論文 / 著書情報
Article / Book Information

題目(和文)	全体凸性条件を満たすDC型非凸正則化モデルに関する研究
Title(English)	A Study on Difference-of-Convex Type Nonconvexly Regularized Convex Models
著者(和文)	ZHANG Yi
Author(English)	Yi Zhang
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12810号, 授与年月日:2024年6月30日, 学位の種別:課程博士, 審査員:山田 功,植松 友彦,府川 和彦,尾形 わかは,山下 真
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12810号, Conferred date:2024/6/30, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	要約
Type(English)	Outline

TOKYO INSTITUTE OF TECHNOLOGY

DOCTORAL THESIS

A Study on Difference-of-Convex Type Nonconvexly Regularized Convex Models

Author:
Yi ZHANG

Supervisor:
Prof. Isao YAMADA

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Engineering

in the

Yamada Lab
Department of Information and Communications Engineering

May 28, 2024

Chapter 1

Introduction

1.1 Regularization Methods for Data Science

1.1.1 What is Regularization?

Regularization methods [1]–[5] stand out as indispensable tools in contemporary data science. In a broad sense, the term "regularization" pertains to the process that imparts greater regularity (or simplicity) to the result of a mathematical problem. However, within the realm of data science, our primary focus lies in the application of regularization methodologies to enhance the performance of mathematical optimization models.

To be precise, in many fields of data science (e.g., signal processing and machine learning), most tasks can be categorized into the following three types:

- **Inference**: given an observation of a mathematical object (e.g., vector, matrix, tensor, function, etc) and the underlying model (i.e., the relation between the object and its observation), the goal is to deduce the true value of the object from its observation.
- **learning**: given a number of instances of two related mathematical objects, the goal is to learn the relation between these two objects from the set of instances.
- **decision making**: given a set of rules that evaluate the consequence of the actions of a player, the goal is to find the optimal actions that the player should take to yield the best consequence.

For each type of aforementioned tasks, the problem can usually be transformed into a mathematical optimization model as follows:

$$\underset{x\in\mathcal{X}}{\text{minimize}} \;\; F(x) \tag{1.1}$$

where x is the value of the expected output (i.e., the target of inference, the relation function between two objects, or the actions that the player plans to take), $\mathcal{X}$ is a set that x may take its value from, and the real-valued function F is a data fidelity term which evaluates the quality of the current value x using the data or knowledge given by the original (inference/learning/decision making) task.

In order for the optimization model (1.1) to yield a meaningful result, we generally hope (1.1) to have some favourable mathematical properties (e.g., the solution set of (1.1) should be nonempty and bounded). Most importantly, we hope that every minimizer of (1.1) can serve as a qualified answer to the original task. However, in practice, such requirements usually cannot be satisfied for the following reasons:

- In inference tasks, the observation is usually noisy, and the amount of data can be small compared to that required for reconstructing the target object [6], [7].

- In learning or decision making tasks, the amount of training data or given knowledge is usually too small[1] compared to the complexity of the task [4], [9].

Accordingly, in practical applications, the plain optimization model (1.1) is usually ill-posed and the minimizer of (1.1) may behave poorly as an answer to the original task. These motivate the use of regularization methods, i.e., to consider the following regularized optimization problem instead:

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}} \;\; J(\boldsymbol{x}) := F(\boldsymbol{x}) + \lambda\Psi(\boldsymbol{x}). \tag{1.2}$$

Compared to the plain optimization model (1.1), the regularization model (1.2) introduces an additional term $\lambda\Psi(\cdot)$ in the cost function $J(\cdot)$. The term $\Psi(\cdot)$ is called a regularizer (or regularization term), which generally incorporates certain human knowledge about what a good answer of the original task should be like. The tuning parameter $\lambda > 0$ is called a regularization parameter, which trades off between the data discrepancy evaluated by $F(\cdot)$ and the prior knowledge of $\Psi(\cdot)$.

By introducing proper human knowledge in the regularizer $\Psi(\cdot)$, one can easily guarantee the nonemptiness and boundedness of the solution set of (1.2), and can rule out undesirable solutions (those that disagree with human intuition) in the plain optimization model (1.1). Hence the regularization model (1.2) usually proves to be much more stable than the plain optimization model (1.1) [3], [5].

