mathematics

MDPI

Article

Generalized Penalized Constrained Regression: Sharp Guarantees in High Dimensions with Noisy Features

Ayed M. Alrashdi [1,*], Meshari Alazmi [2] and Masad A. Alrasheedi [3]

[1] Department of Electrical Engineering, College of Engineering, University of Ha'il, Ha'il 81441, Saudi Arabia
[2] Department of Information and Computer Science, College of Computer Science and Engineering, University of Ha'il, Ha'il 81411, Saudi Arabia; ms.alazmi@uoh.edu.sa
[3] Department of Management Information Systems, College of Business Administration, Taibah University, Madinah 42353, Saudi Arabia; mrshedi@taibahu.edu.sa
* Correspondence: ayed.alrashdi@kaust.edu.sa

Abstract: The generalized penalized constrained regression (G-PCR) is a penalized model for high-dimensional linear inverse problems with structured features. This paper presents a sharp error performance analysis of the G-PCR in the over-parameterized high-dimensional setting. The analysis is carried out under the assumption of a noisy or erroneous Gaussian features matrix. To assess the performance of the G-PCR problem, the study employs multiple metrics such as prediction risk, cosine similarity, and the probabilities of misdetection and false alarm. These metrics offer valuable insights into the accuracy and reliability of the G-PCR model under different circumstances. Furthermore, the derived results are specialized and applied to well-known instances of G-PCR, including ℓ_1-norm penalized regression for sparse signal recovery and ℓ_2-norm (ridge) penalization. These specific instances are widely utilized in regression analysis for purposes such as feature selection and model regularization. To validate the obtained results, the paper provides numerical simulations conducted on both real-world and synthetic datasets. Using extensive simulations, we show the universality and robustness of the results of this work to the assumed Gaussian distribution of the features matrix. We empirically investigate the so-called double descent phenomenon and show how optimal selection of the hyper-parameters of the G-PCR can help mitigate this phenomenon. The derived expressions and insights from this study can be utilized to optimally select the hyper-parameters of the G-PCR. By leveraging these findings, one can make well-informed decisions regarding the configuration and fine-tuning of the G-PCR model, taking into consideration the specific problem at hand as well as the presence of noisy features in the high-dimensional setting.

Keywords: penalized regression; prediction risk; cosine similarity; probability of false alarm; double descent; over-parameterization; constrained ridge regression

MSC: 62J05; 62J07; 60G35; 62E20

Citation: Alrashdi, A.M.; Alazmi, M.; Alrasheedi, M.A. Generalized Penalized Constrained Regression: Sharp Guarantees in High Dimensions with Noisy Features. *Mathematics* **2023**, *11*, 3706. https://doi.org/10.3390/math11173706

Academic Editors: Jinwen Ma and Behzad Pirouz

Received: 25 June 2023
Revised: 3 August 2023
Accepted: 24 August 2023
Published: 28 August 2023

1. Introduction

1.1. Notations and Definitions

To avoid confusion, we start by introducing the notations and definitions used throughout this paper. For any positive integer p, let $[p]$ denote the set $\{1, 2, \cdots, p\}$. Bold face lower case letters (e.g., $\boldsymbol{\theta}$) represent a column vector, and θ_i is its ith entry, while $\|\boldsymbol{\theta}\|_q = \left(\sum_{i=1}^{p} |\theta_i|^q\right)^{\frac{1}{q}}$ is its ℓ_q-norm. The ℓ_∞-norm of a vector is defined as: $\|\boldsymbol{\theta}\|_\infty = \max_i |\theta_i|$. Upper case bold letters such as $\mathbf{X}$ are used to indicate matrices, with $\mathbf{I}_p$ representing the $p \times p$ identity matrix. The symbols $(\cdot)^{-1}$ and $(\cdot)^\top$ are the inversion and transpose operations, respectively. We use $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$ to indicate the probability of an event and the expected value of a random variable, respectively. The notation " $\xrightarrow{P}$ " is

used to represent convergence in probability. We write $X \sim p_X$ to indicate that a random variable X is randomly distributed according to a probability mass (or density) function p_X. Particularly, $\mathbf{v} \sim \mathsf{N}(\mathbf{0}, \mathbf{C_v})$ means that the random vector $\mathbf{v}$ has a normal distribution with $\mathbf{0}$ mean vector and covariance matrix $\mathbf{C_v} = \mathbb{E}[\mathbf{v}\mathbf{v}^\top]$, where $\mathbf{0}$ is the zero vector. For $m \in \mathbb{N}$, a function $\psi : \mathbb{R}^m \to \mathbb{R}$ is said to be pseudo-Lipschitz of order $k \geq 1$ if there exists a constant $L > 0$ such that, for all $\boldsymbol{x}, \mathbf{y} \in \mathbb{R}^m$: $|\psi(\boldsymbol{x}) - \psi(\mathbf{y})| \leq L(1 + \|\boldsymbol{x}\|_2^{k-1} + \|\mathbf{y}\|_2^{k-1})\|\boldsymbol{x} - \mathbf{y}\|_2$.

A function $\mathcal{P} : \mathbb{R}^p \to \mathbb{R}$ is called *separable* if $\mathcal{P}(\boldsymbol{x}) = \sum_{j=1}^{p} \widetilde{\mathcal{P}}(x_j) \ \ \forall \boldsymbol{x} \in \mathbb{R}^p$, where $\widetilde{\mathcal{P}} : \mathbb{R} \to \mathbb{R}$ is a real-valued function. The notation $\mathbf{1}_{\{\mathcal{A}\}}$ is the indicator function, which is defined as

$$\mathbf{1}_{\{\mathcal{A}\}}(x) = \begin{cases} 1, & \text{if} \quad x \in \mathcal{A}, \\ 0, & \text{otherwise.} \end{cases}$$

Finally, we need the following definitions:

- The **generalized Moreau envelope** function of a proper convex function $h : \mathbb{R} \to \mathbb{R}$ is defined as

$$\mathsf{M}_h(a; b, c, d) = \min_{c \leq x \leq d} \frac{1}{2}(x - a)^2 + b\, h(x) \tag{1}$$

  for $a, b, c, d \in \mathbb{R}$, with $b \geq 0, c \leq 0$ and $d \geq 0$. The generalized Moreau envelope given above is an extended version of the well-known Moreau–Yosida envelope function [1].
- The minimizer of the above function is called the **generalized proximal operator**, which is given as

$$\mathrm{prox}_h(a; b, c, d) = \arg\min_{c \leq x \leq d} \frac{1}{2}(x - a)^2 + b\, h(x).$$

1.2. Motivation

Suppose we observe a response vector $\mathbf{y} \in \mathbb{R}^n$ and a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ according to the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_0 + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is a vector of coefficients or parameters, or an unknown signal vector, and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is an error vector. This is also known as a linear inverse problem model [2]. The linear model in (2) appears in many practical problems in engineering and science [3,4]. For example, in *statistics and machine learning* [5–7], $\mathbf{y}$ is the response vector (or the output data); $\mathbf{X}$ is often called the predictor matrix, features matrix, or design matrix, which collects the input data (or features); $\boldsymbol{\theta}_0$ is the so called target vector, which is a vector of some weighting parameters or regression coefficients; and $\boldsymbol{\epsilon}$ is a random noise term. In the context of *compressed sensing* [8,9], $\mathbf{y}$ represents the measured data, $\mathbf{X}$ is a sensing or measurement matrix, $\boldsymbol{\theta}_0$ denotes a signal of interest (to be recovered), and $\boldsymbol{\epsilon}$ is a random noise vector. In addition, in *signal representation* [10,11], $\mathbf{y}$ is a signal of interest, the matrix $\mathbf{X}$ denotes an over-complete dictionary of elementary atoms, the vector $\boldsymbol{\theta}_0$ contains the representation coefficients of the signal $\mathbf{y}$, and $\boldsymbol{\epsilon}$ represents some approximation error. Moreover, in the field of *wireless communications* [12,13], $\mathbf{y}$ represents the received signal, $\mathbf{X}$ is the channel matrix between the transmitter and the receiver, $\boldsymbol{\theta}_0$ is the transmitted signal vector, and $\boldsymbol{\epsilon}$ is the additive thermal noise.

In the past, different computational algorithms have been proposed for recovering (estimating) the unknown vector $\boldsymbol{\theta}_0$. The simplest and most conspicuous approach is the ordinary least squares (OLS) estimator, which finds an estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ by minimizing the residual sum of squares (RSS), i.e.,

$$\widehat{\boldsymbol{\theta}}_{\text{OLS}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2. \tag{3}$$

For the OLS estimator, it is required that $n \geq p$, i.e., $\mathbf{X}$ is a full column rank matrix. In this case, (3) has the following closed-form solution:

$$\widehat{\boldsymbol{\theta}}_{\text{OLS}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{4}$$

In many applications, most of the time, the number of parameters to be recovered p is greater than the number of available samples n, i.e., $p > n$. This scenario is known in the literature as the *over-parameterized* regime [14] (This case, $n < p$, is also called the "compressed measurement" scenario in the compressive sensing context). Such inverse problems are known to be ill-posed unless the unknown vector $\boldsymbol{\theta}_0$ is located in a manifold with a considerably lower dimension than the initial ambient dimension p. These vectors are called *structured vectors* [15]. Examples of structured vectors are vectors with a finite-alphabet structure, sparse and block-sparse structures, low-rankness, etc. [9].

Despite being a popular approach, the OLS estimator performs very poorly when applied to ill-posed or under-determined problems [16]. Thus, to solve ill-posed problems, penalization methods are often used. Examples of these methods include penalized least squares (PLS) [17], least absolute shrinkage and selection operator (LASSO) [18], truncated singular value decomposition (SVD) [19], etc.

For structured vectors, the most widely used approach is the penalized M-estimator [20], which finds an estimate $\widehat{\boldsymbol{\theta}}$ of the unknown vector $\boldsymbol{\theta}_0$ by solving the convex optimization problem

$$\widehat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathfrak{L}(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) + \alpha \, \mathcal{P}(\boldsymbol{\theta}), \tag{5}$$

where $\mathfrak{L} : \mathbb{R}^n \to \mathbb{R}$ is a convex *loss* function that determines how close the estimate $\mathbf{X}\widehat{\boldsymbol{\theta}}$ is to the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_0 + \boldsymbol{\epsilon}$. Furthermore, $\mathcal{P} : \mathbb{R}^p \to \mathbb{R}$ is a convex *penalization* function that enforces the specific structure (the a priori information) of the unknown vector $\boldsymbol{\theta}_0$, and $\alpha > 0$ is a penalization factor that is used to balance the two functions. In addition, we assume that $\mathcal{P}$ is separable, i.e., $\mathcal{P}(\boldsymbol{\theta}) = \sum_{j=1}^{p} \widetilde{\mathcal{P}}(\theta_j)$. Examples of the most popular structure-inducing functions are:

- $\mathcal{P}(\cdot) = \|\cdot\|_1$ induces **sparsity** structure.
- $\mathcal{P}(\cdot) = \|\cdot\|_\star$ encourages **low-rankness** structure, where $\|\cdot\|_\star$ is the nuclear norm of a matrix, which is defined as the sum of its singular values.
- $\mathcal{P}(\cdot) = \|\cdot\|_{1,2}$ induces **block-sparsity** structures, where $\|\cdot\|_{1,2}$ is the mixed $\ell_{1,2}$-norm.
- $\mathcal{P}(\cdot) = \|\cdot\|_\infty$ promotes **finite-alphabet** (i.e., constant-amplitude) signals.

The choice of the loss function $\mathfrak{L}(\cdot)$ depends on the noise distribution [3] as follows:

- If the noise is Gaussian-distributed, then we choose $\mathfrak{L}(\cdot) = (1/2)\|\cdot\|_2^2$ or $\mathfrak{L}(\cdot) = \|\cdot\|_2$, which is related to the maximum likelihood estimation [10].
- If the noise is sparse (e.g., Laplacian distributed), then one can select $\mathfrak{L}(\cdot) = \|\cdot\|_1$.
- If the noise is bounded, then a proper choice is $\mathfrak{L}(\cdot) = \|\cdot\|_\infty$, and so on.

Different popular algorithms that correspond to different choices of $\mathfrak{L}(\cdot)$ and $\mathcal{P}(\cdot)$ include:

- OLS: $\mathcal{P}(\cdot) = 0$, as in (3);
- ℓ_2-penalized LS or ridge regression: $\min_{\boldsymbol{\theta}} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \alpha \, \|\boldsymbol{\theta}\|_2^2$;
- ℓ_1-penalized LS or LASSO: $\min_{\boldsymbol{\theta}} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \alpha \, \|\boldsymbol{\theta}\|_1$;
- group LASSO: $\min_{\boldsymbol{\theta}} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \alpha \|\boldsymbol{\theta}\|_{1,2}$;
- generalized least absolute deviation (LAD): $\min_{\boldsymbol{\theta}} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_1 + \alpha \, \mathcal{P}(\boldsymbol{\theta})$.

The above list is not exhaustive, and many regression, classification, and other statistical learning algorithms can be written in the form of (5).

1.3. Summary of Contributions and Related Work

Since the Gaussian noise is the most widely encountered noise in practice, we focus in this work on optimization problems involving an ℓ_2-norm squared loss and a general

penalization function. We call these problems the *Generalized Penalized Regression (G-PR)*, which solves the following convex optimization problem:

$$\widehat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \alpha \, \mathcal{P}(\boldsymbol{\theta}). \tag{6}$$

In this paper, we provide high-dimensional analysis of a constrained version of the G-PR called the G-PCR (as in Equation (9)) based on the convex Gaussian min–max theorem (CGMT) [20]. This analysis includes studying its general error performance and specializing it to particular cases such as sparse and ridge linear regressions. The derived performance measures, such as the prediction risk and similarity and probability of misdetection, are then used to tune the involved hyper-parameters of the algorithm. Numerical simulations of both synthetic and real data are presented to support the theoretical analysis presented in this work.

Previous works on the high-dimensional performance characterization of convex optimization problems have a very rich history. There are early results that provided order-wise "loose bounds" of the error performance of several penalized regression problems, such as in [14,21–25]. However, the first results that provided a high-dimensional error analysis were derived using the approximate message passing (AMP) algorithm by Bayati et al. [26,27] for the unconstrained (standard) LASSO. Later, Ref. [28] extended the AMP framework to analyze the performance of more general loss functions.

A different approach that is based on the replica method was considered in [29,30] to analyze various problems in the compressed sensing setting.

In addition, another powerful high-dimensional tool called the random matrix theory (RMT) [31] was used in [32–34] to derive asymptotic error analysis of some optimization problems that possess closed-form solutions.

