A Precise Performance Analysis of Support Vector Regression

Houssem Sifaou [1] **Abla Kammoun** [1] **Mohamed-Slim Alouini** [1]

Abstract

In this paper, we study the hard and soft support vector regression techniques applied to a set of n linear measurements of the form $y_i = \boldsymbol{\beta}_\star^T \mathbf{x}_i + n_i$ where $\boldsymbol{\beta}_\star$ is an unknown vector, $\{\mathbf{x}_i\}_{i=1}^n$ are the feature vectors and $\{n_i\}_{i=1}^n$ model the noise. Particularly, under some plausible assumptions on the statistical distribution of the data, we characterize the feasibility condition for the hard support vector regression in the regime of high dimensions and, when feasible, derive an asymptotic approximation for its risk. Similarly, we study the test risk for the soft support vector regression as a function of its parameters. Our results are then used to optimally tune the parameters intervening in the design of hard and soft support vector regression algorithms. Based on our analysis, we illustrate that adding more samples may be harmful to the test performance of support vector regression, while it is always beneficial when the parameters are optimally selected. Such a result reminds a similar phenomenon observed in modern learning architectures according to which optimally tuned architectures present a decreasing test performance curve with respect to the number of samples.

1. Introduction

Motivation. Recent works have demonstrated that the test performance of modern learning architectures exhibits both model-wise and sample-wise double descent phenomena that defy conventional statistical intuition. Model-wise descent, reported in recent works (Belkin et al., 2019a;c; Geiger et al., 2019), suggests that, for very large architectures, performance improves with the number of parameters, thus contradicting the bias-variance trade-off. On the other hand, sample-wise descent, discussed recently in the works of Nakkiran *et al.* (Nakkiran et al., 2020a;b) indicates that more data may harm the performance. One potential solution to avoid such a behavior consists in optimally tuning the involved parameters. In doing so, the test performance in most scenarios decreases with the number of samples.

In this paper, we investigate the sample-wise double descent phenomenon for basic linear models. More precisely, we assume independent data samples $(\mathbf{x}_i, y_i)$, $i = 1, \cdots, n$ distributed as:

$$y_i = \boldsymbol{\beta}_\star^T \mathbf{x}_i + \sigma n_i,$$

where $\mathbf{x}_i \in \mathbb{R}^p$ is the feature vector assumed to have zero mean and covariance $\mathbf{I}_p$, y_i represent the scalar response variables while σn_i stands for zero-mean noise with variance σ^2. To estimate $\boldsymbol{\beta}_\star$, we consider support vector regression techniques: hard-support vector regression (H-SVR), which estimates the regression vector $\boldsymbol{\beta}$ with minimum ℓ_2 norm that satisfies the constraints $y_i = \boldsymbol{\beta}^T \mathbf{x}_i$ up to a maximum error ϵ, and soft-support regression (S-SVR) which uses a regularization constant C that aims to create a trade-off between the minimization of the training error and the minimization of the model complexity. SVR has been extesively used in several applications that vary from biomedical analysis (Hamdi et al., 2018) to financial time series (Ju et al., 2014; Yang et al., 2009; Qu & Zhang) and weather forecasting (Guajardo et al., 2006).

We study the performance of H-SVR and S-SVR when the number of features p and the sample size n grow simultaneously large such that $\frac{n}{p} \to \delta$ with $\delta > 0$ and the norm of $\|\boldsymbol{\beta}_\star\|$ converges to β. One major outcome of the present work is to recover interesting behaviors observed in large-scale machine learning architectures. Particularly, we show numerically that the double descent behavior appears only when the H-SVR or S-SVR parameters are not properly tuned. Such a behavior reminds the recent findings in (Nakkiran et al., 2020b) that suggest that unregularized models often suffer from the sample-wise double descent phenomenon, while optimally tuned models usually present a monotonic risk with respect to the number of samples.

Contributions. This paper investigates the test risk behavior as a function of the sample size for H-SVR and S-SVR techniques. Contrary to least squares regression, which involves explicit form expressions for the solution, H-SVR and S-SVR require solving convex-optimization problems,

[1] Computer, Electrical, and Mathematical Sciences & Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. Correspondence to: Houssem Sifaou <houssem.sifaou@kaust.edu.sa>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

which do not have closed-form solutions. To analyze the test risk, we rely on the Gaussian min-max theorem (CGMT) framework (Thrampoulidis et al., 2018) and more specifically on the extension of this framework recently developed in (Deng et al., 2020), which has been proven to be suitable to analyze functionals of solutions of convex optimizations problems.

Concretely, our results for the H-SVR and S-SVR can be summarized as follows:

1. We derive for a fixed error tolerance ϵ, a sharp phase transition-threshold $\delta_\star$ beyond which the H-SVR becomes infeasible. Interestingly, we illustrate that the transition threshold depends only on ϵ and the noise variance and not on the SNR defined as $\text{SNR} := \frac{\beta^2}{\sigma^2}$. Moreover, we show that $\delta_\star$ is always greater than 1, which should be compared with the condition $\delta < 1$ required for the least square estimator to exist. As a side note, we prove that contrary to hard-margin support vector classifiers, H-SVR can always be feasible through a proper tuning of the tolerance error ϵ. This allows us to study the test risk of the H-SVR as a function of δ when $\delta \in (0, \infty)$ and ϵ carefully tuned to satisfy the feasibility condition.

2. For fixed error tolerance ϵ, we numerically illustrate that for moderate to large SNR the test risk of the H-SVR is a non-monotonic curve presenting a unique minimum that becomes the closest to $\delta_\star$ as the SNR increases. For low SNR values, the test risk is an increasing function of $\delta_\star$ and is always worse than the null risk associated with the null estimator. It is worth mentioning that behavior of the same kind was reported for the min-norm least square estimator in (Hastie et al., 2019). Additionally, we illustrate that when the parameter ϵ is optimally tuned, the test risk becomes a decreasing function of δ and equivalently of the number of data samples.

3. Similarly, we derive the expression for the asymptotic test risk as a function of ϵ, δ, and the regularization constant C. Without optimal tuning of the regularization constant C and factor ϵ, the test curve as a function of the sample test size presents a double descent, which disappears when optimal settings of these constants is considered.