1.1.2 Sparsity-Based Regularization

As one can imagine, the performance of the regularization model (1.2) heavily depends on the selection of the regularizer Ψ. Intuitively, human knowledge about the expected output should vary from task to task, hence the regularizer Ψ is also supposed to be case-specifically designed. In this view, it seems impossible to develop a universal strategy for introducing proper regularizers in (1.2). Nevertheless, in the past three decades [10], [11], it has been demonstrated that many real-world problems exhibit certain "sparse" structures, and by exploiting sparsity, it is possible to develop regularizers that prove effective across a wide range of data science applications.

Mathematically, we say a vector (or matrix, tensor) is "sparse" if most of its elements are zeros, indicating that its nonzero elements are sparsely distributed throughout the vector (or matrix, tensor). While sparsity may appear to be a stringent condition that can only be satisfied by a small family of vectors (or matrices, tensors) at first glance, surprisingly, it turns out that in many data science tasks, the expected output usually either exhibits sparsity itself or can be readily represented sparsely under a simple linear transform [10]–[13].

One typical field of application for sparsity-based regularization is signal processing [10], where many real-world signals are known to exhibit sparsity in certain transform domains (e.g., Fourier domain, wavelet domain, etc.). Notably, it has been established that if a band-limited signal is sparse in its Fourier domain, it can be exactly reconstructed at a sampling rate much lower than the minimum sampling rate required by the Nyquist-Shannon sampling theorem [14], [15]. This groundbreaking

[1] In the era of deep learning, one can easily find large datasets for several important machine learning tasks involving images, speech signals and texts [8]. However, there still remains many problems (e.g., in geological exploration/medical imaging) for which it is difficult to collect a large amount of data due to the high cost of data acquisition or privacy issues.

discovery is now referred to as "compressed sensing", and it has spawned numerous fascinating technologies in signal processing [16]. Additionally, the well-known total variation (TV [17]) model for reconstructing an image from its noisy observation can also be interpreted as a special sparse regularization model that exploits the sparsity of the gradient of the image signal.

Another important application of sparsity-based regularization is statistical learning [11], where the relation function between the covariates and the response variable is usually expected to be a simple function, i.e., the vector of regression coefficients should be sparse. The idea of exploiting sparsity has been a significant driving force behind the development of high-dimensional statistics [9] in the past two decades, leading to the creation of prominent models such as LASSO [18] and its variants [11].

Sparse regularization methods are also widely applied in training deep neural networks [19], [20], control systems and reinforcement learning [21], [22].

1.2 A Long-Standing Difficulty in Regularization Design

As introduced in Sec. 1.1.1, in the regularized optimization problem (1.2), the data fidelity term $F(\cdot)$ is usually given by the task that we want to resolve, whilst the regularizer $\Psi(\cdot)$ needs to be designed by human being. While designing a regularizer, we naturally hope that it possesses both statistical and computational advantages:

1) Statistically, $\Psi(\cdot)$ should correctly represent the prior knowledge that we want to exploit in the regularization model (1.2).

2) Computationally, introducing $\Psi(\cdot)$ into the cost function $J(\cdot)$ should not bring too much difficulty in solving the regularization model (1.2).

However, in practice, the two requirements above usually conflict with each other, and practitioners have to trade off between statistical and computational advantages. This constitutes a long-standing difficulty in regularization design.

In the sequel, as an example, we explain how such conflict happens for conventional sparse regularization models.

1.2.1 Conventional Sparse Regularizers and Their Limitations

Consider the following unconstrained regularization problem with sparse prior:

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \;\; J(\boldsymbol{x}) \coloneqq F(\boldsymbol{x}) + \lambda\Psi(\boldsymbol{x}), \tag{1.3}$$

where we assume F is convex[2] and compare properties of the regularization problem (1.3) with different choices of the sparse regularizer Ψ.

