ARTICLE IN PRESS
JID: EOR [m5G;April 21, 2017;18:31]
European Journal of Operational Research 000 (2017) 1–12
Contents lists available at ScienceDirect
European Journal of Operational Research
journal homepage: www.elsevier.com/locate/ejor

Characterization of the equivalence of robustification and regularization in linear and matrix regression[☆]

Dimitris Bertsimas[a,*], Martin S. Copenhaver[b]

[a] Sloan School of Management and Operations Research Center, MIT, United States
[b] Operations Research Center, MIT, United States

ARTICLE INFO

Article history:
Received 9 November 2016
Accepted 20 March 2017
Available online xxx

Keywords:
Convex programming
Robust optimization
Statistical regression
Penalty methods
Adversarial learning

ABSTRACT

The notion of developing statistical methods in machine learning which are robust to adversarial perturbations in the underlying data has been the subject of increasing interest in recent years. A common feature of this work is that the adversarial *robustification* often corresponds exactly to regularization methods which appear as a loss function plus a penalty. In this paper we deepen and extend the understanding of the connection between robustification and regularization (as achieved by penalization) in regression problems. Specifically,

(a) In the context of linear regression, we characterize precisely under which conditions on the model of uncertainty used and on the loss function penalties robustification and regularization are equivalent.
(b) We extend the characterization of robustification and regularization to matrix regression problems (matrix completion and Principal Component Analysis).

1. Introduction

The development of predictive methods that perform well in the face of uncertainty is at the core of modern machine learning and statistical practice. Indeed, the notion of *regularization*—loosely speaking, a means of controlling the ability of a statistical model to generalize to new settings by trading off with the model's complexity— is at the very heart of such work (Hastie, Tibshirani, & Friedman, 2009). Corresponding regularized statistical methods, such as the Lasso for linear regression (Tibshirani, 1996) and nuclear-norm-based approaches to matrix completion (Candès & Recht, 2012; Recht, Fazel, & Parrilo, 2010), are now ubiquitous and have seen widespread success in practice.

In parallel to the development of such regularization methods, it has been shown in the field of robust optimization that under certain conditions these regularized problems result from the need to immunize the statistical problem against adversarial perturbations in the data (Ben-Tal, Ghaoui, & Nemirovski, 2009; Caramanis, Mannor, & Xu, 2011; Ghaoui & Lebret, 1997; Xu, Caramanis, & Mannor, 2010). Such a *robustification* offers a different perspective on regularization methods by identifying which adversarial perturbations the model is protected against. Conversely, this can help to inform statistical modeling decisions by identifying potential choices of regularizers. Further, this connection between regularization and robustification offers the potential to use sophisticated data-driven methods in robust optimization (Bertsimas, Gupta, & Kallus, 2013; Tulabandhula & Rudin, 2014) to design regularizers in a principled fashion.

With the continuing growth of the adversarial viewpoint in machine learning (e.g. the advent of new deep learning methodologies such as generative adversarial networks (Goodfellow et al., 2014a; Goodfellow, Shlens, & Szegedy, 2014b; Shaham, Yamada, & Negahban, 2015)), it is becoming increasingly important to better understand the connection between robustification and regularization. Our goal in this paper is to shed new light on this relationship by focusing in particular on linear and matrix regression problems. Specifically, our contributions include:

1. In the context of linear regression we demonstrate that in general such a robustification procedure is not equivalent to regularization (via penalization). We characterize precisely under which conditions on the model of uncertainty used and on the loss function penalties one has that robustification is equivalent to regularization.
2. We break new ground by considering problems in the matrix setting, such as matrix completion and Principal Component Analysis (PCA). We show that the nuclear norm, a

☆ Copenhaver is partially supported by the Department of Defense, Office of Naval Research, through the National Defense Science and Engineering Graduate Fellowship.

* Corresponding author.
E-mail addresses: dbertsim@mit.edu (D. Bertsimas), mcopen@mit.edu (M.S. Copenhaver).

http://dx.doi.org/10.1016/j.ejor.2017.03.051
0377-2217/© 2017 Elsevier B.V. All rights reserved.

Table 1
Matrix norms on $\mathbf{\Delta} \in \mathbb{R}^{m\times n}$.