4. We study the robustness of the S-SVR and H-SVR to impulsive noise. We illustrate that contrary to H-SVR, S-SVR, when optimally tuned, is resilient to impulsive noises. Particularly, we show that for mild impulsive noise conditions, S-SVR presents a slightly lower risk than optimally tuned ridge regression estimators but largely outperforms it under moderate to severe impulsive noise conditions.

Related works. The present work is part of the continued efforts to understand the double descent phenomena in large-scale machine learning architectures. An important body of research works focused on establishing the behavior of double descent of the test risk as a function of the model size in a variety of machine learning algorithms (Belkin et al., 2019a; Bös & Opper, 1997; Spigler et al., 2019). Very recently, the work in (Nakkiran et al., 2020a) discovered that double descent occurs not just as a function of the model size but also as a function of the sample size (Nakkiran et al., 2020a). A major consequence of such a behavior is that performance may be degraded as we increase the number of samples.

To further understand the generalization error, several works considered to analyze it as a function of the model size for mathematically tractable settings in regression (Hastie et al., 2019; Belkin et al., 2019b; Muthukumar et al., 2020; Mitra, 2019; Candes & Sur, 2018) and more recently in classification (Deng et al., 2020; Kini & Thrampoulidis, 2020; Montanari et al., 2020), with the goal of investigating under which conditions, the double descent occurs. In this paper, similarly to (Nakkiran et al., 2020b), we instead focus on the effect of the sample size on the test performance, but with the H-SVR and S-SVR as case examples. Moreover, on the technical level, our analysis provides sharp characterizations of the performance using the recently developed extension of the CGMT framework (Deng et al., 2020).

2. Problem formulation

Consider the problem of estimating the scalar response y of a vector $\mathbf{x}$ in $\mathbb{R}^p$ from a set of n data samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ following the linear model:

$$y_i = \boldsymbol{\beta}_\star^T \mathbf{x}_i + \sigma n_i \tag{1}$$

where $\boldsymbol{\beta}_\star$ is an unknown vector, $\{\sigma n_i\}_{i=1}^n$ represent noise samples with zero-mean and variance σ^2. We further assume that $\mathbf{x}_i$ is Gaussian with zero mean and covariance $\mathbf{I}_p$. To estimate $\boldsymbol{\beta}_\star$, we consider support vector regression methods, namely the hard support vector regression denoted by H-SVR and the soft support vector regression referred to as S-SVR. The H-SVR looks for a function $y = \boldsymbol{\beta}^T \mathbf{x}$ such that all data points $(\mathbf{x}_i, \boldsymbol{\beta}^T \mathbf{x}_i)$ deviates at most ϵ from their targets y_i. Formally, this regression problem can be written as:

$$\begin{aligned} \hat{\mathbf{w}}_H := \arg\min_{\mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad & |y_i - \mathbf{w}^T \mathbf{x}_i| \leq \epsilon \end{aligned} \tag{2}$$

where $\|.\|$ denotes the ℓ_2-norm of a vector. It is worth mentioning that when $\epsilon = 0$ and $n \leq p$, the H-SVR boils down to the least square estimator. In this case, it perfectly interpolates the training data, satisfying $y_i = \mathbf{x}_i^T \hat{\mathbf{w}}_H$, $i = 1, \cdots, n$. In general, depending on the value of ϵ, there may not be a solution that satisfies the constraints. As in support vector machines for classification, one solution to deal with such cases is to add slack variables that while relaxing the constraints, penalize, in the objective function, large deviations from them. Applying this approach gives the S-SVR method which involves solving the following optimization problem:

$$\begin{aligned} \hat{\mathbf{w}}_S := \arg\min_{\mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{p}\sum_{i=1}^{n}(\xi_i + \tilde{\xi}_i) \\ \text{s.t.} \quad & y_i - \mathbf{w}^T\mathbf{x}_i \le \epsilon + \xi_i, \ i = 1, \cdots, n \quad (3) \\ & \mathbf{w}^T\mathbf{x}_i - y_i \le \epsilon + \tilde{\xi}_i \\ & \xi_i, \tilde{\xi}_i \ge 0. \end{aligned}$$

The aim of the present work is to characterize analytically the performance of the H-SVR and the S-SVR. The assumption underlying the analysis is to consider that the number of samples and that of features grow with the same pace, and will be made more specific in the sequel.

Prediction risk. The metric of interest in this paper is the prediction risk. For a given estimator $\hat{\boldsymbol{\beta}}$, the prediction risk is defined as:

$$\mathcal{R}(\hat{\boldsymbol{\beta}}) := \mathbb{E}_{\mathbf{x},y}|\mathbf{x}^T\hat{\boldsymbol{\beta}} - \mathbf{x}^T\boldsymbol{\beta}_\star|^2 = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|^2,$$

where $\mathbf{x}$ and y are test points following the model (1) but are independent of the training set. Expressing $\mathcal{R}(\hat{\boldsymbol{\beta}})$ as:

$$\mathcal{R}(\hat{\boldsymbol{\beta}}) = \|\boldsymbol{\beta}_\star\|^2 + \|\hat{\boldsymbol{\beta}}\|^2 - 2\|\boldsymbol{\beta}_\star\|\|\hat{\boldsymbol{\beta}}\|\frac{\boldsymbol{\beta}_\star^T\hat{\boldsymbol{\beta}}}{\|\boldsymbol{\beta}_\star\|\|\hat{\boldsymbol{\beta}}\|},$$

we can easily see that the risk depends on $\hat{\boldsymbol{\beta}}$ through its norm $\|\hat{\boldsymbol{\beta}}\|$ and the cosine similarity between $\boldsymbol{\beta}_\star$ and $\hat{\boldsymbol{\beta}}$:

$$\cos\left(\boldsymbol{\beta}_\star, \hat{\boldsymbol{\beta}}\right) := \frac{\boldsymbol{\beta}_\star^T\hat{\boldsymbol{\beta}}}{\|\boldsymbol{\beta}_\star\|.\|\hat{\boldsymbol{\beta}}\|}.$$

3. Main results

In this paper, we consider studying the performance of the hard and soft support vector regression problems under the asymptotic regime in which n and p grow large at the same pace. More specifically, the following assumption is considered.