We start from a naive choice of Ψ. According to the definition of sparsity (see Sec. 1.1.2), a straightforward design for Ψ is the ℓ_0 pseudo-norm:

$$\Psi(\boldsymbol{x}) \coloneqq \|\boldsymbol{x}\|_0 \coloneqq \left|\{i \in \{1,2,\ldots,n\} \mid x_i \neq 0\}\right|,$$

i.e., the number of nonzero components in the input vector. However, this naive choice generally makes (1.3) a discontinuous (see the black curve in Fig. 1.1 for

[2]We note that the family of convex functions encompasses a large class of data fidelity terms encountered in real-world applications (e.g. generalized linear models, logistic regression; see [11, Sec. 2 and 3] for details), hence the convexity assumption of F does not lead to much limitation in the applicability of (1.3).

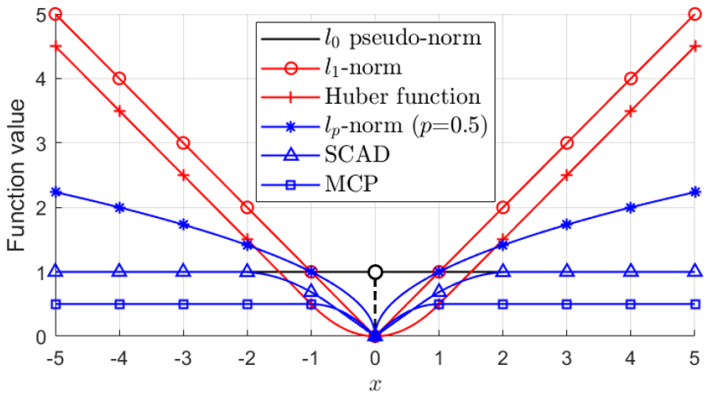

FIGURE 1.1: Conventional convex and nonconvex sparse regularizers (1d case).

illustration) nonconvex optimization problem which is known to be NP-hard [23], even if the data fidelity term F is a simple quadratic function. Accordingly, to yield a solvable regularization model, one has to resort to some continuous approximation of the ℓ_0 pseudo-norm.

Earlier studies typically adopt convex approximations of the ℓ_0 pseudo-norm (e.g., ℓ_1-norm [18], Huber function [24]; see the red curves in Fig. 1.1), which ensure efficient and reliable solution of (1.3). Especially, when F is quadratic and Ψ is the ℓ_1-norm, the resultant regularization model (1.3) reproduces the well-known LASSO model [18], which has achieved great success in a wide range of applications [11]. However, since most convex regularizers are coercive (i.e., $\Psi(\boldsymbol{x})$ goes to $+\infty$ if $\|\boldsymbol{x}\|_2$ goes to $+\infty$; see Fig. 1.1), they usually overpenalize the ith component x_i when $|x_i|$ is large, leading to underestimation of the true solution [25].

To reduce the estimation bias of convex models, continuous nonconvex regularizers (e.g., ℓ_p pseudo-norm with $0 < p < 1$, SCAD [26] and MCP [25]; see the blue curves in Fig. 1.1) have been proposed, which empirically show superior statistical performance [27]. Nevertheless, conventional nonconvex regularizers generally destroy the overall-convexity of (1.3), i.e., convexity of the cost function J. Hence while such nonconvex regularization models are much more tractable compared to the naive ℓ_0 regularization model, when adopted in application, existing optimization algorithms can possibly get stuck in local minima, which poses a concern on the reliability and efficiency of solving nonconvex regularization models.

1.2.2 Dilemma: Overall-Convexity or Representability?

From the discussion above, we can deduce the following characteristics of conventional convex and nonconvex regularizers:

1) Convex regularizers can attain overall-convexity of the regularization model (1.2) when the data fidelity term is convex, which leads to attractive computational properties. However, since the shape of convex functions is not so flexible as that of nonconvex functions, the representability of convex regularizers can be limited, which may yield less satisfactory statistical performance.

2) Nonconvex regularizers enjoy greater representability than convex ones, leading to attractive statistical properties. But they usually sacrifice overall-convexity of the regularization model (1.2), thereby losing computational advantages of convex regularization models.

Accordingly, for conventional approaches of regularization design, practitioners typically face a dilemma of choosing either overall-convexity of the regularization model or representability of the regularizer. In consequence, conventional regularization models generally cannot possess good computational properties and statistical properties at the same time.

