Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2009-060
CBCL-284

December 1, 2009

Sufficient Conditions for Uniform Stability of Regularization Algorithms

Andre Wibisono, Lorenzo Rosasco, and Tomaso Poggio

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Sufficient Conditions for Uniform Stability of Regularization Algorithms

Andre Wibisono[†], Lorenzo Rosasco[†,‡], Tomaso Poggio[†]
[†] *Center for Biological and Computational Learning, MIT, USA*
[‡] *DISI, Università di Genova,Italy*
wibisono@mit.edu, lrosasco@mit.edu, tp@ai.mit.edu

November 23, 2009

Abstract

In this paper, we study the stability and generalization properties of penalized empirical-risk minimization algorithms. We propose a set of properties of the penalty term that is sufficient to ensure uniform β-stability: we show that if the penalty function satisfies a suitable convexity property, then the induced regularization algorithm is uniformly β-stable. In particular, our results imply that regularization algorithms with penalty functions which are strongly convex on bounded domains are β-stable. In view of the results in [3], uniform stability implies generalization, and moreover, consistency results can be easily obtained. We apply our results to show that ℓ_p regularization for $1 < p \leq 2$ and elastic-net regularization are uniformly β-stable, and therefore generalize.

1 Introduction

In supervised learning, we are given a set of n data points $\mathbf{z} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d. from an (unknown) probability distribution ρ on the product space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and we have to estimate a function $f_{\mathbf{z}} \colon \mathcal{X} \to \mathcal{Y}$ that tries to predict $f_{\mathbf{z}}(x) \approx y$ (see [7] and [11]), in a sense that we formalize later in this paper. A *learning algorithm* is an algorithm that takes in as input the training dataset $\mathbf{z}$ and yields as output a solution function $f_{\mathbf{z}}$.

A fundamental class of learning algorithms can be described as the *penalized/regularized empirical-risk minimization* problem, where we find a function f from a hypothesis space $\mathcal{H}$ by minimizing the regularized empirical-risk functional,

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \mathcal{P}(f) \right\}.$$

The first term of the functional above represents the empirical error that the function f makes on the dataset, while the second term represents a regularizer that controls the complexity of the function. The regularization parameter λ is to be chosen according to the problem to adjust the trade-off between the empirical error and the complexity of the function. A special case of the regularized empirical-risk minimization when the penalty function $\mathcal{P}(f)$ is the norm $\|f\|_{\mathcal{H}}^2$ on a reproducing kernel Hilbert space $\mathcal{H}$ [2] is known as *Tikhonov regularization* [5]. It turns out that

many well-known machine learning algorithms, including Regularization Networks, Support Vector Machines, Splines, and Radial Basis Functions, are special cases of Tikhonov regularization with a particular choice of the loss function V and hypothesis space $\mathcal{H}$ (see, for example, [5], [12], and [6]).

Tikhonov regularization has been studied extensively, and has been shown to possess many of the desired properties of learning algorithms, including stability and generalization [3]. The notion of *generalization* can be formalized as the requirement that the empirical risk of the solution be concentrated around the expected risk. In other words, if we observe that the empirical error of the solution function $f_{\mathbf{z}}$ is small, then with high confidence, we know that the expected error is also small. For regularized empirical-risk minimization algorithms, generalization implies *consistency*, that is, the solution approaches the best function in the hypothesis space $\mathcal{H}$ as the size of the dataset grows. In this context, *stability* is the notion that a small change in the training dataset $\mathbf{z}$ does not change much the solution. The connection between generalization and stability is studied in [3], where it is shown that a suitable notion of stability, namely, uniform β-stability, implies generalization. Such a connection is further developed in [10], where weaker notions of stability are studied. In this paper, we consider the former notion of stability.

Besides Tikhonov regularization, there are many algorithms that fall under the regularized empirical-risk minimization setting, but use different penalty functions $\mathcal{P}$. Different choices of penalty functions lead to different properties of the solution functions, that might be desired based on the specific nature of the problems. However, in choosing the penalty function to use, we have to be careful so as to preserve the nice properties of the learning algorithms, such as stability, generalization, and consistency, that Tikhonov regularization enjoys. Nevertheless, checking the uniform stability property of a regularization algorithm directly is difficult, as currently there is no characterization of when an algorithm is stable.

Our goal in this paper is to characterize the conditions on the penalty function that lead to stability. We choose to focus on the penalty function only (rather than including the loss function, or considering the whole regularized empirical-risk functional) because by the construction of the regularized empirical-risk minimization algorithm, the penalty function as the regularizer plays the key role in ensuring that the algorithm is well-behaved, so it is reasonable to expect that there is a condition on the penalty function that leads to the stability of the algorithm. Moreover, a characterization of stability in terms of the penalty function ensures that the property of the algorithm is preserved regardless of the training dataset $\mathbf{z}$ that we receive.

Our main result can be stated (informally) as follows. Let $f_{\mathbf{z}}$ and $f_{\mathbf{z}^j}$ be the solution functions of the regularized empirical-risk minimization algorithm with the training dataset $\mathbf{z}$ and the modified dataset $\mathbf{z}^j$, respectively. If there are some constants $C > 0$ and $\xi > 1$ such that the penalty function $\mathcal{P}$ satisfies

$$\mathcal{P}(f_{\mathbf{z}}) + \mathcal{P}(f_{\mathbf{z}^j}) - 2\mathcal{P}\left(\frac{f_{\mathbf{z}} + f_{\mathbf{z}^j}}{2}\right) \geq C\|f_{\mathbf{z}} - f_{\mathbf{z}^j}\|_{\mathcal{H}}^{\xi}, \tag{1}$$

then the algorithm is uniformly β-stable with $\beta = \mathcal{O}(n^{-\frac{1}{\xi-1}})$. In particular, if $1 < \xi < 3$, then $\beta = o(n^{-1/2})$, and therefore, using results from [3], the algorithm generalizes. As we shall see in this paper, the solution function $f_{\mathbf{z}}$ lies in a ball with finite radius in the hypothesis space $\mathcal{H}$, and therefore, a special case of the result above is when the penalty function $\mathcal{P}$ is *strongly convex* on bounded domains, for then the regularization algorithm satisfies Eq. (1) with $\xi = 2$, and thus generalizes. As an application of our results, we show that ℓ_p regularization for $1 < p \leq 2$ and elastic-net regularization are uniformly β-stable with $\beta = \mathcal{O}(n^{-1})$, and therefore, both algorithms generalize.

The rest of this paper is organized as follows. Section 2 defines the general setting of learning algorithms that we are working with. Section 3 introduces regularization algorithms and proves the main theorem about the stability of regularization algorithms. Section 4 applies the results to ℓ_p regularization for $1 < p \leq 2$ and elastic-net regularization. Section 5 explores the connection between generalization and consistency, and finally, Section 6 discusses the results and possible future work.