Assumption 1. *Our study is based on the following set of assumptions:*

- *n and p grow to infinity with $\frac{n}{p} \to \delta$.*
- *The noise variance σ^2 is a fixed positive constant.*
- *$\|\boldsymbol{\beta}_\star\| \to \beta$ where β is a certain positive scalar.*
- **Data model***: The training $\{\mathbf{x}_i\}_{i=1}^n$ are independent and identically distributed following standard normal distribution. Moreover, $\{n_i\}$ are independent and drawn from a zero-mean unit-variance symmetric distribution p_N.*

3.1. Hard SVR

As mentioned earlier, the H-SVR problem is not always feasible. We provide in Theorem 1 a sharp characterization of the feasibility region of the H-SVR in the asymptotic regime defined in Assumption 1.

Theorem 1 (Feasibility of the H-SVR)**.** *Let $\delta_\star$ be defined as:*

$$\delta_\star = \frac{1}{\inf_{t\in\mathbb{R}} \mathbb{E}\left(|G + tN| - t\frac{\epsilon}{\sigma}\right)_+^2}, \tag{4}$$

where $(x)_+ \triangleq \max(x, 0)$ and the expectation is taken over the distribution of G and N where $G \sim \mathcal{N}(0, 1)$ and $N \sim p_N$. Consider the asymptotic regime and data model in Assumption 1. Then, the following statements hold true:

$$\delta > \delta_\star \Rightarrow \mathbb{P}\left[\text{The H-SVR is feasible for suff. large } n\right] = 0 \tag{5}$$

$$\delta < \delta_\star \Rightarrow \mathbb{P}\left[\text{The H-SVR is feasible for suff. large } n\right] = 1 \tag{6}$$

The proofs can be found in the supplementary material.

Remark 1. **(Feasibility of the H-SVR depends on the noise variance but not on the SNR.)** The above result establishes that the existence of the H-SVR undergoes a sharp transition phenomenon. Particularly, in the limit of large sample size n and number of features p such that $\frac{n}{p} \to \delta$, the H-SVR is almost surely unfeasible when $\delta > \delta_\star$ and always feasible when $\delta < \delta_\star$. The obtained expression is reminiscent of other previously established result for the existence of the hard-margin SVM for classification established in a series of recent works (Kammoun & Alouini, 2020) and (Deng et al., 2020). However, contrary to the expressions obtained in these works, the separability boundary curve captured by $\delta_\star$ does not depend on the Euclidean norm of $\boldsymbol{\beta}_\star$, or equivalently on the SNR defined as $\frac{\|\boldsymbol{\beta}_\star\|^2}{\sigma^2}$ but only on the noise variance. The reason behind this is that feasibility is essentially related to how much the data samples deviate from the hyperplane defined as $\boldsymbol{\beta}_\star^T\mathbf{x} = y$. We note that as σ approaches 0 and $\epsilon \neq 0$, $\delta_\star \to \infty$, which implies that the H-SVR is always feasible in this case. This is because in the noiseless case, all data samples $(\mathbf{x}_i, y_i)$ belong to the hyperplane $y_i = \boldsymbol{\beta}_\star^T\mathbf{x}_i$ and thus $\boldsymbol{\beta}_\star$ is in the feasibility set of the H-SVR regardless of the value of ϵ and also on β. On the other hand, as σ increases, $\delta_\star$ decreases, which suggests that the H-SVR becomes less feasible since it is more difficult to find a hyperplane that contains all data samples with a reasonable error tolerance ϵ.

Remark 2. (The H-SVR can be feasible when the least square estimator is not) One can easily check that if $\epsilon = 0$, then $\delta_\star = 1$. This result makes sense since as long as $\delta < 1$, the linear system $\boldsymbol{\beta}^T \mathbf{x}_i = y_i$ is under determined and as such a solution $\boldsymbol{\beta}$ exists. However, when $\delta_\star > 1$, the linear system becomes over-determined, and as such it is impossible to find a solution to this linear system. Moreover, since $\delta_\star$ is an increasing function of ϵ, we conclude that $\delta_\star > 1$ for all $\epsilon > 0$. This particularly shows that the H-SVR provides a larger feasibility region than the least square estimator, for which δ should be less than 1 to exist. We can even push this result further and claim that *every δ is feasible once ϵ is appropriately tuned.* To see this, it suffices to note that $\epsilon \mapsto \delta_\star$ is an increasing function establishing a one-to-one map from $(0, \infty)$ to $(1, \infty)$. Hence, for any $\delta \in (1, \infty)$ there exists $\epsilon^\star(\delta)$ such that for all $\epsilon > \epsilon^\star(\delta)$, the H-SVR is almost surely feasible.

For the sake of illustration, Figure 1 displays $\delta_\star$ as a function of ϵ for several values of σ. As can be seen, $\delta_\star$ is an increasing function growing to infinity with ϵ. The value of ϵ plays a fundamental role to remediate the effect of the noise and ensure the feasibility of the H-SVR. One can note as expected that $\delta_\star$ for $\epsilon = 0.1$ and $\sigma = 0.1$ is the same as the one obtained when $\epsilon = 0.2$ and $\sigma = 0.2$. This finding can be easily concluded from (4). Moreover, in agreement with our previous discussion, we can easily see that $\delta_\star$ decreases significantly as the noise variance increases. Such a scenario can be fixed by adapting the value of ϵ.

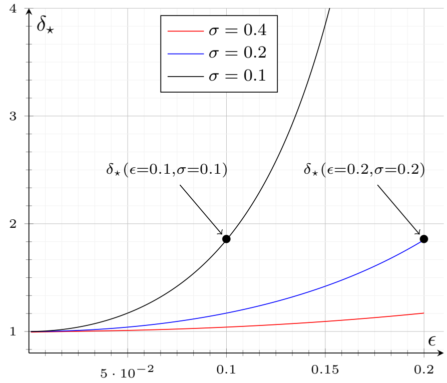


Figure 1. Theoretical predictions of $\delta_\star$ as a function of ϵ for different noise variance values. The figure shows that every δ can be forced to be in the feasibility region by appropriately choosing ϵ.

Having characterized the feasibility region of the H-SVR, we are now ready to provide sharp asymptotics of its performance in terms of the prediction risk and the cosine similarity.