1.3 Nonconvexly Regularized Convex Models

In order to resolve the dilemma between overall-convexity and representability, particular convexity-preserving (CP) regularizers have been proposed to yield improved convex regularization [28]–[33]. The so-called CP regularizer is a special parameterized nonconvex regularizer. Although the CP regularizer itself is nonconvex, its shape can be adjusted by certain tuning parameter so as to induce the overall-convexity[3] of the regularization problem (1.2), leading to an unusual nonconvexly regularized convex (NRC) model. Therefore, different from conventional convex and nonconvex regularization models, one can imagine that NRC models enjoy both computational and statistical advantages.

In the sequel, we briefly introduce the development of CP regularizers and NRC models, and elucidate their underlying mechanism. Our analysis will show that most CP regularizers and cost functions of NRC models are difference-of-convex (DC [34]) functions, i.e., they can be decomposed as the difference between two convex functions (see Sec. 2.2.1 for details). This observation naturally leads to a special interest in such DC type NRC models, i.e., our central object of study in this thesis.

1.3.1 Early Studies

The idea of CP regularizers and NRC models dates back to over three decades ago [28]–[30]. However, early studies [28]–[30], [35], [36] usually assume the presence of a strongly convex term in the regularization problem (1.2) (e.g., a strongly convex data fidelity term F or ℓ_2-norm contained in the regularization term Ψ), hence are fundamentally limited. For example, in [29], the author proposed the following nonconvex regularizer:

$$\Psi_{\mathrm{BI}}\left(\boldsymbol{x};\alpha\right)=\sum_{i\sim j}\beta_{i,j}|x_i-x_j|-\alpha\sum_i\left(x_i-\frac{1}{2}\right)^2 \tag{1.4}$$

for estimating binary images[4], where $\boldsymbol{x}\coloneqq\left[x_1,\ldots,x_n\right]^\top\in\mathbb{R}^n$ is the estimate of the unknown binary image, $i\sim j$ means that x_i and x_j are neighbouring pixels, $\boldsymbol{\beta}\coloneqq\left(\beta_{i,j}\right)_{i\sim j}\subset\mathbb{R}_+$, $\alpha>0$ is the shape-controlling tuning parameter.

Since the first term in (1.4) is convex and the second is concave, Ψ_{BI} is the difference between two convex functions, i.e., Ψ_{BI} is a DC function. Let us consider the quadratic data fidelity $F_{\mathrm{quad}}(\boldsymbol{x})\coloneqq\frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\right\|_2^2$ and the following cost function

$$J_{\mathrm{BI}}(\boldsymbol{x};\alpha)\coloneqq F_{\mathrm{quad}}(\boldsymbol{x})+\lambda\Psi_{\mathrm{BI}}(\boldsymbol{x};\alpha).$$

then since F_{quad} is convex, it is evident that J_{BI} is also a DC function.

[3] More precisely, given a proper data fidelity term F and weight parameter λ in (1.2), one can always find some proper tuning parameter value such that the resultant cost function J in (1.2) is convex.

[4] One can verify that the first term of Ψ_{BI} is a convex term which promotes the correlated structure of the image, and the second term is a concave term which promotes the binarity of pixels.

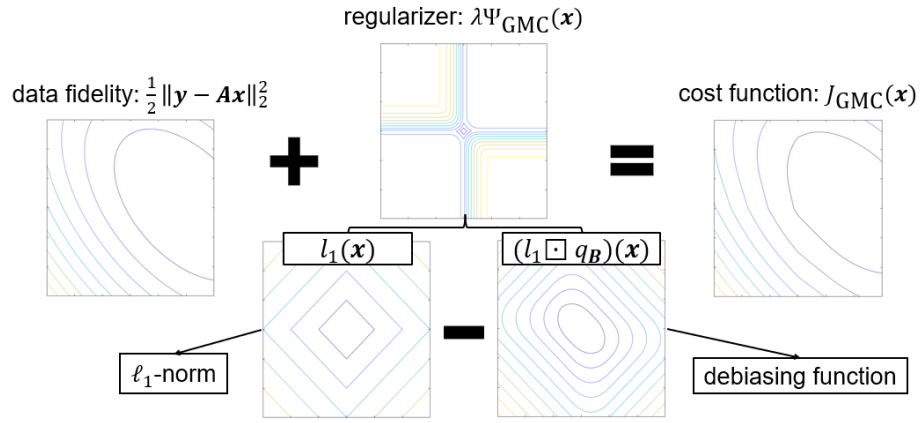

FIGURE 1.2: Illustration of the GMC model (2d case, where the contour graph of each component function is plotted).