2 Problem Setting

In this section, we establish the framework of the problem and introduce our notation. First, we describe the setting on the probability and function spaces that we consider. Then, we define the notions of stability and generalization of learning algorithms, and how these two notions relate.

2.1 Probability and Function Spaces

Let $\mathcal{X}$ be a separable metric space and $\mathcal{Y}$ be a real separable Hilbert space. We denote the inner product and norm in $\mathcal{Y}$ by $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$ and $|\cdot|_{\mathcal{Y}}$, respectively. We assume that there is an underlying unknown probability distribution ρ on the product space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\mathbf{z} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a dataset that we observe, where the points $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ are drawn identically and independently from ρ. Our objective is to find a function $f \colon \mathcal{X} \to \mathcal{Y}$ from a hypothesis space $\mathcal{H}$ that "best" explains the values $y \in \mathcal{Y}$ via the variables $x \in \mathcal{X}$.

To describe the hypothesis space $\mathcal{H}$, we consider the setting introduced in [4]. Let Γ be a countable set. We start with a collection of feature functions $(\varphi_\gamma)_{\gamma \in \Gamma}$, where each $\varphi_\gamma \colon \mathcal{X} \to \mathcal{Y}$ is a measurable function satisfying the property that for every $x \in \mathcal{X}$,

$$\sum_{\gamma \in \Gamma} |\varphi_\gamma(x)|_{\mathcal{Y}}^2 \leq \kappa$$

for some $0 < \kappa < +\infty$. Let $\ell_2(\Gamma)$ denote the set of sequences $\alpha = (\alpha_\gamma)_{\gamma \in \Gamma}$, where $\alpha_\gamma \in \mathbb{R}$ and $\sum_{\gamma \in \Gamma} \alpha_\gamma^2 < +\infty$. The usual inner product and norm in $\ell_2(\Gamma)$ are denoted by $\langle \cdot, \cdot \rangle_{\ell_2}$ and $\| \cdot \|_{\ell_2}$, respectively.

The *hypothesis space* that we consider is the linear span of the feature functions $(\varphi_\gamma)_{\gamma \in \Gamma}$,

$$\mathcal{H} = \{f_\alpha \colon \mathcal{X} \to \mathcal{Y} \mid f_\alpha = \sum_{\gamma \in \Gamma} \varphi_\gamma \alpha_\gamma,\ \alpha \in \ell_2(\Gamma)\}.$$

Note that the mapping $\alpha \mapsto f_\alpha$ is a linear operator, i.e. $f_{\alpha_1 - \alpha_2} = f_{\alpha_1} - f_{\alpha_2}$. The construction of $\mathcal{H}$ ensures that for every $f_\alpha \in \mathcal{H}$, the pointwise evaluation $f_\alpha(x)$ is bounded. In fact, $f_\alpha(x) = \sum_{\gamma \in \Gamma} \varphi_\gamma(x) \alpha_\gamma$ converges absolutely, because the Cauchy-Schwarz inequality gives us

$$\sum_{\gamma \in \Gamma} |\varphi_\gamma(x) \alpha_\gamma|_{\mathcal{Y}} \leq \left(\sum_{\gamma \in \Gamma} |\varphi(x)|_{\mathcal{Y}}^2 \right)^{1/2} \left(\sum_{\gamma \in \Gamma} \alpha_\gamma^2 \right)^{1/2} \leq \kappa^{1/2} \|\alpha\|_{\ell_2}.$$

In particular, the inequality above gives us the following bound on the infinity norm of f_α,

$$\|f_\alpha\|_\infty = \sup_{x \in \mathcal{X}} |f_\alpha(x)|_{\mathcal{Y}} \leq \kappa^{1/2} \|\alpha\|_{\ell_2}. \tag{2}$$

Our definition of $\mathcal{H}$ actually implies that it is a reproducing kernel Hilbert space. For more detail on the reproducing property of $\mathcal{H}$, see [4].

To measure the "quality" of a function f, we introduce a *loss function* $V\colon \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$. We assume that V satisfies the following three properties:

(i) for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the functional $\alpha \mapsto V(f_\alpha(x), y)$ is convex;

(ii) there exists $B > 0$ such that for all $y \in \mathcal{Y}$, we have $V(0, y) \leq B$;

(iii) V is L-Lipschitz in the first variable, so that for all $y, y_1, y_2 \in \mathcal{Y}$, we have

$$|V(y_1, y) - V(y_2, y)| \leq L|y_1 - y_2|_{\mathcal{Y}}.$$

We will see in the following that in the case of regularization algorithms, condition (i) allows us to characterize the stability property of the algorithm in terms of the penalty function. Properties (ii) and (iii) ensure that the loss function is well-behaved. We note that the three properties listed above are reasonable and are satisfied by most commonly used loss functions. For example, if $\mathcal{Y} = \mathbb{R}$, then the hinge loss $V(y_1, y_2) = (1 - y_1 y_2)_+$ satisfies properties (i)–(iii), and if $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$, then the square loss $V(y_1, y_2) = (y_1 - y_2)^2$ also satisfies properties (i)–(iii).

Given a function $f \in \mathcal{H}$, we can define the *empirical risk* of f (with respect to $\mathbf{z}$) to be

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i),$$

and the *expected risk* of f to be

$$\mathcal{E}(f) = \int_{\mathcal{X} \times \mathcal{Y}} V(f(x), y)\, d\rho(x, y).$$

The empirical risk can be seen as a proxy for the expected risk. The definition of generalization in the following section will make the previous statement precise.

2.2 Stability and Generalization

We now define the notions of stability and generalization of learning algorithms, and explain how these concepts interact.

Stability. The notion of stability that we use is the *uniform β-stability* definition introduced in [3]. First, given a training dataset $\mathbf{z}$, we define $\mathbf{z}^j$ to be the modified dataset obtained by switching the jth point of $\mathbf{z}$ with a new point (x', y'), so that $\mathbf{z}^j = (S \setminus \{(x_j, y_j)\}) \cup \{(x', y')\}$, for $1 \leq j \leq n$. Let $f_{\mathbf{z}}$ and $f_{\mathbf{z}^j}$ denote the solution functions of the learning algorithm with dataset $\mathbf{z}$ and $\mathbf{z}^j$, respectively. Then we say that a learning algorithm is *uniformly β-stable* if

$$\sup_{x \in \mathcal{X},\, y \in \mathcal{Y}} \left|V(f_{\mathbf{z}}(x), y) - V(f_{\mathbf{z}^j}(x), y)\right| \leq \beta.$$

Intuitively, the definition above means that a small change in the input dataset should not much change the solution function $f_{\mathbf{z}}$, as measured by the loss function V. Note that the bound β will in general depend on n, the size of the dataset $\mathbf{z}$, and λ, the regularization parameter.