Theorem 2 (Asymptotic prediction risk of H-SVR). *Define function $\mathcal{D} : \mathbb{R}^2 \to \mathbb{R}$ as:*

$$\mathcal{D}(\tilde{\gamma}_1, \tilde{\gamma}_2) = \sqrt{\delta}\sqrt{\mathbb{E}\left(|\sqrt{\tilde{\gamma}_1^2 + \tilde{\gamma}_2^2}G + N| - \frac{\epsilon}{\sigma}\right)_+^2} - \tilde{\gamma}_1$$

where the expectation is taken over the distribution of the independent random variables G and N drawn respectively from the standard normal distribution $\mathcal{N}(0,1)$ and p_N. Let $\tilde{\gamma}_1^\star$ and $\tilde{\gamma}_2^\star$ be the unique solutions to the following optimization problem:

$$(\tilde{\gamma}_1^\star, \tilde{\gamma}_2^\star) = \arg \min_{\substack{\tilde{\gamma}_1, \tilde{\gamma}_2 \\ \mathcal{D}(\tilde{\gamma}_1, \tilde{\gamma}_2) \leq 0}} \frac{1}{2}\left(\tilde{\gamma}_2 - \frac{\beta}{\sigma}\right)^2 + \frac{1}{2}\tilde{\gamma}_1^2. \quad (7)$$

Let $\hat{\mathbf{w}}_H$ be the solution to the H-SVR. Then, under Assumption 1 and assuming $\delta < \delta_\star$, the following convergences hold true:

$$\mathcal{R}(\hat{\mathbf{w}}_H) - \overline{R}_H \xrightarrow{a.s.} 0$$

where $\overline{R}_H = \sigma^2((\tilde{\gamma}_1^\star)^2 + (\tilde{\gamma}_2^\star)^2)$ and

$$\frac{\hat{\mathbf{w}}_H^T \boldsymbol{\beta}_\star}{\|\hat{\mathbf{w}}_H\| \|\boldsymbol{\beta}_\star\|} \xrightarrow{a.s.} \frac{\frac{\beta}{\sigma} - \tilde{\gamma}_2^\star}{\sqrt{(\tilde{\gamma}_1^\star)^2 + (\tilde{\gamma}_2^\star - \frac{\beta}{\sigma})^2}}.$$

Remark 3. (Behavior of the H-SVR as $\delta \to 0$). As δ tending to zero, one can check after careful investigation of the asymptotic expressions that $\tilde{\gamma}_2^\star \to \frac{\beta}{\sigma}$ and $\tilde{\gamma}_1^\star \to 0$. In this case, the asymptotic risk is thus given by β^2. To see this, it suffices to note that $\mathcal{D}(\delta^{\frac{1}{4}}, \frac{\beta}{\sigma})$ converges from below to zero as $\delta \downarrow 0$. By continuity of $\mathcal{D}$, we may find η sufficiently small such that, for all $(\tilde{\gamma}_1, \tilde{\gamma}_2) \in \mathcal{C}(\eta)$ where

$$\mathcal{C}(\eta) = \left\{(\tilde{\gamma}_1, \tilde{\gamma}_2) \mid \tilde{\gamma}_1 \in (0, \eta), \ \tilde{\gamma}_2 \in (\frac{\beta}{\sigma} - \eta, \frac{\beta}{\sigma} + \eta)\right\},$$

we have $\mathcal{D}(\tilde{\gamma}_1, \tilde{\gamma}_2) \leq 0$. Moreover, it is easy to see that the objective in (7) can be bounded by η^2 when $(\tilde{\gamma}_1, \tilde{\gamma}_2) \in \mathcal{C}(\eta)$. Evaluation of this objective when $|\tilde{\gamma}_2 - \frac{\beta}{\sigma}| \geq \sqrt{2}\eta$ or when $|\tilde{\gamma}_1| \geq \sqrt{2}\eta$ yields values greater than η^2. Hence, necessarily, $(\tilde{\gamma}_1, \tilde{\gamma}_2) \in \mathcal{C}(\sqrt{2}\eta)$, which proves the desired. Finally, observing that the risk of the null estimator $\hat{\boldsymbol{\beta}} = 0$ is also β^2, we conclude that the H-SVR is worse than the null estimator when a small number of samples is employed.

Remark 4. (Behavior of the H-SVR as $\delta \to \delta_\star$). As $\delta \to \delta_\star$, the set $\{(\tilde{\gamma}_1, \tilde{\gamma}_2) \mid \mathcal{D}(\tilde{\gamma}_1, \tilde{\gamma}_2) \leq 0\}$ becomes the unit set $\{(\tilde{\gamma}_1^\circ, 0)\}$ where $\tilde{\gamma}_1^\circ$ is the smallest solution to the equation $\mathcal{D}(\tilde{\gamma}_1, 0) = 0$. The asymptotic risk thus becomes equal to $\sigma^2(\tilde{\gamma}_1^\circ)^2$ and is as such independent of the SNR $\frac{\beta^2}{\sigma^2}$. As a result, when δ approaches $\delta_\star$, it is the noise variance and the value of ϵ that determines the performance of the H-SVR and not the SNR. This is in opposition to the behavior in the operation region $\delta \to 0$, for which the risk tends to β^2.

3.2. Soft SVR

In this section, the soft SVR problem is considered and the convergence of the corresponding prediction risk is established.

Theorem 3 (Asymptotic prediction risk of S-SVR)**.** *Under Assumption 1, we have*

$$\mathcal{R}(\hat{\mathbf{w}}_S) \xrightarrow{a.s} \sigma^2(\tilde{\gamma}_1^{*2} + \tilde{\gamma}_2^{*2})$$

where $\tilde{\gamma}_1^*$ and $\tilde{\gamma}_2^*$ are the solutions of the following scalar optimization problem

$$\overline{\phi} = \min_{\tilde{\gamma}_1, \tilde{\gamma}_2} \sup_{\chi > 0} \overline{D}(\tilde{\gamma}_1, \tilde{\gamma}_2, \chi)$$

with

$$\overline{D}(\tilde{\gamma}_1, \tilde{\gamma}_2, \chi) = \frac{1}{2}\tilde{\gamma}_1^2 + \frac{1}{2}\left(\tilde{\gamma}_2 - \frac{\beta}{\sigma}\right)^2 - \frac{\tilde{\gamma}_1 \chi}{2\sigma} + \frac{\delta}{\sigma}\mathbb{E}\left\{C\left[X - \frac{C\tilde{\gamma}_1}{2\chi}\right]\mathbb{1}_{\{X\chi > \tilde{\gamma}_1 C\}} + \frac{\chi}{2\tilde{\gamma}_1}X^2 \mathbb{1}_{\{X\chi \leq \tilde{\gamma}_1 C\}}\right\}$$

where $X = \left(\left|\sqrt{\tilde{\gamma}_1^2 + \tilde{\gamma}_2^2}G + N\right| - \epsilon/\sigma\right)_+$ and the expectation is with respect to the distributions of G and N with $G \sim \mathcal{N}(0,1)$ and $N \sim p_N$.