Recently, Thrampoulidis et al. developed a new high-dimensional analysis framework that is based on the convex Gaussian min–max theorem (CGMT). First, this framework was used in [35–38] to provide precise error analysis of the LASSO and square-root LASSO. Then, in [20], it was extended to obtain asymptotic error performance analysis of unconstrained penalized M-estimator regression problems. The first CGMT-based results on constrained regression models were derived in [39–41] for the box relaxation optimization (BRO) and its regularized variant. This BRO method is used to promote constant-amplitude structures. The authors in [42–44] extended the previous CGMT results to obtain sharp error performance characterization of constrained versions of the popular LASSO and Elastic-Net (EN) problems. These extended versions are called the Box-LASSO and Box-EN, respectively. Furthermore, the authors in [45,46] extended the above results to derive symbol error rate performance of a more general method called the sum of absolute values (SOAV) optimization and its constrained pair (Box-SOAV) for discrete-valued binary and sparse signal recovery.

Even though the focus of this paper is on regression problems, we should highlight that the CGMT framework was also applied to characterize the high-dimensional error performance of classification problems as in [47–49], phase retrieval problems [50,51], and various statistical learning problems [52–54].

In most of these works, the features matrix is considered to be fully known, but in practice, data are always noisy and contain different types of errors. This motivates the analysis considered in this paper to be performed under uncertainties in the design matrix (see Section 2.2). As compared to related work, such as [41,43,44] which considered the imperfect design matrix assumption, this work differs in multiple ways.

- The proposed constrained G-PCR problem in (9) considers a general penalization function $\mathcal{P}(\cdot)$ instead of the specific penalties used in previous works.
- This work derives a general performance measure (Theorem 1) that is more broad and useful than the particular metrics previously taken into consideration, such as the mean square error (MSE), symbol error probability, etc.

- This work generalizes these previous results, as they can be obtained as special cases of the results of this paper.
- In Appendix B, we highlight the use of the same machinery developed in this work to analyze a closely related class of problems known as Square-Root Generalized Penalized Constrained Regression.

2. Problem Setup

2.1. Dataset Model

Consider the problem of estimating a scalar response y_i from a set of n independent *training* data samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the feature vector, following the linear model

$$y_i = \boldsymbol{\theta}_0^\top \boldsymbol{x}_i + \epsilon_i, \; i \in [n], \tag{7}$$

where $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is an unknown *structured* target vector, and $\{\epsilon_i\}_{i=1}^n$ denotes the noise samples with zero mean and variance σ_ϵ^2. Furthermore, the feature vectors $\boldsymbol{x}_i$ are assumed to independent and identically distributed (i.i.d.) random normal vectors with zero mean and covariance matrix $\frac{1}{p}\mathbf{I}_p$.

The model in (7) can be compactly written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_0 + \boldsymbol{\epsilon},$$

where $\mathbf{y} = [y_1, y_2, \cdots, y_n]^\top$, $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^\top$, and $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \cdots, \epsilon_n]^\top$.

2.2. Main Assumptions

Our study is based on the following set of assumptions:

- The unknown target vector $\boldsymbol{\theta}_0$ is assumed to be a structured vector, with entries Θ_0 that are sampled i.i.d. from a probability distribution function p_Θ, which has zero mean, and variance $\mathbb{E}[\Theta_0^2] = \sigma_\theta^2$, where $0 < \sigma_\theta^2 < \infty$.
- The noise variance $\sigma_\epsilon^2 < \infty$ is a fixed positive constant.
- As discussed above, the data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a Gaussian matrix with i.i.d. $\mathsf{N}(0, \frac{1}{p})$ elements. The choice of the $1/p$ as the variance level in $\mathbf{X}$ is commonly used in the literature; see [20,27]. This is done to ensure that $\|\mathbf{X}\boldsymbol{\theta}\|_2^2/n$ and $\|\boldsymbol{\theta}\|_2^2/p$ are of the same order.

Furthermore, in this work, we assume that the data matrix, $\mathbf{X}$, is not perfectly known, and we only have an *erroneous* copy of it, $\widehat{\mathbf{X}}$, which is given as:

$$\widehat{\mathbf{X}} = \mathbf{X} + \mathbf{E}, \tag{8}$$

where $\widehat{\mathbf{X}}$, and $\mathbf{E} \in \mathbb{R}^{n \times p}$ are independent matrices which have i.i.d. entries drawn from $\mathsf{N}(0, \frac{1-\sigma_e^2}{p})$, and $\mathsf{N}(0, \frac{\sigma_e^2}{p})$, respectively. (This uncertainty notion is widely encountered in practice. For example, it could be used to represent model mismatch, errors, and noises from the data collection process, noise in the sensors used to gather the measurements, etc.) Here, $\mathbf{E}$ represents the unknown error matrix, and $\sigma_e^2 \in [0, 1]$ is the variance of the error.

- n and p grow to infinity with $\frac{n}{p} \longrightarrow \zeta \in (0, \infty)$.

2.3. Generalized Penalized Constrained Regression (G-PCR)

In this paper, we refer to (5) as the *standard* G-PR, but we analyze a modified version that we call the *Generalized Penalized Constrained Regression (G-PCR)*, which solves the following optimization instead:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{V}^p} \|\widehat{\mathbf{X}}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \alpha \mathcal{P}(\boldsymbol{\theta}), \tag{9}$$

where $\mathbb{V} = [-L, U]$, and $L, U \in \mathbb{R}_+ \cup \{0\}$.

When compared to (6), the constraint $\mathbb{V}$ is used instead of $\mathbb{R}$, and $\widehat{\mathbf{X}}$ is used instead of $\mathbf{X}$. This is due to the fact that $\mathbf{X}$ is not perfectly known and we only have its noisy estimate $\widehat{\mathbf{X}}$.

When we compare (9) to (6), we can see that there is only a slight difference between them, which is the constraint set $\mathbb{V}$ instead of $\mathbb{R}$. This small change assures significant performance improvements in the algorithm in many practical applications, such as image and signal processing [55], wireless communications [41,43,56], etc. These improvements are shown for several cases in Sections 4 and 5.

3. Sharp Asymptotics

3.1. Measures of Performance

This paper considers the following measures used to assess the high-dimensional performance of the G-PCR.

- **Prediction Risk**: One of the most extensively used measures of performance is the prediction risk. For a given estimator $\widehat{\boldsymbol{\theta}}$, the prediction risk is defined as

$$\mathsf{R}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) := \frac{1}{p}\mathbb{E}_{x,y}\left[\left|\boldsymbol{x}^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right|^2\right] = \frac{1}{p}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2^2, \tag{10}$$

  where $\boldsymbol{x}$ and y are new *test* points following the linear model in (7) but are independent of the training data.
- **Similarity**: Another metric that is used to quantify the degree of *alignment* between the target vector $\boldsymbol{\theta}_0$ and its estimate $\widehat{\boldsymbol{\theta}}$ is the (dis)similarity. It is a measure of orientation rather than magnitude. It is defined as

$$\varrho(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) := \frac{\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\theta}_0}{\|\widehat{\boldsymbol{\theta}}\|_2 \|\boldsymbol{\theta}_0\|_2} \in [-1, 1]. \tag{11}$$

This similarity measure could also be thought of as the correlation between the estimated and true target vectors. Essentially, we desire estimates that *maximize* this similarity. Note that this metric is also known as the *cosine similarity* in machine learning literature, since $\varrho(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) = \cos(\angle(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0))$.

Note that these two measures are related as

$$\mathsf{R}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) = \frac{1}{p}\Big[\|\widehat{\boldsymbol{\theta}}\|_2^2 + \|\boldsymbol{\theta}_0\|_2^2 - 2\|\widehat{\boldsymbol{\theta}}\|_2\|\boldsymbol{\theta}_0\|_2\varrho(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)\Big].$$

3.2. High-Dimensional Performance Evaluation

In this subsection, we provide the main results of the paper, namely, the sharp analysis of the asymptotic performance of the G-PCR convex program. We start by analyzing the estimation performance via a general pseudo-Lipschitz function as in Theorem 1 below, which sharply characterizes the *general* asymptotic behavior of the error. Then, we use this theorem to compute particular performance measures such as the prediction risk, similarity, etc.

Theorem 1 (General Performance Metric)**.** *Consider the high-dimensional setup of Section 2.2, and let the assumptions therein hold. Moreover, let $\widehat{\boldsymbol{\theta}}$ be a minimizer of the G-PCR program in (9) for a fixed $\alpha > 0$. Let $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a pseudo-Lipschitz function. Then, in the limit of $p \to \infty$, it holds that*

$$\frac{1}{p}\sum_{j=1}^{p}\psi(\widehat{\theta}_j, \theta_{0,j}) \xrightarrow{P} \mathbb{E}\left[\psi\left(\mathrm{prox}_{\widetilde{\mathcal{P}}}\left(\Theta_0 + \frac{H}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}}; \frac{\alpha}{q_\star\gamma_\star\sqrt{\zeta}(1-\sigma_e^2)}, -L, U\right), \Theta_0\right)\right], \tag{12}$$

where $(q_\star, \gamma_\star)$ is the unique optimal solution to the following objective,

$$\begin{aligned}\sup_{q\geq 0}\inf_{\gamma>0}\ \mathcal{O}_{\widetilde{\mathcal{P}}}(q,\gamma) \ &:= \tfrac{q\sqrt{\zeta}}{2\gamma}+\tfrac{q\gamma\sqrt{\zeta}}{2}\left(\sigma_\epsilon^2+\sigma_e^2\sigma_\theta^2\right)-\tfrac{q}{2\gamma\sqrt{\zeta}}-\tfrac{q^2}{4}\\ &\quad+q\gamma\sqrt{\zeta}(1-\sigma_e^2)\,\mathbb{E}\left[\mathsf{M}_{\widetilde{\mathcal{P}}}\left(\Theta_0+\tfrac{H}{\gamma\sqrt{\zeta(1-\sigma_e^2)}};\tfrac{\alpha}{q\gamma\sqrt{\zeta}(1-\sigma_e^2)},-L,U\right)\right],\end{aligned} \tag{13}$$

and the expectation is taken with respect to independent random variables $\Theta_0 \sim p_\Theta$ *and* $H\sim\mathsf{N}(0,1)$.

Proof. The proof is given in Appendix A. □

Remark 1 (Choice of $\psi(\cdot,\cdot)$)**.** *The performance metric in Theorem 1 is computed in terms of evaluation of a pseudo-Lipschitz function,* $\psi(\cdot,\cdot)$*. As an example,* $\psi(a,b)=(a-b)^2$ *can be used to compute the prediction risk, or* $\psi(a,b)=|a-b|$ *can be used to evaluate the mean absolute error (MAE). We will appeal to this theorem later with various choices of* $\psi(\cdot,\cdot)$ *to evaluate different performance measures on* $\widehat{\boldsymbol{\theta}}$.

Remark 2 (Optimal Solutions)**.** *Note that* $q_\star$ *and* $\gamma_\star$ *can be calculated by any search technique such as the golden-section search method and the ternary search* [57].

Remark 3 (Decoupling Property)**.** *Theorem 1 shows that*

$$\frac{1}{p}\sum_{j=1}^{p}\psi(\widehat{\theta}_j,\theta_{0,j})\xrightarrow{P}\mathbb{E}\Big[\psi(\widehat{\Theta},\Theta_0)\Big], \tag{14}$$

where

$$\widehat{\Theta}:=\operatorname{prox}_{\widetilde{\mathcal{P}}}\left(\underbrace{\Theta_0+\frac{H}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}}}_{:=Y};\frac{\alpha}{q_\star\gamma_\star\sqrt{\zeta}(1-\sigma_e^2)},-L,U\right). \tag{15}$$

This provides some insights for the structured signal recovery with the G-PCR optimization. From (14), it can be seen that the random variable $\widehat{\Theta}$ *shares the same statistical properties as the estimate* $\widehat{\boldsymbol{\theta}}$ [46]*. Thus, Equation (15) can be considered as a* decoupled *scalar version of the original system as depicted in Figure 1. Particularly, in the original system (Figure 1a), the true target vector* $\boldsymbol{\theta}_0$ *is first mixed by the design matrix* $\mathbf{X}$*, and then the additive white Gaussian noise (AWGN) vector* $\boldsymbol{\epsilon}$ *is added to form the measurement vector* $\mathbf{y}$*. On the other hand, in the decoupled system (Figure 1b), the unknown variable* Θ_0 *is only mixed by the Gaussian vector* $\frac{1}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}}H$*, where* $Y:=\Theta_0+\frac{1}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}}H$*. Furthermore, letting* $B:=\frac{1}{q_\star\gamma_\star\sqrt{\zeta}(1-\sigma_e^2)}$*, it can be observed that the generalized proximal operator solution*

$$\widehat{\Theta}=\operatorname{prox}_{\widetilde{\mathcal{P}}}(Y;\alpha B,-L,U)=\operatorname*{arg\,min}_{-L\leq\theta\leq U}\frac{1}{2}(Y-\theta)^2+\alpha B\,\widetilde{\mathcal{P}}(\theta)$$

has a decoupled scalar form of the original G-PCR in (9), which can be expressed as

$$\widehat{\boldsymbol{\theta}}=\operatorname*{arg\,min}_{-L\leq\theta_j\leq U}\frac{1}{2}\|\mathbf{y}-\widehat{\mathbf{X}}\boldsymbol{\theta}\|_2^2+\alpha B\,\mathcal{P}(\boldsymbol{\theta}),\ \ j\in[p]$$

up to a scaling of B. This suggests that, in the high-dimensional asymptotic setting, one can use the decoupled scalar system to characterize the probabilistic properties of the G-PCR recovery problem.

This decoupling property was also shown for similar problems, such as box-constrained sum of absolute values (Box-SOAV) optimization for sparse recovery [46]*, sparse logistic regression* [58]*, and the approximate message passing (AMP) algorithm* [59].

As a first application of Theorem 1, we provide a sharp high-dimensional performance evaluation of the prediction risk as given in the following corollary.

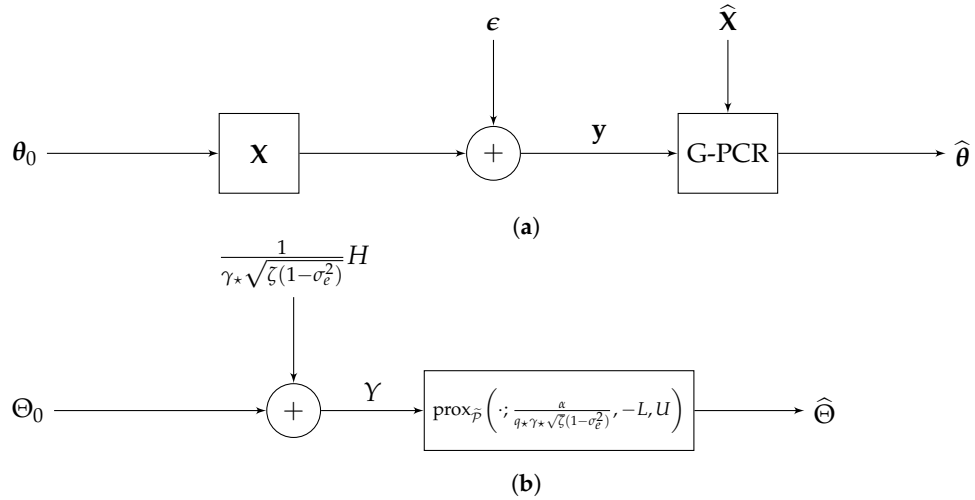


Figure 1. A system model comparison between the original G-PCR recovery algorithm and its scalar decoupled version. (**a**) Original system. (**b**) Scalar decoupled system.

