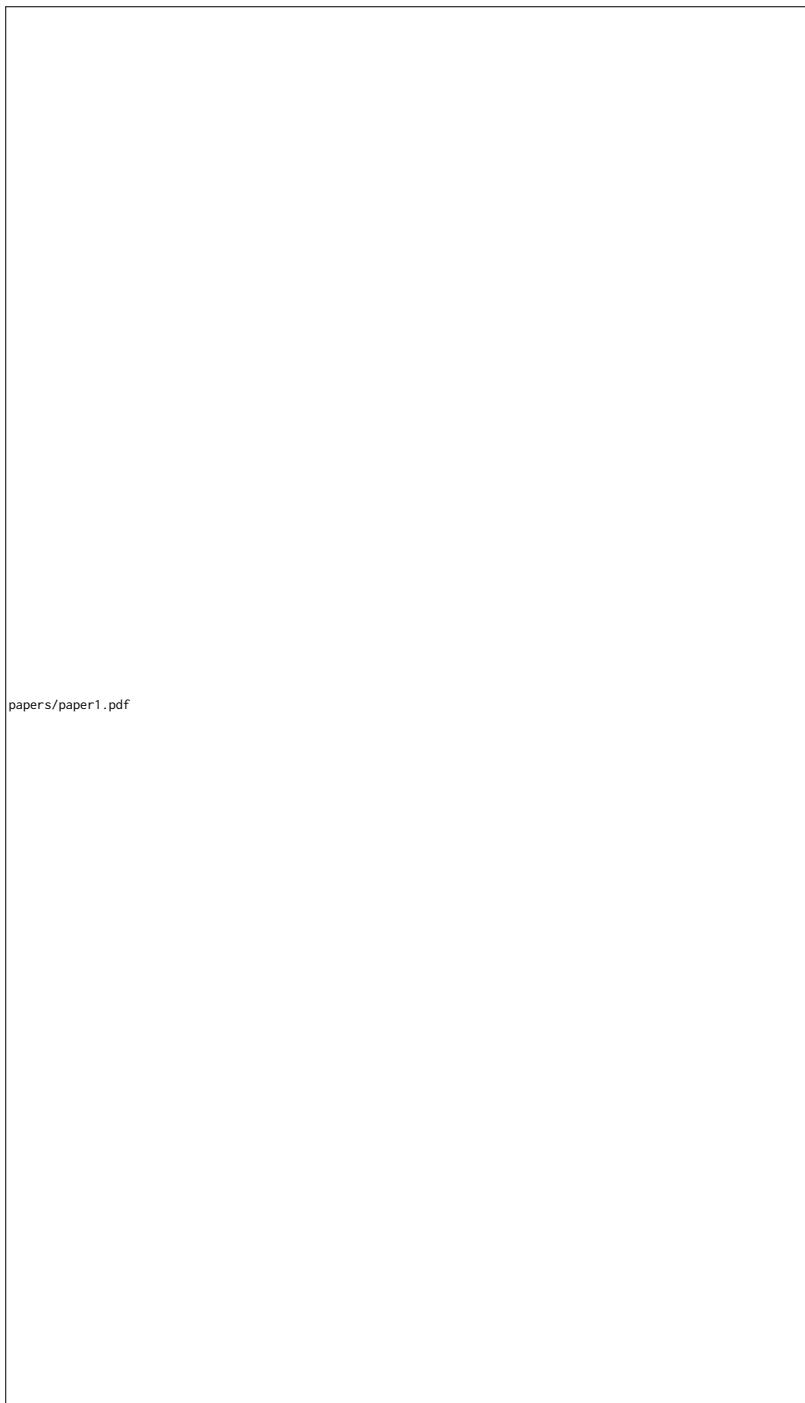
papers/paper1.pdf



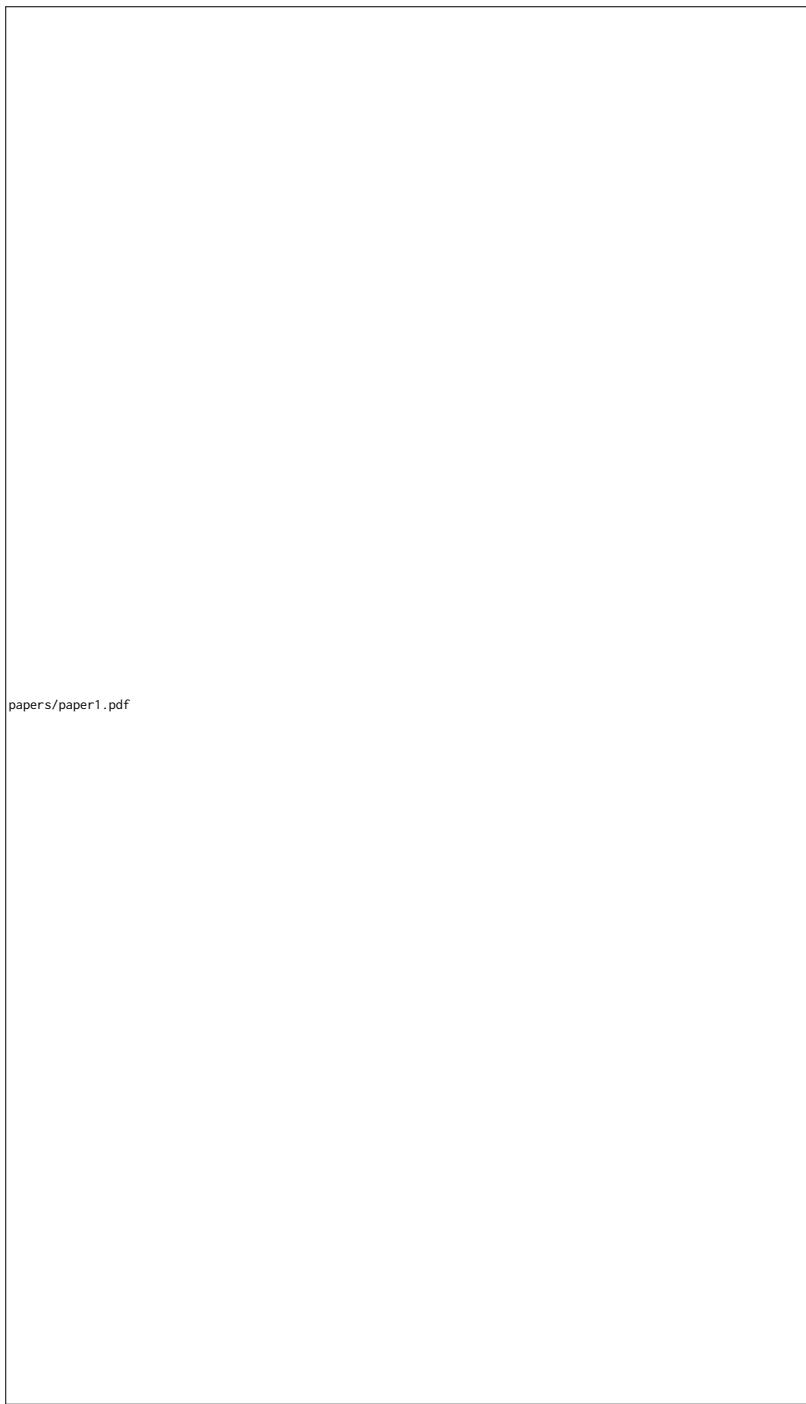