Remark 5. Theorem 3 can be used to optimally tune the parameters (ϵ, C) so that they minimize the asymptotic test risk. For that, one is required to estimate the noise variance σ^2 and the signal power β^2. These can be easily estimated through the following approach. Let $\mathbf{y} = [y_1, \cdots, y_n]^T$ be the vector of the responses and $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ the matrix stacking all training samples. Then,

$$\mathbf{y} = \mathbf{X}^T \boldsymbol{\beta}_\star + \sigma \mathbf{n},$$

where $\mathbf{n} = [n_1, \cdots, n_n]^T$. From the strong law of large numbers,

$$\frac{1}{n}\mathbf{y}^T\mathbf{y} \xrightarrow{a.s.} \sigma^2 + \beta^2. \tag{8}$$

On the other hand, assuming $\delta > 1$,

$$\frac{1}{n}\mathbf{y}^T(\mathbf{I}_n - \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X})\mathbf{y} \xrightarrow{a.s.} \sigma^2(1 - \frac{1}{\delta}). \tag{9}$$

Combining (8) and (9), consistent estimators for the noise variance σ^2 and for β^2 are given by:

$$\hat{\sigma}^2 = \frac{\frac{1}{n}\mathbf{y}^T(\mathbf{I}_n - \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X})\mathbf{y}}{1 - \frac{1}{\delta}}, \tag{10}$$

$$\hat{\beta}^2 = \frac{1}{n}\mathbf{y}^T\mathbf{y} - \frac{\frac{1}{n}\mathbf{y}^T(\mathbf{I}_n - \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X})\mathbf{y}}{1 - \frac{1}{\delta}}. \tag{11}$$

In case $\delta < 1$, the problem of estimating the noise variance becomes more challenging, and there are, to the best of our knowledge, no general unbiased estimators with the same statistical guarantees as in the case $\delta > 1$. To address this issue, some other techniques may be used (Cherkassky & Ma, 2004) but they are not guaranteed to lead to consistent estimators.

Remark 6. Behavior of the S-SVR when $\delta \to 0$. The test risk of the S-SVR is much more involved than that of the H-SVR. Nevertheless, it can be easily seen that when δ goes to zero,

$$\lim_{\delta \to 0} \sup_{\chi \geq 0} \overline{D}(\tilde{\gamma}_1, \tilde{\gamma}_2, \chi) \overset{(a)}{=} \inf_{\delta \geq 0} \sup_{\chi \geq 0} \overline{D}(\tilde{\gamma}_1, \tilde{\gamma}_2, \chi) \tag{12}$$

$$\overset{(b)}{=} \sup_{\chi \geq 0} \inf_{\delta \geq 0} \overline{D}(\tilde{\gamma}_1, \tilde{\gamma}_2, \chi) \tag{13}$$

$$= \sup_{\chi \geq 0} \lim_{\delta \to 0} \overline{D}(\tilde{\gamma}_1, \tilde{\gamma}_2, \chi) \tag{14}$$

$$= \frac{1}{2}\tilde{\gamma}_1^2 + \frac{1}{2}(\tilde{\gamma}_2 - \frac{\beta}{\sigma})^2 \tag{15}$$

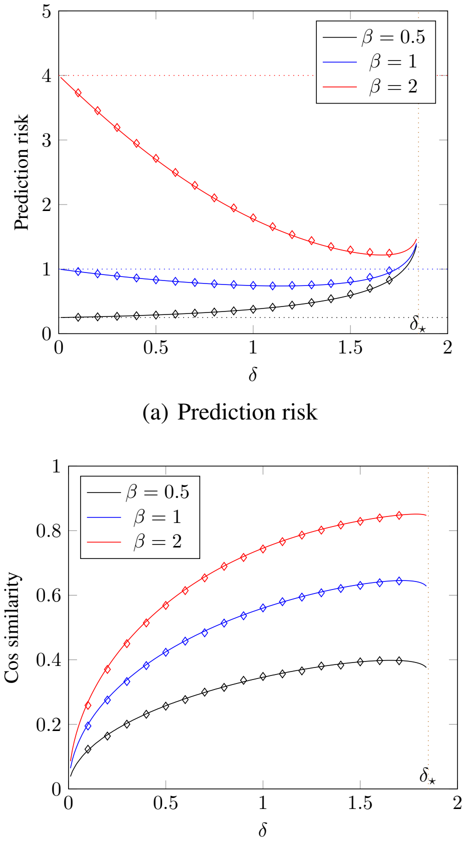
where (a) follows from the fact that the objective function is an increasing function in δ and (b) from the fact that the objective function is convex in δ and concave in χ. The asymptotic limit of $\overline{D}$ in (15) has a unique minimum given by $\tilde{\gamma}_2 = \frac{\beta}{\sigma}$ and $\tilde{\gamma}_1 = 0$. Plugging these values into that of the test risk, we conclude that when δ goes to zero, the test risk of the S-SVR converges to that of the null estimator.