Generalization. We say that an algorithm *generalizes* if the empirical risk of the solution function $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})$ converges in probability to the expected risk $\mathcal{E}(f_{\mathbf{z}})$ as the size of the dataset $\mathbf{z}$ increases. More specifically, an algorithm generalizes if we have a probabilistic bound on the generalization error $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathbf{z}})$, and the error term in the bound tends to 0 as n tends to $+\infty$. As stated in the introduction, a generalization property of an algorithm gives us a belief that when we observe a small empirical error, the expected error is also small. Intuitively, generalization also means predictivity: if an algorithm generalizes, then the solution function can predict the values $y \in \mathcal{Y}$ corresponding to the variables $x \in \mathcal{X}$ well, even for unseen data points (x, y).

The relation between stability and generalization comes from the following result by Bousquet and Elisseeff [3]:

Theorem 2.1. *Suppose that the regularization algorithm is uniformly β-stable. Then for every $0 < \delta \leq 1$, the following bound holds with probability at least $1 - \delta$,*

$$\mathcal{E}(f_{\mathbf{z}}) \leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \beta + \left(2n\beta + L\kappa^{1/2}\tau + B\right)\sqrt{\frac{\log(1/\delta)}{2n}},$$

where $\tau \geq 0$ is an upper bound to $\|f_{\mathbf{z}}\|_{\mathcal{H}}$ (see Corollary 3.2).

This theorem above essentially states that stability implies generalization. In particular, if $\beta = o(n^{-1/2})$, that is,

$$\lim_{n \to +\infty} \frac{\beta(n)}{n^{-1/2}} = \lim_{n \to +\infty} \beta(n) \cdot n^{1/2} = 0,$$

then the error term in the bound above tends to 0 as n tends to $+\infty$.

3 Stability of Regularization Algorithms

In this section, we consider a class of learning algorithms known as regularization algorithms. We first give some preliminary results, and then prove the main theorem about the stability of regularization algorithms.

3.1 Regularization Algorithms

Let $\mathcal{P}\colon \ell_2(\Gamma) \to [0, +\infty)$ be a penalty function that regularizes the complexity of the function f_α corresponding to $\alpha \in \ell_2(\Gamma)$. As mentioned in the introduction, the learning algorithm that we consider in this paper is the *regularized empirical-risk minimization* algorithm, also called regularization algorithm. In our setting, the regularization algorithm corresponds to the following minimization problem,

$$\min_{\alpha \in \ell_2(\Gamma)} \left\{ \frac{1}{n} \sum_{i=1}^{n} V(f_\alpha(x_i), y_i) + \lambda \mathcal{P}(\alpha) \right\}.$$

The regularization parameter $\lambda > 0$ controls the trade-off between the empirical error and the complexity of the function f_α.

Note that we currently do not make any assumption on the penalty function $\mathcal{P}$. However, since we are working in an unbounded domain $\ell_2(\Gamma)$, to ensure that the minimizer exists, we assume that $\mathcal{P}$ is *coercive*, that is,

$$\lim_{\|\alpha\|_{\ell_2} \to +\infty} \mathcal{P}(\alpha) = +\infty.$$

For example, for $1 < p \leq 2$, the function

$$\mathcal{P}(\alpha) = \|\alpha\|_{\ell_p}^p = \sum_{\gamma \in \Gamma} |\alpha_\gamma|^p$$

is coercive, since the ℓ_p-norm is stronger than the ℓ_2-norm for $1 < p \leq 2$, i.e. $\|\alpha\|_{\ell_2} \leq \|\alpha\|_{\ell_p}$. On the other hand, if $p > 2$ and the index set Γ is infinite, then $\mathcal{P}(\alpha) = \|\alpha\|_{\ell_p}^p$ is not coercive. To see that, take $\Gamma = \mathbb{N}$, and consider a sequence $(\alpha_n)_{n \in \mathbb{N}}$ of points in $\ell_2(\mathbb{N})$, where the jth component of α_n is $1/\sqrt{j}$, if $1 \leq j \leq n$, and 0 otherwise. Then we have

$$\lim_{n \to +\infty} \|\alpha_n\|_{\ell_2} = \lim_{n \to +\infty} \left(\sum_{j=1}^{n} \frac{1}{j} \right)^{1/2} = +\infty,$$

but since $p > 2$,

$$\lim_{n \to +\infty} \|\alpha_n\|_{\ell_p}^p = \lim_{n \to +\infty} \sum_{j=1}^{n} \frac{1}{j^{p/2}} < +\infty.$$

In fact, $\mathcal{P}(\alpha) = \|\alpha\|_{\ell_p}^p$ for $1 < p \leq 2$ will be one of the examples that we consider in Section 3.2, corresponding to the algorithm ℓ_p regularization.

Let $\alpha_{\mathbf{z}}$ denote a minimizer of the objective functional with dataset $\mathbf{z}$. Note that we do not require that the minimizer be unique; the statements about $\alpha_{\mathbf{z}}$ that we make in this paper hold for all points α in the set of minimizers of the objective functional. Henceforth, we will refer to $\alpha_{\mathbf{z}}$ as *the* minimizer. Similarly, let $\alpha_{\mathbf{z}^j}$ denote the minimizer of the objective functional with the modified dataset $\mathbf{z}^j$. For clarity of notation, we denote the function $f_{\alpha_{\mathbf{z}}}$ corresponding to $\alpha_{\mathbf{z}}$ by $f_{\mathbf{z}}$, and similarly, we denote the function $f_{\alpha_{\mathbf{z}^j}}$ corresponding to $\alpha_{\mathbf{z}^j}$ by $f_{\mathbf{z}^j}$.

An important property of the minimizer $\alpha_{\mathbf{z}}$ is the following simple result.

Lemma 3.1. *If $\lambda > 0$, then*

$$\mathcal{P}(\alpha_{\mathbf{z}}) \leq \frac{B + \mathcal{P}(\mathbf{0})}{\lambda}.$$

Proof. First we note that the function $f_{\mathbf{0}}$ corresponding to the zero element $\mathbf{0} \in \ell_2(\Gamma)$ is an identically zero function. Therefore, invoking the property that $\alpha_{\mathbf{z}}$ is the minimizer of the objective functional, we obtain

$$\begin{aligned} \lambda \mathcal{P}(\alpha_{\mathbf{z}}) &\leq \frac{1}{n} \sum_{i=1}^{n} V(f_{\mathbf{z}}(x_i), y_i) + \lambda \mathcal{P}(\alpha_{\mathbf{z}}) \\ &\leq \frac{1}{n} \sum_{i=1}^{n} V(f_{\mathbf{0}}(x_i), y_i) + \mathcal{P}(\mathbf{0}) \\ &\leq B + \mathcal{P}(\mathbf{0}). \end{aligned}$$

Dividing both sides by λ yields the desired inequality. □

Since $\mathcal{P}$ is coercive, Lemma 3.1 yields the following corollary.