It has been proven in [29] that if the shape-controlling parameter α satisfies[5]

$$\lambda_{\min}(\boldsymbol{A}^\top\boldsymbol{A}) \geq \lambda\alpha > 0, \tag{1.5}$$

then the concave second term of $\lambda\Psi_{\mathrm{BI}}$ would be overpowered by F_{quad}, whereby the cost function $J_{\mathrm{BI}}(\cdot;\alpha)$ is convex.

This example reveals an important fact, that is: *if there exists a strongly convex term in the cost function J, then we can introduce some concave terms into the regularizer Ψ to improve its regularizing properties while maintaining the overall-convexity of the cost function*. However, (1.5) implies the nonsingularity of $\boldsymbol{A}^\top\boldsymbol{A}$, which usually fails to hold in applications such as sparse signal reconstruction [14], [15].

1.3.2 The Generalized Minimax Concave (GMC) Model

The first CP regularizer that does not require strong convexity of F_{quad} (i.e., nonsingularity of $\boldsymbol{A}^\top\boldsymbol{A}$) is the generalized minimax concave (GMC) penalty [31] proposed by I. Selesnick in 2017:

$$\Psi_{\mathrm{GMC}}(\boldsymbol{x};\boldsymbol{B}) := l_1(\boldsymbol{x}) - (l_1 \square q_{\boldsymbol{B}})(\boldsymbol{x}), \tag{1.6}$$

where $l_1(\boldsymbol{x}) := \|\boldsymbol{x}\|_1$ is the ℓ_1-norm, $q_{\boldsymbol{B}}(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{B}\boldsymbol{x}\|_2^2$ is a quadratic function with $\boldsymbol{B} \in \mathbb{R}^{p\times n}$ being the shape-controlling tuning parameter, $\square$ is the infimal convolution operator[6]. Indeed, the subtrahend function in (1.6) can be regarded as a convex smooth approximation of the ℓ_1-norm, where $q_{\boldsymbol{B}}$ serves as a smoothing function (see Sec. 2.3.1 for details), hence as Ψ_{BI} in (1.4), Ψ_{GMC} is also a DC function. From a similar discussion as above, one can verify that the following cost function

$$J_{\mathrm{GMC}}(\boldsymbol{x};\boldsymbol{B}) := F_{\mathrm{quad}}(\boldsymbol{x}) + \lambda\Psi_{\mathrm{GMC}}(\boldsymbol{x};\boldsymbol{B}) := \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\Psi_{\mathrm{GMC}}(\boldsymbol{x};\boldsymbol{B}) \tag{1.7}$$

is also a DC function.

We note that the GMC penalty is a nonseparable[7] multidimensional generalization of the minimax concave penalty (MCP [25]), more precisely, if $\boldsymbol{B}^\top\boldsymbol{B}$ is diagonal,

[5] $\lambda_{\min}(\cdot)$ is the smallest eigenvalue of the input matrix.

[6] For $f, g : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$, $(f\square g)(\boldsymbol{x}) := \inf_{\boldsymbol{z}\in\mathbb{R}^n}(f(\boldsymbol{z}) + g(\boldsymbol{x}-\boldsymbol{z}))$.

[7] A multivariate function $f : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is separable if there exists n univariate functions $f_1, f_2, \ldots, f_n$ such that $f(\boldsymbol{x}) = \sum_{i=1}^n f_i(x_i)$.

then Ψ_{GMC} reproduces a weighted sum of MCP, which accounts for the name of the GMC penalty. In contrast to the standard MCP, the shape of $\Psi_{\mathrm{GMC}}(\cdot;\boldsymbol{B})$ can be adjusted flexibly via changing $\boldsymbol{B}$. Particularly, for the cost function J_{GMC} in (1.7), it is proved in [31] that if $\boldsymbol{B}$ satisfies

$$\boldsymbol{A}^\top\boldsymbol{A} \succeq \lambda\boldsymbol{B}^\top\boldsymbol{B}, \tag{1.8}$$

then the concave term $-\lambda(l_1\square q_{\boldsymbol{B}})(\boldsymbol{x})$ is overpowered by $F_{\mathrm{quad}}(\boldsymbol{x})$ (see Fig. 1.2 for illustration), and the cost function $J_{\mathrm{GMC}}(\cdot;\boldsymbol{B})$ is convex.