4. Numerical illustration

4.1. H-SVR

4.1.1. Test risk as a function of the number of samples

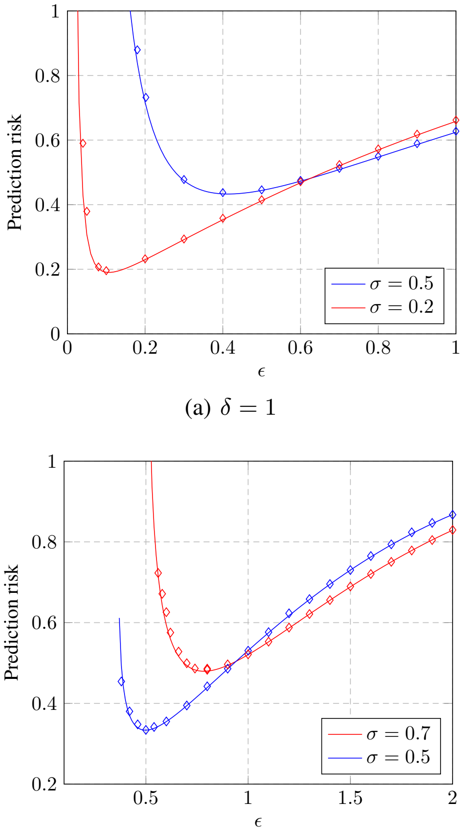
In a first experiment, we investigate the behavior of the test risk of H-SVR as a function of the number of samples for different values of the signal power $\beta^2 = \|\boldsymbol{\beta}_\star\|^2$. Particularly, for each $\beta \in \{0.5, 1, 2\}$, we fix the noise variance σ^2 and ϵ and plot the test risk and cosine similarity over the range $[0, \delta_\star]$ over which the H-SVR is feasible. Fig. 2 represents the theoretical results along with their empirical averages obtained for $p = 200$ and $n = \lfloor \delta p \rfloor$. As can be seen, this figure's results validate the accuracy of the theoretical predictions: a perfect match is noted over the whole range of δ and for all signal power values. We also corroborate our predictions in Remark 3 and Remark 4: the risk tends to β^2 which is the null estimator's risk when $\delta \to 0$, while it tends to the same limit irrespective of β^2 as $\delta \to \delta_\star$. Away from these limiting cases, we note that for moderate to high signal powers, the test risk presents a non-monotonic behavior with respect to δ and as such with respect to the number of samples. The minimal risk corresponds to a δ that becomes the nearest to $\delta_\star$ as the signal power β^2 increases. However, for low signal powers, the test risk is an increasing function of the number of samples and is always larger than β^2, which is the null estimator's risk. Such behavior is similar to that of the min-norm least square estimator, which becomes worse than the null estimator when the SNR is less than 1 (Hastie et al., 2019).



(a) Prediction risk

(b) Cos similarity

Figure 2. Performance of H-SVR as a function of δ when $\sigma = 1$, $\epsilon = 1$ for different values of β. The continuous line curves correspond to the asymptotic performance while the points denote finite-sample performance when $p = 200$ and $n = \lfloor \delta p \rfloor$. The null risks (corresponding to $\hat{\mathbf{w}}_H = \mathbf{0}$) are also reported by the dotted lines for the different values of β.



(a) $\delta = 1$

(b) $\delta = 1.4$

Figure 3. Performance of H-SVR vs ϵ when $\beta = 1$ for different values of the noise variance σ^2 and for different values of δ. The continuous line curves correspond to the theoretical predictions while the points denote finite-sample performance when $p = 200$ and and $n = \lfloor \delta p \rfloor$.

4.1.2. Impact of choice of ϵ on the test performance

Fig. 3 displays the test risk with respect to ϵ for fixed signal power and noise variance and oversampling ratios $\delta = 1$ and $\delta = 1.4$, respectively. As can be noted, the test performance is sensitive to the choice of ϵ. An arbitrary choice of ϵ may lead to a significant loss in test performance. Indeed, a small ϵ tolerates less deviation from the plane $y = \hat{\mathbf{w}}_H^T \mathbf{x}$, which becomes inappropriate when the noise variance increases. On the other hand, a larger ϵ tolerates more deviation, and as such, tends to give less credit on the information from the training samples. We can also note that the optimal ϵ increases when more training samples are used. This can be explained by the fact that when using more training samples, it becomes harder to fit them into the insensitivity tube. Moreover, as can be seen from this figure, a right choice for the value ϵ is essential in practice, as arbitrary choices may lead to severe risk performance degradation. Several previous works addressed this question (Cherkassky & Ma, 2004; 2002). However, they do not rely on theoretical analysis but instead on cross-validation approaches. Finally, with reference to **Remark 2**, we plot in Fig. 4 the risk of H-SVR with respect to δ when at each δ, the optimal ϵ that minimizes the optimal risk is used. The obtained results show that the test risk becomes, in this case, a decreasing function of δ. This is in agreement with the fact that in optimally regularized learning architectures, there is always a gain from using more training samples.

4.2. S-SVR

4.2.1. Impact of the parameters ϵ and C

In Fig. 5 and Fig. 6, we investigate the effect of the hyperparameters C and ϵ on the performance of soft SVR. As shown in these figures, arbitrary choices for the pair (ϵ, C) may lead to a significant degradation in the test risk per-

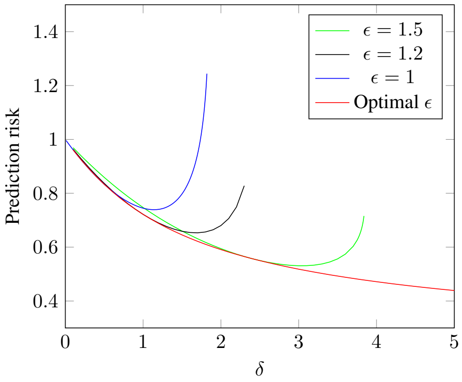


Figure 4. Performance of H-SVR as a function of δ when $\sigma = 1$ and $\beta = 1$ for different values of ϵ. All the curves correspond to the asymptotic performance predictions.

formance compared to the optimal performance associated with optimal selection of (ϵ, C). Thus, our results again emphasize the importance of the appropriate selection of these parameters and suggest the practical relevance of theoretical aided approaches to select these parameters that may complement existing cross-validation techniques.

4.2.2. Sample-wise descent phenomenon

In Fig. 7, we plot the test risk with respect to δ for the S-SVR for different choices of C and fixed ϵ. As can be seen, for δ tending to zero, the S-SVR test risk converges to that of the null risk, which was also predicted by our analysis in **Remark 6**. Moreover, depending on the value of C, the test risk may manifest a cusp at $\delta \sim \delta_\star$ that becomes more pronounced as C increases. This can be explained by the fact that as C increases, the behavior of S-SVR approaches that of H-SVR, for which the problem becomes unfeasible when $\delta > \delta_\star$. We also note that the choice of the parameter C plays a fundamental role in the test risk performance. Optimal values of C always guarantee that the test risk performance decreases with more training samples being used, while arbitrary choices can lead to the test risk increasing for more training samples. This emphasizes the importance of the appropriate selection of the parameter C to avoid the double descent phenomenon.