Corollary 3.2. *There exists $\tau \geq 0$ such that $\|\alpha_{\mathbf{z}}\|_{\ell_2} \leq \tau$.*

We note that the constant τ might depend on λ since the bound of $\mathcal{P}(\alpha_{\mathbf{z}})$ depends on λ. When needed, we will refer to τ as $\tau(\lambda)$. Note also that this constant τ appears in the generalization bound in Theorem 2.1.

Furthermore, we have the following bound on the value of the loss function.

Lemma 3.3. *For all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have $V(f_{\mathbf{z}}(x), y) \leq L\kappa^{1/2}\tau + B$.*

Proof. Using the triangle inequality and the Lipschitz property of V, we obtain

$$\begin{aligned} V(f_{\mathbf{z}}(x), y) &\leq |V(f_{\mathbf{z}}(x), y) - V(0, y)| + V(0, y) \\ &\leq L|f_{\mathbf{z}}(x)|_{\mathcal{Y}} + B. \end{aligned}$$

Then, using the bound on the infinity-norm on $f_{\mathbf{z}}$ (Eq. (2)), we have

$$|f_{\mathbf{z}}(x)|_{\mathcal{Y}} \leq \|f_{\mathbf{z}}\|_{\infty} \leq \kappa^{1/2}\|\alpha_{\mathbf{z}}\|_{\ell_2} \leq \kappa^{1/2}\tau.$$

Combining the two inequalities above yields the lemma. □

3.2 Stability Induced by the Penalty Function

In this section, we prove that regularization algorithms whose penalty functions satisfy a certain convexity property on the minimizers are uniformly β-stable. To do that, we need the following result from [3], which exploits the fact that the dataset $\mathbf{z}$ and the modified dataset $\mathbf{z}^j$ differ at precisely one point. Note that in order to apply this result, we need the assumption that the functional $\alpha \mapsto V(f_\alpha(x), y)$ is convex, as listed in property (i) of V.

Lemma 3.4. *For $0 \leq t \leq 1$, we have the following inequality*[1]

$$\mathcal{P}(\alpha_{\mathbf{z}}) - \mathcal{P}(t\alpha_{\mathbf{z}} + (1-t)\alpha_{\mathbf{z}^j}) + \mathcal{P}(\alpha_{\mathbf{z}^j}) - \mathcal{P}((1-t)\alpha_{\mathbf{z}} + t\alpha_{\mathbf{z}^j}) \leq \frac{2tL}{n\lambda}\|f_{\mathbf{z}} - f_{\mathbf{z}^j}\|_{\infty}.$$

Now we are ready to state the theorem.

Theorem 3.5. *Suppose that for some constants $C > 0$ and $\xi > 1$, the penalty function satisfies*

$$\mathcal{P}(\alpha_{\mathbf{z}}) + \mathcal{P}(\alpha_{\mathbf{z}^j}) - 2\mathcal{P}\left(\frac{\alpha_{\mathbf{z}} + \alpha_{\mathbf{z}^j}}{2}\right) \geq C\|\alpha_{\mathbf{z}} - \alpha_{\mathbf{z}^j}\|_{\ell_2}^{\xi}. \tag{3}$$

Then the regularization algorithm is uniformly β-stable with

$$\beta = \left(\frac{L^{\xi}\,\kappa^{\xi/2}}{n\lambda C}\right)^{\frac{1}{\xi-1}}.$$

Proof. Using the Lipschitz property of V and Eq. (2), we have

$$\sup_{x\in\mathcal{X},\, y\in\mathcal{Y}} |V(f_{\mathbf{z}}(x), y) - V(f_{\mathbf{z}^j}(x), y)| \leq L\|f_{\mathbf{z}} - f_{\mathbf{z}^j}\|_{\infty} \leq L\kappa^{1/2}\|\alpha_{\mathbf{z}} - \alpha_{\mathbf{z}^j}\|_{\ell_2}.$$

[1] The coefficient on the right hand side of the inequality is slightly different from what appears in Lemma 20 in [3], because the setting in [3] considers $\mathbf{z}^{\setminus j} = \mathbf{z} \setminus \{(x_j, y_j)\}$ as the modified dataset, while we consider $\mathbf{z}^j = \big(\mathbf{z} \setminus \{(x_j, y_j)\}\big) \cup \{(x', y')\}$.

Now we need to bound $\|\alpha_{\mathbf{z}} - \alpha_{\mathbf{z}^j}\|_{\ell_2}$. Applying Lemma 3.4 with $t = 1/2$, we have the following inequality

$$\mathcal{P}(\alpha_{\mathbf{z}}) + \mathcal{P}(\alpha_{\mathbf{z}^j}) - 2\mathcal{P}\left(\frac{\alpha_{\mathbf{z}} + \alpha_{\mathbf{z}^j}}{2}\right) \leq \frac{L}{n\lambda}\|f_{\mathbf{z}} - f_{\mathbf{z}^j}\|_\infty.$$

Combining the inequality above with the given property of $\mathcal{P}$ and Eq. (2), we obtain

$$\|\alpha_{\mathbf{z}} - \alpha_{\mathbf{z}^j}\|_{\ell_2} \leq \left(\frac{L}{n\lambda C}\right)^{1/\xi} \|f_{\mathbf{z}} - f_{\mathbf{z}^j}\|_\infty^{1/\xi} \leq \left(\frac{L\kappa^{1/2}}{n\lambda C}\right)^{1/\xi} \|\alpha_{\mathbf{z}} - \alpha_{\mathbf{z}^j}\|_{\ell_2}^{1/\xi}.$$

Since $\xi > 1$, this gives us

$$\|\alpha_{\mathbf{z}} - \alpha_{\mathbf{z}^j}\|_{\ell_2} \leq \left(\frac{L\kappa^{1/2}}{n\lambda C}\right)^{\frac{1}{\xi - 1}}.$$

Finally, substituting this result to our first inequality above yields

$$\sup_{x \in \mathcal{X},\, y \in \mathcal{Y}} |V(f_{\mathbf{z}}(x), y) - V(f_{\mathbf{z}^j}(x), y)| \leq \left(\frac{L^\xi\, \kappa^{\xi/2}}{n\lambda C}\right)^{\frac{1}{\xi - 1}},$$

as desired. □

As noted in Section 2.2, we want $\beta = o(n^{-1/2})$ to obtain a generalization bound where the error term tends to 0 as n tends to $+\infty$. Equivalently, for the regularization algorithm to generalize, we need the power ξ in Eq. (3) to be such that $1 < \xi < 3$.

