Hypothesis Transfer Learning with Surrogate Classification Losses: Generalization Bounds through Algorithmic Stability

Anass Aghbalou [1] **Guillaume Staerman** [2]

Abstract

Hypothesis transfer learning (HTL) contrasts domain adaptation by allowing for a previous task leverage, named the source, into a new one, the target, without requiring access to the source data. Indeed, HTL relies only on a hypothesis learnt from such source data, relieving the hurdle of expansive data storage and providing great practical benefits. Hence, HTL is highly beneficial for real-world applications relying on big data. The analysis of such a method from a theoretical perspective faces multiple challenges, particularly in classification tasks. This paper deals with this problem by studying the learning theory of HTL through algorithmic stability, an attractive theoretical framework for machine learning algorithms analysis. In particular, we are interested in the statistical behaviour of the regularized empirical risk minimizers in the case of binary classification. Our stability analysis provides learning guarantees under mild assumptions. Consequently, we derive several complexity-free generalization bounds for essential statistical quantities like the training error, the excess risk and cross-validation estimates. These refined bounds allow understanding the benefits of transfer learning and comparing the behaviour of standard losses in different scenarios, leading to valuable insights for practitioners.

1. Introduction

Traditional supervised machine learning methods share the common assumption that training data and test data are drawn from the same underlying distribution. However, this assumption is often too restrictive to hold in practice. In many real-world applications, a hypothesis is learnt and deployed in different environments that exhibit a distributional shift. A more realistic assumption is that the marginal distributions of training (*source*) and testing (*target*) domains are different but related. This is the framework of *domain adaptation* (DA), where the learner is provided little or no labeled data from the target domain but a large amount of data from the source domain. This problem arises in various real-world applications like natural language processing (Dredze et al., 2007; Ruder et al., 2019), sentiment analysis (Blitzer et al., 2007b; Liu et al., 2019), robotics (Zhang et al., 2012; Bousmalis et al., 2018) and many other areas.

Several works shed light on the theory of DA (Blitzer et al., 2007a; Mansour et al., 2009; Ben-David et al., 2010; Zhang et al., 2012; Cortes et al., 2015; Zhang et al., 2019) and suggest schemes that generally rely on minimizing some similarity distances between the source and the target domains. However, the theoretical analysis shows that a DA procedure needs many unlabeled data from both domains to be efficient. Besides, even when unlabeled data are abundant, minimizing a similarity distance can be time-consuming in many scenarios.

To tackle this practical limitation, a new framework that relies only on the source hypothesis was introduced, the so-called *hypothesis transfer learning* (HTL) (Li & Bilmes, 2007; Orabona et al., 2009; Kuzborskij & Orabona, 2013; Perrot & Habrard, 2015; Kuzborskij & Orabona, 2017; Du et al., 2017). HTL is tailored to the scenarios where the user has no direct access to the source domain nor to the relatedness between the source and target environments. As a direct consequence, HTL does not introduce any assumptions about the similarity between the source and target distributions. It has the advantage of not storing abundant source data in practice.

In this work, we analyze HTL through Regularized Empirical Risk Minimization (RERM) in the binary classification framework. Our working assumptions encompass many widely used *surrogate* losses, such as the exponential loss used by several boosting algorithms like AdaBoost (Freund & Schapire, 1997), the logistic loss, the softplus loss, which serves as a smooth approximation of the hinge loss

[1]Télécom Paris, Institut Polytechnique de Paris, LTCI, Palaiseau, France. [2]Université Paris-Saclay, Inria, CEA, ´ Palaiseau, 91120, France. Correspondence to: Anass Aghbalou <anass.aghablou@telecom-paris.fr>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

(Dugas et al., 2000), the mean squared error (MSE) and the squared hinge that represents the default losses for least squares/modified least squares algorithms (Rifkin et al., 2003). The attractive quality of these surrogate losses is that they are *classification calibrated* (Zhang, 2004; Bartlett et al., 2006). In other words, they represent a convex upper bound for the classification error and minimizing the expected risk regarding a surrogate loss yields a predictor with sound accuracy.

This paper's theoretical analysis uses the notion of *algorithmic stability*. Formally, assuming that one has access to a small labeled set, we derive many complexity-free generalisation bounds that depend only on the source hypothesis's quality. In particular, such an analysis allows us to compare the behavior of different losses in different scenarios and to answer some practical questions such as: *which surrogate loss is recommended when the source and target domains are related? Which surrogate loss is robust to heavy distribution shift?*

The notion of algorithmic stability and its consequences in learning theory has received much attention since its introduction in (Devroye & Wagner, 1979). It allows obtaining complexity-free generalization bounds for a large class of learning algorithms such as k-nearest-neighbours (Devroye & Wagner, 1979), empirical risk minimizers (Kearns & Ron, 1999), Support Vector Machine (Bousquet & Elisseeff, 2002), Bagging (Elisseeff et al., 2005), RERM (Zhang, 2004; Wibisono et al., 2009), stochastic gradient descent (Hardt et al., 2016), neural networks with a simple architecture (Charles & Papailiopoulos, 2018), to name but a few. For an exhaustive review of the different notions of *stability* and their consequences on the generalization risk of a learning algorithm, the reader is referred to (Kutin & Niyogi, 2002).

Only a few works derive theoretical guarantees for RERM in the HTL framework and are all formalized in a regression setting. A stability analysis has been provided for the HTL algorithm in the case of RLS for regression in Kuzborskij & Orabona (2013) limited to the least-squares loss. Later, Kuzborskij & Orabona (2017) considered the class of smooth losses and obtained statistical rates on the empirical risk, being a particular case of the stability guarantees. However, this smoothness assumption may be considered strong since it is not satisfied for hypotheses learnt from the exponential loss or vacuously satisfied for hypotheses learnt from the softplus loss. Besides, Du et al. (2017) proposed a novel algorithm to adapt the source hypothesis to the target domain. Nonetheless, the theoretical guarantees they derived are obtained with several strong assumptions, unverifiable in practice. The obtained bounds depend on many unknown parameters (for further details, see Section 3, where all these assumptions are explicitly listed and discussed). Other theoretical results studying HTL outside the framework of RERM can be found (Li & Bilmes, 2007; Morvant et al., 2012; Perrot & Habrard, 2015; Dhouib & Redko, 2018). However, most of these theoretical results depend on a complexity/distance measure or/and are valid on a different framework than classification. For example, Perrot & Habrard (2015) explores the notion of algorithmic stability in *metric learning* with Lipschitz loss functions to study the excess risk of some algorithms. The obtained bounds are not intuitive as they depend on the Lipschitz constant and cannot be easily extended to many usual classification losses. Furthermore, the proof techniques in the latter work are far from ours.

On the other hand, when the source is known, many theoretical guarantees can be found in the domain adaptation literature, see e.g. Mansour et al. (2009); Ben-David et al. (2010); Zhang et al. (2012); Cortes et al. (2015) and Zhang et al. (2019), among others. Their rates involve the complexity of the hypothesis class and the distance between the source and the target distribution that may be unknown in practice and drastically deteriorate the rates.

Another related subject is *meta learning*, broadly described as leveraging data from pre-existing tasks to derive algorithms or representations that yield superior results on unencountered tasks. Many theoretical works such as (Khodak et al., 2019; Balcan et al., 2019; Denevi et al., 2019) or (Denevi et al., 2020) have studied this problem. Yet, the obtained theoretical guarantees in the latter works depend on the smoothness parameters of the loss function and the regularizers. The proof techniques from the present paper can be incorporated into the proof of the latter references to obtain more sharp and intuitive learning bounds, that is, bounds exclusively depending on the quality of the source hypothesis.

Contributions In this paper, we investigate the statistical risk of the hypothesis transfer learning procedure dedicated to the binary classification task. To that end, we adopt the angle of algorithmic stability that offers an appealing theoretical framework to analyze such a method. This is the first work exploring algorithmic stability for HTL with the usual classification loss functions. In this paper, we provide a (pointwise) hypothesis stability analysis of the HTL in the classification framework for any losses satisfying mild conditions. Furthermore, we show that our main assumptions are valid for the most popular classification losses and derive their associated constants. Based on these stability results, we investigate the statistical behavior of the generalization gap and the excess risk of the HTL procedure. We provide an intuitive finite-sample analysis of these quantities and highlight the statistical behavior of common losses.

2. Background and Preliminaries

In this section, we start by recalling the framework of Hypothesis transfer learning and describe the concept of stability.

2.1. Hypothesis Transfer Learning

Considering the source and target domains, hypothesis transfer learning leverages the learnt hypothesis with the source dataset, without having access to the raw source data or any information between source and target domains, to solve a machine learning task on the target domain. Formally, we denote by $\mathcal{Z}_S$ and $\mathcal{Z}_T$ the source and target domains and assume that we have access to $n \in \mathbb{N}, n \geq 1$ i.i.d. observations $\mathcal{D}_T = Z_1, \ldots, Z_n \in \mathcal{Z}_T$ with a distribution P_T lying in the target domain and a source hypothesis h_S learnt from $m \in \mathbb{N}, m \geq 1$ i.i.d. observations $\mathcal{D}_S = Z_1^S, \ldots, Z_m^S \in \mathcal{Z}_S$ drawn from the source distribution P_S. In the HTL framework, we do not have access to the source observations but only to the resulting source hypothesis h_S. It is worth noting that $n \ll m$ in many practical scenarios. In this paper, we focus on the binary classification task. Therefore, our domains consist of a Cartesian product of a source/target covariate space $\mathcal{X}_S/\mathcal{X}_T$ and the set $\{-1, 1\}$, i.e. $\mathcal{Z}_S = \mathcal{X}_S \times \{-1, 1\}$ and $\mathcal{Z}_T = \mathcal{X}_T \times \{-1, 1\}$. In addition, we assume that $\mathcal{X}_T \subset \mathcal{X}_S \subset \mathbb{R}^d$. Consider two classes of hypotheses $\mathcal{H}_S$ and $\mathcal{H}_T$, an HTL algorithm aims to use a source hypothesis $h_S \in \mathcal{H}_S$ learnt on $\mathcal{D}_S$ to improve the performance of a classification algorithm over $\mathcal{D}_T$. Precisely, it is defined as a map

$$\begin{aligned} \mathcal{A} : (\mathcal{Z}_T)^n \times \mathcal{H}_S &\to \mathcal{H}_T \\ (\mathcal{D}_T, h_S) &\mapsto h_T. \end{aligned}$$

Throughout the paper, we assume that h_S is given and fixed, and we use the shorthand notation $\mathcal{A}(\mathcal{D}_T)$ instead of $\mathcal{A}(\mathcal{D}_T, h_S)$ for the sake of clarity.