Corollary 1 (Prediction Risk)**.** *Under the same assumptions of Theorem 1, and for* $\Theta_0 \sim p_\Theta$ *that is independent of* $H \sim \mathsf{N}(0,1)$*, it holds that*

$$\begin{aligned}\mathsf{R}(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}_0) \xrightarrow{P} \quad & \mathbb{E}_{\Theta_0,H}\left[\left(\text{prox}_{\widetilde{\mathcal{P}}}\left(\Theta_0+\tfrac{H}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}};\tfrac{\alpha}{q_\star\gamma_\star\sqrt{\zeta}(1-\sigma_e^2)},-L,U\right)-\Theta_0\right)^2\right] \\ & = \tfrac{1}{1-\sigma_e^2}\left(\tfrac{1}{\gamma_\star^2}-\sigma_\theta^2\sigma_e^2-\sigma_\epsilon^2\right),\end{aligned} \tag{16}$$

where $q_\star$ *and* $\gamma_\star$ *are the unique optimal solutions to the objective function in* (13)*.*

Proof. Using Theorem 1 with $\psi(a,b)=(a-b)^2$, we can obtain the above expression of the prediction risk. Details are deferred to Appendix A.2.4. □

Remark 4 (Optimal Hyper-parameters)**.** *Corollary 1 allows us to determine the optimal hyper-parameters, such as* α, L*, and* U*, that minimize the prediction risk. To do so, it is first required to estimate some variances, such as* $\sigma_\theta^2, \sigma_\epsilon^2$ *and* σ_e^2*, from the available data. Those can be easily estimated by using existing algorithms such as* [60,61]*.*

It should be noted that this theoretical hyper-parameter optimal tuning as discussed above avoids the traditional time/data-consuming practice of cross-validation used to tune the hyper parameters.

The following corollary sharply characterizes the similarity measure defined earlier in (11).

Corollary 2 (Similarity)**.** *Under the same assumptions and settings of Theorem 1, and in the limit of* $p \to \infty$*, it holds that*

$$\varrho(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}_0) \xrightarrow{P} \frac{\mathbb{E}_{\Theta_0,H}\left[\text{prox}_{\widetilde{\mathcal{P}}}\left(\Theta_0+\tfrac{H}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}};\tfrac{\alpha}{q_\star\gamma_\star\sqrt{\zeta}(1-\sigma_e^2)},-L,U\right)\Theta_0\right]}{\sqrt{\sigma_\theta^2\cdot\mathbb{E}_{\Theta_0,H}\left[\text{prox}^2_{\widetilde{\mathcal{P}}}\left(\Theta_0+\tfrac{H}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}};\tfrac{\alpha}{q_\star\gamma_\star\sqrt{\zeta}(1-\sigma_e^2)},-L,U\right)\right]}}, \tag{17}$$

where $q_\star$ *and* $\gamma_\star$ *are the unique optimal solutions to* (13)*.*

Proof. The proof follows from Theorem 1 and from the continuous mapping Theorem [62]. Details are given in Appendix A.2.5. □

In the subsequent sections, we consider various instants of (9), such as ℓ_1-norm and ℓ_2-norm penalization, to illustrate the application of the theoretical asymptotic expressions derived in this section.

4. Sparse Linear Regression

In this section, we study the performance of the G-PCR with an ℓ_1-norm penalization. As indicated in the introduction, this penalty function is used to promote sparse solutions. In contemporary machine learning applications, it is common to encounter a significantly large number of features, p. To prevent the problem of over-fitting, it becomes crucial to engage in feature selection, which involves eliminating irrelevant variables from the regression model [18]. A popular technique for accomplishing this is by introducing an ℓ_1-norm penalty to the loss function. This approach is widely adopted and used for feature selection tasks.

Therefore, we specialize Theorem 1 to analyze the asymptotic performance of the G-PCR with an ℓ_1-norm penalization. Particularly, for an s-sparse vector, we study the performance of the following optimization problem:

$$\widehat{\boldsymbol{\theta}} = \arg \min_{-L \le \theta_j \le U} \|\widehat{\mathbf{X}}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \alpha \|\boldsymbol{\theta}\|_1, \; j \in [p]. \tag{18}$$

(We say that a vector $\mathbf{v} \in \mathbb{R}^p$ is an s-sparse vector if only s of its p elements are non-zero (on average), and most of its elements are zeros, where $s \ll p$.)

4.1. Asymptotic Behavior of Sparse G-PCR

To analyze (18), we specialize Theorem 1 with $\widetilde{\mathcal{P}}(\cdot) = |\cdot|$. Then, the generalized proximal operator and Moreau envelope functions can be expressed, respectively, in the following closed forms:

$$\text{prox}_{|\cdot|}(x; b, c, d) := \eta_1(x; b, c, d) = \begin{cases} d & \text{, if } x \ge d + b \\ x - b & \text{, if } b < x < d + b \\ 0 & \text{, if } |x| \le b \\ x + b & \text{, if } c - b < x < -b \\ c & \text{, if } x \le c - b, \end{cases} \tag{19}$$

and

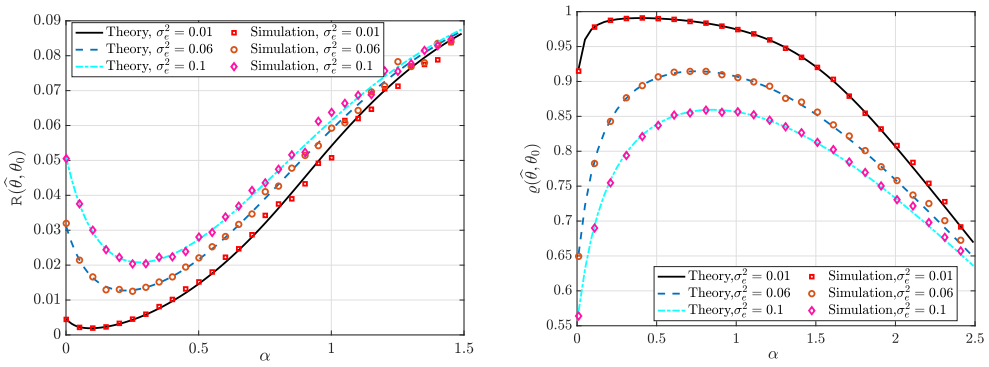
$$\mathsf{M}_{|\cdot|}(x; b, c, d) = \begin{cases} \frac{1}{2}(d - x)^2 + bd & \text{, if } x \ge d + b \\ bx - \frac{1}{2}b^2 & \text{, if } b < x < d + b \\ \frac{1}{2}x^2 & \text{, if } |x| \le b \\ -bx - \frac{1}{2}b^2 & \text{, if } c - b < x < -b \\ \frac{1}{2}(c - x)^2 - bc & \text{, if } x \le c - b. \end{cases} \tag{20}$$

Note that this proximal operator is a generalization of the well-known soft-thresholding operator, i.e., $\eta(x; b) = \text{sign}(x)\ \text{ReLU}(|x| - b)$, where the Rectified Linear Unit (ReLU) is defined as $\text{ReLU}(t) = \max(0, t)$.

These expressions can be used to solve the scalar optimization in (13) of Theorem 1 and to simplify the similarity expression in (17). Specifically, (13) becomes

$$\begin{aligned} \sup_{q \ge 0} \inf_{\gamma > 0} \; \mathcal{O}_{|\cdot|}(q, \gamma) \quad &= \tfrac{q\sqrt{\zeta}}{2\gamma} + \tfrac{q\gamma\sqrt{\zeta}}{2}\left(\sigma_\epsilon^2 + \sigma_e^2\sigma_\theta^2\right) - \tfrac{q}{2\gamma\sqrt{\zeta}} - \tfrac{q^2}{4} \\ &\quad + q\gamma\sqrt{\zeta}(1 - \sigma_e^2)\, \mathbb{E}\left[\mathsf{M}_{|\cdot|}\left(\Theta_0 + \tfrac{H}{\gamma\sqrt{\zeta(1-\sigma_e^2)}}; \tfrac{\alpha}{q\gamma\sqrt{\zeta}(1-\sigma_e^2)}, -L, U\right)\right]. \end{aligned} \tag{21}$$

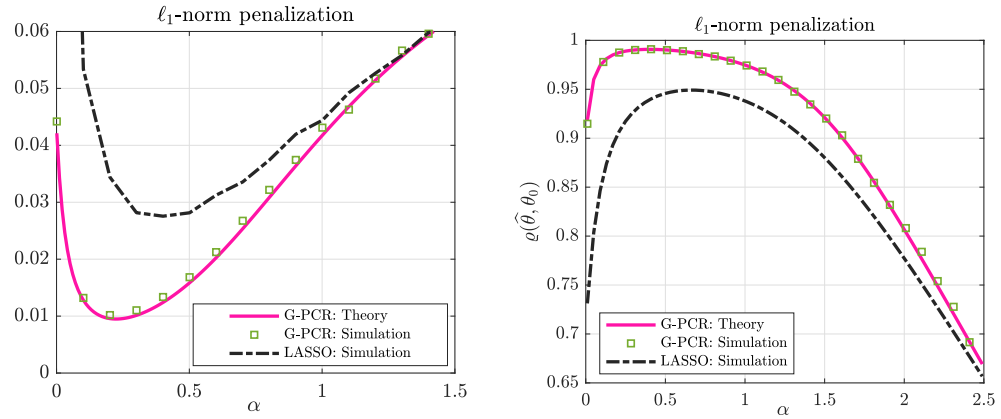
Figure 2 illustrates the performance of the ℓ_1-penalized G-PCR as a function of α for different levels of error variance σ_e^2. We generated the target vector $\boldsymbol{\theta}_0$ randomly with elements in $\{-1, 0, +1\}$ and $\mathbb{P}(\theta_{0,j} = -1) = \mathbb{P}(\theta_{0,j} = +1) = 0.2$. This means that the sparsity factor $\rho := \frac{s}{p}$ is 0.2 in these simulations. From these figures, we can also see that the theoretical expressions of the prediction risk and the similarity match the empirical simulations very well. Furthermore, it can be noted in Figure 2a that, for different values of σ_e^2, there exists an optimal value $\alpha_\star$ that achieves the minimum possible prediction risk. Similarly, notice in Figure 2b the optimal $\alpha_\star$ that maximizes the similarity metric. It can also be observed that increasing α beyond $\alpha_\star$ reduces the similarity ϱ between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. In these simulations, we set $-L = \min(\boldsymbol{\theta}_0) = -1$ and $U = \max(\boldsymbol{\theta}_0) = +1$.



(**a**) (**b**)

Figure 2. Performance of the G-PCR vs. the penalization factor for a sparse linear regression. The parameters are set as follows: $p = 300, \zeta = 1.5, \rho = 0.2, \sigma_\epsilon^2 = 0.1, L = 1$, and $U = 1$. The simulation results are averaged over 50 independent Monte Carlo trials. (**a**) The prediction risk. (**b**) The cosine similarity.

In Figure 3, we compare the unconstrained G-PR in (6) (which is equivalent to a standard LASSO formulation in this case) to the proposed G-PCR for an over-parameterized setting with $\zeta = 0.85$. As we can see from this figure, the G-PCR clearly outperforms the unconstrained one in both metrics. Moreover, despite the fact that our theoretical results are assumed to be asymptotic in the problem dimensions (i.e., $n \to \infty$ and $p \to \infty$), we can see from all of the above figures that our rigorous results are accurate even for problems with a few hundred variables, e.g., $p = 300$.



(**a**) (**b**)

Figure 3. Performance comparison between the G-PCR and G-PR. For the numerical simulations, the results are averaged over 100 independent trials, with $p = 128, \zeta = 0.85, \rho = 0.1, \sigma_\epsilon^2 = 0.2$, $\sigma_e^2 = 0.05, L = 1$, and $U = 1$. (**a**) The prediction risk. (**b**) The cosine similarity.

4.2. Support Recovery

In this section, we analyze the so-called support recovery of the sparse G-PCR. As discussed earlier, a sparse vector means that it has few non-zero elements. We define the support of $\boldsymbol{\theta}_0$ as follows:

$$\Omega := \{j \in [p] \mid \theta_{0,j} \neq 0\} \subseteq [p].$$

Here, we are interested in computing the probability that an element on the support of $\boldsymbol{\theta}_0$ has been recovered correctly. Let $\widehat{\boldsymbol{\theta}}$ be a solution to the optimization problem in (18). Let us fix $\xi > 0$ as a user-predefined hard threshold based on whether an entry of $\widehat{\boldsymbol{\theta}}$ is decided to be on the support or not. Formally, we construct the following set as the estimate of the support given $\widehat{\boldsymbol{\theta}}$:

$$\widehat{\Omega}_\xi := \{j \in [p] \mid |\widehat{\theta}_j| > \xi\}.$$

In order to analyze the support recovery correctness, we consider the following error metrics, which are known as the probability of misdetection (MD) and the probability of false alarm (FA), respectively.

$$\mathsf{P}_{\mathrm{MD}}(\xi) = \mathbb{P}(j \notin \widehat{\Omega}_\xi \mid j \in \Omega), \text{ and } \mathsf{P}_{\mathrm{FA}}(\xi) = \mathbb{P}(j \in \widehat{\Omega}_\xi \mid j \notin \Omega).$$

In the following lemma, we study the asymptotic performance of both of these measures.

Lemma 1. *Let $\widehat{\boldsymbol{\theta}}$ be a solution to (18), and assume that $\boldsymbol{\theta}_0$ is a sparse signal. Fix $\alpha > 0$ and $\xi > 0$. Then, in the limit of $p \to \infty$, it holds that*

$$\mathsf{P}_{\mathrm{MD}}(\xi) \xrightarrow{P} \mathbb{P}\left(\left|\eta_1\left(\Theta_0 + \frac{H}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}}; \frac{\alpha}{q_\star\gamma_\star\sqrt{\zeta}(1-\sigma_e^2)}, -L, U\right)\right| \leq \xi\right), \tag{22}$$

and

$$\mathsf{P}_{\mathrm{FA}}(\xi) \xrightarrow{P} \mathbb{P}\left(\left|\eta_1\left(\frac{H}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}}; \frac{\alpha}{q_\star\gamma_\star\sqrt{\zeta}(1-\sigma_e^2)}, -L, U\right)\right| \geq \xi\right), \tag{23}$$

where $\eta_1(\cdot;\cdot,\cdot,\cdot)$ is as defined in (19), $(q_\star, \gamma_\star)$ is the unique optimal solution to (21), and the probabilities are taken with respect to the randomness of Θ_0 and $H \sim \mathsf{N}(0,1)$.