4.3. Comparison with ridge regression estimators under impulsive noises:

In this experiment, we investigate S-SVR and H-SVR's resilience when optimally designed (optimal ϵ for H-SVR and optimal C and ϵ for S-SVR) to impulsive noises and compare them to the ridge regression with optimal regularization. Particularly, we consider the case in which the noise is sampled from the distributional model:

$$n = \sqrt{\tau}\mathcal{N}(0, 1)$$

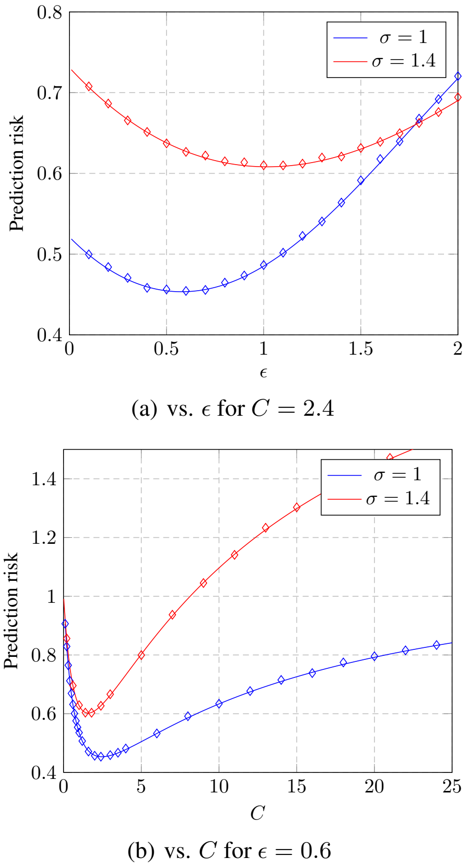


Figure 5. Prediction risk of S-SVR vs ϵ and C when $\delta = 2$, $\beta = 1$ and $\sigma = 1$. The continuous line curves correspond to the theoretical predictions while the points denote finite-sample performance when $p = 200$ and and $n = \lfloor \delta p \rfloor$.

where τ follows an inverse Gamma distribution with shape $\frac{d}{2}$ and scale $\frac{2}{d}$, that is $\tau \sim \frac{d}{\chi_d^2}$ with χ_d^2 being the chi-square distribution with d degrees of freedom.

Fig. 8 represents the test risk performance of the aforementioned estimators for $d = 3$ and $d = 10$. As can be seen, H-SVR is very sensitive to impulsive noises with a performance approaching that of the null estimator for highly impulsive noises. This behavior can be explained by the fact that in highly impulsive noises (small d), H-SVR needs a very large ϵ to guarantee that outliers satisfy the feasibility conditions. However, with a large ϵ, the constraints in (2) becomes irrelevant for the remaining well-behaved observations. This favors the H-SVR to select the null estimator as it would minimize the objective in (2) while satisfying the constraints. On the other hand, the S-SVR overcomes such behavior since it does not have to satisfy the constraints of (2). By selecting the parameter C to the value that minimizes the test risk, it will adaptively control the effect of outliers by relaxing the most unlikely constraints in (2). Moreover,

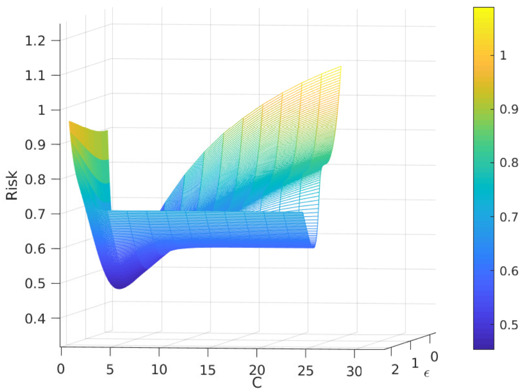
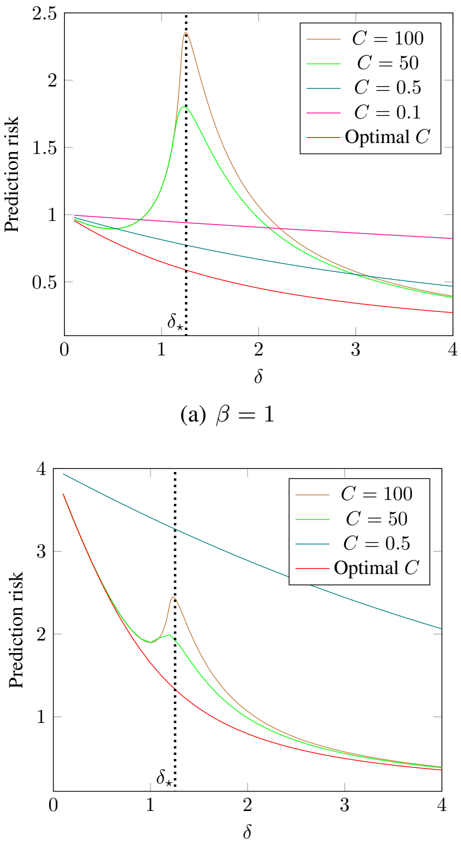


Figure 6. Prediction risk vs (C,ϵ) for $\delta = 2$, $\beta = 1$ and $\sigma = 1$.

it can be seen that, although S-SVR is slightly less efficient than the ridge regression estimator in mild impulsive noises, it presents a much lower risk under highly impulsive noises. While the robustness of regularized support vector machine methods for both regression and classification is a well-known fact reported in many previous works (Xu et al., 2009; Hable & Christmann, 2011), our result contributes to quantitatively assess such robustness by measuring the test risk under impulsive noises.