Remarkably, (1.8) does not require $\boldsymbol{A}^\top\boldsymbol{A}$ to be nonsingular as (1.5) does. Instead, by (1.8), Ψ_{GMC} is able to exploit the "partially strong convexity" of F_{quad}, more precisely, *if the data fidelity term is strongly convex in certain direction (e.g. the eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$ with nonzero eigenvalues), then one can introduce concavity into Ψ_{GMC} in this direction to achieve better approximation of the ℓ_0 pseudo-norm.*

1.3.3 Extensions of the GMC Model

Certain efforts have been made to broaden applicability of Ψ_{GMC}. One notable extension is the linearly involved generalized Moreau enhanced (LiGME) model [32], [37]:

$$\Psi_{\mathrm{LiGME}}(\boldsymbol{x};\boldsymbol{B}) = \psi(\boldsymbol{L}\boldsymbol{x}) - (\psi\square q_{\boldsymbol{B}})(\boldsymbol{L}\boldsymbol{x}), \tag{1.9}$$

where $\boldsymbol{L} \in \mathbb{R}^{q\times n}$ is the analysis matrix which encodes the sparsifying domain of the interested signal (e.g., Fourier matrix, wavelet matrix, discrete difference operator), $\boldsymbol{B} \in \mathbb{R}^{p\times q}$ is the shape-controlling tuning parameter; $\psi \in \Gamma_0(\mathbb{R}^q)$ is a convex[8] kernel function which is no longer restricted to the ℓ_1-norm, but can be any proximable[9] function. Accordingly, the LiGME model allows applying the construction technique of the GMC penalty to more general convex kernel functions. A variant of the LiGME model with split feasibility type constraints [10] is studied in [39].

Another useful extension of GMC is the sharpening sparse regularizers (SSR) framework [33]:

$$\Psi_{\mathrm{SSR}}(\boldsymbol{x};\boldsymbol{B}) := l_1(\boldsymbol{x}) - ((l_1\circ\boldsymbol{L})\square(\Phi\circ\boldsymbol{B}))(\boldsymbol{x}), \tag{1.10}$$

where $\boldsymbol{L} \in \mathbb{R}^{q\times n}$ is the analysis matrix which is embedded at a different position from the LiGME model, $\boldsymbol{B} \in \mathbb{R}^{p\times n}$ is the shape-controlling tuning parameter, $\Phi(\boldsymbol{z}) := \sum_{i=1}^{p}\phi(z_i)$ with $\phi \in \Gamma_0(\mathbb{R})$ is an isotropic smoothing function which is not restricted to the ℓ_2-norm. While the SSR model does not consider variability of the kernel function, it allows adopting a different smoothing function Φ, thus can adjust the shape of the regularizer more delicately.

So far, overall-convexity conditions and proximal splitting type [40] algorithms have been developed independently with respect to the GMC [31], LiGME [32] and SSR [33] models.

---

[8] More precisely, $\Gamma_0(\mathbb{R}^q)$ is the set of all proper, lower semicontinuous convex function from $\mathbb{R}^q$ to $\mathbb{R}\cup\{+\infty\}$; see Sec. 2.1.2 for details.

[9] For f in $\Gamma_0(\mathbb{R}^n)$, we say that f is *proximable* if its proximity operator $\mathrm{Prox}_{\gamma f}(\boldsymbol{x}) := \arg\min_{\boldsymbol{z}\in\mathbb{R}^n}\left[\gamma f(\boldsymbol{z}) + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2\right]$ can be computed to high precision efficiently for every $\gamma > 0$.