Now recall that a function $\mathcal{P}\colon \ell_2(\Gamma) \to \mathbb{R}$ is *strongly convex* if there exists a constant $C > 0$ such that for every $\alpha_1, \alpha_2 \in \ell_2(\Gamma)$,

$$\mathcal{P}(\alpha_1) + \mathcal{P}(\alpha_2) - 2\mathcal{P}\left(\frac{\alpha_1 + \alpha_2}{2}\right) \geq C\|\alpha_1 - \alpha_2\|_{\ell_2}^2.$$

Clearly, a strongly convex penalty function satisfies Eq. (3) with $\xi = 2$, and hence we can apply Theorem 3.5 to conclude that the regularization algorithm is uniformly β-stable with $\beta = \mathcal{O}(n^{-1})$, and therefore generalizes. In fact, we can make a slightly stronger statement, since Eq. (3) only requires $\mathcal{P}$ to be strongly convex on the minimizers, and by Corollary 3.2, we know that the minimizers $\alpha_{\mathbf{z}}$ and $\alpha_{\mathbf{z}^j}$ have bounded ℓ_2-norms. Therefore, the conclusion of Theorem 3.5 still holds for penalty functions that are strongly convex on the bounded domain $\{\alpha \in \ell_2(\Gamma) \mid \|\alpha\|_{\ell_2} \leq \tau\}$, where τ is the bound on the ℓ_2-norm of the minimizers, as in Corollary 3.2. For example, the function $\mathcal{P}(\alpha) = \|\alpha\|_{\ell_p}^p$, for $1 < p \leq 2$ is not strongly convex over $\ell_2(\Gamma)$. However, it is strongly convex over bounded domains, as shown in Section 4.1, which allows us to apply Theorem 3.5.

Moreover, we note that while the condition in Eq. (3) is in principle weaker than strong convexity, it is nonconstructive, as it involves the minimizers $\alpha_{\mathbf{z}}$ and $\alpha_{\mathbf{z}^j}$. In practice, it is more convenient to check that the penalty function $\mathcal{P}$ is strongly convex (or strongly convex on bounded domains, as stated in the preceding paragraph), rather than showing that $\mathcal{P}$ satisfies Eq. (3) on the minimizers. To verify strong convexity, we can proceed from the definition directly, as we shall do in Sections 4.1 and 4.2, or use the following result from [8], which reduces the problem into checking strong convexity in one dimension[2].

[2]The original formulation of the theorem in [8] is for functions with domains in finite-dimensional Euclidean spaces $\mathbb{R}^m$, but the result extends naturally to $\ell_2(\Gamma)$.

Lemma 3.6. *A function $\mathcal{P}\colon \ell_2(\Gamma) \to \mathbb{R}$ is strongly convex on a convex set $A \subset \ell_2(\Gamma)$ if and only if the function*

$$g(t) = \mathcal{P}\left(\alpha_1 + \frac{t}{\|\alpha_1 - \alpha_2\|_{\ell_2}}(\alpha_2 - \alpha_1)\right)$$

is strongly convex on the interval $0 \le t \le \|\alpha_1 - \alpha_2\|_{\ell_2}$ with the same coefficient for all $\alpha_1, \alpha_2 \in A$, $\alpha_1 \neq \alpha_2$.

We summarize the discussion above in the following corollary.

Corollary 3.7. *A regularization algorithm with a penalty function $\mathcal{P}$ that is strongly convex over the bounded domain $\|\alpha\|_{\ell_2} \le \tau$ is uniformly β-stable with $\beta = \mathcal{O}(n^{-1})$, and therefore generalizes.*

4 Examples

In this section, we consider some specific regularization algorithms, characterized by different penalty functions $\mathcal{P}$, and apply Theorem 3.5 to show that they are uniformly β-stable.

4.1 Stability of ℓ_p Regularization, for $1 < p \le 2$

The first regularization algorithm that we consider is ℓ_p regularization for $1 < p \le 2$, which is characterized by the penalty function

$$\mathcal{P}(\alpha) = \|\alpha\|_{\ell_p}^p = \sum_{\gamma \in \Gamma} |\alpha_\gamma|^p.$$

The algorithm ℓ_p regularization is also known as *bridge regression* in statistics, and can be interpreted as a middle ground between $p = 2$ (which yields smooth solutions) and $p = 1$ (which yields sparse solutions). Note that Tikhonov regularization is a particular case when $p = 2$.

As noted in 3, the function $\mathcal{P}$ is coercive because for every $\alpha \in \ell_2(\Gamma)$, we have $\|\alpha\|_{\ell_2} \le \|\alpha\|_{\ell_p}$. Moreover, since we have the following bound

$$\|\alpha_{\mathbf{z}}\|_{\ell_2} \le \|\alpha_{\mathbf{z}}\|_{\ell_p} \le \left(\frac{B}{\lambda}\right)^{1/p},$$

$\mathcal{P}$ satisfies Corollary 3.2 with $\tau(\lambda) = (B/\lambda)^{1/p}$.

Now to apply Theorem 3.5, we need to show that the penalty function $\mathcal{P}(\alpha) = \|\alpha\|_{\ell_p}^p$ satisfies Eq. (3).

Proposition 4.1. *For $1 < p \le 2$, ℓ_p regularization satisfies Eq. (3) with $\xi = 2$ and*

$$C = \frac{1}{4}p(p-1)\left(\frac{B}{\lambda}\right)^{\frac{p-2}{p}}.$$

Proof. Consider a function $g\colon \mathbb{R} \to \mathbb{R}$ defined as $g(\theta) = |\theta|^p$. Notice that we have $g'(\theta) = p \cdot \operatorname{sign}(\theta) \cdot |\theta|^{p-1}$ and $g''(\theta) = p(p-1)|\theta|^{p-2}$. Clearly the first derivative of g exists and is continuous for all

$\theta \in \mathbb{R}$. On the other hand, the second derivative of g exists and is continuous for all $\theta \neq 0$, while for $\theta = 0$, we have

$$g''(0) = \lim_{\epsilon \to 0} \frac{g(\epsilon) - 2g(0) + g(-\epsilon)}{\epsilon^2} = \lim_{\epsilon \to 0} \frac{2|\epsilon|^p}{\epsilon^2} = \begin{cases} 2, & \text{if } p = 2, \\ +\infty, & \text{if } p < 2. \end{cases}$$

Then by the second-order Mean Value Theorem, which holds even when the derivative of the function is $+\infty$ or $-\infty$ [1], for every $a, b \in \mathbb{R}$, we have

$$|a|^p + |b|^p - 2\left|\frac{a+b}{2}\right|^p = \frac{1}{4}(b-a)^2 p(p-1)|c|^{p-2},$$

where $\min\{a, b\} \leq c \leq \max\{a, b\}$.

In particular, for every $\gamma \in \Gamma$, we can set a and b to be the components $\alpha_{\mathbf{z},\gamma}$ and $\alpha_{\mathbf{z}^j,\gamma}$ of $\alpha_{\mathbf{z}}$ and $\alpha_{\mathbf{z}^j}$, respectively, to obtain

$$|\alpha_{\mathbf{z},\gamma}|^p + |\alpha_{\mathbf{z}^j,\gamma}|^p - 2\left|\frac{\alpha_{\mathbf{z},\gamma} + \alpha_{\mathbf{z}^j,\gamma}}{2}\right|^p = \frac{1}{4}(\alpha_{\mathbf{z},\gamma} - \alpha_{\mathbf{z}^j,\gamma})^2 p(p-1)|c_\gamma|^{p-2}, \tag{4}$$

where $\min\{\alpha_{\mathbf{z},\gamma}, \alpha_{\mathbf{z}^j,\gamma}\} \leq c_\gamma \leq \max\{\alpha_{\mathbf{z},\gamma}, \alpha_{\mathbf{z}^j,\gamma}\}$.