Let $\ell : \mathcal{H}_T \times \mathcal{Z}_T \mapsto \mathbb{R}_+$ denote a loss function so that $\ell(h_T, Z)$ is the error of $h_T \in \mathcal{H}_T$ on the observation $Z = (X, Y) \in \mathcal{Z}_T$. In this work, we assume that $\ell(h_T, Z) = \phi\left(h_T(X)Y\right)$ for some non negative convex function ϕ. The generalization risk of the predictor $\mathcal{A}(D_T)$ is denoted by

$$\begin{aligned} \mathcal{R}\big[\mathcal{A}\left(\mathcal{D}_T\right)\big] &= \mathbb{E}_{Z\sim P_T}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right)\right] \\ &= \mathbb{E}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) \mid \mathcal{D}_T\right]. \end{aligned}$$

Notice that the randomness in the latter expectation stems from the novel observation Z only while the trained algorithm $\mathcal{A}(\mathcal{D}_T)$ is fixed. Its empirical counterpart, the *training* error of $\mathcal{A}\left(\mathcal{D}_T\right)$ writes as

$$\widehat{\mathcal{R}}\big[\mathcal{A}(\mathcal{D}_T)\big] = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}(\mathcal{D}_T), Z_i).$$

The latter estimate is known to be optimistic since most learning algorithms are conceived to minimize the training loss. Thus, a more reliable estimate would be the *deleted* estimate or the so-called leave-one-out (*l.o.o.*) estimate:

$$\widehat{\mathcal{R}}_{\mathrm{loo}}\big[\mathcal{A}(\mathcal{D}_T)\big] = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z_i\right), \tag{2.1}$$

where $\mathcal{D}_T^{\setminus i} = \mathcal{D}_T \setminus \{Z_i\}$ denotes the dataset $\mathcal{D}_T$ with the i'th element removed.

Remark 2.1 (ACCELERATED *l.o.o.*)**.** *At first sight, one can notice that computing the* l.o.o. *risk measure is a heavy task in practice since one needs to train the algorithm n times. However, in our case, one can use the closed form formula of the* l.o.o. *estimate for* RERM *algorithms derived in Wang et al. (2018).*

2.2. Algorithmic Stability

In this part, we briefly recall important notions of stability that will be used in the paper. The notion of *stability* was first introduced in Devroye & Wagner (1979) to derive non-asymptotic guarantees for the leave-one-out estimate. Let denote by $[n]$ the set of indices $\{1, \ldots, n\}$. The algorithm $\mathcal{A}$ is called stable if removing a training point Z_i, $i \in [n]$, from the $\mathcal{D}_T$ or replacing Z_i with an independent observation Z' drawn from the same distribution does not alter the risk of the output. Later, Bousquet & Elisseeff (2002) introduced the strongest notion of stability, namely *uniform stability*, an assumption used to derive probability upper bounds for the training error and the *l.o.o.* estimate (Bousquet & Elisseeff, 2002; Elisseeff et al., 2005; Hardt et al., 2016; Bousquet et al., 2020; Klochkov & Zhivotovskiy, 2021). Equipped with the above notations, *uniform* stability, also called *leave-one-out* stability, can be defined as follows.

Definition 2.1. The algorithm $\mathcal{A}$ is said to be $\beta(n)$-*uniformly* stable with respect to a loss function ℓ if, for any $i \in [n]$ and $Z \in \mathcal{Z}_T$, it holds:

$$\left|\ell\left(\mathcal{A}(\mathcal{D}_T), Z\right) - \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z\right)\right| \leq \beta(n).$$

In practice, uniform stability may be too restrictive since the bound above must hold for all Z, irrespective of its marginal distribution. While weaker, the following notion of stability is still enough to control the leave-one-out deviations (Devroye & Wagner, 1979; Bousquet & Elisseeff, 2002; Elisseeff et al., 2005; Kuzborskij & Orabona, 2013).

Definition 2.2. The algorithm $\mathcal{A}$ has a *hypothesis* stability $\beta(n)$ with respect to a loss function ℓ if, for any $i \in [n]$, it holds:

$$\left\|\ell\left(\mathcal{A}(\mathcal{D}_T), Z\right) - \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z\right)\right\|_1 \leq \beta(n),$$

where $\|X\|_q = (\mathbb{E}[|X|^q])^{1/q}$ is the L_q norm of X.

We now recall a direct analogue of hypothesis stability: the *pointwise hypothesis stability*. The latter property is used to derive PAC learning bounds for the training error (Bousquet & Elisseeff, 2002; Elisseeff et al., 2005; Charles & Papailiopoulos, 2018).

Definition 2.3. The algorithm $\mathcal{A}$ has a *pointwise hypothesis* stability $\gamma(n)$ with respect to a loss function ℓ if, for any $i \in [n]$, it holds:

$$\left\| \ell\left(\mathcal{A}(\mathcal{D}_T), Z_i\right) - \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z_i\right) \right\|_1 \leq \gamma(n).$$

Note that the approach based on stability does not refer to a complexity measure like the VC dimension or the Rademacher complexity. There is no need to prove uniform convergence, and the generalization error (cf. Equation 4.1 below) depends directly on the stability parameter. Our work aims to use the notion of algorithmic stability to derive sharper bounds for the HTL problem. More precisely, the magnitude of the obtained bounds is directly related to the quality of h_S on the target domain $\big($represented by $\mathcal{R}[h_S]\big)$ instead of the complexity of the hypothesis class (Ben-David et al., 2010; Zhang et al., 2012; Cortes et al., 2015; Zhang et al., 2019).

2.3. Working Framework

This paper analyses hypothesis transfer learning through regularised empirical risk minimization (RERM). In particular, it includes the popular Regularized Least Squares (RLS) with biased regularization (Orabona et al., 2009) that has been analyzed in Kuzborskij & Orabona (2013) and Kuzborskij & Orabona (2017). Formally, we consider the following algorithm $\mathcal{A}$ such that:

$$\mathcal{A}(\mathcal{D}_T, h_S) = \hat{h}(\cdot\,; \mathcal{D}_T) + h_S(\cdot), \tag{2.2}$$

where the function $\hat{h} : \mathbb{R}^d \to \mathbb{R}$ is obtained from the target set of data via the minimization problem:

$$\begin{aligned}\hat{h} &= \operatorname*{arg\,min}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \phi\left(\left(h\left(X_i\right) + h_S\left(X_i\right)\right) Y_i\right) + \lambda \|h\|_k^2 \\ &= \operatorname*{arg\,min}_{h \in \mathcal{H}} \widehat{\mathcal{R}}(h + h_S) + \lambda \|h\|_k^2, \end{aligned} \tag{2.3}$$

with the family of hypotheses $\mathcal{H}$ being a reproducing kernel Hilbert space (RKHS) endowed with a kernel k, an inner product $\langle \cdot, \cdot \rangle$ and a norm $\|\cdot\|_k$. The resulting map arising from the HTL is the sum of the source hypothesis h_S and the target hypothesis $\hat{h}$ where $\hat{h}$ is learnt involving the source map.

It is worth noting that our analysis encompasses the least square with biased regularization (Schölkopf et al., 2001; Orabona et al., 2009) commonly studied in transfer learning (Kuzborskij & Orabona, 2013; 2017), briefly recalled below.

Remark 2.2 (LINK WITH RLS)**.** *The* RLS *with biased regularization is a particular case of the proposed algorithm 2.2. Indeed, by choosing k as the linear kernel $k(x_1, x_2) = x_1^\top x_2$ and the loss $\phi(x) = (1-x)^2$, it is equivalent to*

$$\mathcal{A} = \hat{h} + h_S,$$

with $\hat{h}(x) = \hat{u}^\top x$ and

$$\hat{u} = \operatorname*{arg\,min}_{u \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(u^\top X_i + h_S(X_i) - Y_i\right)^2 + \lambda \|u\|_2^2. \tag{2.4}$$

Furthermore, if $h_S(x) = v^\top x$ is a linear classifier with $v \in \mathbb{R}^d$, then

$$\hat{u} = \operatorname*{arg\,min}_{u \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(u^\top X_i - Y_i\right)^2 + \lambda \|u - v\|_2^2,$$

which is the original form of biased regularisation algorithms (Schölkopf et al., 2001; Orabona et al., 2009). See Appendix A.1 for technical details.

3. Stability Analysis

The subsequent analysis requires technical assumptions, listed below. We assume that the source hypothesis and the kernel k are bounded, as stated in the following assumptions.