[10] A constraint set C is of split feasibility type [38] if C can be rewritten as

$$C := \{\boldsymbol{x}\in\mathbb{R}^n \mid \boldsymbol{A}_i\boldsymbol{x}\in C_i, \text{ for } i = 1,\dots,s\}.$$

where for $i = 1,\dots,s$, $\boldsymbol{A}_i$ is a linear operator and C_i is a "simple" nonempty closed convex set (by "simple", we mean the projection onto C_i can be computed to high precision efficiently).

1.3.4 DC Type Nonconvexly Regularized Convex Models

From aforementioned examples of CP regularizers and NRC models, one can verify that all of them are DC functions and their DC structure plays a critical role in attaining overall-convexity. This naturally raises our interest in studying the class of NRC models that have DC type regularizers and cost functions.

Here we would like to present a brief summary on how a DC type NRC model works, which gives a clearer image of the central object of study in this thesis. For a given real-world problem, it generally takes three steps to construct a DC type NRC model for solving the target problem:

1) We need to find a proper regularizer Ψ to encode the prior information about the true solution. Notably, a DC type CP regularizer Ψ should be a parameterized DC function of the following form:

$$\Psi(\boldsymbol{x};\mathcal{P}) \coloneqq \Psi_1(\boldsymbol{x}) - \Psi_2(\boldsymbol{x};\mathcal{P}), \tag{1.11}$$

where $\Psi_1(\cdot)$ and $\Psi_2(\cdot;\mathcal{P})$ are convex, $\mathcal{P}$ is a tuning parameter (which can be a vector, matrix or tensor) that can adjust the shape of $\Psi_2(\cdot;\mathcal{P})$ flexibly.

2) Combining the data fidelity term F (which is assumed to be convex) and the regularizer Ψ, we can obtain the cost function J of the regularization model (1.2), which again is a parameterized DC function:

$$J(\boldsymbol{x};\mathcal{P}) \coloneqq F(\boldsymbol{x}) + \lambda\Psi(\boldsymbol{x}) = (F(\boldsymbol{x}) + \lambda\Psi_1(\boldsymbol{x})) - \lambda\Psi_2(\boldsymbol{x};\mathcal{P}). \tag{1.12}$$

3) For a given regularization parameter $\lambda > 0$ to work with, we need to find a proper value of the shape-controlling parameter $\mathcal{P}$ such that the subtrahend function $\lambda\Psi(\cdot;\mathcal{P})$ in (1.12) can be dominated by the minuend function $(F(\cdot) + \lambda\Psi_1(\cdot))$, whereby the overall-convexity of the cost function $J(\cdot;\mathcal{P})$ can be attained.

From the analysis above, one can imagine that the idea of DC type NRC models is indeed not limited to sparse regularization, but can be applied to any regularization problem in principle.

1.4 Organization

The favourable properties of DC type NRC models motivate the current study, aiming to address fundamental problems encountered in the design, solution and practical application of this unusual and attractive class of regularization models. More precisely, we consider the following four issues:

1) **Design of DC type NRC models:** for an arbitrary regularization problem, is there a general approach to design proper DC type CP regularizers and NRC models for solving the target problem? Moreover, given the value of the regularization parameter λ in a DC type NRC model, how should we select the shape-controlling parameter to yield overall-convexity?

2) **Optimization algorithm for DC type NRC models:** is there a unified approach for solving general DC type NRC models? If such approach exists, how reliable and how efficient is it?

3) **Statistical analysis of DC type NRC models:** what do we know about statistical performance of DC type NRC models? Interesting questions in this aspect include, e.g., does every solution of a DC type NRC model serve as a good answer to the original problem we want to solve? How does the performance of a DC type NRC model change with the regularization parameter λ and how should we select the optimal λ?

4) **Extensions for stochastic regularization problems:** if the data fidelity term is a random function (e.g., adaptive filtering problems [41]), can we apply the idea of DC type NRC models to such stochastic regularization problems?

This study provides very encouraging results for every issue mentioned above:

1) In Chapter 3, we propose a general framework for designing DC type CP regularizers and NRC models, which provides a widely useful approach for resolving the issue 1). Additionally, we propose a unified solution algorithm based on DC optimization theory (cf. Sec. 2.2) for solving the proposed class of DC type NRC models, which partially addresses the issue 2).