Proof. The proof can be obtained from Theorem 1 with some approximations of these metrics to Lipschitz functions. Details are omitted for briefness. See [39] for a similar proof. □

Next, we give an example to illustrate this lemma.

Example: Sparse-Binary Target Vectors

For an s-sparse target vector, define $\rho := \frac{s}{p} \in (0,1]$ as the sparsity factor. Then, as an example, let us assume that each element $\theta_{0,j}$, for $j \in [p]$, is i.i.d. drawn from the following distribution (this model has been widely adopted in the relevant literature; see, for example, [59,63,64]):

$$p_\Theta(\theta) = (1-\rho)\,\delta_0(\theta) + \rho\,\delta_0(\theta - \mathcal{E}), \tag{24}$$

for some $\mathcal{E} > 0$, and $\delta_0(\cdot)$ indicates a Dirac delta function (i.e., a point-mass distribution). In other words, the elements of $\boldsymbol{\theta}_0$ are zero with probability $1-\rho$, and the non-zero elements all have the value $\mathcal{E}$. Figure 4 illustrates this distribution.

For a Θ_0 that follows the distribution in (24), and for $\xi \in (0, \mathcal{E})$, with $L = 0$ and $U = \mathcal{E}$, the error measures in Lemma 1 simplify to the following:

$$\mathsf{P}_{\mathrm{MD}} \xrightarrow{P} \Phi\left((\xi - \mathcal{E})\gamma_\star\sqrt{\zeta(1-\sigma_e^2)} + \frac{\alpha}{q_\star\sqrt{1-\sigma_e^2}}\right), \tag{25}$$

and

$$\mathrm{P_{FA}} \xrightarrow{P} 1 - \Phi\left(\xi\gamma_\star\sqrt{\zeta(1-\sigma_e^2)} + \frac{\alpha}{q_\star\sqrt{1-\sigma_e^2}}\right), \tag{26}$$

where $\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-u^2}\mathrm{d}u$ is the cumulative distribution function (CDF) of the standard normal distribution.

Figure 5 shows the accuracy of the above-derived theoretical expressions as compared to empirical simulations for the considered sparse-binary vector example.

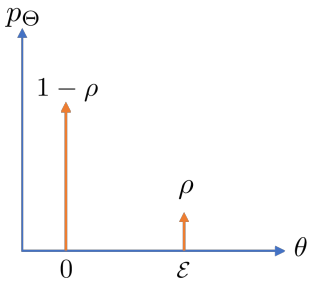
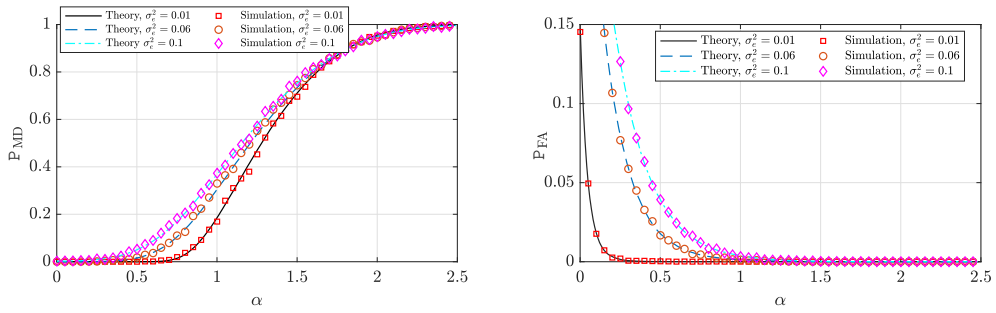


Figure 4. Probability mass function (PMF) of a sparse-binary distribution.



(**a**) (**b**)

Figure 5. Support recovery performance of the G-PCR versus the penalization factor for a sparse-binary signal recovery. The parameters are set as follows: $\mathcal{E} = 1, p = 300, \zeta = 0.85, \rho = 0.2$, $\sigma_e^2 = 0.05, \xi = 0.1, L = 0$, and $U = 1$. The simulations are averaged over 100 independent Monte Carlo trials. (**a**) Probability of misdetection. (**b**) Probability of false alarm.

5. G-PCR with ℓ_2^2-Norm Penalization

Even though it does not promote a particular structure, ℓ_2-norm penalization is used in many signal processing, statistics, and machine learning applications to stabilize the model when we have ill-conditioned or under-determined systems [17]. Adding this penalization will shrink all the coefficients toward zero and hence decrease the variance of the resultant model; therefore, it can be used to avoid over-fitting. Within the Bayesian framework, the incorporation of this penalization implies that the regression coefficients are assumed to follow a Gaussian distribution. This assumption is often justifiable in numerous applications, in which the regression coefficients are typically taken from a random process. In this section, we provide high-dimensional asymptotic performance analysis of the G-PCR with ℓ_2^2-norm penalization; that is:

$$\widehat{\boldsymbol{\theta}} = \arg \min_{-L\leq\theta_j\leq U} \|\widehat{\mathbf{X}}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \alpha\|\boldsymbol{\theta}\|_2^2,\ j \in [p]. \tag{27}$$

To analyze (27), we use Theorem 1. However, here the generalized proximal operator and Moreau envelope functions of $\widetilde{\mathcal{P}}(\cdot) = (\cdot)^2$ can be expressed, respectively, in the following closed-forms:

$$\mathrm{prox}_{(\cdot)^2}(x; b, c, d) = \begin{cases} \frac{x}{1+2b}, & \text{if } c \leq x \leq d \\ c, & \text{if } x < c \\ d, & \text{if } x > d, \end{cases} \tag{28}$$

and

$$\mathsf{M}_{(\cdot)^2}(x;b,c,d) = \begin{cases} \frac{bx^2}{1+2b}, & \text{if } c \le x \le d \\ \frac{1}{2}(x-c)^2 + bc^2, & \text{if } x < c \\ \frac{1}{2}(x-d)^2 + bd^2, & \text{if } x > d. \end{cases} \tag{29}$$

Letting $b = \frac{\alpha}{q\gamma\sqrt{\zeta(1-\sigma_e^2)}}$ and $\lambda = \gamma\sqrt{\zeta(1-\sigma_e^2)}$, the scalar optimization in (13) of Theorem 1 reduces to:

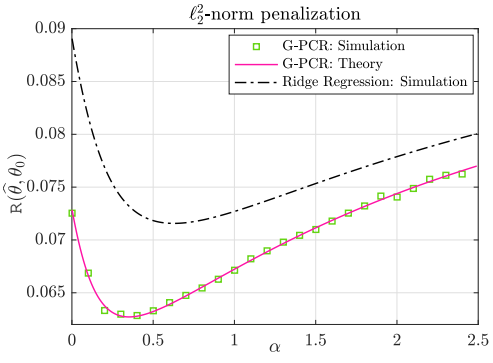
$$\begin{aligned} \sup_{q\ge 0}\inf_{\gamma>0}\ \mathcal{O}_{(\cdot)^2}(q,\gamma) \ &= \tfrac{q\sqrt{\zeta}}{2\gamma} + \tfrac{q\gamma\sqrt{\zeta}}{2}\left(\sigma_e^2+\sigma_e^2\sigma_\theta^2\right) - \tfrac{q}{2\gamma\sqrt{\zeta}} - \tfrac{q^2}{4} \\ &+ q\gamma\sqrt{\zeta}(1-\sigma_e^2)\left\{\mathbb{E}\left[\tfrac{b}{1+2b}\left(\Theta_0+\tfrac{H}{\lambda}\right)^2 \mathbf{1}_{\{-L\le\Theta_0+\frac{H}{\lambda}\le U\}}\right.\right. \\ &+ \left[\tfrac{1}{2}\left(\Theta_0+\tfrac{H}{\lambda}+L\right)^2 + bL^2\right]\mathbf{1}_{\{\Theta_0+\frac{H}{\lambda}\le -L\}} \\ &+ \left.\left.\left[\tfrac{1}{2}\left(\Theta_0+\tfrac{H}{\lambda}-U\right)^2 + bU^2\right]\mathbf{1}_{\{\Theta_0+\frac{H}{\lambda}\ge U\}}\right]\right\}, \end{aligned} \tag{30}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

In the same manner, we can simplify the similarity expression in (17) using the closed-form expression of the generalized proximal operator of the ℓ_2^2-norm in (28). The prediction risk and the similarity metric are given by (16) and (17), respectively. However, now $(q_\star, \gamma_\star)$ is the unique solution to (30). To illustrate the ideal let us consider the next examples.

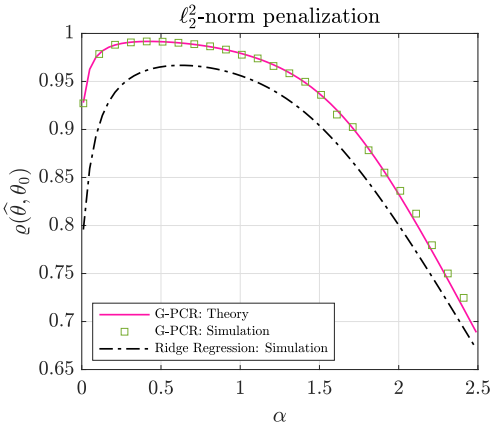
5.1. Numerical Illustration

As stated at the beginning of this section, ℓ_2-norm penalization can be used for Gaussian distributed target vectors. Therefore, as a first illustration, let us assume that $\theta_{0,j} \sim \mathsf{N}(0,1)\ \ \forall j \in [p]$. Figure 6 depicts the risk/similarity performance of the G-PCR with an ℓ_2-norm penalization for several levels of the penalization factor α. It also shows that the G-PCR outperforms the unconstrained G-PR, which is equivalent to a ridge regression formulation here. Again, Figure 6a illustrates that there exists an optimal value $\alpha_\star$ that minimizes prediction risk, while Figure 6b shows that there is an optimal value of the penalization factor $\alpha_\star$ that gives the maximum similarity. Both figures show the high accuracy of the derived asymptotic expressions as compared to Monte Carlo simulations.



(a)

Figure 6. *Cont.*



(**b**)

Figure 6. Performance of the ℓ_2^2-norm G-PCR vs. the penalization factor for a Gaussian target vector. The parameters are set as follows: $p = 500, \zeta = 1.3, \sigma_\theta^2 = 1, \sigma_\epsilon^2 = 0.01, \sigma_\epsilon^2 = 0.01, L = 1$, and $U = 1$. The results are averaged over 50 independent realizations. (**a**) The prediction risk. (**b**) The cosine similarity.

5.2. Binary Target Vector Estimation

Let us assume that $\boldsymbol{\theta}_0 = \{\pm 1\}^p$, i.e., it takes only one of two possible values, $+1$ or -1, with equal probability, i.e.,

$$p_\Theta(\theta) = \frac{1}{2}[\delta_0(\theta - 1) + \delta_0(\theta + 1)]. \tag{31}$$

Such vectors are widely encountered in many practical applications, such as the detection of wireless communication signals [12,41]. We use (27) as our estimation method, with $L = U = 1$. For this vector, $\sigma_\theta^2 = 1$; i.e., the covariance matrix of $\boldsymbol{\theta}_0$ is $\mathbf{C}_\theta = \mathbf{I}_p$.

The task of estimating $\boldsymbol{\theta}_0$ here is equivalent to a binary classification task, with the two classes being $+1$ and -1. After obtaining the estimates using (27), we can map (decode) them to the relative class using the following link function:

$$\bar{\boldsymbol{\theta}} = \text{sign}(\widehat{\boldsymbol{\theta}}).$$

We can use the prediction risk and similarity to measure the performance. However, a more suitable performance measure for this kind of target vector is the so-called "**classification error rate**", which is defined as:

$$\mathcal{C}_{\text{err}} := \frac{1}{p} \sum_{j=1}^{p} \mathbf{1}_{\{\bar{\theta}_j \neq \theta_{0,j}\}}. \tag{32}$$

The next lemma derives an asymptotic expression for this metric.

Lemma 2. *Let $\widehat{\boldsymbol{\theta}}$ be a solution to (27), and assume that $\boldsymbol{\theta}_0 = \{\pm 1\}^p$, with a PMF p_Θ that follows (31). Fix $\zeta > 0$. Then, in the limit of $p \to \infty$, it holds that*

$$\mathcal{C}_{\text{err}} \xrightarrow{P} 1 - \Phi\left(\gamma_\star \sqrt{\zeta(1 - \sigma_\epsilon^2)}\right), \tag{33}$$

where $\gamma_\star$ is the optimal solution of (30) in γ.

Proof. The proof is similar to that of Lemma 1. Please refer to [39,47]. Details are skipped for brevity. □

Figure 7 illustrates the sharpness of Lemma 2 as compared to empirical simulations. Similar to previous figures, this figure demonstrates the perfect match between numerical simulations and theory.

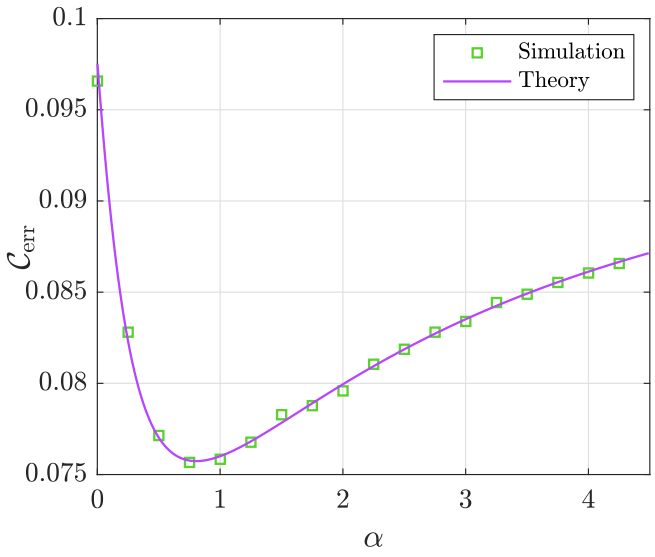


Figure 7. Classification error rate ($\mathcal{C}_{\text{err}}$) of the G-PCR vs.the penalization factor for a binary target vector. The parameters are set as follows: $p = 500, \zeta = 0.85, \sigma_e^2 = 0.01, \sigma_\epsilon^2 = 0.02$, and $L = U = 1$. The results are averaged over 100 independent Monte Carlo trials.

5.3. Unpenalized Regression

When $\alpha = 0$ in (9), we have an optimization problem with no penalization. The resulting algorithm is known as the box relaxation optimization (BRO) [39], which has been extensively studied in the literature. It is used to promote boundedness structure. In fact, when $L = U$, the BRO is equivalent to an ℓ_∞-norm penalization. Setting $\alpha = 0$ in (30), and after some mapping of the involved variables, we can obtain the same results as in [39] for binary vectors.