Name	Notation	Definition	Description
p-Frobenius	F_p	$\left(\sum_{ij} \lvert\Delta_{ij}\rvert^p\right)^{1/p}$	Entrywise ℓ_p norm
p-spectral (Schatten)	σ_p	$\lVert\boldsymbol{\mu}(\mathbf{\Delta})\rVert_p$	ℓ_p norm on the singular values
Induced	(h, g)	$\max_{\boldsymbol{\beta}} \frac{g(\mathbf{\Delta}\boldsymbol{\beta})}{h(\boldsymbol{\beta})}$	Induced by norms g, h

popular penalty function used throughout this setting, arises directly through robustification. As with the case of vector regression, we characterize under which conditions on the model of uncertainty there is equivalence of robustification and regularization in the matrix setting.

The structure of the paper is as follows. In Section 2, we review background on norms and consider robustification and regularization in the context of linear regression, focusing both on their equivalence and non-equivalence. In Section 3, we turn our attention to regression with underlying matrix variables, considering in depth both matrix completion and PCA. In Section 4, we include some concluding remarks.

2. A robust perspective of linear regression

2.1. Norms and their duals

In this section, we introduce the necessary background on norms which we will use to address the equivalence of robustification and regularization in the context of linear regression. Given a vector space $V \subseteq \mathbb{R}^n$ we say that $\|\cdot\| : V \to \mathbb{R}$ is a *norm* if for all $\mathbf{v}, \mathbf{w} \in V$ and $\alpha \in \mathbb{R}$

1. If $\|\mathbf{v}\| = 0$, then $\mathbf{v} = 0$,
2. $\|\alpha\mathbf{v}\| = |\alpha|\|\mathbf{v}\|$ (absolute homogeneity), and
3. $\|\mathbf{v}+\mathbf{w}\| \le \|\mathbf{v}\| + \|\mathbf{w}\|$ (triangle inequality).

If $\|\cdot\|$ satisfies conditions 2 and 3, but not 1, we call it a *seminorm*. For a norm $\|\cdot\|$ on $\mathbb{R}^n$ we define its dual, denoted $\|\cdot\|_*$, to be

$$\|\boldsymbol{\beta}\|_* := \max_{\mathbf{x}\in\mathbb{R}^n} \frac{\mathbf{x}'\boldsymbol{\beta}}{\|\mathbf{x}\|},$$

where $\mathbf{x}'$ denotes the transpose of $\mathbf{x}$ (and therefore $\mathbf{x}'\boldsymbol{\beta}$ is the usual inner product). For example, the ℓ_p norms $\|\boldsymbol{\beta}\|_p := (\Sigma_i|\beta_i|^p)^{1/p}$ for $p \in [1, \infty)$ and $\|\boldsymbol{\beta}\|_\infty := \max_i|\beta_i|$ satisfy a well-known duality relation: ℓ_{p^*} is dual to ℓ_p, where $p^* \in [1, \infty]$ with $1/p + 1/p^* = 1$. We call p^* the *conjugate* of p. More generally for matrix norms[1] $\|\cdot\|$ on $\mathbb{R}^{m\times n}$ the dual is defined analogously:

$$\|\mathbf{\Delta}\|_* := \max_{\mathbf{A}\in\mathbb{R}^{m\times n}} \frac{\langle\mathbf{A}, \mathbf{\Delta}\rangle}{\|\mathbf{A}\|},$$

where $\mathbf{\Delta} \in \mathbb{R}^{m\times n}$ and $\langle\cdot,\cdot\rangle$ denotes the trace inner product: $\langle\mathbf{A}, \mathbf{\Delta}\rangle = \mathrm{Tr}(\mathbf{A}'\mathbf{\Delta})$, where $\mathbf{A}'$ denotes the transpose of $\mathbf{A}$. We note that the dual of the dual norm is the original norm (Boyd & Vandenberghe, 2004).

Three widely used choices for matrix norms (see Horn & Johnson, 2013) are Frobenius, spectral, and induced norms. The definitions for these norms are given below for $\mathbf{\Delta} \in \mathbb{R}^{m\times n}$ and summarized in Table 1 for convenient reference.