Using Corollary 3.2, with $\tau(\lambda) = (B/\lambda)^{1/p}$, as noted in the beginning of this section, we can derive a bound c_γ in Eq. (4) as follows

$$|c_\gamma| \leq \max\{|\alpha_{\mathbf{z},\gamma}|, |\alpha_{\mathbf{z}^j,\gamma}|\} \leq \max\{\|\alpha_{\mathbf{z}}\|_{\ell_2}, \|\alpha_{\mathbf{z}^j}\|_{\ell_2}\} \leq \left(\frac{B}{\lambda}\right)^{1/p}.$$

Furthermore, since $1 < p \leq 2$, this gives us

$$|c_\gamma|^{p-2} \geq \left(\frac{B}{\lambda}\right)^{\frac{p-2}{p}}.$$

Substituting the bound on c_γ to Eq. (4), we obtain

$$|\alpha_{\mathbf{z},\gamma}|^p + |\alpha_{\mathbf{z}^j,\gamma}|^p - 2\left|\frac{\alpha_{\mathbf{z},\gamma} + \alpha_{\mathbf{z}^j,\gamma}}{2}\right|^p \geq \frac{1}{4}p(p-1)\left(\frac{B}{\lambda}\right)^{\frac{p-2}{p}} (\alpha_{\mathbf{z},\gamma} - \alpha_{\mathbf{z}^j,\gamma})^2.$$

Finally, summing over $\gamma \in \Gamma$ gives us

$$\|\alpha_{\mathbf{z}}\|_{\ell_p}^p + \|\alpha_{\mathbf{z}^j}\|_{\ell_p}^p - 2\left\|\frac{\alpha_{\mathbf{z}} + \alpha_{\mathbf{z}^j}}{2}\right\|_{\ell_p}^p \geq \frac{1}{4}p(p-1)\left(\frac{B}{\lambda}\right)^{\frac{p-2}{p}} \|\alpha_{\mathbf{z}} - \alpha_{\mathbf{z}^j}\|_{\ell_2}^2.$$

This completes the proof of the proposition. □

A consequence of Proposition 4.1 and Theorem 3.5 is the following corollary.

Corollary 4.2. *For $1 < p \leq 2$, ℓ_p regularization is uniformly β-stable with*

$$\beta = \frac{1}{p(p-1)}\left(\frac{B}{\lambda}\right)^{\frac{2-p}{p}} \frac{4L^2\kappa}{n\lambda}.$$

This implies that with probability at least $1-\delta$, the following generalization bound holds

$$\mathcal{E}(f_{\mathbf{z}}) \leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \frac{1}{p(p-1)}\left(\frac{B}{\lambda}\right)^{\frac{2-p}{p}}\frac{4L^2\kappa}{n\lambda} + \frac{1}{p(p-1)}\left(\frac{B}{\lambda}\right)^{\frac{2-p}{p}}\frac{8L^2\kappa}{\lambda}\sqrt{\frac{\log(1/\delta)}{2n}} + \left(L\kappa^{1/2}\left(\frac{B}{\lambda}\right)^{1/p} + B\right)\sqrt{\frac{\log(1/\delta)}{2n}},$$

and thus we conclude that ℓ_p regularization for $1 < p \leq 2$ generalizes.

4.2 Stability of Elastic-Net Regularization

Elastic-net regularization uses the penalty function

$$\mathcal{P}(\alpha) = \|\alpha\|_{\ell_1,w} + \varepsilon\|\alpha\|_{\ell_2}^2 = \sum_{\gamma\in\Gamma}(w_\gamma|\alpha_\gamma| + \varepsilon\alpha_\gamma^2),$$

for some weights $w_\gamma \geq 0$ and a parameter $\varepsilon \geq 0$. Elastic-net regularization was first introduced in [14], and then analyzed in [4]. Similar to ℓ_p regularization, elastic-net regularization can also be interpreted as an interpolation of ℓ_1 regularization, which enforces sparsity, and ℓ_2 regularization, which enforces smoothness. The coefficient ε controls the degree in which these two components interact. In this section, we assume that $\varepsilon > 0$.

Clearly, $\mathcal{P}$ is coercive, because $\mathcal{P}(\alpha) \geq \varepsilon\|\alpha\|_{\ell_2}^2$. Moreover, using Lemma 3.1, we obtain

$$\|\alpha_{\mathbf{z}}\|_{\ell_2} \leq \left(\frac{1}{\varepsilon}\big(\|\alpha_{\mathbf{z}}\|_{\ell_1,w} + \varepsilon\|\alpha_{\mathbf{z}}\|_{\ell_2}^2\big)\right)^{1/2} \leq \left(\frac{B}{\lambda\varepsilon}\right)^{1/2},$$

so that $\mathcal{P}$ satisfies Corollary 3.2 with $\tau(\lambda) = \big(B/\lambda\varepsilon\big)^{1/2}$.

Now to apply Theorem 3.5, we have the following proposition.

Proposition 4.3. *Elastic-net regularization satisfies **(P2)** with $\xi = 2$ and $C = \varepsilon/2$.*

Proof. From the convexity of the function $|\cdot|$, for every component $\alpha_{\mathbf{z},\gamma}$ and $\alpha_{\mathbf{z}^j,\gamma}$ of $\alpha_{\mathbf{z}}$ and $\alpha_{\mathbf{z}^j}$, respectively, we have

$$|\alpha_{\mathbf{z},\gamma}| + |\alpha_{\mathbf{z}^j,\gamma}| - 2\left|\frac{\alpha_{\mathbf{z},\gamma} + \alpha_{\mathbf{z}^j,\gamma}}{2}\right| \geq 0.$$

By taking the weighted sum of the inequality above over $\gamma \in \Gamma$ with weights $w_\gamma \geq 0$, we obtain

$$\|\alpha_{\mathbf{z}}\|_{\ell_1,w} + \|\alpha_{\mathbf{z}^j}\|_{\ell_1,w} - 2\left\|\frac{\alpha_{\mathbf{z}} + \alpha_{\mathbf{z}^j}}{2}\right\|_{\ell_1,w} \geq 0.$$

Now, noticing that we have the following identity,

$$\|\alpha_{\mathbf{z}}\|_{\ell_2}^2 + \|\alpha_{\mathbf{z}^j}\|_{\ell_2}^2 - 2\left\|\frac{\alpha_{\mathbf{z}} + \alpha_{\mathbf{z}^j}}{2}\right\|_{\ell_2}^2 = \frac{1}{2}\|\alpha_{\mathbf{z}} - \alpha_{\mathbf{z}^j}\|_{\ell_2}^2,$$

we conclude that

$$\mathcal{P}(\alpha_{\mathbf{z}}) + \mathcal{P}(\alpha_{\mathbf{z}^j}) - 2\mathcal{P}\left(\frac{\alpha_{\mathbf{z}} + \alpha_{\mathbf{z}^j}}{2}\right) \geq \frac{\varepsilon}{2}\|\alpha_{\mathbf{z}} - \alpha_{\mathbf{z}^j}\|_{\ell_2}^2.$$

□

We summarize the result in the following corollary.