Assumption 1. *The source hypothesis is bounded on the target space:*

$$\|h_S\|_\infty = \sup_{x \in \mathcal{X}_T} |h_S(x)| < \infty.$$

Assumption 2. *The kernel k is bounded:*

$$\sup_{x_1, x_2 \in \mathcal{X}_T} k(x_1, x_2) \leq \kappa.$$

The boundness of the kernel is a common and mild assumption (see e.g. Bousquet & Elisseeff, 2002; Zhang, 2004; Wibisono et al., 2009). It is satisfied by many usual kernels like the Gaussian kernel and the sigmoid kernel. Furthermore, when $\mathcal{X}_T$ is bounded, then polynomial kernels are also bounded.

We now investigate the accuracy of the HTL proposed framework and provide general stability results under slight assumptions. Furthermore, we show that these assumptions are satisfied by most of the popular ML surrogate losses used in practice and derive precisely the associated constants involved in our theoretical results.

3.1. Hypothesis Stability

This section analyzes the hypothesis stability of general surrogate ML losses for the proposed HTL framework. To study the stability of Algorithm 2.2, we start by showing that the solution of the optimization problem 2.3 lies in the sphere with a data-driven radius, as stated in the following lemma.

Lemma 3.1. *Suppose that Assumptions 1 and 2 are satisfied. Then the solution of Equation* (2.3) *lies in the set* $\{h \in \mathcal{H},\ \|h\|_\infty \leq \hat{r}_\lambda\}$ *with*

$$\hat{r}_\lambda = \kappa\sqrt{\alpha\widehat{\mathcal{R}}\left[h_S\right]},$$

where $\alpha = \kappa/\lambda$.

Proof. The proof is postponed in the Appendix B.1. □

This lemma ensures that the norm of the solution of the optimisation problem 2.3 decreases when the quality of h_S increases. In the rest of the paper, for a given index $i \in [n]$, we denote by $\hat{r}^i_\lambda = \kappa\sqrt{\alpha\widehat{\mathcal{R}}^{\setminus i}\left[h_S\right]}$, $\widehat{\mathcal{R}}^{\setminus i}$ the training error with the i'th sample removed and $\hat{\rho}^i_\lambda = \max\left(\hat{r}_\lambda, \hat{r}^i_\lambda\right)$.

Before stating our main theorem, we first require an additional assumption involving the empirical radius obtained in Lemma 3.1.

Assumption 3. *The function* ϕ *is differentiable and convex. Furthermore,* $\forall i \in [n]$*, it holds:*

$$\mathbb{E}\left[\sup_{|y'|,|y|\leq\hat{\rho}^i_\lambda} |\phi'(h_S(X')Y' + y')\phi'(h_S(X)Y + y)|\right] \leq \Psi_1\left(\mathcal{R}\left[h_S\right]\right),$$

where $Z = (X, Y)$*,* $Z' = (X', Y')$ *are two samples drawn from* P_T *independent of* $\mathcal{D}^{\setminus i}$ *and* Ψ_1 *is a decreasing function verifying* $\Psi_1(0) = 0$*.*

The bound stated in the theorem below reveals the generalisation properties of the presented HTL procedure through the stability framework.

Proposition 3.1. *Suppose that Assumptions 1, 2 and 3 are satisfied. Then the algorithm* $\mathcal{A}$ *(cf. Equation* (2.2)*) is hypothesis stable with parameter*

$$\beta(n) = \frac{\alpha\left(\Psi_1\left(\mathcal{R}\left[h_S\right]\right) \wedge \|\phi'\|^2_\infty\right)}{n}.$$

Proof. The proof is postponed to the Appendix B.2. □

We obtain a stability rate of order $\mathcal{O}\left(\frac{\Psi_1(\mathcal{R}[h_S])\alpha}{n}\right)$ for any losses satisfying Assumption 3. It naturally depends on the risk of the source classifier, where the expectation is taken on the target data distribution. Therefore, the source task directly influences the rate of the HTL classifier. The standard stability rate of RERM without transfer learning (without source) is of order $\mathcal{O}(\alpha/n)$, see Theorem 4.3 in Zhang (2004) or Theorem 3.5 in Wibisono et al. (2009). A relevant source hypothesis allows us to obtain faster rates than in standard RERM. Thus, one can directly notice the benefits of using a *good* source hypothesis on the stability of RERM. The negative transfer, i.e. the source hypothesis has a negative effect and deteriorates the target learner, is analyzed and discussed in Section 4.1.

Remark 3.1 (RELATED WORK)**.** *The only existing result studying hypothesis stability in HTL is in Kuzborskij & Orabona (2013). However, the analysis is only in a regression framework with the mean squared error loss. The proof techniques in Kuzborskij & Orabona (2013) rely heavily on the closed-form formulas of the ordinary least square estimate, which does not hold in a general setting like ours. Furthermore, we obtain equivalent (up to constants) stability rates as in Kuzborskij & Orabona (2013). More details are given in Section 3.3 where we explicit constants* Ψ_1 *for most of popular losses.*

Existing assumptions in DA and HTL literature Statistical guarantees obtained in these fields generally assume that the loss function verifies a smoothness condition. For example, in Mansour et al. (2009) and Cortes et al. (2015), their analysis supposes that ℓ verifies the triangle inequality, which holds only for the MSE and squared hinge. Moreover, the obtained upper bounds in these works depend on the complexity of $\mathcal{H}$ and some discrepancy distances between the source and target distributions P_S and P_T, which deteriorates the statistical rates. In Kuzborskij & Orabona (2017), they suppose that the derivative of the loss is Lipschitz which is not the case for the exponential. Furthermore, even if the loss satisfies this smoothness assumption, their constants depend heavily on the smoothness parameter, and it would yield vacuous bounds in many practical situations. For example, the softplus function $\psi_s(x) = s\log(1+e^{\frac{1-x}{s}})$ with small values of s serves as an approximation of the hinge loss $\max(0, 1-x)$ and is $1/s$ Lipschitz. This function converges to the Hinge loss when $s \to 0$ and usual choices of s are usually close to 0. Therefore, the Lipschitz constant of the derivative $1/s$ verifies $1/s \gg 1$, and the bounds from Kuzborskij & Orabona (2017) become vacuous. Besides, Du et al. (2017) made several assumptions about the true regression function of both the source and target domains. To clarify, by the true regression function, f, we refer to the actual model denoted by $Y = f(X)$. However, these assumptions are challenging to empirically confirm due to their reliance on the real source and target distributions, which generally remain unknown. Moreover, the theoretical

guarantees achieved depend on several constants, also derived from the true distribution, that makes quantifying the bounds magnitude a complex task.

To our best knowledge, the vast majority of existing theoretical results from the HTL literature have similar assumptions to those discussed above. However, in this work, our assumptions are flexible: we only require the differentiability of the loss and a *local* majorant of the derivative, which will make the analysis more flexible and more suited for the usual classification losses.

To understand the intuition behind Assumption 3 notice that, when $\mathcal{R}[h_S] \to 0$, $\phi(h_S(X)Y)$ approaches the minimum then $\phi'(h_S(X)Y)$ approaches 0 (in expectation). Thus, the function Ψ_1 can be seen as a function that dictates the rate of convergence of the derivative to 0 as h_S approaches the optimal hypothesis. One must note that the latter assumption is verified for many loss functions, namely any loss satisfying the following inequality $|\phi'(x)| \leq \Psi(\phi(x))$ for some concave loss function Ψ. The function Ψ effectively mediates between ϕ and ϕ'. As an example, in the context of Mean Squared Error (MSE) loss, it is straightforwardly observable that $|\phi'(x)| \leq \sqrt{\phi(x)}$. Thus ϕ' is directly linked to $\phi(x)$ via the square root function.

Remark 3.2 (SCORE SCALING)**.** RERM *for regression (cf. Equation 2.4) is equivalent to fitting a predictor on the residuals $Y_i - h_S(X_i)$. However, in the classification case, if we follow the standard approach that $h_S : \mathcal{X} \mapsto \mathcal{Y} = \{-1, 1\}$ is a binary classifier (Mansour et al., 2009; Cortes et al., 2015), then latter residuals are either* 1 *or* 0*. Thus, this won't provide enough information for many losses to improve the training. To see this, see the example of the logistic loss and notice that $\phi(1) = \log(1 + e^{-1})$ and $\phi(-1) = \log(1 + e^{1})$. Therefore, in the best case scenario, $\mathcal{R}[h_S] = \log(1+e^{-1})$, which is far from the minimum (that is zero). To tackle this problem, we suggest taking the score learned on the source, which is more informative, especially when the loss function used to train the algorithm on the source has the same minimum as the loss used to train on the target. Note that one can also think of transforming the score, for example, if ϕ is the logistic loss $\phi(x) = \log(1 + e^{-x})$ and $h_S \in]-1, 1[$ we can use an increasing transformation function to an interval $]-C, C[$ with $C >> 1$ in order to adapt to the target loss which is nearly 0 for large values x.*

3.2. Pointwise Hypothesis Stability

To go further than the widely used hypothesis stability, we analyze our HTL problem through the angle of pointwise hypothesis stability. Results presented in this part will be the cornerstone of those shown in Section 4. To analyze the pointwise hypothesis stability of Algorithm 2.2, we require a direct analogue of Assumption 3, involving the data-driven radius provided in Lemma 3.1.

Assumption 4. *The function ϕ is differentiable and convex. Furthermore, $\forall i \in [n]$, it holds:*

$$\mathbb{E}\left[\sup_{|y'|,|y| \leq \hat{\rho}^i_\lambda} |\phi'(h_S(X)Y + y')\phi'(h_S(X)Y + y)|\right] \leq \Psi_2\left(\mathcal{R}\left[h_S\right]\right).$$

where $Z = (X, Y)$ is a sample drawn from P_T independent of $\mathcal{D}^{\setminus i}$ and Ψ_2 is a decreasing function verifying $\Psi_2(0) = 0$.