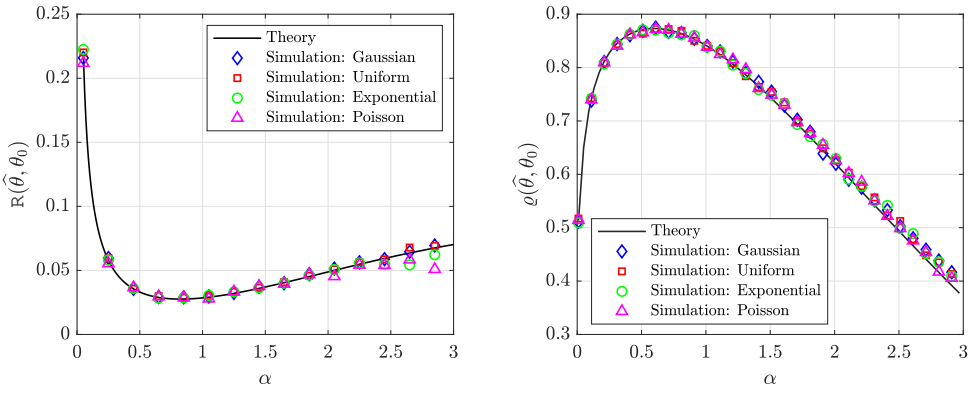
6. Additional Numerical Experiments

In this section, we provide additional numerical experiments to validate our results. These experiments are performed on synthetic data beyond the Gaussian ensemble and on real data as well. In addition, we empirically discuss the double descent phenomenon.

6.1. Synthetic Data: Universality of the Gaussian Design

Theorem 1 assumes that the elements of matrix $\mathbf{X}$ are i.i.d. Gaussian distributed. However, we expect the asymptotic results derived in this paper (prediction risk, similarity, etc.) to be robust and hold for a larger class of random matrices. Rigorous proofs are presented in [65–69], where the asymptotic prediction is shown to have a universal limit (as $p \to \infty$) with respect to random matrices with i.i.d. entries.

To validate the above claims, see Figure 8, where we plotted the prediction risk and the similarity for a *sparse-Gaussian* target vector with i.i.d. entries that follow the distribution: $\theta_{0,j} \sim (1-\rho)\ \delta_0 + \rho\ \mathsf{N}(0,2)$. We used the G-PCR with an ℓ_1-norm in (18) to obtain the estimates. In addition to the Gaussian design matrix, we simulated the performance using other random matrices with i.i.d. entries drawn from a uniform distribution $\sqrt{\frac{3}{p}}\,\mathcal{U}[-1,1]$, an exponential distribution $\frac{1}{\sqrt{p}}\text{Exp}(1)$, and from a Poisson distribution $\frac{1}{\sqrt{p}}\text{Poiss}(1)$. Note that the normalization of these matrices is used to satisfy the high-dimensionality assumptions in Section 2.2. From this figure, we can see that the behavior seems to be nearly identical for all distributions, suggesting that our results enjoy a universality property.



(**a**) (**b**)

Figure 8. Performance the G-PCR for a sparse-Gaussian target vector. We set the parameters as $p = 180, \zeta = 0.95, \sigma_e^2 = 0.1, \sigma_\epsilon^2 = 0.2, \rho = 0.1$, and $L = U = \sqrt{2}$. The results are averaged over 50 independent trials. (**a**) The prediction risk. (**b**) The similarity.

6.2. Real-World Data

In previous section, we showed the robustness of our results to the distribution of the i.i.d. entries of the data matrix $\mathbf{X}$. In this section, we take it a step further and consider real-world datasets instead of the synthetic data discussed earlier. These datasets are essentially not random and do not have i.i.d. elements. However, as seen in the numerical simulations below, they match our theoretical results to a great extent.

As an illustration, we present in Figure 9 the outcomes of these simulations for three real datasets. Each of these datasets consists of a small number of samples (n) and a high-dimensional feature space (p), which is consistent with the over-parameterized setting ($p > n$). These datasets are mainly for detecting several diseases and cancer samples. We generated the target vector, $\boldsymbol{\theta}_0$, randomly with entries following the distribution $\theta_{0,j} \sim (1-\rho)\,\delta_0 + \rho\,\mathsf{N}(0,1)$. The noise vector $\boldsymbol{\epsilon}$ was generated using i.i.d. $\mathsf{N}(0, 0.2)$ elements. We generated the observations as $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_0 + \boldsymbol{\epsilon}$. The G-PCR with ℓ_1-norm penalty in (18) is used to obtain $\widehat{\boldsymbol{\theta}}$ with $\widehat{\mathbf{X}} = \mathbf{X} + \mathbf{E}$.

The three figures correspond to the following datasets:

- Figure 9a: For this figure, we used *breast cancer* data [70] (available at: https://github.com/kivancguckiran/microarray-data (accessed on 27 May 2023)). This dataset has been used in [71] for DNA microarray gene expression classification using the LASSO. It consists of 22,215 gene expressions (features) and 118 samples. From this matrix, we took a sub-matrix $\mathbf{X}$ of aspect ratio $\zeta = 0.75$. We standardized all columns of matrix $\mathbf{X}$ to have mean 0 and variance 1.
- Figure 9b: In this figure, *glioma* disease data [72] were used (available at: https://github.com/kivancguckiran/microarray-data (accessed on 27 May 2023)). This dataset includes 54,613 features and 180 samples. Similar to the breast cancer data, the sub-matrix $\mathbf{X}$ with the same aspect ratio ζ was selected and standardized.
- Figure 9c: The dataset used in this figure includes *colon cancer* data [73] (available at: http://www.weizmann.ac.il/mcb/UriAlon/download/downloadable-data (accessed on 27 May 2023)). This dataset was used in [74] for a sparse-group LASSO model. It includes 2000 genes and 62 samples (22 normal tissues and 40 colon tumor tissues). Similar to the previous datasets, we selected a sub-matrix $\mathbf{X}$ with aspect ratio ζ and standardized it.

For all figures, we can see that the agreement between theory and simulations is remarkably good.

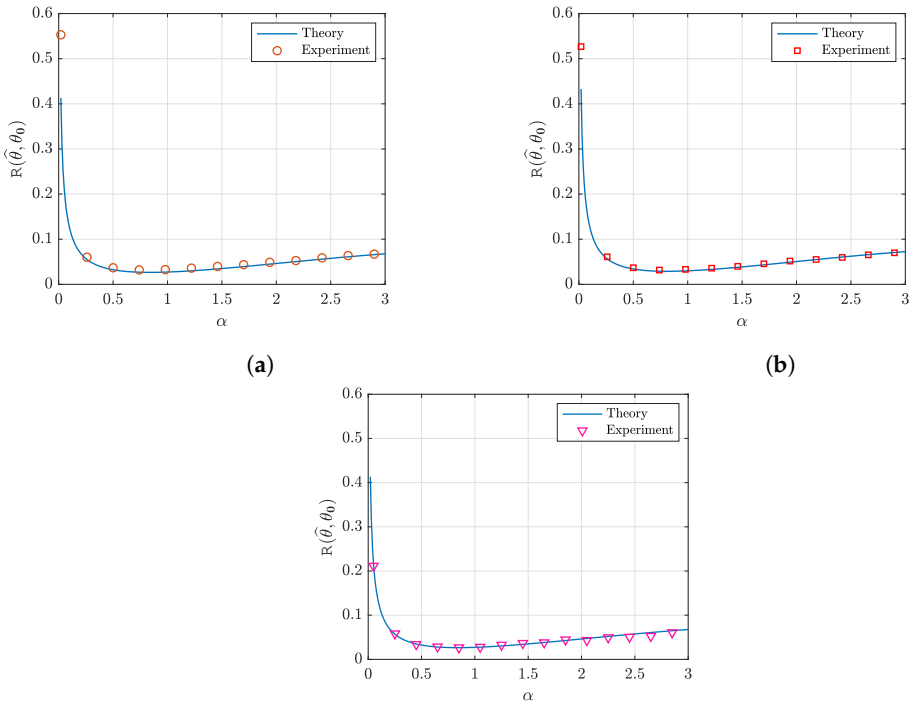


Figure 9. Prediction risk as a function of the penalization factor α. Here, the data matrix $\mathbf{X}$ is a standardized real dataset. We used a sparse-Gaussian vector $\boldsymbol{\theta}_0$, and we generated the observations as $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_0 + \boldsymbol{\epsilon}$. The G-PCR with ℓ_1-norm penalty is used to obtain $\widehat{\boldsymbol{\theta}}$ with $\widehat{\mathbf{X}} = \mathbf{X} + \mathbf{E}$. The parameters are set as $\zeta = 0.75, \rho = 0.1, \sigma_\epsilon^2 = 0.2, \sigma_e^2 = 0.1$, and $L = U = 1$. The results are averaged over 200 independent trials. (**a**) Breast cancer data. (**b**) Glioma disease data. (**c**) Colon cancer data.

6.3. Double Descent Phenomenon

In Figure 10, we plotted the prediction risk as a function of ζ for different choices of the penalization factor α. As can be seen, for an arbitrary choice of α, the prediction risk of the G-PCR first decreases for small values of ζ, then increases until it reaches a peak known as the interpolation peak. After that, the prediction risk decreases monotonically with respect to ζ. This is known as the double descent phenomenon [75]. On the other hand, optimal values of the penalization factor $\alpha_\star$ always guarantee that the prediction risk decreases with more training samples being used (i.e., with increasing ζ). This emphasizes the important role of the optimal tuning of α to mitigate the double descent phenomenon and to give the best performance.

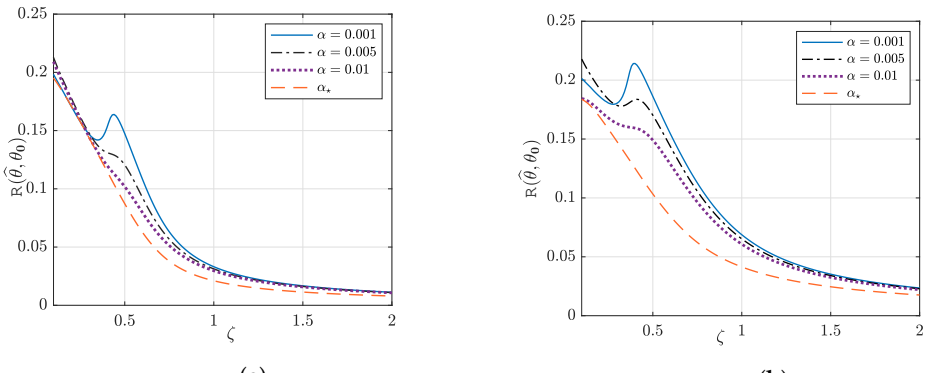


Figure 10. Prediction risk as a function of the aspect ratio ζ. We used G-PCR with an ℓ_1-norm penalty and sparse-binary vector with $\rho = 0.2, \sigma_e^2 = 0.1$, $L = 0$, and $U = 1$. (**a**) $\sigma_\epsilon^2 = 0.1$. (**b**) $\sigma_\epsilon^2 = 0.3$. Illustration of the double descent and how optimal penalization can mitigate it.

7. Conclusions

In this paper, we studied the high-dimensional error performance of the generalized penalized constrained regression (G-PCR) optimization with noisy features. Several analytical expressions were derived to measure the performance, such as the prediction risk, similarity, probability of misdetection, and probability of false alarm. Different popular instances of this optimization, such as ℓ_1-norm penalized regression and ℓ_2-norm penalization, were considered. We presented numerical simulations to validate these expressions based on both synthetic and real data. These results can be used to tune the involved hyper-parameters efficiently.

Furthermore, we empirically investigated the so-called double descent phenomenon and showed that optimal penalization can mitigate its effect. We also illustrated through several simulations the universality of our results beyond the assumed Gaussian distribution.

Finally, we note that numerical simulations have shown that our rigorous results are accurate even for problems with a few hundred variables, despite the fact that these results are assumed to be asymptotic in the problem dimensions.

Author Contributions: Conceptualization, A.M.A.; Methodology, A.M.A.; Software, A.M.A.; Validation, A.M.A.; Formal analysis, A.M.A.; Investigation, A.M.A. and M.A.; Resources, M.A.; Data curation, A.M.A. and M.A.A.; Writing—original draft, A.M.A.; Writing—review & editing, A.M.A., M.A. and M.A.A.; Visualization, A.M.A. and M.A.; Supervision, A.M.A.; Project administration, M.A.A.; Funding acquisition, M.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deputyship for Research & Innovation, Ministry of Education, Saudi Arabia through project number 445-9-196.

Data Availability Statement: The data presented in this study are available within the article.

Acknowledgments: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number 445-9-196. Also, the authors would like to extend their appreciation to Taibah University for its supervision support.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of the Main Results

In this appendix, we provide an outline of the proof of the high-dimensional analysis of the prediction risk of the considered G-PCR learning algorithm. Our main analysis framework is the convex Gaussian min–max theorem (CGMT). For the reader's convenience, we firstly recall the CGMT.

Appendix A.1. Main Analysis Framework: CGMT

The CGMT is an extension of Gordon's comparison lemma [76]. Gordon's lemma was used in the analysis of some high-dimensional inference problems, such as the study of sharp phase-transitions in noiseless compressed sensing. The CGMT was initiated first in [36] and further developed in [20]. It uses convexity to compare the min–max values of two Gaussian processes.