2) In Chapter 4, we point out a practical problem (i.e., the inner loop terminating issue; see Sec. 4.3 for details) that may be encountered in applying DC optimization algorithms, and we propose a novel DC algorithm for resolving it. The proposed DC algorithm further refines our solution algorithm for DC type NRC models in Chapter 3 and completes our answer to the issue 2).

3) In Chaper 5, we focus on a representative DC type NRC model termed sGMC model (which is a significant instance of the GMC model introduced in Sec. 1.3.2; see Sec. 2.4.1 for details) for sparse linear regression, and we analyze its solution-set geometry and regularization path. Our study indicates that despite nonconvexity of the sGMC penalty, the sGMC model preserves all the celebrated properties of the conventional LASSO [18] model, hence can serve as a less biased surrogate of LASSO. Notably, our study reveals a couterintuitive fact: while the sGMC penalty is a nonconvex extension of the LASSO penalty, the minimum ℓ_2-norm sGMC regularization path[11] remains to be piecewise linear in the regularization parameter λ. Based on this finding, we propose an efficient iterative algorithm for computing the entire minimum ℓ_2-norm regularization path, which is useful in finding the optimal λ for the sGMC model. Chapter 5 addresses the issue 3) for the sGMC model.

4) In Chapter 6, we further discuss extensions of DC type NRC models to sparse adaptive filtering problems. It turns out that in the realm of stochastic regularization problems, it can be difficult to exploit the overall-convexity of an NRC model. However, the DC optimization techniques studied in this thesis is still useful, based on which we propose a less biased sparse adaptive filtering algorithm exploiting the DC structure of a novel sparse regularizer. Chapter 6 addresses issue 4) for sparse adaptive filtering problems.

Finally, Chapter 7 concludes the results obtained in this thesis.

[11] A regularization path (or solution path) of a regularization model (1.2) is the solution of (1.2) as a function of the regularization parameter λ.

List of Publications

Journal Papers

Published:

1. Yi Zhang and Isao Yamada, "A Unified Framework for Solving a General Class of Nonconvexly Regularized Convex Models," in *IEEE Transactions on Signal Processing*, vol. 71, pp. 3518-3533, 2023.

2. Yi Zhang and Isao Yamada, "An Inexact Proximal Linearized DC Algorithm with Provably Terminating Inner Loop," in *Optimization*, pp. 1-33, jan 2024 (published online).

   doi: 10.1080/02331934.2024.2314241

Preprint:

3. Y. Zhang and I. Yamada, "Solution-Set Geometry and Regularization Path of a Nonconvexly Regularized Convex Sparse Model." arXiv, Mar. 22, 2024. Available: http://arxiv.org/abs/2311.18438

International Conference

1. Yi Zhang and Isao Yamada, "DC-LiGME: An Efficient Algorithm for Improved Convex Sparse Regularization," in *Proceedings of 55th Asilomar Conference on Signals, Systems, and Computers*, 2021.

2. Yi Zhang and Isao Yamada, "A Unified Class of DC-type Convexity-Preserving Regularizers for Improved Sparse Regularization," in *Proceedings of 30th European Signal Processing Conference (EUSIPCO)*, 2022.

3. Yi Zhang and Isao Yamada, "A Compensated Shrinkage Affine Projection Algorithm for Debiased Sparse Adaptive Filtering," in *Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

4. Yi Zhang and Isao Yamada, "Computing an Entire Solution Path of a Nonconvexly Regularized Convex Sparse Model," in *Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Domestic Conference

1. Yi Zhang and Isao Yamada, "A DC-programming based Fast Iterative Algorithm for Generalized Moreau Enhanced Models", in *Proceedings of 36th IEICE SIP Symposium*, 2021.

2. Yi Zhang and Isao Yamada, "A Fast DC Algorithm for a Unified Class of Convexity-Preserving Sparse Regularizers,"日本オペレーションズ・リサーチ学会 2022年秋季研究発表会&シンポジウム, 2022.

3. Y. Zhang and I. Yamada, "A Debiased Sparseness-Promoting Affine Projection Algorithm Based on Nonconvex Proximal Gradient Method", in *Proceedings of 37th IEICE SIP Symposium*, 2022.

Award

1. 電子情報通信学会 信号処理研究専門委員会 SIP 若手奨励賞 (36th IEICE SIP Symposium, Nov. 2021).