Under the latter assumption, the following proposition is obtained in a similar manner to Proposition 3.1.

Proposition 3.2. *Suppose that Assumptions 1, 2 and 4 are satisfied. Then the algorithm $\mathcal{A}$ (cf. Equation* (2.2)*) is pointwise hypothesis stable with parameter*

$$\gamma(n) = \frac{\alpha\left(\Psi_2\left(\mathcal{R}\left[h_S\right]\right) \wedge \|\phi'\|^2_\infty\right)}{n}.$$

Proof. The proof is postponed to the Appendix B.3. □

Again, this result shows the benefits of using a good hypothesis on the pointwise hypothesis stability of RERM. This stability result, combined with that of Proposition 3.1, can be leveraged to propose new convergence results on the generalisation gap and the excess risk of this HTL problem for a wide class of losses, as shown in Section 4. In the sequel, we explicitly compute the functions Ψ_1 and Ψ_2 for many widely used classification losses.

3.3. Deriving Constants for Popular Losses

As the results of Propositions 3.1 and 3.2 are general and stated for any losses satisfying Assumptions 3 and 4, it is the purpose of this part to investigate our results with widespread machine learning losses. To that end, we first show that these Assumptions are satisfied for the most popular losses. Second, we derive constants involved in these two statistical rates. In particular, we focus on the five following losses:

- Exponential: $\phi(x) = e^{-x}$.
- Logistic: $\phi(x) = \log\left(1 + e^{-x}\right)$.
- Mean Squared Error: $\phi(x) = (1 - x)^2$.
- Squared Hinge: $\phi(x) = \max(0, 1 - x)^2$.
- Softplus: $\phi_s(x) = s \log\left(1 + e^{\frac{1-x}{s}}\right)$, for some $s > 0$.

In the next proposition, we show that most of classical losses verifies Assumptions 3, 4 and we detail their associated functions Ψ_1 and Ψ_2.

Loss	$\Psi_1(x)$	$\Psi_2(x)$
Sq. hinge	$8x(4\alpha+1)$	$8x(4\alpha+1)$
MSE	$8x(4\alpha+1)$	$8x(4\alpha+1)$
Exponential	$C_S x^2 e^{2\alpha x}$	$M_S C_S x e^{2\alpha x}$
Logistic	$C_S e^{2\alpha x}(e^{\sqrt{x}}-1)^2$	$C_S e^{2\alpha x}(e^{\sqrt{x}}-1)$
Softplus	$C_S e^{2\alpha x}(e^{\sqrt{\frac{x}{s}}}-1)^2$	$C_S e^{2\alpha x}(e^{\sqrt{\frac{x}{s}}}-1)$

Table 1. Examples of losses verifying Assumptions 3, 4 and their corresponding functions. The constants M_S and C_S are given by $M_S = \sup_{z\in\mathcal{Z}_T} \ell(h_S, z)$, $C_S = \exp\left\{2 + \frac{2\alpha M_S}{n} + \frac{4\alpha^2 M_S^2}{n-1}\right\}$.

Proposition 3.3. *The exponential, logistic, squared hinge,* MSE *and softplus losses satisfy Assumptions 3 and 4 with corresponding functions Ψ_1 and Ψ_2 listed in Table 1.*

Proof. The proof is postponed to the Appendix B.4. □

This result shows that bounds derived in Propositions 3.1 and 3.2 are therefore valid under mild assumptions. Indeed, our results only require the kernel and the source hypothesis to be bounded, classical in the HTL framework. Thus, we obtain the first stability result in HTL without limiting assumptions, which remains valid in a practical setting.

As shown in Table 1, functions Ψ_1 and Ψ_2 are linear for the square hinge and the MSE losses. Besides, for the softplus and logistic losses, we have $\|\phi'\|_\infty = 1$ and their stability parameters capped by α/n. Thus, the impact of an irrelevant source hypothesis h_S with large $\mathcal{R}[h_S]$ remains negligible on the stability of RERM when using these losses. In contrast, for the exponential loss, the functions Ψ_1 and Ψ_2 are roughly exponential, and the corresponding convergence rate deteriorates quickly as $\mathcal{R}[h_S]$ increases. This is indeed not surprising since a prediction in the wrong direction ($\text{sign}(h_S(X)) \neq Y$) would increase the loss $e^{-h_S(X)Y}$ exponentially fast. In the particular case of the MSE, we obtain the same stability rate $\mathcal{O}\left(\frac{\alpha\mathcal{R}[h_S]}{n}\right)$ as in the regression framework (Kuzborskij & Orabona, 2013). In the next section, we shall discuss the implications of these stability rates on the *generalization gap* (Hardt et al., 2016; Charles & Papailiopoulos, 2018), cross-validation schemes and the excess risk of Algorithm 2.2.

4. Generalisation Guarantees for HTL with Surrogate Losses

In this part, we leverage the stability results provided in Section 3 in several statistical errors commonly used.

4.1. Generalization Gap

Here we investigate the accuracy of the algorithm $\mathcal{A}$ through the generalization gap. Precisely, this gap is defined as the expected error between the empirical risk and the theoretical risk of the algorithm $\mathcal{A}$:

$$\mathcal{E}_{\text{gen}} = \left|\mathbb{E}\left[\widehat{\mathcal{R}}\left[\mathcal{A}(\mathcal{D}_T)\right] - \mathcal{R}\left[\mathcal{A}(\mathcal{D}_T)\right]\right]\right|.$$

To discuss the impact of h_S on the generalization gap, it suffices to analyse the stability parameters $\beta(n)$ and $\gamma(n)$. Indeed, $\mathcal{E}_{\text{gen}}$ is directly linked to these quantities, as stated in the following theorem.

Theorem 4.1. *Suppose that $\mathcal{A}$ has a hypothesis stability $\beta(n)$ and a pointwise hypothesis stability $\gamma(n)$. Then, it holds:*

$$\mathcal{E}_{gen} \leq \beta(n) + \gamma(n).$$

Furthermore, suppose that Assumptions 1, 2, 3 and 4 are satisfied. Thus, $\beta(n)$ and $\gamma(n)$ are given by Propositions 3.1 and 3.2 and the generalization gap of $\mathcal{A}$ (cf. Equation (2.2)*) is upper-bounded as:*

$$\mathcal{E}_{gen} \leq \alpha \frac{\left(\Psi_1\left(\mathcal{R}\left[h_S\right]\right) + \Psi_2\left(\mathcal{R}\left[h_S\right]\right)\right) \wedge \left(2\left\|\phi'\right\|_\infty^2\right)}{n}.$$

Proof. The proof is postponed to the Appendix B.5. □

When the source hypothesis is relevant, the risk $\mathcal{R}[h_S]$ is close to zero so that $e^{\mathcal{R}[h_S]} - 1 \approx \mathcal{R}[h_S]$ and $e^{\alpha\mathcal{R}[h_S]} \approx 1$. Equipped with Table 1, this theorem yields the following upper bounds for $\mathcal{E}_{\text{gen}}$:

- MSE, Sq. hinge: $\mathcal{E}_{\text{gen}} = \mathcal{O}\left(\frac{\alpha\mathcal{R}[h_S]}{n}\right)$.
- Logistic: $\mathcal{E}_{\text{gen}} = \mathcal{O}\left(\alpha\frac{\sqrt{\mathcal{R}[h_S]}\wedge 2}{n}\right)$.
- Softplus: $\mathcal{E}_{\text{gen}} = \mathcal{O}\left(\alpha\frac{\left(\sqrt{\mathcal{R}[h_S]/s}\right)\wedge 2}{n}\right)$.
- Exponential: $\mathcal{E}_{\text{gen}} = \mathcal{O}\left(\frac{\alpha M_S\mathcal{R}[h_S]}{n}\right)$.

Thus, if $\mathcal{R}[h_S]$ is small, the exponential, the squared hinge and the MSE losses have the fastest generalization gap rate. Therefore, our analysis suggests that the user should privilege using the latter losses if one disposes of a good hypothesis h_S.

Negative learning The phenomenon of negative transfer occurs when the hypothesis h_S learned from the source domain has a detrimental effect on the target learner. In such a case, training without using h_S on the target domain would yield a better learner. We refer the reader to

Weiss et al. (2016) and Wang et al. (2019) for further details about this topic. For the softplus and the logistic losses, the generalization gap remains bounded by $\mathcal{O}(\alpha/n)$ even if $\mathcal{R}[h_S] \to \infty$. As a consequence, Algorithm 2.2 with the sofplus and logistic losses is robust to negative learning since the generalization gap still achieves the same rate of convergence $\mathcal{O}(\alpha/n)$ as a standard RERM algorithm with no source information *i.e.* $h_S = 0$ (see *e.g.* Zhang, 2004; Wibisono et al., 2009). Finally, we must highlight that one should avoid using the exponential loss when the source and target domains are unrelated due to the presence of the term $e^{\alpha\mathcal{R}[h_S]}$ in the corresponding upper bound.