To illustrate the main ideas of the CGMT, let us first consider the following doubly indexed Gaussian random processes:

$$\mathcal{X}_{\mathbf{r},\mathbf{w}} := \mathbf{w}^\top \mathbf{G}\mathbf{r} + \Xi(\mathbf{r},\mathbf{w}), \tag{A1a}$$

$$\mathcal{Y}_{\mathbf{r},\mathbf{w}} := \|\mathbf{r}\|_2 \mathbf{h}_1^\top \mathbf{w} + \|\mathbf{w}\|_2 \mathbf{h}_2^\top \mathbf{r} + \Xi(\mathbf{r},\mathbf{w}), \tag{A1b}$$

where $\mathbf{G} \in \mathbb{R}^{n\times p}$, $\mathbf{h}_1 \in \mathbb{R}^n$, $\mathbf{h}_2 \in \mathbb{R}^p$, they all have i.i.d. standard Gaussian elements, and $\Xi : \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}$. For these two processes, consider the following min–max optimization programs, which are referred to as the primal optimization (PO) and the auxiliary optimization (AO):

$$F(\mathbf{G}) := \min_{\mathbf{r}\in\mathcal{S}_\mathbf{r}} \max_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \mathcal{X}_{\mathbf{r},\mathbf{w}}, \tag{A2a}$$

$$f(\mathbf{h}_1,\mathbf{h}_2) := \min_{\mathbf{r}\in\mathcal{S}_\mathbf{r}} \max_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \mathcal{Y}_{\mathbf{r},\mathbf{w}}, \tag{A2b}$$

where the sets $\mathcal{S}_\mathbf{r} \subset \mathbb{R}^p$ and $\mathcal{S}_\mathbf{w} \subset \mathbb{R}^n$ are assumed to be compact and convex sets. In addition, if the function $\Xi(\mathbf{r},\mathbf{w})$ is continuous and convex–concave on $\mathcal{S}_\mathbf{r} \times \mathcal{S}_\mathbf{w}$, then, according to the CGMT formulation in Theorem 6 in [20], for any $\chi \in \mathbb{R}$ and $\mu > 0$:

$$\mathbb{P}(|F(\mathbf{G}) - \chi| > \mu) \leq 2\mathbb{P}(|f(\mathbf{h}_1,\mathbf{h}_2) - \chi| > \mu). \tag{A3}$$

The above result states that if we can show that the optimal AO cost is $f(\mathbf{h}_1,\mathbf{h}_2) \xrightarrow{P} c_\star$ asymptotically, $c_\star \in \mathbb{R}$, then it can be concluded that the optimal PO cost is $F(\mathbf{G}) \xrightarrow{P} c_\star$. The premise is that it is usually much easier to analyze the AO instead of the PO. In addition, the CGMT (Theorem 6.1(iii) in [20]) shows that concentration of the optimal solution to the AO problem implies concentration of the optimal solution of the PO around the same value. In other words, if minimizers of (A2b) satisfy that $\|\widehat{\mathbf{r}}_f(\mathbf{h}_1,\mathbf{h}_2)\|_2 \xrightarrow{P} \nu_\star$, where $\nu_\star > 0$, then the same holds true for minimizers of (A2a), i.e., $\|\widehat{\mathbf{r}}_F(\mathbf{G})\|_2 \xrightarrow{P} \nu_\star$. In addition, we make use of the following corollary that holds true in the high-dimensional asymptotic regime.

Corollary A1 (Asymptotic CGMT [20])**.** *Using the same notations and assumptions as in the above discussion, let $\mathcal{S} \subset \mathcal{S}_\mathbf{r}$ and $\mathcal{S}^c := \mathcal{S}_\mathbf{r}/\mathcal{S}$. Define $F_{\mathcal{S}^c}(\mathbf{G})$ and $f_{\mathcal{S}^c}(\mathbf{h}_1,\mathbf{h}_2)$ as the optimal costs in (A2a) and (A2b), respectively, given that we now constrain the optimization over $\mathbf{r} \in \mathcal{S}^c$. Suppose there exist constants $\bar{J} < \bar{J}_{\mathcal{S}^c}$, such that $f_{\mathcal{S}^c}(\mathbf{h}_1,\mathbf{h}_2) \xrightarrow{P} \bar{J}_{\mathcal{S}^c}$ and $f(\mathbf{h}_1,\mathbf{h}_2) \xrightarrow{P} \bar{J}$. Then,*

$$\lim_{p\to\infty} \mathbb{P}(\widehat{\mathbf{r}}_{F(\mathbf{G})} \in \mathcal{S}) = 1. \tag{A4}$$

For more details about the framework of CGMT, the reader is advised to see [20].

Next, we use the CGMT to provide a proof outline of the general error asymptotic behavior provided in Theorem 1.

Appendix A.2. Sharp Analysis of the G-PCR

Appendix A.2.1. Primal and Auxiliary Problems of the G-PCR

To obtain the main asymptotic result using CGMT, we first need to rewrite the G-PCR learning problem in (9) as a PO problem. For convenience, define the vector $\mathbf{r} := \boldsymbol{\theta} - \boldsymbol{\theta}_0$, and the following set

$$\mathbb{B} := \left\{ r_j \in \mathbb{R} \mid -L - \theta_{0,j} \leq r_j \leq U - \theta_{0,j}, j \in [p] \right\}; \tag{A5}$$

then, the problem in (9) can be reformulated as

$$\widehat{\mathbf{r}} := \arg\min_{\mathbf{r}\in\mathbb{B}^p} \|\widehat{\mathbf{X}}\mathbf{r} + \mathbf{E}\boldsymbol{\theta}_0 - \boldsymbol{\epsilon}\|_2^2 + \alpha\mathcal{P}(\mathbf{r} + \boldsymbol{\theta}_0). \tag{A6}$$

Introducing the Convex Conjugate: Any convex function $h : \mathbb{R}^n \to \mathbb{R}$ can be expressed in terms of its convex conjugate $h^\star : \mathbb{R}^n \to \mathbb{R}$ as:

$$h(\mathbf{t}) = \sup_{\bar{\mathbf{w}}\in\mathbb{R}^n} \bar{\mathbf{w}}^\top\mathbf{t} - h^\star(\bar{\mathbf{w}}) = \sup_{\mathbf{w}\in\mathbb{R}^n} \sqrt{p}\mathbf{w}^\top\mathbf{t} - h^\star(\sqrt{p}\mathbf{w}).$$

Using the above definition, we can express the ℓ_2^2-norm loss function in (A6) as

$$\|\mathbf{t}\|_2^2 = \sup_{\mathbf{w}\in\mathbb{R}^n} \sqrt{p}\mathbf{w}^\top\mathbf{t} - \frac{p}{4}\|\mathbf{w}\|_2^2. \tag{A7}$$

Hence, (A6) becomes equivalent to the following:

$$\min_{\mathbf{r}\in\mathbb{B}^p}\sup_{\mathbf{w}\in\mathbb{R}^n}\sqrt{p}\mathbf{w}^\top\widehat{\mathbf{X}}\mathbf{r}+\sqrt{p}\mathbf{w}^\top\mathbf{E}\boldsymbol{\theta}_0-\sqrt{p}\mathbf{w}^\top\boldsymbol{\epsilon}-\frac{p}{4}\|\mathbf{w}\|_2^2+\alpha\mathcal{P}(\mathbf{r}+\boldsymbol{\theta}_0). \tag{A8}$$

To apply the CGMT, we need the optimization sets to be compact. This is true for $\mathcal{S}_\mathbf{r}=\mathbb{B}^p$, but $\mathcal{S}_\mathbf{w}=\mathbb{R}^n$ is not. This issue can be treated in a similar way to the method in Appendix A in [20]. We introduce an artificial compact set $\mathcal{S}_\mathbf{w}=\{\mathbf{w}\in\mathbb{R}^n \mid \|\mathbf{w}\|_2\le R_w\}$ for a sufficiently large constant $R_w>0$ that is independent of p. The optimization problem is unaffected by this constraint set asymptotically. After that, we obtain

$$\min_{\mathbf{r}\in\mathbb{B}^p}\sup_{\mathbf{w}\in\mathcal{S}_\mathbf{w}}\sqrt{p(1-\sigma_e^2)}\mathbf{w}^\top\widetilde{\mathbf{X}}\mathbf{r}+\sigma_e\sqrt{p}\mathbf{w}^\top\widetilde{\mathbf{E}}\boldsymbol{\theta}_0-\sqrt{p}\mathbf{w}^\top\boldsymbol{\epsilon}-\frac{p}{4}\|\mathbf{w}\|_2^2+\alpha\mathcal{P}(\mathbf{r}+\boldsymbol{\theta}_0), \tag{A9}$$

where $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{E}}$ are independent Gaussian matrices that have i.i.d. $\mathsf{N}(0,1/p)$ elements each. Now, the above problem is in the format of a PO with

$$\Xi(\mathbf{r},\mathbf{w})=\sigma_e\sqrt{p}\mathbf{w}^\top\widetilde{\mathbf{E}}\boldsymbol{\theta}_0-\sqrt{p}\mathbf{w}^\top\boldsymbol{\epsilon}-\frac{p}{4}\|\mathbf{w}\|_2^2+\alpha\mathcal{P}(\mathbf{r}+\boldsymbol{\theta}_0).$$

Therefore, the corresponding AO problem is

$$\min_{\mathbf{r}\in\mathbb{B}^p}\sup_{\mathbf{w}\in\mathcal{S}_\mathbf{w}}\sqrt{1-\sigma_e^2}\|\mathbf{r}\|_2\mathbf{h}_1^\top\mathbf{w}+\sqrt{1-\sigma_e^2}\|\mathbf{w}\|_2\mathbf{h}_2^\top\mathbf{r}+\Xi(\mathbf{r},\mathbf{w}), \tag{A10}$$

where $\mathbf{h}_1\sim\mathsf{N}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{h}_2\sim\mathsf{N}(\mathbf{0},\mathbf{I}_p)$ are independent standard Gaussian vectors.

Appendix A.2.2. Simplifying the Auxiliary Problem

The next step is to reduce (simplify) the AO into a scalar problem, i.e., a problem that has only scalar variables. To do so, first, let

$$\widetilde{\mathbf{h}}=\sqrt{1-\sigma_e^2}\|\mathbf{r}\|_2\mathbf{h}_1-\sqrt{p}\boldsymbol{\epsilon}+\sqrt{p\sigma_e^2}\widetilde{\mathbf{E}}\boldsymbol{\theta}_0. \tag{A11}$$

Using standard probability theory results, one can show that $\widetilde{\mathbf{h}}\sim\mathsf{N}(\mathbf{0},\mathbf{C}_{\widetilde{\mathbf{h}}})$ with a covariance matrix that is given by

$$\mathbf{C}_{\widetilde{\mathbf{h}}}=\left((1-\sigma_e^2)\|\mathbf{r}\|_2^2+p\sigma_\epsilon^2+\sigma_e^2\|\boldsymbol{\theta}_0\|_2^2\right)\mathbf{I}_n.$$

Thus, the AO in (A10) holds that

$$\min_{\mathbf{r}\in\mathbb{B}^p}\sup_{\mathbf{w}\in\mathcal{S}_\mathbf{w}}\widetilde{\mathbf{h}}^\top\mathbf{w}+\sqrt{1-\sigma_e^2}\|\mathbf{w}\|_2\mathbf{h}_2^\top\mathbf{r}-\frac{p}{4}\|\mathbf{w}\|_2^2+\alpha\mathcal{P}(\mathbf{r}+\boldsymbol{\theta}_0). \tag{A12}$$

In order to further simplify the AO, we fix the norm of $\mathbf{w}$ to $q:=\|\mathbf{w}\|_2$. In this case, one can simply optimize over the direction of $\mathbf{w}$, which reduces the AO problem to

$$\min_{\mathbf{r}\in\mathbb{B}^p}\sup_{q\ge0}q\|\widetilde{\mathbf{h}}\|_2+\sqrt{1-\sigma_e^2}q\mathbf{h}_2^\top\mathbf{r}-\frac{pq^2}{4}+\alpha\mathcal{P}(\mathbf{r}+\boldsymbol{\theta}_0). \tag{A13}$$

Moreover, to have the proper convergence, we have to normalize the above cost function by factor of $\frac{1}{p}$. Then, we obtain

$$\sup_{q\ge0}\min_{\mathbf{r}\in\mathbb{B}^p}q\sqrt{\frac{1}{p}\left[(1-\sigma_e^2)\|\mathbf{r}\|_2^2+p\sigma_\epsilon^2+\sigma_e^2\|\boldsymbol{\theta}_0\|_2^2\right]}\ \frac{\|\mathbf{g}\|_2}{\sqrt{p}}+\sqrt{1-\sigma_e^2}q\frac{1}{p}\mathbf{h}_2^\top\mathbf{r}-\frac{q^2}{4}+\frac{\alpha}{p}\mathcal{P}(\mathbf{r}+\boldsymbol{\theta}_0), \tag{A14}$$

where $\mathbf{g}\sim\mathsf{N}(\mathbf{0},\mathbf{I}_n)$. Note the change in the order of the min–sup, which can be justified using (Appendix A in [20]). Next, we wish to write the above optimization as a separable problem by using the following (for $u\ge0$):

$$\sqrt{u} = \inf_{\gamma>0} \frac{\gamma u}{2} + \frac{1}{2\gamma}. \tag{A15}$$

Note that the optimal solution to (A15) is $\hat{\gamma} = \frac{1}{\sqrt{u}}$. Using this identity with

$$u = (1-\sigma_e^2)\frac{1}{p}\|\mathbf{r}\|_2^2 + \sigma_\epsilon^2 + \sigma_e^2\frac{1}{p}\|\boldsymbol{\theta}_0\|_2^2, \tag{A16}$$

we can write the problem in (A14) as

$$\begin{aligned} \sup_{q\geq 0}\inf_{\gamma>0} \quad & \tfrac{q\|\mathbf{g}\|_2}{2\gamma\sqrt{p}} + \tfrac{\gamma q\|\mathbf{g}\|_2}{2\sqrt{p}}\left(\sigma_\epsilon^2 + \tfrac{\sigma_e^2}{p}\|\boldsymbol{\theta}_0\|_2^2\right) - \tfrac{q^2}{4} \\ & + \min_{\mathbf{r}\in\mathbb{B}^p} \tfrac{\gamma q\|\mathbf{g}\|_2}{2\sqrt{p}}\tfrac{(1-\sigma_e^2)}{p}\|\mathbf{r}\|_2^2 + q\sqrt{1-\sigma_e^2}\tfrac{\mathbf{h}_2^\top\mathbf{r}}{p} + \tfrac{\alpha}{p}\mathcal{P}(\mathbf{r}+\boldsymbol{\theta}_0). \end{aligned} \tag{A17}$$

By the weak law of large numbers (WLLN), we can show that $\frac{\|\mathbf{g}\|_2}{\sqrt{p}} \xrightarrow{P} \sqrt{\zeta}$ and $\frac{1}{p}\|\boldsymbol{\theta}_0\|_2^2 \xrightarrow{P} \sigma_\theta^2$. Now, let us work with the initial variable $\boldsymbol{\theta}$ rather than $\mathbf{r}$; then, the above optimization problem converges to

$$\begin{aligned} \sup_{q\geq 0}\inf_{\gamma>0} \quad & \tfrac{q\sqrt{\zeta}}{2\gamma} + \tfrac{q\gamma\sqrt{\zeta}}{2}\left(\sigma_\epsilon^2 + \sigma_e^2\sigma_\theta^2\right) - \tfrac{q^2}{4} \\ & + \tfrac{1}{p}\textstyle\sum_{j=1}^p \min_{-L\leq\theta_j\leq U}\left\{\tfrac{q\gamma\sqrt{\zeta}}{2}(1-\sigma_e^2)(\theta_j-\theta_{0,j})^2 + \sqrt{1-\sigma_e^2}qh_{2,j}(\theta_j-\theta_{0,j}) + \alpha\widetilde{\mathcal{P}}(\theta_j)\right\}. \end{aligned} \tag{A18}$$