[1] We treat a matrix norm as any norm on $\mathbb{R}^{m\times n}$ which satisfies the three conditions of a usual vector norm, although some authors reserve the term "matrix norm" for a norm on $\mathbb{R}^{m\times n}$ which also satisfies a submultiplicativity condition (see Horn and Johnson, 2013, pg. 341).

1. The p-Frobenius norm, denoted $\|\cdot\|_{F_p}$, is the entrywise ℓ_p norm on the entries of $\mathbf{\Delta}$:
$$\|\mathbf{\Delta}\|_{F_p} := \left(\sum_{ij} |\Delta_{ij}|^p\right)^{1/p}.$$
Analogous to before, F_{p^*} is dual to F_p, where $1/p + 1/p^* = 1$.
2. The p-spectral (Schatten) norm, denoted $\|\cdot\|_{\sigma_p}$, is the ℓ_p norm on the singular values of the matrix $\mathbf{\Delta}$:
$$\|\mathbf{\Delta}\|_{\sigma_p} := \|\boldsymbol{\mu}(\mathbf{\Delta})\|_p,$$
where $\boldsymbol{\mu}(\mathbf{\Delta})$ denotes the vector containing the singular values of $\mathbf{\Delta}$. Again, σ_{p^*} is dual to σ_p.
3. Finally we consider the class of induced norms. If $g : \mathbb{R}^m \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$ are norms, then we define the induced norm $\|\cdot\|_{(h,\,g)}$ as
$$\|\mathbf{\Delta}\|_{(h,g)} := \max_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{g(\mathbf{\Delta}\boldsymbol{\beta})}{h(\boldsymbol{\beta})}.$$
An important special case occurs when $g = \ell_p$ and $h = \ell_q$. When such norms are used, (q, p) is used as shorthand to denote (ℓ_q, ℓ_p). Induced norms are sometimes referred to as operator norms. We reserve the term operator norm for the induced norm $(\ell_2, \ell_2) = (2, 2) = \sigma_\infty$, which measures the largest singular value.

2.2. Uncertain regression

We now turn our attention to uncertain linear regression problems and regularization. The starting point for our discussion is the standard problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} g(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

where $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{X} \in \mathbb{R}^{m\times n}$ are data and g is some convex function, typically a norm. For example, $g = \ell_2$ is least squares, while $g = \ell_1$ is known as least absolute deviation (LAD). In favor of models which mitigate the effects of overfitting these are often replaced by the *regularization* problem

$$\min_{\boldsymbol{\beta}} g(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + h(\boldsymbol{\beta}),$$

where $h : \mathbb{R}^n \to \mathbb{R}$ is some penalty function, typically taken to be convex. This approach often aims to address overfitting by penalizing the complexity of the model, measured as $h(\boldsymbol{\beta})$. (For a more formal treatment using Hilbert space theory, (see Bauschke & Combettes, 2011; Bousquet, Boucheron, & Lugosi, 2004). For example, taking $g = \ell_2^2$ and $h = \ell_2^2$, we recover the so-called regularized least squares (RLS), also known as ridge regression (Hastie et al., 2009). The choice of $g = \ell_2^2$ and $h = \ell_1$ leads to Lasso, or least absolute shrinkage and selection operator, introduced in Tibshirani (1996). Lasso is often employed in scenarios where the solution $\boldsymbol{\beta}$ is desired to be sparse, i.e., $\boldsymbol{\beta}$ has very few nonzero entries. Broadly speaking, regularization can take much more general forms; for our purposes, we restrict our attention to regularization that appears in the penalized form above.

In contrast to this approach, one may alternatively wish to reexamine the nominal regression problem $\min_{\boldsymbol{\beta}} g(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ and instead attempt to solve this taking into account adversarial noise in the data matrix $\mathbf{X}$. As in Ghaoui and Lebret (1997), Lewis (2002), Lewis and Pang (2009), Ben-Tal et al. (2009), Xu et al. (2010), this approach may take the form

$$\min_{\boldsymbol{\beta}} \max_{\mathbf{\Delta}\in\mathcal{U}} g(\mathbf{y} - (\mathbf{X}+\mathbf{\Delta})\boldsymbol{\beta}), \tag{1}$$

where the set $\mathcal{U} \subseteq \mathbb{R}^{m\times n}$ characterizes the user's belief about uncertainty on the data matrix $\mathbf{X}$. This set $\mathcal{U}$ is known in the