Remark 4.1 (CROSS VALIDATION PROCEDURES)**.** *The notion of stability has many attractive qualities. In particular, it yields complexity-free bounds for cross-validation methods. (see* e.g. *Bousquet & Elisseeff, 2002; Kumar et al., 2013; Celisse & Mary-Huard, 2018). For example, one can easily show that*

$$\mathbb{E}\left[\left|\widehat{\mathcal{R}}_{\text{loo}}\left[\mathcal{A}\left(\mathcal{D}_T\right)\right] - \mathcal{R}\left[\mathcal{A}\left(\mathcal{D}_T\right)\right]\right|\right] \leq \beta(n).$$

Proposition 3.1 shows that the quality of risk estimation with l.o.o. *depends directly on the quality of the source predictor h_S. Note that the same conclusion holds for model selection with* l.o.o. *cross-validation: Given a family of source hypotheses, the quality of the model selection procedure depends directly on the quality of the provided learners independently of the complexity of $\mathcal{H}_T$. Besides, using the same proof techniques, we can show that Algorithm 2.2 is L_2 stable with stability parameter depending on $\Psi\left(\mathcal{R}\left[h_S\right]\right)$. L_2 stability is similar to hypothesis stability, where the L_1 moment is replaced by the L_2 moment in Definition 2.2. The latter notion allows obtaining theoretical guarantees regarding the K-fold and the* l.o.o. *schemes. It also derives asymptotic confidence intervals for cross-validation procedures in risk estimation and model selection (Bayle et al., 2020; Austern & Zhou, 2020). In our particular case, Proposition 3.1 implies that the tightness of the confidence intervals of cross-validation methods depends only on the quality of h_S.*

4.2. Excess Risk

In this section we analyse the excess risk of Algorithm 2.2 defined as:

$$\mathcal{E}_{\text{ex}} = \mathbb{E}\left[\mathcal{R}\left[\mathcal{A}\right] - \mathcal{R}\left[h^* + h_S\right]\right],$$

where $h^* = \arg\min_{h\in\mathcal{H}} \mathcal{R}\left[h_S + h\right]$. To this end, we start by showing that $\mathcal{E}_{\text{ex}}$ depends on the upper bounds of the *(pointwise) hypothesis stability* and the regularization parameter λ. Further, we derive precise finite-sample rates for the surrogate losses introduced in Section 3.3.

Theorem 4.2. *Suppose that $\|h^*\|_k < \infty$. Then, the excess risk of algorithm 2.2 verifies,*

$$\mathcal{E}_{ex} \leq \gamma(n) + \beta(n) + \lambda \left\|h^*\right\|_k^2.$$

Making λ varying with the sample size n, we obtain various consistent bounds for different losses. In the sequel, we assume that $\kappa \leq 1$ and $M_S \leq 1$ to avoid notional burden. When ϕ is either the MSE *or the squared hinge and $\lambda = \sqrt{\frac{\mathcal{R}[h_S]}{\sqrt{n}}}$, it holds:*

$$\mathcal{E}_{ex} \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{R}[h_S]}{\sqrt{n}}}\right).$$

Furthermore, if ϕ is the exponential loss and $n \geq \frac{M_S^2 \ln(n)^2}{\mathcal{R}[h_S]}$, picking $\lambda = 4\frac{\sqrt{\mathcal{R}[h_S]} \wedge 1}{\ln(n)}$ yields:

$$\mathcal{E}_{ex} \leq \mathcal{O}\left(\frac{\sqrt{\mathcal{R}[h_S]} \wedge 1}{\ln(n)}\right),$$

otherwise picking $\lambda = \frac{\ln(n)^2}{\sqrt{n}}$ gives:

$$\mathcal{E}_{ex} \leq \mathcal{O}\left(\frac{\ln(n)^2}{\sqrt{n}}\right).$$

Suppose that the function ϕ is the logistic loss or the softplus. Then the choice $\lambda = \frac{1}{\sqrt{n}}$ yields:

$$\mathcal{E}_{ex} \leq \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

In particular, Theorem 4.2 yields the consistency of RERM. Furthermore, the Remark 4.1 regarding the generalization gap still holds for the excess risk. First, when $\mathcal{R}[h_S]$ is small, Algorithm 2.3 with MSE or squared hinge would have the fastest convergence rate. Second, when $\mathcal{R}[h_S]$ is large compared to the sample size n, then the safest option is to use the logistic or the softplus losses with $\lambda = \frac{1}{\sqrt{n}}$. Note that, if $\mathcal{R}[h_S]$ is small an improved convergence rate $\left(1/\sqrt{-n\ln\left(\mathcal{R}[h_s]\right)}\right)$ can be achieved for the latter losses (see Appendix B.6 for further details). Finally, Algorithm 2.2 with the exponential loss is likely to suffer from negative learning. Indeed, if $\mathcal{R}[h_S]$ is large, one needs a large amount of data to ensure the non-triviality of the rate $\mathcal{R}[h_S]/\ln(n)$. It is worth noting that the rate of convergence with the exponential loss is naturally logarithmic even without a source hypothesis; see, for instance, Corollary 4.1 and Theorem 4.4 in Zhang (2004). To conclude, using a good source hypothesis improves convergence rates of RERM compared to those derived without transfer (Zhang, 2004).

Remark 4.2 (ON THE UNIVERSAL CONSISTENCY)**.** *If we assume that the kernel k is non-polynomial, h_S is continuous and the distribution of $X \in \mathcal{X}_T$ is* regular *(see* e.g. *Definition 4.2 in Zhang, 2004). Then, one can use any* universal approximation theorem *(see for instance Theorem 4.1 in Zhang, 2004) to obtain*

$$h^* = \underset{h \in \mathcal{H}}{\arg\min}\, \mathcal{R}\left[h_S + h\right] = \underset{h \in \mathcal{L}(\mathcal{X}_T, \mathbb{R})}{\arg\min}\, \mathcal{R}\left[h_S + h\right],$$

where $\mathcal{L}(\mathcal{X}_T, \mathbb{R})$ is the space of real-valued functions defined on $\mathcal{X}_T$. The universal consistency of $\mathcal{A}$ follows immediately from Theorem 4.2. Further, all the losses presented in this paper are classification calibrated *(Bartlett et al., 2006) meaning that:*

$$\underset{h \in \mathcal{L}(\mathcal{X}_T, \mathbb{R})}{\arg\min}\, \mathcal{R}\left[h\right] = \underset{h \in \mathcal{L}(\mathcal{X}_T, \mathbb{R})}{\arg\min}\, \mathcal{R}^{0\text{-}1}\left[h\right],$$

where $\mathcal{R}^{0\text{-}1}\left[h\right] = P_T(\operatorname{sign}\left(h(X)\right) \neq Y)$ is the usual classification accuracy. Thus, minimizing the excess risk would likely yield a classifier with good accuracy.

5. Numerical experiments

We illustrate our analysis by providing some results using simulated data that aim to underscore the robustness of each loss to negative learning scenarios. The experiment is conducted as follows. A source domain is considered with random variables $(X_S, Y_S) \in \mathbb{R}^2 \times \{-1, 1\}$, where the positive and negative classes are respectively drawn from two multivariate t-distributions $\mathcal{T}((r, 0), 3I_2, 2.5)$ and $\mathcal{T}((-r, 0), 3I_2, 2.5)$. We train a linear classifier h_S on a source dataset of size 10000 using the SVM algorithm.

To emphasize the impact of negative learning on each loss, we generate a smaller target dataset of size 100. The distributions for positive and negative classes are given by $\mathcal{T}(((r+d)cos(\theta), (r+d)sin(\theta)), 3I_2, 2.5)$ and $\mathcal{T}((-(r+d)cos(\theta), -(r+d)sin(\theta)), 3I_2, 2.5)$, respectively. For different values of θ, the target risk $\mathcal{R}\left[\hat{h} + h_s\right]$ of the analyzed RERM algorithm (with $\lambda = 1$) trained on the small size dataset is estimated using a test set of size 10000.

It is important to note that when $\theta = 0$, it corresponds to the scenario of positive learning since the decision boundaries of both domains are similar. On the other hand, the case where $\theta = \pi$ corresponds to negative learning since the true decision functions of the source and the target domain are pointing to opposite directions.

Figure 1 presents the median true risk of the HTL algorithm (cf. Equation 2.3) as a function of θ for $(r, d) = (5, 5)$ computed over 1000 simulations. The parameter s of the softplus loss is set to 0.1. Consistent with our theoretical analysis, the softplus and logistic functions exhibit significant robustness to negative transfer.

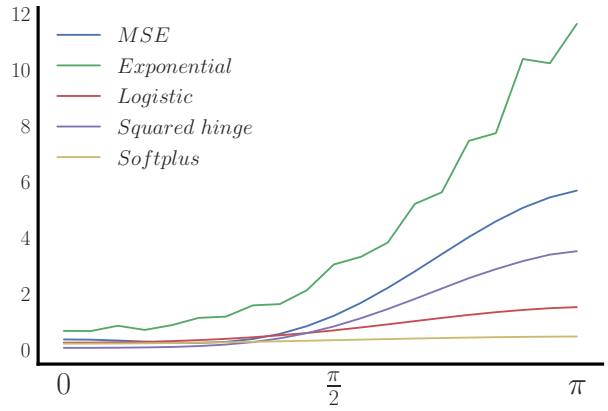

Figure 1. Target risk of Algorithm 2.2 as a function of θ.

6. Conclusion

In this paper, we study hypothesis transfer learning through the angle of Algorithmic Stability. Following the work of Kuzborskij & Orabona (2013), where hypothesis stability is shown for the MSE in the regression setting, we derive similar hypothesis stability rates in classification with general losses under slight assumptions. Furthermore, we show that our assumptions are satisfied for the most popular machine learning losses, making our work valuable for practitioners. Moreover, we leverage our stability results to provide finite-sample analysis on the generalization gap and the excess risk. We show that HTL framework is efficient and explicit (fast) rates for these popular losses. Our theoretical analysis will help practitioners better understand the benefits of HTL and give insight into the loss choices.