Completing the squares in θ_j in the last minimization of the above problem, and using the fact that $\frac{1}{p}\mathbf{h}_2^\top\boldsymbol{\theta}_0 \xrightarrow{P} 0$, we obtain

$$\begin{aligned} \sup_{q\geq 0}\inf_{\gamma>0} \quad & \tfrac{q\sqrt{\zeta}}{2\gamma} + \tfrac{q\gamma\sqrt{\zeta}}{2}\left(\sigma_\epsilon^2 + \sigma_e^2\sigma_\theta^2\right) - \tfrac{q^2}{4} - \tfrac{1}{p}\textstyle\sum_{j=1}^p \tfrac{q}{2\gamma\sqrt{\zeta}}h_{2,j}^2 \\ & + q\gamma\sqrt{\zeta}(1-\sigma_e^2)\tfrac{1}{p}\textstyle\sum_{j=1}^p \min_{-L\leq\theta_j\leq U}\tfrac{1}{2}\left(\theta_j - \left(\theta_{0,j} + \tfrac{h_{2,j}}{\gamma\sqrt{\zeta(1-\sigma_e^2)}}\right)\right)^2 + \tfrac{\alpha}{q\gamma\sqrt{\zeta}(1-\sigma_e^2)}\widetilde{\mathcal{P}}(\theta_j). \end{aligned} \tag{A19}$$

Note that the last summation term in (A19) can be expressed by the generalized Moreau envelope function $\mathcal{M}_{\widetilde{\mathcal{P}}}(\cdot)$ defined in (1). Hence, we obtain the following problem:

$$\begin{aligned} \sup_{q\geq 0}\inf_{\gamma>0} \quad & \tfrac{q\sqrt{\zeta}}{2\gamma} + \tfrac{q\gamma\sqrt{\zeta}}{2}\left(\sigma_\epsilon^2 + \sigma_e^2\sigma_\theta^2\right) - \tfrac{q^2}{4} - \tfrac{1}{p}\textstyle\sum_{j=1}^p \tfrac{q}{2\gamma\sqrt{\zeta}}h_{2,j}^2 \\ & + q\gamma\sqrt{\zeta}(1-\sigma_e^2)\tfrac{1}{p}\textstyle\sum_{j=1}^p \mathsf{M}_{\widetilde{\mathcal{P}}}\left(\theta_{0,j} + \tfrac{h_{2,j}}{\gamma\sqrt{\zeta(1-\sigma_e^2)}}; \tfrac{\alpha}{q\gamma\sqrt{\zeta}(1-\sigma_e^2)}, -L, U\right). \end{aligned} \tag{A20}$$

Next, by the WLLN, $\frac{1}{p}\sum_{j=1}^p h_{2,j}^2 \xrightarrow{P} 1$, and for all $q>0$ and $\gamma>0$, we have

$$\begin{aligned} & \frac{1}{p}\sum_{j=1}^p \mathsf{M}_{\widetilde{\mathcal{P}}}\left(\theta_{0,j} + \frac{h_{2,j}}{\gamma\sqrt{\zeta(1-\sigma_e^2)}}; \frac{\alpha}{q\gamma\sqrt{\zeta}(1-\sigma_e^2)}, -L, U\right) \\ & \quad \xrightarrow{P} \mathbb{E}\left[\mathsf{M}_{\widetilde{\mathcal{P}}}\left(\Theta_0 + \frac{H}{\gamma\sqrt{\zeta(1-\sigma_e^2)}}; \frac{\alpha}{q\gamma\sqrt{\zeta}(1-\sigma_e^2)}, -L, U\right)\right], \end{aligned}$$

where the expectation is taken with respect to the independent scalar random variables $\Theta_0 \sim p_\Theta$ and $H \sim \mathsf{N}(0,1)$.

Finally, (A20) converges to the following scalar problem:

$$\sup_{q\geq 0}\inf_{\gamma>0} \quad \frac{q\sqrt{\zeta}}{2\gamma}+\frac{q\gamma\sqrt{\zeta}}{2}\left(\sigma_\epsilon^2+\sigma_e^2\sigma_\theta^2\right)-\frac{q^2}{4}-\frac{q}{2\gamma\sqrt{\zeta}} \\ +q\gamma\sqrt{\zeta}(1-\sigma_e^2)\mathbb{E}\left[\mathsf{M}_{\widetilde{\mathcal{P}}}\left(\Theta_0+\frac{H}{\gamma\sqrt{\zeta(1-\sigma_e^2)}};\frac{\alpha}{q\gamma\sqrt{\zeta}(1-\sigma_e^2)},-L,U\right)\right]. \tag{A21}$$

Appendix A.2.3. General Performance Metric: Proof of Theorem 1

Now that we derived the scalar optimization problem, we proceed to prove Theorem 1. Recall that in the process of scalarizing the AO, we introduced the generalized Moreau envelope function in (A20). It can be shown that the optimizer of this function gives the AO solution in $\boldsymbol{\theta}$. Let $(q_\star,\gamma_\star)$ be the unique solution to (A21). Then, the AO solution can be presented as

$$\widehat{\theta}_j^{\mathrm{AO}}=\widehat{\Theta}:=\mathrm{prox}_{\widetilde{\mathcal{P}}}\left(\Theta_0+\frac{H}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}};\frac{\alpha}{q_\star\gamma_\star\sqrt{\zeta}(1-\sigma_e^2)},-L,U\right),j\in[p],$$

where H is a standard normal random variable and $\Theta_0\sim p_\Theta$ independent of H.

The last step is to show the convergence of any pseudo-Lipschitz function $\psi(\cdot,\cdot)$. Using the weak law of large numbers and the fact that the elements of $\boldsymbol{\theta}_0$ are i.i.d. sampled from a density p_θ, we obtain

$$\frac{1}{p}\sum_{j=1}^{p}\psi(\widehat{\theta}_j^{\mathrm{AO}},\theta_{0,j})\xrightarrow{P}\mathbb{E}\Big[\psi(\widehat{\Theta},\Theta_0)\Big], \tag{A22}$$

where the expectation is taken over $H\sim\mathsf{N}(0,1)$ and $\Theta_0\sim p_\Theta$ independent of H. To use the CGMT (Corollary A1), we introduce the following set:

$$\mathcal{S}_\eta=\left\{\mathbf{v}\in\mathbb{R}^p\mid\left|\frac{1}{p}\sum_{j=1}^{p}\psi(v_j,\theta_{0,j})-\mathbb{E}\Big[\psi(\widehat{\Theta},\Theta_0)\Big]\right|<\eta\right\},$$

for $\eta>0$.

The convergence result in (A22) establishes that $\lim_{p\to\infty}\mathbb{P}(\widehat{\boldsymbol{\theta}}^{\mathrm{AO}}\in\mathcal{S}_\eta)=1$. Hence, using the CGMT (Corollary A1), $\lim_{p\to\infty}\mathbb{P}(\widehat{\boldsymbol{\theta}}\in\mathcal{S}_\eta)=1$, where $\widehat{\boldsymbol{\theta}}$ is the solution to the original G-PCR in (9). This concludes the proof of Theorem 1.

Appendix A.2.4. Prediction Risk Analysis: Proof of Corollary 1

The objective of this part is to analyze the prediction risk of the G-PCR asymptotically. To begin with, for any $\eta>0$, define the following set:

$$\check{\mathcal{S}}_\eta=\left\{\mathbf{r}\in\mathbb{R}^p\mid\left|\frac{1}{p}\|\mathbf{r}\|_2^2-\frac{1}{1-\sigma_e^2}\left(\frac{1}{\gamma_\star^2}-\sigma_\theta^2\sigma_e^2-\sigma_\epsilon^2\right)\right|<\eta\right\},$$

where $\gamma_\star$ is the solution to (A21). Recall from (A16) that $\hat{\gamma}_p=\frac{1}{\sqrt{\hat{u}}}$ and $\hat{u}=(1-\sigma_e^2)\frac{1}{p}\|\widetilde{\mathbf{r}}\|_2^2$ $+\sigma_\epsilon^2+\sigma_e^2\frac{1}{p}\|\boldsymbol{\theta}_0\|_2^2$. Hence,

$$\frac{1}{p}\|\widetilde{\mathbf{r}}\|_2^2=\frac{1}{1-\sigma_e^2}\left(\frac{1}{\hat{\gamma}_p^2}-\frac{\sigma_e^2}{p}\|\boldsymbol{\theta}_0\|_2^2-\sigma_\epsilon^2\right),$$

where $\widetilde{\mathbf{r}}$ is the optimal solution to (A14) and $\hat{\gamma}_p$ is the solution to (A17). Using the uniform convergence of the cost functions, we can show that $\hat{\gamma}_p\xrightarrow{P}\gamma_\star$. Hence, using the WLLN, $\frac{1}{p}\|\boldsymbol{\theta}_0\|_2^2\xrightarrow{P}\sigma_\theta^2$ and, therefore,

$$\mathsf{R}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}_0)=\frac{1}{p}\|\widetilde{\mathbf{r}}\|_2^2=\frac{1}{p}\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2^2\xrightarrow{P}\frac{1}{1-\sigma_e^2}\left(\frac{1}{\gamma_\star^2}-\sigma_\theta^2\sigma_e^2-\sigma_\epsilon^2\right). \tag{A23}$$

Remember that $\widetilde{\mathbf{r}} = \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$, where $\widetilde{\boldsymbol{\theta}}$ is the AO solution in $\boldsymbol{\theta}$. From Equation (A23), we can see that, for all $\eta > 0$, $\mathsf{R}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \in \check{\mathcal{S}}_\eta$, with probability approaching 1. Then, an application of the CGMT yields that $\mathsf{R}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \in \check{\mathcal{S}}_\eta$ with high probability.

Furthermore, Corollary 1 can also be proven as an immediate result of Theorem 1 with $\psi(a, b) = (a - b)^2$ therein. Hence,

$$\mathsf{R}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \xrightarrow{P} \mathbb{E}_{\Theta_0, H}\left[\left(\mathrm{prox}_{\widetilde{\mathcal{P}}}\left(\Theta_0 + \frac{H}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}}; \frac{\alpha}{q_\star\gamma_\star\sqrt{\zeta}(1-\sigma_e^2)}, -L, U\right) - \Theta_0\right)^2\right]. \tag{A24}$$

Combining the results in (A23) and (A24) concludes the proof of Corollary 1.

Appendix A.2.5. Similarity Analysis: Proof of Corollary 2

The proof of Corollary 2 is based on the CGMT to derive asymptotic predictions of the numerator and the denominator of the similarity expression $\varrho(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ in (11) separately, and then to use the continuous mapping Theorem [62] to arrive at the desired result. For the sake of brevity, we only highlight the main steps of the proof.

The similarity expression in (11) can be rewritten as

$$\varrho(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) = \frac{\frac{1}{p}\sum_{j=1}^{p}\widehat{\theta}_j\theta_{0,j}}{\sqrt{\frac{1}{p}\sum_{j=1}^{p}\widehat{\theta}_j^2}\cdot\sqrt{\frac{1}{p}\sum_{j=1}^{p}\theta_{\star,j}^2}}. \tag{A25}$$

For the numerator, we use Theorem 1 with $\psi(a, b) = a \cdot b$, to obtain the following convergence:

$$\text{Numerator} = \frac{1}{p}\sum_{j=1}^{p}\widehat{\theta}_j\theta_{0,j} \xrightarrow{P} \mathbb{E}_{\Theta_0, H}\left[\mathrm{prox}_{\widetilde{\mathcal{P}}}\left(\Theta_0 + \frac{H}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}}; \frac{\alpha}{q_\star\gamma_\star\sqrt{\zeta}(1-\sigma_e^2)}, -L, U\right)\Theta_0\right].$$

The denominator consists of two terms, each of which converges as well. For the first term, $\sqrt{\frac{1}{p}\sum_{j=1}^{p}\widehat{\theta}_j^2}$, use Theorem 1 with $\psi(a, a) = a^2$ and the continuous mapping Theorem to obtain

$$\sqrt{\frac{1}{p}\sum_{j=1}^{p}\widehat{\theta}_j^2} \xrightarrow{P} \sqrt{\mathbb{E}_{\Theta_0, H}\left[\mathrm{prox}_{\widetilde{\mathcal{P}}}^2\left(\Theta_0 + \frac{H}{\gamma_\star\sqrt{\zeta(1-\sigma_e^2)}}; \frac{\alpha}{q_\star\gamma_\star\sqrt{\zeta}(1-\sigma_e^2)}, -L, U\right)\right]}.$$

For the second term in the denominator, $\sqrt{\frac{1}{p}\sum_{j=1}^{p}\theta_{\star,j}^2}$, using the WLLN, we have

$$\sqrt{\frac{1}{p}\sum_{j=1}^{p}\theta_{\star,j}^2} \xrightarrow{P} \sqrt{\mathbb{E}[\Theta_0^2]} = \sqrt{\sigma_\theta^2}.$$

Putting together all of the above convergence results coupled with an application of the continuous mapping Theorem [62], we obtain the asymptotic expression of the similarity measure in (17).

Appendix B. A Note on the Square-Root Generalized Penalized Constrained Regression

Sqrt G-PCR Learning Algorithm

Let us consider the following optimization problem:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{V}^p} \|\widehat{\mathbf{X}}\boldsymbol{\theta} - \mathbf{y}\|_2 + \frac{\alpha}{\sqrt{p}}\mathcal{P}(\boldsymbol{\theta}), \tag{A26}$$

where

$$\mathbb{V} = [-L, U], \text{ and } \quad L, U \in \mathbb{R}_+ \cup \{0\}.$$

Problems of the above type are known as regularized square-root regression problems [77]. Here, instead of the ℓ_2^2-norm squared loss in (9), there is a non-squared ℓ_2-norm loss. This leads to optimization problems with a loss function that is not separable. Examples of this algorithm include the square-root LASSO [24] and the square-root group LASSO [78]. Please see [20,64,77] for the motivations for using the non-squared loss. Furthermore, the scaling of the penalization factor α by a factor of $\frac{1}{\sqrt{p}}$ is just for convergence issues of the analysis of the CGMT (see [20] for further justification). The analysis of the above optimization, which we call the **Square-root G-PCR (Sqrt G-PCR)** is very similar to the one provided in the previous sections of this paper for the G-PCR problem. The only difference, however, is that, instead of (A7), we have

$$\|\mathbf{t}\|_2 = \max_{\|\mathbf{w}\|_2 \leq 1} \mathbf{w}^\top \mathbf{t}. \tag{A27}$$

Following the same analysis (with some normalization adjustments) as in Appendix A, but using (A27) instead of (A7), we finally arrive at the following deterministic scalar max–min optimization problem:

$$\begin{aligned} \sup_{0\leq q\leq 1} \inf_{\gamma>0} \widetilde{\mathcal{O}}_{\widetilde{\mathcal{P}}}(q,\gamma): \quad &= \tfrac{q\sqrt{\zeta}}{2\gamma} + \tfrac{q\gamma\sqrt{\zeta}}{2}\left(\sigma_\epsilon^2 + \sigma_e^2\sigma_\beta^2\right) - \tfrac{q}{2\gamma\sqrt{\zeta}} \\ &\quad + q\gamma\sqrt{\zeta}(1-\sigma_e^2)\mathbb{E}\left[\mathsf{M}_{\widetilde{\mathcal{P}}}\left(\Theta_0 + \tfrac{H}{\gamma\sqrt{\zeta(1-\sigma_e^2)}}; \tfrac{\alpha}{q\gamma\sqrt{\zeta}(1-\sigma_e^2)}, -L, U\right)\right]. \end{aligned} \tag{A28}$$

Comparing $\widetilde{\mathcal{O}}_{\widetilde{\mathcal{P}}}(q,\gamma)$ to $\mathcal{O}(q,\gamma)_{\widetilde{\mathcal{P}}}$ in (13), we can see two main differences, which are the absence of the $-\frac{q^2}{4}$ term and the presence of the constraint $0 \leq q \leq 1$ in $\widetilde{\mathcal{O}}_{\widetilde{\mathcal{P}}}(q,\gamma)$.