Corollary 4.4. *Elastic-net regularization is uniformly β-stable with*

$$\beta = \frac{2L^2\kappa}{\varepsilon n\lambda}.$$

Therefore, with probability at least $1-\delta$, the following generalization bound holds

$$\mathcal{E}(f_{\mathbf{z}}) \leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \frac{2L^2\kappa}{\varepsilon n\lambda} + \left(\frac{4L^2\kappa}{\varepsilon\lambda} + L\kappa^{1/2}\left(\frac{B}{\lambda\varepsilon}\right)^{1/2} + B\right)\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Then we conclude that elastic-net regularization generalizes.

We add the following remark. In general, if the penalty function $\mathcal{P}$ is convex, but not strictly convex, then the minimization problem is ill-posed, since the solution might be neither unique nor stable. Nonetheless, we can add an extra term $\mu\|\alpha\|_{\ell_2}^2$ to the objective functional, for some $\mu > 0$, and consider a modified minimization problem with the composite penalty function $\tilde{\mathcal{P}}(\alpha) = \mathcal{P}(\alpha) + \mu\|\alpha\|_{\ell_2}^2$. Following the proof of the stability of elastic-net regularization, we observe that the composite penalty function $\tilde{\mathcal{P}}$ is strongly convex, so that our results apply. The new minimization problem is well-posed: the solution exists, is unique, and is stable. The parameter $\mu > 0$ can be adjusted to control the extent in which the extra term $\mu\|\alpha\|_{\ell_2}^2$ affects the properties of the solution. In fact, using standard variational techniques, we can show that as μ tends to 0, the solution function of the modified minimization problem tends to the solution of the original minimization problem with minimum ℓ_2-norm.

5 Generalization and Consistency

A property of a learning algorithm related to generalization is the notion of *consistency*. Consistency requires that when the size of the training data increases, the expected risk of the solution function converges in probability to the minimum risk achievable by the functions in the hypothesis space.

In our context, a regularization algorithm is parametrized by a regularization parameter λ, and therefore, we say that the algorithm is *consistent* if there is a sequence $(\lambda_n)_{n\in\mathbb{N}}$, $\lambda_n > 0$, such that $(\lambda_n)_{n\in\mathbb{N}}$ converges to 0 as n tends to $+\infty$, and for every $\epsilon > 0$,

$$\lim_{n\to+\infty} \Pr\left(\mathcal{E}(f_{\mathbf{z}}^{\lambda_n}) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) > \epsilon\right) = 0,$$

where $f_{\mathbf{z}}^{\lambda_n}$ is the solution function of the regularized empirical-risk minimization problem with regularization parameter $\lambda = \lambda_n$.

It turns out that for regularization algorithms, generalization implies consistency, as we see in the following theorem.

Theorem 5.1. *Suppose that the regularization algorithm has generalization property. Let $e_\delta(\lambda, n)$ denote the error term in the generalization bound. If the following three conditions hold for some $k_1, k_2, k_3 > 0$,*

1. *$e_\delta(\lambda, n) = \mathcal{O}(\lambda^{-k_1})$ as $\lambda \to 0$;*

2. $e_\delta(\lambda, n) = \mathcal{O}(n^{-k_2})$ *as* $n \to +\infty$*;*

3. *the bound* $\tau(\lambda)$ *in Corollary 3.2 has order of growth* $\tau(\lambda) = \mathcal{O}(\lambda^{-k_3})$ *as* $\lambda \to 0$*;*

then the regularization algorithm is consistent.

Proof. First, we define

$$f^\lambda = \arg\min_{f \in \mathcal{H}} \{\mathcal{E}(f) + \lambda \mathcal{P}(f)\}.$$

Now fix $\delta > 0$. We decompose the sample error into three components as follows,

$$\mathcal{E}(f_{\mathbf{z}}^\lambda) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) \leq (\mathcal{E}(f_{\mathbf{z}}^\lambda) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^\lambda)) + (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^\lambda) + \lambda \mathcal{P}(f_{\mathbf{z}}^\lambda) - \mathcal{E}(f^\lambda)) + \left(\mathcal{E}(f^\lambda) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)\right).$$

Since $f_{\mathbf{z}}^\lambda$ is the minimizer of the objective functional,

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^\lambda) + \lambda \mathcal{P}(f_{\mathbf{z}}^\lambda) \leq \mathcal{E}_{\mathbf{z}}(f^\lambda) + \lambda \mathcal{P}(f^\lambda),$$

and therefore, we have the following,

$$\mathcal{E}(f_{\mathbf{z}}^\lambda) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) \leq (\mathcal{E}(f_{\mathbf{z}}^\lambda) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^\lambda)) + (\mathcal{E}_{\mathbf{z}}(f^\lambda) - \mathcal{E}(f^\lambda)) + \left(\mathcal{E}(f^\lambda) + \mathcal{P}(f^\lambda) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)\right).$$

The first term in the inequality above is the generalization error, and the generalization property of the regularization algorithm implies that with probability at least $1 - \delta$, we have the bound

$$\mathcal{E}(f_{\mathbf{z}}^\lambda) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^\lambda) \leq e_\delta(n, \lambda) = \mathcal{O}\left(\frac{1}{\lambda^{k_1} n^{k_2}}\right).$$

For the second term, we first note that following the proof of Lemma 3.1, we can show that $\mathcal{P}(f^\lambda) \leq (B + \mathcal{P}(f_{\mathbf{0}}))/\lambda$, so that by Corollary 3.2, the norm of f^λ is bounded, $\|f^\lambda\|_{\mathcal{H}} \leq \tau(\lambda)$. Then we can apply Lemma 3.3 to obtain the bound $0 \leq V(f^\lambda(x), y) \leq L\kappa^{1/2}\tau(\lambda) + B$. Now, for every $t > 0$, Hoeffding's inequality gives us

$$\Pr\left(\mathcal{E}_{\mathbf{z}}(f^\lambda) - \mathcal{E}(f^\lambda) \geq t\right) \leq \exp\left(-\frac{2nt^2}{\left(L\kappa^{1/2}\tau(\lambda) + B\right)^2}\right).$$

Equivalently, for every $\delta > 0$, the following bound holds with probability at least $1 - \delta$,

$$\mathcal{E}_{\mathbf{z}}(f^\lambda) - \mathcal{E}(f^\lambda) \leq \left(L\kappa^{1/2}\tau(\lambda) + B\right)\left(\frac{\log(1/\delta)}{2n}\right)^{1/2} = \mathcal{O}\left(\frac{1}{\lambda^{k_3} n^{1/2}}\right).$$

Finally, let $A(\lambda)$ denote the third term,

$$A(\lambda) = \mathcal{E}(f^\lambda) + \lambda \mathcal{P}(f^\lambda) - \inf_{f \in \mathcal{H}} \mathcal{E}(f).$$

Then we note that $A(\lambda) \to 0$ as $\lambda \to 0$.