The proposed work is general and may fit with many other domains. Future work may involve our analysis for different Machine Learning tasks where transfer learning procedures can be beneficial such as robust learning (Shafahi et al., 2020; Laforgue et al., 2021; Staerman et al., 2021a), anomaly detection (Andrews et al., 2016; Chandola et al., 2009; Staerman et al., 2020; 2022a), speech (Campi et al., 2021; 2023), automatic language generation (Staerman et al., 2021b; Golovanov et al., 2019), knowledge distillation (Cho & Hariharan, 2019), events-based modelling (Staerman et al., 2022b), fairness (Colombo et al., 2022b) or general neural-networks based tasks (Colombo et al., 2022a; Picot et al., 2023; Darrin et al., 2023).

References

Andrews, J., Tanay, T., Morton, E. J., and Griffin, L. D. Transfer representation-learning for anomaly detection.

JMLR, 2016.

Austern, M. and Zhou, W. Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111*, 2020.

Balcan, M.-F., Khodak, M., and Talwalkar, A. Provable guarantees for gradient-based meta-learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 424–433. PMLR, 09–15 Jun 2019.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Bayle, P., Bayle, A., Janson, L., and Mackey, L. Cross-validation confidence intervals for test error. *arXiv preprint arXiv:2007.12671*, 2020.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20, 2007a.

Blitzer, J., Dredze, M., and Pereira, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 440–447, 2007b.

Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., Downs, L., Ibarz, J., Pastor, P., Konolige, K., et al. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 4243–4250. IEEE, 2018.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Bousquet, O., Klochkov, Y., and Zhivotovskiy, N. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pp. 610–626. PMLR, 2020.

Campi, M., Peters, G. W., Azzaoui, N., and Matsui, T. Machine learning mitigants for speech based cyber risk. *IEEE Access*, 9:136831–136860, 2021.

Campi, M., Peters, G. W., and Toczydlowska, D. Ataxic speech disorders and parkinson's disease diagnostics via stochastic embedding of empirical mode decomposition. *Plos one*, 18(4):e0284667, 2023.

Celisse, A. and Mary-Huard, T. Theoretical analysis of cross-validation for estimating the risk of the k-nearest neighbor classifier. *The Journal of Machine Learning Research*, 19(1):2373–2426, 2018.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3): 1–58, 2009.

Charles, Z. and Papailiopoulos, D. Stability and generalization of learning algorithms that converge to global optima. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 745–754. PMLR, 10–15 Jul 2018.

Cho, J. H. and Hariharan, B. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4794–4802, 2019.

Colombo, P., Dadalto, E., Staerman, G., Noiry, N., and Piantanida, P. Beyond mahalanobis distance for textual ood detection. *Advances in Neural Information Processing Systems*, 35:17744–17759, 2022a.

Colombo, P., Staerman, G., Noiry, N., and Piantanida, P. Learning disentangled textual representations via statistical measures of similarity. *arXiv preprint arXiv:2205.03589*, 2022b.

Cortes, C., Mohri, M., and Muñoz Medina, A. Adaptation algorithm and theory based on generalized discrepancy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 169–178, 2015.

Darrin, M., Staerman, G., Gomes, E. D. C., Cheung, J. C., Piantanida, P., and Colombo, P. Unsupervised layer-wise score aggregation for textual ood detection. *arXiv preprint arXiv:2302.09852*, 2023.

Denevi, G., Ciliberto, C., Grazzi, R., and Pontil, M. Learning-to-learn stochastic gradient descent with biased regularization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1566–1575. PMLR, 09–15 Jun 2019.

Denevi, G., Pontil, M., and Ciliberto, C. The advantage of conditional meta-learning for biased regularization and fine tuning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 964–974. Curran Associates, Inc., 2020.

Devroye, L. and Wagner, T. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979. doi: 10.1109/TIT.1979.1056087.

Dhouib, S. and Redko, I. Revisiting similarity learning for domain adaptation. *Advances in Neural Information Processing Systems*, 31, 2018.

Dredze, M., Blitzer, J., Talukdar, P. P., Ganchev, K., Graça, J. V., and Pereira, F. Frustratingly hard domain adaptation for dependency parsing. 2007.

Du, S. S., Koushik, J., Singh, A., and Póczos, B. Hypothesis transfer learning via transformation functions. *Advances in neural information processing systems*, 30, 2017.

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., and Garcia, R. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 13, 2000.

Elisseeff, A., Evgeniou, T., and Pontil, M. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(3):55–79, 2005.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Golovanov, S., Kurbanov, R., Nikolenko, S., Truskovskyi, K., Tselousov, A., and Wolf, T. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6053–6058, 2019.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Kearns, M. and Ron, D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation*, 11(6):1427–1453, 1999.

Khodak, M., Balcan, M.-F. F., and Talwalkar, A. S. Adaptive gradient-based meta-learning methods. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Klochkov, Y. and Zhivotovskiy, N. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34: 5065–5076, 2021.

Kumar, R., Lokshtanov, D., Vassilvitskii, S., and Vattani, A. Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning*, pp. 27–35. PMLR, 2013.

Kutin, S. and Niyogi, P. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pp. 275–282, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608974.

Kuzborskij, I. and Orabona, F. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pp. 942–950. PMLR, 2013.

Kuzborskij, I. and Orabona, F. Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106(2): 171–195, 2017.

Laforgue, P., Staerman, G., and Clémençon, S. Generalization bounds in the presence of outliers: a median-of-means study. In *International Conference on Machine Learning*, pp. 5937–5947. PMLR, 2021.

Li, X. and Bilmes, J. A bayesian divergence prior for classiffier adaptation. In *Artificial Intelligence and Statistics*, pp. 275–282. PMLR, 2007.

Liu, R., Shi, Y., Ji, C., and Jia, M. A survey of sentiment analysis based on transfer learning. *IEEE Access*, 7: 85401–85412, 2019.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Morvant, E., Habrard, A., and Ayache, S. Parsimonious unsupervised and semi-supervised domain adaptation with good similarity functions. *Knowledge and Information Systems*, 33(2):309–349, 2012.

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Orabona, F., Castellini, C., Caputo, B., Fiorilla, A. E., and Sandini, G. Model adaptation with least-squares svm for adaptive hand prosthetics. In *2009 IEEE International Conference on Robotics and Automation*, pp. 2897–2903. IEEE, 2009.

Perrot, M. and Habrard, A. A theoretical analysis of metric hypothesis transfer learning. In *International Conference on Machine Learning*, pp. 1708–1717. PMLR, 2015.

Picot, M., Granese, F., Staerman, G., Romanelli, M., Messina, F., Piantanida, P., and Colombo, P. A halfspace-mass depth-based method for adversarial attack detection. *Transactions on Machine Learning Research*, 2023.

Rifkin, R., Yeo, G., Poggio, T., et al. Regularized least-squares classification. *Nato Science Series Sub Series III Computer and Systems Sciences*, 190:131–154, 2003.

Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pp. 15–18, 2019.

Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In *International conference on computational learning theory*, pp. 416–426. Springer, 2001.

Shafahi, A., Saadatpanah, P., Zhu, C., Ghiasi, A., Studer, C., Jacobs, D., and Goldstein, T. Adversarially robust transfer learning. In *8th International Conference on Learning Representations (ICLR 2020)(virtual)*. International Conference on Learning Representations, 2020.

Staerman, G., Mozharovskyi, P., Clémen, S., et al. The area of the convex hull of sampled curves: a robust functional statistical depth measure. In *International Conference on Artificial Intelligence and Statistics*, pp. 570–579. PMLR, 2020.

Staerman, G., Laforgue, P., Mozharovskyi, P., and d'Alché Buc, F. When ot meets mom: Robust estimation of wasserstein distance. In *International Conference on Artificial Intelligence and Statistics*, pp. 136–144. PMLR, 2021a.

Staerman, G., Mozharovskyi, P., Colombo, P., Clémençon, S., and d'Alché Buc, F. A pseudo-metric between probability distributions based on depth-trimmed regions. *arXiv preprint arXiv:2103.12711*, 2021b.

Staerman, G., Adjakossa, E., Mozharovskyi, P., Hofer, V., Sen Gupta, J., and Clémençon, S. Functional anomaly detection: a benchmark study. *International Journal of Data Science and Analytics*, pp. 1–17, 2022a.

Staerman, G., Allain, C., Gramfort, A., and Moreau, T. Fadin: Fast discretized inference for hawkes processes with general parametric kernels. *arXiv preprint arXiv:2210.04635*, 2022b.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Wang, S., Zhou, W., Lu, H., Maleki, A., and Mirrokni, V. Approximate leave-one-out for fast parameter tuning in high dimensions. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5228–5237. PMLR, 10–15 Jul 2018.

Wang, Z., Dai, Z., Póczos, B., and Carbonell, J. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11293–11302, 2019.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Wibisono, A., Rosasco, L., and Poggio, T. Sufficient conditions for uniform stability of regularization algorithms. *Computer Science and Artificial Intelligence Laboratory Technical Report, MIT-CSAIL-TR-2009-060*, 2009.

Zhang, C., Zhang, L., and Ye, J. Generalization bounds for domain adaptation. *Advances in neural information processing systems*, 25, 2012.

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Zhang, Y., Liu, T., Long, M., and Jordan, M. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pp. 7404–7413. PMLR, 2019.

A. Preliminary Results

In this section, we show some useful technical lemmas used in the subsequent proofs.