5. Conclusion

In this paper, we studied the asymptotic test risk of hard and soft support vector regression techniques with isotropic Gaussian features and under symmetric noise distributions in the regime of high dimensions. We used these results to illustrate the impact of the intervening parameters on the test risk behavior. Particularly, we demonstrate that arbitrary choices of the parameters of the hard SVR and the soft SVR may lead to the test risk presenting a non-monotonic behavior as a function of the number of samples, which illustrates the fact that adding more samples may be harmful to the performance. On the contrary, we show that optimally-tuned hard SVR and soft SVR present a decreasing test risk curve, which shows the importance of carefully selecting their parameters to minimize the test risk and guarantee the positive impact of more data on the test risk performance. Our findings are consistent with similar results obtained for linear regression and neural networks (Nakkiran et al., 2020b). However, as compared to linear regression, we demonstrate that soft support vector regression with optimal regularization is more robust to the presence of outliers, corroborating similar previous findings in earlier works in (Xu et al., 2009; Hable & Christmann, 2011). Several extensions of our work are worth investigating. One important research direction is to understand the effect of correlated features on the test risk of hard and support regression techniques. Some recent works have investigated the role of correlation and regularization on the test risk for linear regression models (Lolas, 2020; Kobak et al., 2020). A significant advantage of such an analysis is that it can illustrate the importance of investing efforts in theoretically-aided approaches to assist in setting the regularization parameters. Another important research direction is investigating the use of kernel support vector regression methods and understanding their underlying mechanisms to handle involved non-linear data models.



(a) $\beta = 1$

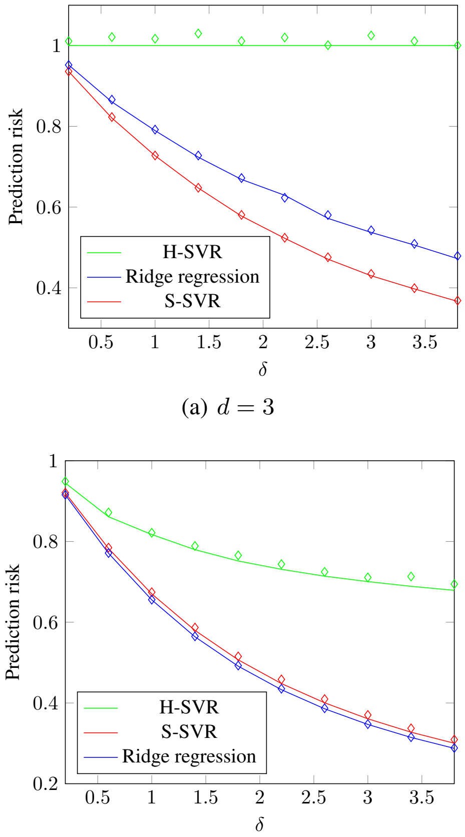
(b) $\beta = 2$

Figure 7. Prediction risk of S-SVR vs. δ for different values of C when $\epsilon = 0.6$ and $\sigma = 1$. Illustration of the sample-wise double decent and how optimal regularization mitigates it.

References

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a. ISSN 0027-8424. doi: 10.1073/pnas.1903070116.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *ArXiv*, abs/1903.07571, 2019b.



(a) $d = 3$

(b) $d = 10$

Figure 8. Prediction risk of S-SVR and ridge regression vs. δ when optimal regularization is used. Comparison of S-SVR and ridge regression with elliptic noise. The continuous line curves correspond to the asymptotic performance while the points denote finite-sample performance when $p = 200$ and $n = \lfloor \delta p \rfloor$.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality? In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1611–1619. PMLR, 16–18 Apr 2019c.

Bös, S. and Opper, M. Dynamics of training. *Advances in Neural Information Processing Systems*, pp. 141–147, 1997.

Candes, E. J. and Sur, P. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv e-prints*, art. arXiv:1804.09753, April 2018.

Cherkassky, V. and Ma, Y. Selection of meta-parameters for support vector regression. In *Artificial Neural Networks — ICANN 2002*, pp. 687–693, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

Cherkassky, V. and Ma, Y. Practical selection of svm parameters and noise estimation for svm regression. *Neural networks*, 17(1):113–126, 2004.

Deng, Z., Kammoun, A., and Thrampoulidis, C. A model of Double Descent for High-dimensional Binary Linear Classification. *submitted to Information and Inference Journal of the IMA*, 2020.

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *ArXiv*, abs/1901.01608, 2019.

Guajardo, J., Weber, R., and Miranda, J. A forecasting methodology using support vector regression and dynamic feature selection. *Journal of Information & Knowledge Management*, 5(04):329–335, 2006.

Hable, R. and Christmann, A. On qualitative robustness of support vector machines. *Journal of Multivariate Analysis*, 102(6):993–1007, 2011.

Hamdi, T., Ali, J. B., Di Costanzo, V., Fnaiech, F., Moreau, E., and Ginoux, J.-M. Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm. *Biocybernetics and Biomedical Engineering*, 38(2):362–372, 2018.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Ju, X., Cheng, M., Xia, Y., Quo, F., and Tian, Y. Support vector regression and time series analysis for the forecasting of bayannur's total water requirement. *Procedia Computer Science*, 31:523–531, 2014.

Kammoun, A. and Alouini, M.-S. On the precise error analysis of support vector machines. *Submitted to Open journal of signal processing*, 2020.

Kini, G. R. and Thrampoulidis, C. Analytic study of double descent in binary classification: The impact of loss. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2527–2532, 2020. doi: 10.1109/ISIT44484.2020.9174344.

Kobak, D., Lomond, J., and Sanchez, B. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *J. Mach. Learn. Res.*, 21:169:1–169:16, 2020.

Lolas, P. Regularization in high-dimensional regression and classification via random matrix theory. *arXiv: Statistics Theory*, 2020.

Mitra, P. P. Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation. *arXiv:1906.03667*, 2019.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime, 2020.

Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1 (1):67–83, 2020. doi: 10.1109/JSAIT.2020.2984716.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=B1g5sA4twr.

Nakkiran, P., Venkat, P., Kakade, S. M., and Ma, T. Optimal regularization can mitigate double descent. *CoRR*, abs/2003.01897, 2020b. URL https://arxiv.org/abs/2003.01897.

Qu, H. and Zhang, Y. A new kernel of support vector regression for forecasting high-frequency stock returns. *Mathematical Problems in Engineering*, 2016.

Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. A jamming transition from under- to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52, 10 2019. doi: 10.1088/1751-8121/ab4c8b.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. Precise Error Analysis of Regularized M-Estimators in High Dimensions. *IEEE Transactions on Information Theory*, 64(8), August 2018.

Xu, H., Caramanis, C., and Mannor, S. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(7), 2009.

Yang, H., Huang, K., King, I., and Lyu, M. R. Localized support vector regression for time series prediction. *Neurocomputing*, 72(10-12):2659–2669, 2009.