This means that the prediction risk and the similarity of the Sqrt G-PCR in (A26) converge to the same asymptotic limits in (16) and (17), respectively, but now with $q_\star$ and $\gamma_\star$, which are solutions to (A28) instead of (13).

References

1. Parikh, N.; Boyd, S. Proximal algorithms. *Found. Trends Optim.* **2014**, *1*, 127–239. [CrossRef]
2. Tarantola, A. *Inverse Problem Theory and Methods for Model Parameter Estimation*; SIAM: Philadelphia, PA, USA, 2005.
3. Kailath, T.; Sayed, A.H.; Hassibi, B. *Linear Estimation*; Prentice Hall: Hoboken, NJ, USA, 2000.
4. Groetsch, C.W.; Groetsch, C. *Inverse Problems in the Mathematical Sciences*; Springer: Berlin/Heidelberg, Germany, 1993; Volume 52.
5. Bishop, C.M. Pattern recognition. *Mach. Learn.* **2006**, *128*, 1–58.
6. Rencher, A.C.; Schaalje, G.B. *Linear Models in Statistics*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
7. Dhar, V. Data science and prediction. *Commun. ACM* **2013**, *56*, 64–73. [CrossRef]
8. Donoho, D.L. Compressed sensing. *IEEE Trans. Inf. Theory* **2006**, *52*, 1289–1306. [CrossRef]
9. Duarte, M.F.; Eldar, Y.C. Structured compressed sensing: From theory to applications. *IEEE Trans. Signal Process.* **2011**, *59*, 4053–4085. [CrossRef]
10. Poor, H.V. *An Introduction to Signal Detection and Estimation*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1998.
11. Fadili, J.M.; Bullmore, E. Penalized partially linear models using sparse representations with an application to fMRI time series. *IEEE Trans. Signal Process.* **2005**, *53*, 3436–3448. [CrossRef]
12. Goldsmith, A. *Wireless Communications*; Cambridge University Press: Cambridge, UK, 2005.
13. Marzetta, T.L.; Yang, H. *Fundamentals of Massive MIMO*; Cambridge University Press: Cambridge, UK, 2016.
14. Candes, E.; Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.* **2007**, *35* , 2313–2351.
15. Bach, F. Structured sparsity-inducing norms through submodular functions. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–9 December 2010; Volume 23.
16. Aster, R.C.; Borchers, B.; Thurber, C.H. *Parameter Estimation and Inverse Problems*; Elsevier: Amsterdam, The Netherlands, 2018.
17. McDonald, G.C. Ridge regression. *Wiley Interdiscip. Rev. Comput. Stat.* **2009**, *1*, 93–100. [CrossRef]
18. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
19. Varah, J.M. Pitfalls in the numerical solution of linear ill-posed problems. *SIAM J. Sci. Stat. Comput.* **1983**, *4*, 164–176. [CrossRef]
20. Thrampoulidis, C.; Abbasi, E.; Hassibi, B. Precise error analysis of regularized *M*-estimators in high dimensions. *IEEE Trans. Inf. Theory* **2018**, *64*, 5592–5628. [CrossRef]

21. Bickel, P.J.; Ritov, Y.; Tsybakov, A.B. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.* **2009**, *37*, 1705–1732. [CrossRef]
22. Negahban, S.; Yu, B.; Wainwright, M.J.; Ravikumar, P.K. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; pp. 1348–1356.
23. Wainwright, M.J. Sharp thresholds for high-dimensional and noisy sparsity recovery using -constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **2009**, *55*, 2183–2202. [CrossRef]
24. Belloni, A.; Chernozhukov, V.; Wang, L. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **2011**, *98*, 791–806. [CrossRef]
25. Li, Y.H.; Hsieh, Y.P.; Zerbib, N.; Cevher, V. A geometric view on constrained M-estimators. *arXiv* **2015**, arXiv:1506.08163.
26. Bayati, M.; Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory* **2011**, *57*, 764–785. [CrossRef]
27. Bayati, M.; Montanari, A. The LASSO risk for Gaussian matrices. *IEEE Trans. Inf. Theory* **2012**, *58*, 1997–2017. [CrossRef]
28. Donoho, D.; Montanari, A. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Relat. Fields* **2016**, *166*, 935–969. [CrossRef]
29. Rangan, S.; Goyal, V.; Fletcher, A.K. Asymptotic analysis of map estimation via the replica method and compressed sensing. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; Volume 22.
30. Kabashima, Y.; Wadayama, T.; Tanaka, T. Statistical mechanical analysis of a typical reconstruction limit of compressed sensing. In Proceedings of the 2010 IEEE International Symposium on Information Theory, Austin, TX, USA, 13–18 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 1533–1537.
31. Couillet, R.; Debbah, M. *Random Matrix Methods for Wireless Communications*; Cambridge University Press: Cambridge, UK, 2011.
32. Karoui, N.E. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: Rigorous results. *arXiv* **2013**, arXiv:1311.2445.
33. Liao, Z.; Couillet, R. Random matrices meet machine learning: A large dimensional analysis of ls-svm. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 2397–2401.
34. El Karoui, N. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Relat. Fields* **2018**, *170*, 95–175. [CrossRef]
35. Stojnic, M. Recovery thresholds for ℓ_1 optimization in binary compressed sensing. In Proceedings of the 2010 IEEE International Symposium on Information Theory, Austin, TX, USA, 13–18 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 1593–1597.
36. Stojnic, M. A framework to characterize performance of lasso algorithms. *arXiv* **2013**, arXiv:1303.7291.
37. Thrampoulidis, C.; Oymak, S.; Hassibi, B. Regularized Linear Regression: A Precise Analysis of the Estimation Error. In Proceedings of the COLT, Paris, France, 3–6 July 2015; pp. 1683–1709.
38. Thrampoulidis, C.; Panahi, A.; Guo, D.; Hassibi, B. Precise error analysis of the lasso. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 3467–3471.
39. Thrampoulidis, C.; Xu, W.; Hassibi, B. Symbol error rate performance of box-relaxation decoders in massive MIMO. *IEEE Trans. Signal Process.* **2018**, *66*, 3377–3392. [CrossRef]
40. Atitallah, I.B.; Thrampoulidis, C.; Kammoun, A.; Al-Naffouri, T.Y.; Hassibi, B.; Alouini, M.S. Ber analysis of regularized least squares for bpsk recovery. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 4262–4266.
41. Alrashdi, A.M.; Kammoun, A.; Muqaibel, A.H.; Al-Naffouri, T.Y. Asymptotic Performance of Box-RLS Decoders under Imperfect CSI with Optimized Resource Allocation. *IEEE Open J. Commun. Soc.* **2022**, *3*, 2051–2075. [CrossRef]
42. Atitallah, I.B.; Thrampoulidis, C.; Kammoun, A.; Al-Naffouri, T.Y.; Alouini, M.S.; Hassibi, B. The BOX-LASSO with application to GSSK modulation in massive MIMO systems. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1082–1086.
43. Alrashdi, A.M.; Alrashdi, A.E.; Alghadhban, A.; Eleiwa, M.A. Optimum GSSK Transmission in Massive MIMO Systems Using the Box-LASSO Decoder. *IEEE Access* **2022**, *10*, 15845–15859. [CrossRef]
44. Alrashdi, A.M.; Atitallah, I.B.; Al-Naffouri, T.Y. Precise performance analysis of the box-elastic net under matrix uncertainties. *IEEE Signal Process. Lett.* **2019**, *26*, 655–659. [CrossRef]
45. Hayakawa, R.; Hayashi, K. Binary vector reconstruction via discreteness-aware approximate message passing. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1783–1789.
46. Hayakawa, R.; Hayashi, K. Asymptotic Performance of Discrete-Valued Vector Reconstruction via Box-Constrained Optimization With Sum of ℓ_1 Regularizers. *IEEE Trans. Signal Process.* **2020**, *68*, 4320–4335. [CrossRef]
47. Deng, Z.; Kammoun, A.; Thrampoulidis, C. A model of double descent for high-dimensional binary linear classification. *Inf. Inference J. IMA* **2022**, *11*, 435–495. [CrossRef]

48. Kini, G.R.; Thrampoulidis, C. Analytic study of double descent in binary classification: The impact of loss. In Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 2527–2532.
49. Salehi, F.; Abbasi, E.; Hassibi, B. The performance analysis of generalized margin maximizers on separable data. In Proceedings of the International Conference on Machine Learning, Virtual Online, 13–18 July 2020; PMLR: Mc Kees Rocks, PA, USA, 2020; pp. 8417–8426.
50. Dhifallah, O.; Thrampoulidis, C.; Lu, Y.M. Phase retrieval via linear programming: Fundamental limits and algorithmic improvements. In Proceedings of the 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 3–6 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1071–1077.
51. Salehi, F.; Abbasi, E.; Hassibi, B. A precise analysis of phasemax in phase retrieval. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 976–980.
52. Bosch, D.; Panahi, A.; Hassibi, B. Precise Asymptotic Analysis of Deep Random Feature Models. *arXiv* **2023**, arXiv:2302.06210.
53. Dhifallah, O.; Lu, Y.M. A precise performance analysis of learning with random features. *arXiv* **2020**, arXiv:2008.11904.
54. Dhifallah, O.; Lu, Y.M. Phase transitions in transfer learning for high-dimensional perceptrons. *Entropy* **2021**, *23*, 400. [CrossRef]
55. Ting, M.; Raich, R.; Hero, A.O., III. Sparse image reconstruction for molecular imaging. *IEEE Trans. Image Process.* **2009**, *18*, 1215–1227. [CrossRef]
56. Gui, G.; Peng, W.; Wang, L. Improved sparse channel estimation for cooperative communication systems. *Int. J. Antennas Propag.* **2012**, *2012*, 476509. [CrossRef]
57. Luenberger, D.G.; Ye, Y. *Linear and Nonlinear Programming*; Springer: Berlin/Heidelberg, Germany, 1984; Volume 2.
58. Salehi, F.; Abbasi, E.; Hassibi, B. The impact of regularization on high-dimensional logistic regression. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
59. Donoho, D.L.; Maleki, A.; Montanari, A. Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 18914–18919. [CrossRef]
60. Hayakawa, R. Noise variance estimation using asymptotic residual in compressed sensing. *arXiv* **2020**, arXiv:2009.13678.
61. Suliman, M.A.; Alrashdi, A.M.; Ballal, T.; Al-Naffouri, T.Y. SNR estimation in linear systems with Gaussian matrices. *IEEE Signal Process. Lett.* **2017**, *24*, 1867–1871. [CrossRef]
62. Kobayashi, H.; Mark, B.L.; Turin, W. *Probability, Random Processes, and Statistical Analysis: Applications to Communications, Signal Processing, Queueing Theory and Mathematical Finance*; Cambridge University Press: Cambridge, UK, 2011.
63. Donoho, D.L.; Maleki, A.; Montanari, A. The noise-sensitivity phase transition in compressed sensing. *IEEE Trans. Inf. Theory* **2011**, *57*, 6920–6941. [CrossRef]
64. Thrampoulidis, C.; Abbasi, E.; Hassibi, B. Lasso with non-linear measurements is equivalent to one with linear measurements. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, USA, 7–12 December 2015; pp. 3420–3428.
65. Abbasi, E.; Salehi, F.; Hassibi, B. Universality in learning from linear measurements. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
66. Hu, H.; Lu, Y.M. Universality laws for high-dimensional learning with random features. *IEEE Trans. Inf. Theory* **2022**, *69*, 1932–1964. [CrossRef]
67. Han, Q.; Shen, Y. Universality of regularized regression estimators in high dimensions. *arXiv* **2022**, arXiv:2206.07936.
68. Dudeja, R.; Bakhshizadeh, M. Universality of linearized message passing for phase retrieval with structured sensing matrices. *IEEE Trans. Inf. Theory* **2022**, *68*, 7545–7574. [CrossRef]
69. Gerace, F.; Krzakala, F.; Loureiro, B.; Stephan, L.; Zdeborová, L. Gaussian Universality of Perceptrons with Random Labels. *arXiv* **2023**, arXiv:2205.13303.
70. Chin, K.; DeVries, S.; Fridlyand, J.; Spellman, P.T.; Roydasgupta, R.; Kuo, W.L.; Lapuk, A.; Neve, R.M.; Qian, Z.; Ryder, T.; et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **2006**, *10*, 529–541. [CrossRef]
71. Güçkiran, K.; Cantürk, İ.; Özyilmaz, L. DNA microarray gene expression data classification using SVM, MLP, and RF with feature selection methods relief and LASSO. *Süleyman Demirel Üniv. Fen Bilim. Enstitüsü Derg.* **2019**, *23*, 126–132. [CrossRef]
72. Sun, L.; Hui, A.M.; Su, Q.; Vortmeyer, A.; Kotliarov, Y.; Pastorino, S.; Passaniti, A.; Menon, J.; Walling, J.; Bailey, R.; et al. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell* **2006**, *9*, 287–300. [CrossRef]
73. Alon, U.; Barkai, N.; Notterman, D.A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 6745–6750. [CrossRef] [PubMed]
74. Li, J.; Dong, W.; Meng, D. Grouped gene selection of cancer via adaptive sparse group lasso based on conditional mutual information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *15*, 2028–2038. [CrossRef] [PubMed]
75. Belkin, M.; Hsu, D.; Ma, S.; Mandal, S. Reconciling modern machine-learning practice and the classical bias—Variance trade-off. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 15849–15854. [CrossRef] [PubMed]
76. Gordon, Y. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric Aspects of Functional Analysis*; Lecture Notes in Mathematics ; Springer: Berlin/Heidelberg, Germany, 1988; pp. 84–106.

77. Chu, H.T.; Toh, K.C.; Zhang, Y. On Regularized Square-root Regression Problems: Distributionally Robust Interpretation and Fast Computations. *J. Mach. Learn. Res.* **2022**, *23*, 13885–13923.
78. Bunea, F.; Lederer, J.; She, Y. The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Trans. Inf. Theory* **2013**, *60*, 1313–1325. [CrossRef]