Lemma A.1. *Suppose that X, Y, Z are three mutually independent random variables such that $E(X) = E(Y)$. Then it holds:*

$$\mathbb{E}\left[(X+Z)(Y+Z)\right] \leq 2\left(\mathbb{E}\left[X\right]^2 + \mathbb{E}\left[Z^2\right]\right).$$

Proof. Since X, Y, Z are mutually independent one has the following identities,

$$\begin{aligned}\mathbb{E}\left[(X+Z)(Y+Z)\right] &= \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] + \mathbb{E}\left[X\right]\mathbb{E}\left[Z\right] + \mathbb{E}\left[Z\right]\mathbb{E}\left[Y\right] + \mathbb{E}\left[Z^2\right] \\ &= \mathbb{E}\left[X\right]^2 + 2\mathbb{E}\left[X\right]\mathbb{E}\left[Z\right] + \mathbb{E}\left[Z^2\right].\end{aligned}$$

Now, noticing that $\left(E[Z]^2 \leq E[Z^2]\right)$ we get:

$$\begin{aligned}\mathbb{E}\left[X\right]^2 + 2\mathbb{E}\left[X\right]\mathbb{E}\left[Z\right] + \mathbb{E}\left[Z^2\right] &\leq 2\mathbb{E}\left[X\right]^2 + \mathbb{E}\left[Z\right]^2 + \mathbb{E}\left[Z^2\right] \\ &\leq 2\left(\mathbb{E}\left[X\right]^2 + \mathbb{E}\left[Z^2\right]\right),\end{aligned}$$

which is the desired result. □

In the sequel, we shall provide an upper bound for the exponential of $\hat{\tau}_\lambda^i$ defined as:

$$\hat{\tau}_\lambda^i = \sqrt{\alpha\left(\widehat{\mathcal{R}}^{\setminus i}[h_S] + \frac{M_S}{n}\right)}, \tag{A.1}$$

with $M_S = \sup_{z \in \mathcal{Z}_T} \ell(h_S, z)$ and

$$\widehat{\mathcal{R}}^{\setminus i}[h] = \frac{1}{n-1}\sum_{j \neq i} \ell(h, Z_j), \tag{A.2}$$

the training error of a hypothesis h with the i'th datum removed. The quantity $\hat{\tau}_\lambda^i$ will serve as an upper bound of $\hat{\rho}_\lambda^i = \max(\hat{r}_\lambda, \hat{r}_\lambda^i)$ independent of the observation $Z_i \in \mathcal{D}_T$. Indeed, by definition:

$$\hat{\tau}_\lambda^i \geq \hat{r}_\lambda^i = \sqrt{\alpha\left(\widehat{\mathcal{R}}^{\setminus i}[h_S]\right)}.$$

Moreover, it holds:

$$\widehat{\mathcal{R}}\left[h\right] \leq \widehat{\mathcal{R}}^{\setminus i}\left[h\right] + \frac{\ell(h, Z_i)}{n} \leq \widehat{\mathcal{R}}^{\setminus i}\left[h\right] + \frac{M_S}{n},$$

so that $\hat{\tau}_\lambda^i \geq \hat{r}_\lambda^i$. Thus, we have $\hat{\tau}_\lambda^i \geq \hat{\rho}_\lambda^i$.

Lemma A.2. *Let $W_1, W_2, \ldots, W_n$ be a sequence of $i.i.d.$ random variables bounded by $C > 0$. Then one has*

$$\mathbb{E}\left[e^{\hat{\mu}}\right] \leq e^{\mu + \frac{C^2}{n}},$$

where $\mu = \mathbb{E}\left[W_1\right]$ and $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n W_i$.

Proof. The proof follows in two steps. First, we apply Hoeffding's inequality to obtain:

$$\mathbb{P}\left(|\hat{\mu} - \mu| \geq t\right) \leq e^{\frac{-nt^2}{C^2}}.$$

Second, applying Theorem 2.5.2 in Vershynin (2018) yields:

$$\mathbb{E}\left[e^{\hat{\mu} - \mu}\right] \leq e^{\frac{C^2}{n}},$$

which leads to the desired result. □

Lemma A.3. *For all $i \in [n]$ and $p \in \mathbb{N}$, the quantity $e^{\hat{\tau}_\lambda^i}$ verifies:*

$$\mathbb{E}\left[e^{p\hat{\tau}_\lambda^i}\right] \leq e^{p+\frac{\alpha p M_S}{n}+\frac{\alpha^2 p^2 M_S^2}{n-1}} e^{p\alpha\mathcal{R}[g]}.$$

Proof. First, using the fact that $\sqrt{x} \leq x+1$, one has:

$$e^{\hat{\tau}_\lambda^i} \leq e^{p\alpha\widehat{\mathcal{R}}^{\setminus i}[h_S]+\alpha\frac{pM_S}{n}+p}.$$

Since $p\alpha\widehat{\mathcal{R}}^{\setminus i}[h_S] = \frac{1}{n-1}\sum_{i\neq j} p\alpha\ell(h, Z_i)$, applying Lemma A.2 with $W_i = p\alpha\ell(h, Z_i)$ and $C = p\alpha M_S$ yields the desired result.

□

To prove Propositions 3.1 and 3.2, we extend Theorem 4.3 in Zhang (2004), that gives an upper bound for standard RERM to the HTL framework. This extension leads to the next lemma.

Lemma A.4. *The leave one out deviations of the algorithm $\mathcal{A}$ (cf. Equation* (2.2)*) verifies:*

$$\left\|\mathcal{A}\left(\mathcal{D}_T\right) - \mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right)\right\|_k \leq \frac{k\left(X_i, X_i\right)^{1/2}\left|\phi'\left(\mathcal{A}\left(\mathcal{D}_T, X_i\right)Y_i\right)\right|}{\lambda n}.$$

Proof. Since ϕ is convex, the Bregman divergence of ϕ is non negative. More precisely,

$$d_\phi(x, y) = \phi(x) - \phi(y) - (x-y)\phi'(y) \geq 0,$$

so that, for any $Z_i = (X_i, Y_i) \in \mathcal{D}_T$ one has:

$$\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z_i\right) - d_\phi\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}, X_i\right)Y_i, \mathcal{A}\left(\mathcal{D}_T, X_i\right)Y_i\right) \leq \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z_i\right),$$

where $\mathcal{A}\left(\mathcal{D}_T, X_i\right)$ is the prediction of the input X_i by the algorithm $\mathcal{A}$. Also, the term on the left side in the above inequality can be written as follows:

$$\begin{aligned}\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z_i\right) - & d_\phi\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}, X_i\right)Y_i, \mathcal{A}\left(\mathcal{D}_T, X_i\right)Y_i\right) = \ell\left(\mathcal{A}(\mathcal{D}_T), Z_i\right) \\ & + \phi'\left(\mathcal{A}(\mathcal{D}_T, X_i)Y_i\right)\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}, X_i\right) - \mathcal{A}\left(\mathcal{D}_T, X_i\right)\right)Y_i,\end{aligned}$$

so that:

$$\ell\left(\mathcal{A}(\mathcal{D}_T), Z_i\right) + \phi'\left(\mathcal{A}(\mathcal{D}_T, X_i)Y_i\right)\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}, X_i\right) - \mathcal{A}\left(\mathcal{D}_T, X_i\right)\right)Y_i \leq \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z_i\right).$$

Thus, we get:

$$\widehat{\mathcal{R}}^{\setminus i}\left[\mathcal{A}\left(\mathcal{D}_T\right)\right] + S_i \leq \widehat{\mathcal{R}}^{\setminus i}\left[\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right)\right], \tag{A.3}$$

where $\widehat{\mathcal{R}}^{\setminus i}$ is defined previously in Equation (A.2) and $S_i = \frac{1}{n}\sum_{j\neq i}\phi'\left(\mathcal{A}(\mathcal{D}_T, X_j)Y_j\right)\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}, X_j\right) - \mathcal{A}(\mathcal{D}_T, X_j)\right)$.

Let $\hat{h}^{\setminus i}$ denote the solution of the optimization problem 2.3 with the i'th datum removed. One gets by definition of $\mathcal{A}$ (cf. Equation (2.2)),

$$\widehat{\mathcal{R}}^{\setminus i}\left[\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right)\right] + \lambda\|\hat{h}^{\setminus i}\|_k^2 \leq \widehat{\mathcal{R}}^{\setminus i}\left[\mathcal{A}\left(\mathcal{D}_T\right)\right] + \lambda\|\hat{h}\|_k^2.$$

Using (A.3), it yields:

$$\begin{aligned}S_i &\leq \lambda\left(\left\|\hat{h}\right\|_k^2 - \left\|\hat{h}^{\setminus i}\right\|_k^2\right) \\ &\leq -\lambda\|\hat{h} - \hat{h}^{\setminus i}\|_k^2 - 2\lambda\langle\hat{h}, \hat{h}^{\setminus i} - \hat{h}\rangle,\end{aligned}$$

where the second line follows from $\|x\| - \|y\| = \|x-y\|^2 + 2\langle x-y, y\rangle$. Reverting the inequality leads to:

$$\begin{aligned}\lambda\|\hat{h} - \hat{h}^{\setminus i}\|_k^2 &\leq -\frac{1}{n_T}\sum_{j\in T^{\setminus i}} \phi'\left(\mathcal{A}(\mathcal{D}_T, X_j)Y_j\right)\langle \hat{h}^{\setminus i} - \hat{h}, k\left(X_i, \cdot\right)\rangle - 2\lambda\langle \hat{h}, \hat{h}^{\setminus i} - \hat{h}\rangle \\ &\leq \left\|\frac{1}{n_T}\sum_{j\in T^{\setminus i}} \phi'\left(\mathcal{A}(\mathcal{D}_T, X_j)Y_j\right) k\left(X_i, \cdot\right) + 2\lambda g\right\|_k \|\hat{h}^{\setminus i} - \hat{h}\|_k. \end{aligned} \tag{A.4}$$