papers/paper1.pdf



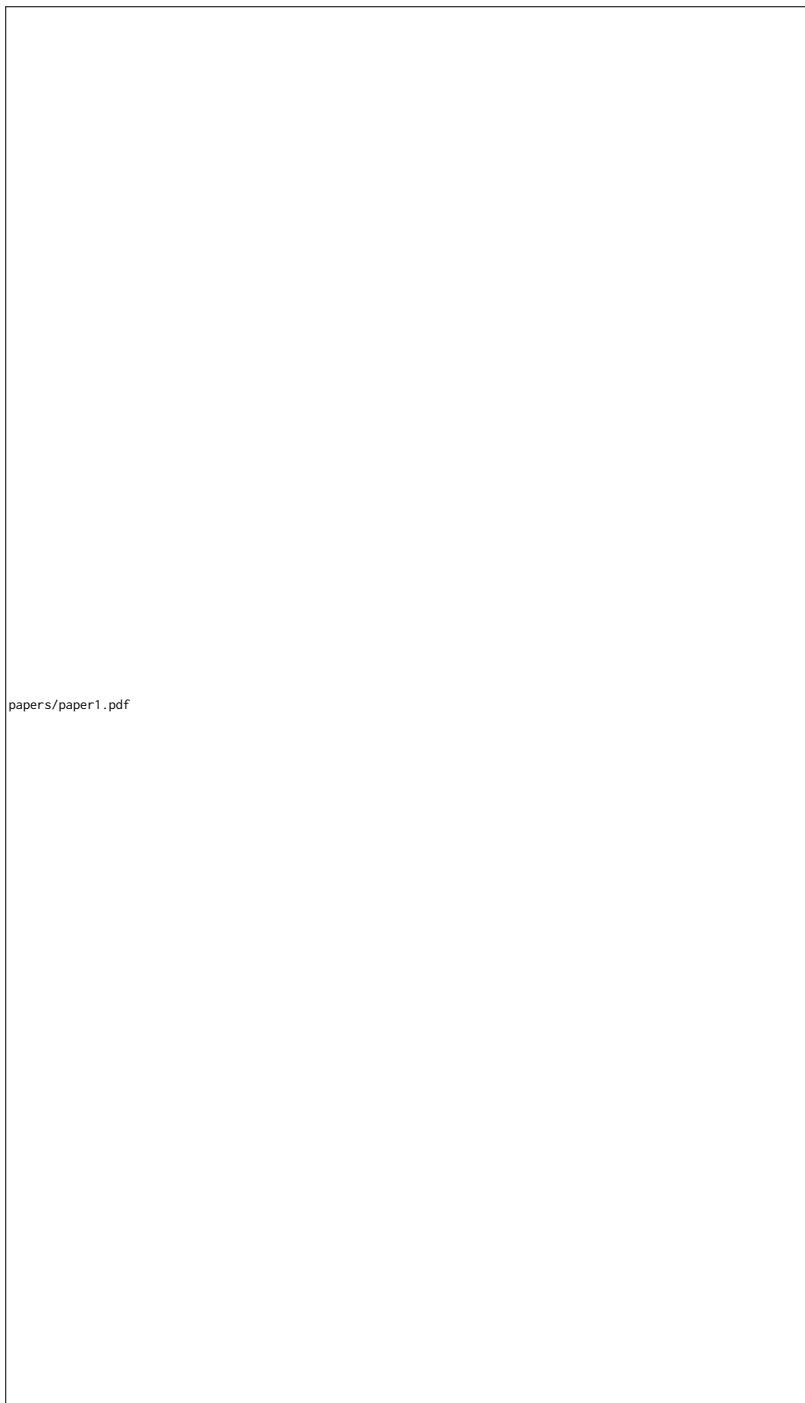
papers/paper1.pdf