Combining the three components above, we have the following bound, which holds with probability at least $1 - 2\delta$,

$$\mathcal{E}(f_{\mathbf{z}}^\lambda) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) \leq \mathcal{O}\left(\frac{1}{\lambda^{k_1} n^{k_2}}\right) + \mathcal{O}\left(\frac{1}{\lambda^{k_3} n^{1/2}}\right) + A(\lambda).$$

Now choose a sequence $(\lambda_n)_{n\in\mathbb{N}}$, $\lambda_n > 0$, such that

$$\lim_{n\to+\infty} \lambda_n^{k_1} n^{k_2} = \lim_{n\to+\infty} \lambda_n^{k_3} n^{1/2} = +\infty.$$

We can do this, for example, by choosing $\lambda_n = n^{-k}$, for some $0 < k < \min\{\frac{k_2}{k_1}, \frac{1}{2k_3}\}$. Then as $n \to +\infty$, each term of the right hand side of the inequality above vanishes. Therefore, for every $\epsilon > 0$,

$$\lim_{n\to+\infty} \Pr\left(\mathcal{E}(f_{\mathbf{z}}^{\lambda_n}) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) > \epsilon\right) = 0,$$

which means that the algorithm is consistent, as desired. □

Now we can apply the computations in Section 4 to show that ℓ_p regularization for $1 < p \leq 2$ and elastic-net regularization are consistent.

Corollary 5.2. *For ℓ_p regularization, $1 < p \leq 2$, the bound $\tau(\lambda)$ has order of growth $\tau(\lambda) = \mathcal{O}(\lambda^{-1/p})$, and the error term $e_\delta(\lambda, n)$ in the generalization bound has order of growth $e_\delta(\lambda, n) = \mathcal{O}(\lambda^{-2/p} n^{-1/2})$. Therefore, with the choice $\lambda_n = n^{-p/8}$, ℓ_p regularization for $1 < p \leq 2$ is consistent.*

Corollary 5.3. *For elastic-net regularization, the bound $\tau(\lambda)$ has order of growth $\tau(\lambda) = \mathcal{O}(n^{-1/2})$, and the error term $e_\delta(\lambda, n)$ has order of growth $e_\delta(\lambda, n) = \mathcal{O}(\lambda^{-1} n^{-1/2})$. Then with the choice $\lambda_n = n^{-1/4}$, elastic-net regularization is consistent.*

6 Discussion

In this paper, we investigated the stability property of regularization algorithms in terms of the conditions on the penalty functions. We have shown that regularization algorithms whose penalty functions satisfy Eq. (3) are uniformly β-stable. In particular, if the penalty function is strongly convex over bounded domains, then the regularization algorithm is uniformly β-stable with $\beta = \mathcal{O}(n^{-1})$, and hence generalizes. The strong convexity characterization that we propose is quite practical to verify, and therefore should be considered when designing new algorithms to enforce stability and generalization. As an application of our results, we showed that ℓ_p regularization and elastic-net regularization are uniformly β-stable.

As noted in Sections 4.1 and 4.2, ℓ_p regularization and elastic-net regularization can be interpreted as interpolations of ℓ_1 regularization and ℓ_2 regularization. Interestingly, our result does not apply to ℓ_1 regularization, as the ℓ_1-norm is not strongly convex. If the strong convexity characterization that we proposed in this paper turns out to be a natural requirement, then this suggests that ℓ_1 regularization is not uniformly β-stable. In fact, a recent result in [13] indicates that ℓ_1 regularization cannot be stable. This naturally raises a question of whether the set of conditions that we propose is not only sufficient, but also necessary for uniform stability. Another question is whether ℓ_1 regularization is stable under a weaker notion of stability, for example, the $\text{CVEEE}_{\text{loo}}$ stability proposed in [9], that is still sufficient for generalization.

Acknowledgments. This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Science & Artificial Intelligence Laboratory (CSAIL). This research was sponsored by grants from DARPA (IPTO and DSO) and National Science Foundation. Additional support was provided by: Adobe, Honda Research Institute USA, King Abdullah University Science and Technology grant to B. DeVore, NEC, Sony and especially by the Eugene McDermott Foundation. Lorenzo Rosasco was also partially supported by the FIRB project LEAP RBIN04PARL and by the EU Integrated Project Health-e-Child IST-2004-027749.

References

[1] Tom M. Apostol, *Mathematical analysis*, Addison-Wesley Publishing Company, Inc., 1957.

[2] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc. **68** (1950), 337–404. MR MR0051437 (14,479c)

[3] Olivier Bousquet and André Elisseeff, *Stability and generalization*, J. Mach. Learn. Res. **2** (2002), no. 3, 499–526. MR MR1929416 (2003h:68112)

[4] Christine De Mol, Ernesto De Vito, and Lorenzo Rosasco, *Elastic-net regularization in learning theory*, CBCL Paper 273, Massachusetts Institute of Technology, Cambridge, MA, 2008.

[5] Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio, *Regularization networks and support vector machines*, Adv. Comput. Math. **13** (2000), no. 1, 1–50. MR MR1759187 (2001f:68053)

[6] Federico Girosi, Michael Jones, and Tomaso Poggio, *Regularization theory and neural networks architectures*, Neural Computation **7** (1995), 219–269.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning*, Springer-Verlag, 2001.

[8] M. V. Ĭovanovich, *A remark on strongly convex and quasiconvex functions*, Mat. Zametki **60** (1996), no. 5, 778–779. MR MR1619858

[9] Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin, *Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization*, Adv. Comput. Math. **25** (2006), no. 1-3, 161–193. MR MR2231700 (2007j:68077)

[10] Tomaso Poggio, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi, *General conditions for predictivity in learning theory*, Nature **428** (2004), 419–422.

[11] Vladimir N. Vapnik, *Statistical learning theory*, Wiley-Interscience, 1998.

[12] Grace Wahba, *Spline models for observational data*, Society for Industrial and Applied Mathematics, 1990.

[13] Huan Xu, Shie Mannor, and Constantine Caramanis, *Sparse algorithms are not stable: a no-free-lunch theorem*, Proceedings of the Allerton Conference on Communication, Control, and Computing, 2008.

[14] Hui Zou and Trevor Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B Stat. Methodol. **67** (2005), no. 2, 301–320. MR MR2137327