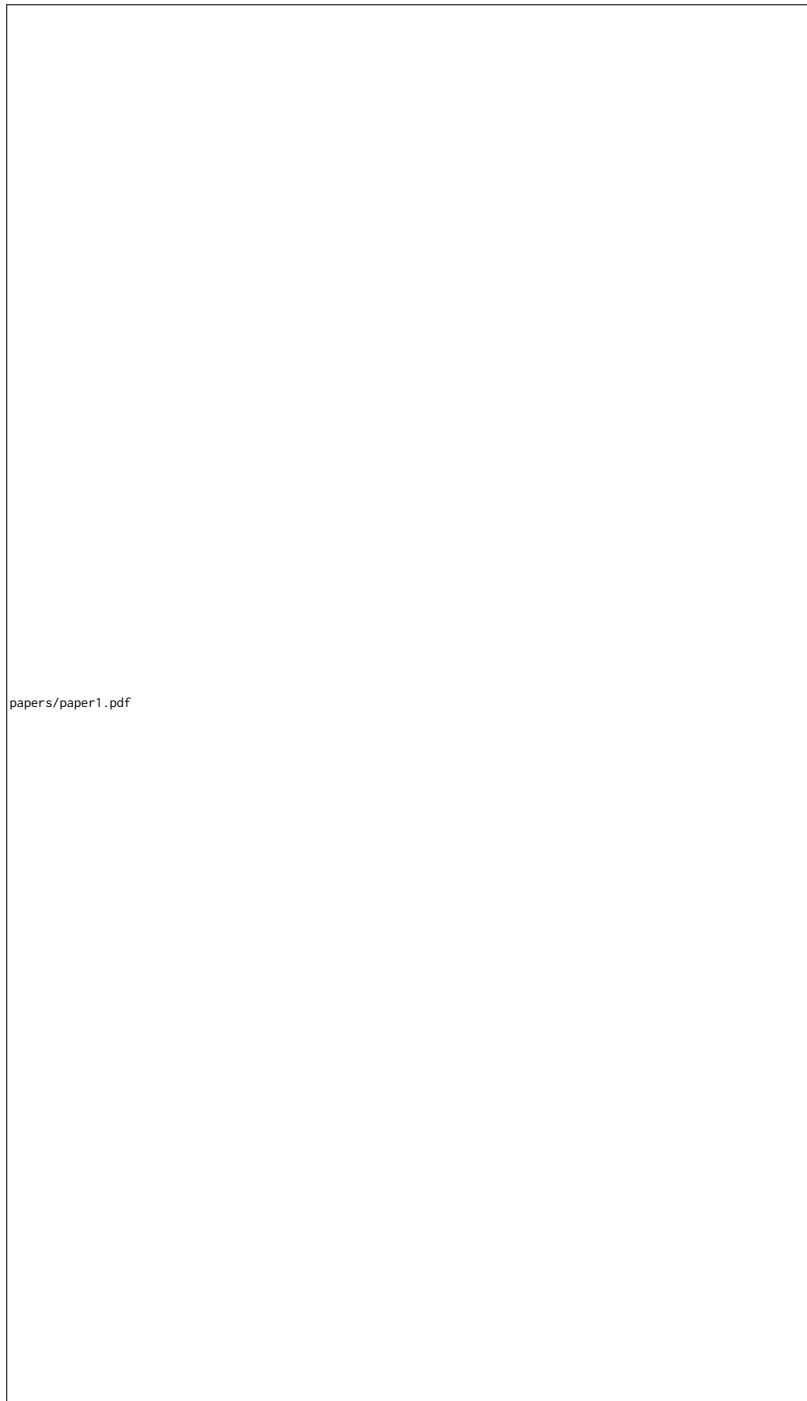
papers/paper1.pdf



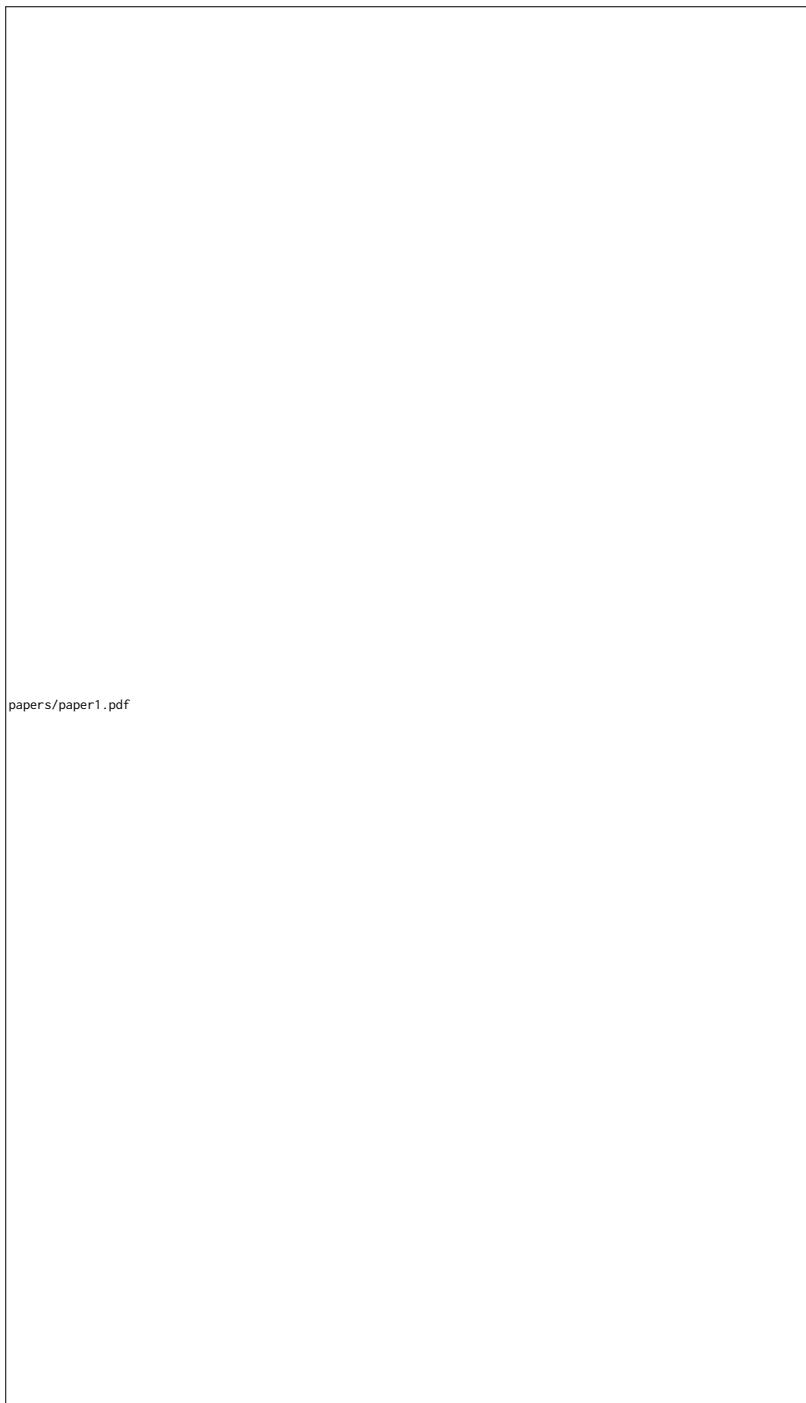
papers/paper1.pdf



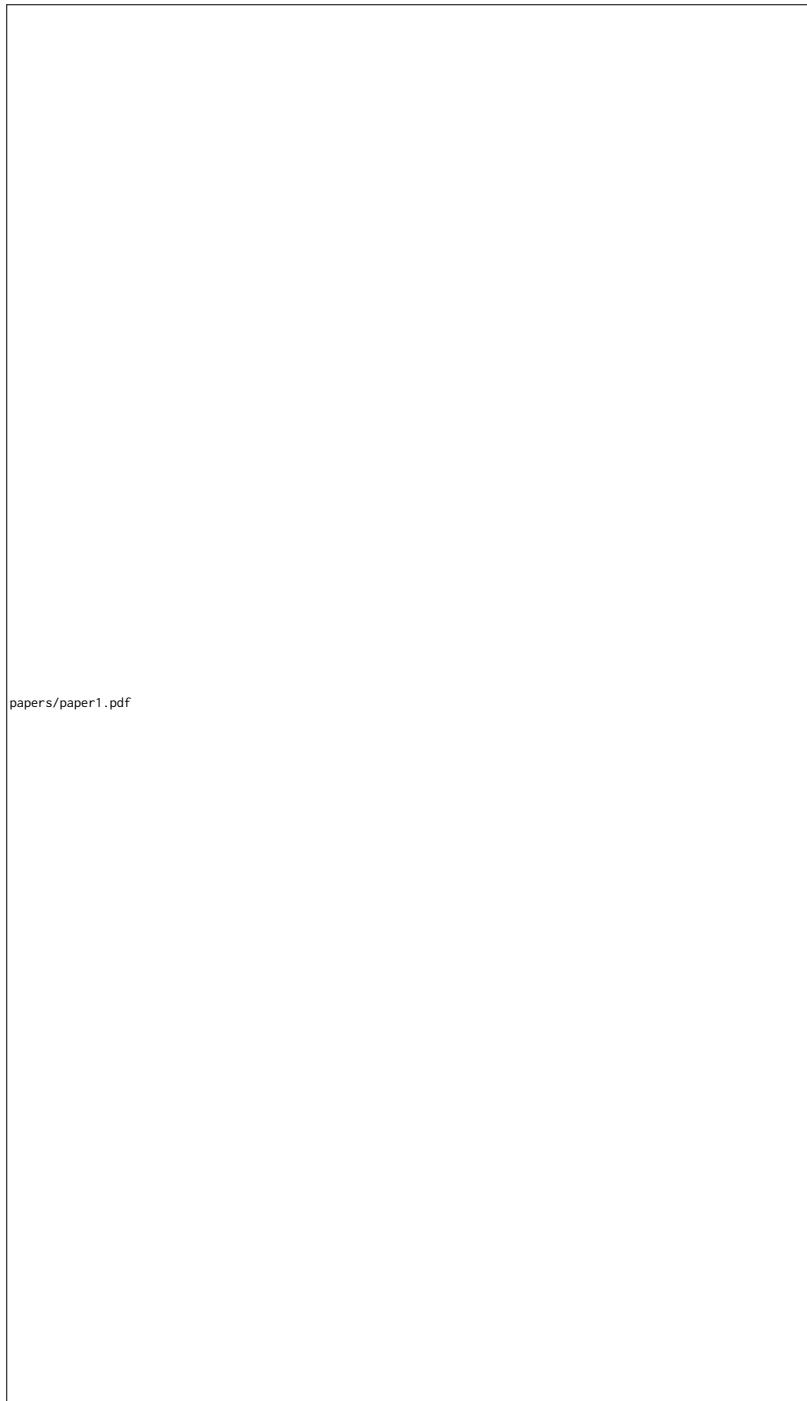
papers/paper1.pdf



papers/paper1.pdf



papers/paper1.pdf



papers/paper1.pdf

**II**



III



**IV**



V