The last inequalities hold because of the definition of S_i:

$$\begin{aligned} S_i &= \frac{1}{n}\sum_{j\neq i} \phi'\left(\mathcal{A}(\mathcal{D}_T, X_j)Y_j\right)\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}, X_j\right) - \mathcal{A}(\mathcal{D}_T, X_j)\right) \\ &= \frac{1}{n}\sum_{j\neq i} \phi'\left(\mathcal{A}(\mathcal{D}_T, X_j)Y_j\right)\left(\hat{h}^{\setminus i}\left(X_j\right) - \hat{h}\left(X_j\right)\right) \\ &= \frac{1}{n}\sum_{j\neq i} \phi'\left(\mathcal{A}(\mathcal{D}_T, X_j)Y_j\right)\langle \hat{h}^{\setminus i} - \hat{h}, k(X_j, \cdot)\rangle. \end{aligned}$$

On the other hand, since $A(\mathcal{D}_T, X_j) = h_S(X_j) + \langle \hat{h}, k(X_j, \cdot)\rangle$ and by Theorem 3.1.20 in Nesterov et al. (2018), we know that the following optimality condition holds:

$$\frac{1}{n}\sum_{j=1}^{n} \phi'\left(\mathcal{A}(\mathcal{D}_T, X_j)Y_j\right) k\left(X_j, \cdot\right) + 2\lambda\hat{h} = 0.$$

Therefore Inequality (A.4) becomes:

$$\lambda\|\hat{h} - \hat{h}^{\setminus i}\|^2 \leq \left\|\frac{1}{n}\phi'\left(\mathcal{A}(T, X_i)Y_i\right)\right\|_k \|k(X_i, \cdot)\|_k \|\hat{h}^{\setminus i} - \hat{h}\|_k.$$

it remains to remind that $\|k\left(X_i, \cdot\right)\|^2 = k\left(X_i, X_i\right)$ and $\left\|\mathcal{A}\left(\mathcal{D}_T\right) - \mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right)\right\|_k = \left\|\hat{h}^{\setminus i} - \hat{h}\right\|_k$ to complete the proof. □

Before highlighting the link between Algorithm 2.2 with RLS, let's remind a useful lemma (representer theorem) that allows simplifying the optimization problem 2.3 in practice.

Lemma A.5. *The learning rule $\hat{h}$ (cf. Equation 2.3) lies in the linear span in $\mathcal{H}$ of the vectors $\left(k\left(X_i, \cdot\right)\right)_{1\leq i\leq n}$,* i.e.

$$\hat{h} \in \mathcal{H}_{\mathcal{D}},$$

with $\mathcal{H}_{\mathcal{D}} = \{\sum_1^n \alpha_i k\left(X_i, \cdot\right) \mid \alpha_1, \ldots, \alpha_n \in \mathbb{R}\}$.

Proof. Since $\mathcal{H}_{\mathcal{D}}$ is a finite dimensionnal subspace of $\mathcal{H}$, any $h \in \mathcal{H}$ can be decomposed as:

$$h = h_{\mathcal{D}} + h^{\perp},$$

with $h_{\mathcal{D}} \in \mathcal{H}_{\mathcal{D}}$ and $h^{\perp} \perp \mathcal{H}_{\mathcal{D}}$. Furthermore using the fact that $h(x) = \langle h, k(x, \cdot)\rangle_k$, for all $i \in [n]$, one obtains:

$$h(X_i) = \langle h, k(X_i, \cdot)Y_i\rangle = \langle h_{\mathcal{D}}, k(X_i, \cdot)Y_i\rangle = h_{\mathcal{D}}(X_i)Y_i.$$

Thus, for any $Z_i \in \mathcal{D}_T$, it holds:

$$\ell(h + h_S, Z_i) = \phi\left(\left(h(X_i) + h_S(X_i)\right)Y_i\right) = \phi\left(\left(h_{\mathcal{D}}\left(X_i\right) + h_S\left(X_i\right)\right)Y_i\right) = \ell(h_{\mathcal{D}} + h_S, Z_i),$$

which gives

$$\widehat{\mathcal{R}}(h + h_S) = \widehat{\mathcal{R}}(h_{\mathcal{D}} + h_S).$$

On the other hand, by the Pythagorean theorem,

$$\|h_{\mathcal{D}}\|_k^2 \leq \|h\|_k^2,$$

and

$$\widehat{\mathcal{R}}(h + h_S) + \lambda\|h_{\mathcal{D}}\|_k^2 \leq \widehat{\mathcal{R}}(h_{\mathcal{D}} + h_S) + \lambda\|h_{\mathcal{D}}\|_k^2.$$

Thus, the solution of the minimization problem 2.3 must lie in $\mathcal{H}_{\mathcal{D}}$. □

A.1. Link with Least Squares with Biased Regularisation

To begin, it is a well know fact that, when the kernel k is linear then the RKHS space consists of the set of linear classifiers:

$$\mathcal{H} = \left\{h(x) = u^\top x \mid u \in \mathbb{R}^d\right\}.$$

In this case, the solution of the optimization problem with the mean square loss $\ell(h, Z) = (1 - h(X)Y)^2$, writes as $\hat{h} = \hat{u}^\top x$ with

$$\begin{aligned}
\hat{u} &= \underset{u\in\mathbb{R}^d}{\arg\min}\, \frac{1}{n}\sum_{i=1}^n \left(u^\top X_i Y_i + h_S(X_i)Y_i - 1\right)^2 + \lambda\|u\|_2^2 \\
&= \underset{u\in\mathbb{R}^d}{\arg\min}\, \frac{1}{n}\sum_{i=1}^n Y_i^2 \left(u^\top X_i + h_S(X_i) - \frac{1}{Y_i}\right)^2 + \lambda\|u\|_2^2 \\
&= \underset{u\in\mathbb{R}^d}{\arg\min}\, \frac{1}{n}\sum_{i=1}^n \left(u^\top X_i + h_S(X_i) - Y_i\right)^2 + \lambda\|u\|_2^2,
\end{aligned}$$

where the last inequality follows from the facts that $Y_i^2 = 1$ and $\frac{1}{Y_i} = Y_i$. Furthermore, if $h_S(x) = v^T x$ for some $v \in \mathbb{R}^d$ one has:

$$\begin{aligned}
\hat{u} &= \underset{u\in\mathbb{R}^d}{\arg\min}\, \frac{1}{n}\sum_{i=1}^n \left((u+v)^\top X_i - Y_i\right)^2 + \lambda\|u\|_2^2 \\
&= \underset{u\in\mathbb{R}^d}{\arg\min}\, \frac{1}{n}\sum_{i=1}^n \left(u^\top X_i - Y_i\right)^2 + \lambda\|u - v\|_2^2.
\end{aligned}$$

This is the original form of biased regularisation algorithms.

B. Technical Proofs of the Main Results

Before starting the proof of our main results, we remind two properties of RKHS spaces that are:

$$\forall x, y \in \mathcal{X}_T\ , \quad \langle k(y,\cdot), k(x,\cdot)\rangle = k(x,y),$$

and

$$\forall h \in \mathcal{H}\ ,\ \forall x \in \mathcal{X}_T\ , \quad h(x) = \langle h, k(x,\cdot)\rangle.$$

Under Assumption 2, using Cauchy Schwartz-inequality yields:

$$\forall h \in \mathcal{H}\ , \quad \|h\|_\infty \leq \sqrt{\kappa}\,\|h\|_k\,.$$

B.1. Proof of Lemma 3.1

This lemma follows from our assumptions and a simple fact. Indeed, notice that by definition of $\hat{h}$

$$\widehat{\mathcal{R}}(\hat{h} + h_S) + \lambda\|\hat{h}\|^2 \leq \widehat{\mathcal{R}}(\mathbf{0} + h_S).$$

Furthermore, $\widehat{\mathcal{R}}(h_S + \hat{h})$ is non-negative since ϕ is non-negative which concludes the proof.

B.2. Proof of Proposition 3.1

Let $Z = (X, Y) \in \mathcal{Z}_T$ and remind that, by definition of $\mathcal{A}$, one has:

$$\left|\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right| = \left|\phi\left(\left(\hat{h}(X) + h_S(X)\right)Y\right) - \phi\left(\left(\hat{h}^{\setminus i}(X) + h_S(X)\right)Y\right)\right|,$$

where $\hat{h}$ is the solution of the optimization problem 2.3. Moreover, since ϕ is differentiable, one can apply the mean value theory to obtain:

$$\begin{aligned}\left|\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right| &= \left|\phi'\left(\left(y_{\mathcal{D}} + h_S(X)\right)Y\right)\right| \left|\hat{h}(X) - \hat{h}^{\setminus i}(X)\right| \\ &\leq \sqrt{\kappa}\left|\phi'\left(\left(y_{\mathcal{D}} + h_S(X)\right)Y\right)\right| \left\|\hat{h} - \hat{h}^{\setminus i}\right\|_k \\ &= \sqrt{\kappa}\left|\phi'\left(\left(y_{\mathcal{D}} + h_S(X)\right)Y\right)\right| \left\|\mathcal{A}(\mathcal{D}_T) - \mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right)\right\|_k,\end{aligned}$$

for some $|y_{\mathcal{D}}| \leq \max\left(\hat{h}(X), \hat{h}^{\setminus i}(X)\right)$. By Lemma 3.1, we have $|y_{\mathcal{D}}| \leq \hat{\rho}_\lambda^i = \max\left(\hat{r}_\lambda, \hat{r}_\lambda^{\setminus i}\right)$. Now, Using Theorem A.4 with Assumption 2 yields:

$$\left|\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right| \leq \kappa \frac{\left|\phi'\left(\left(y_{\mathcal{D}} + h_S(X)\right)Y\right)\phi'\left(\left(\hat{h}(X_i) + h_S(X_i)\right)Y_i\right)\right|}{\lambda n}, \tag{B.1}$$

which gives using the fact that $\left\|\hat{h}\right\|_\infty \leq \hat{r}_\lambda \leq \hat{\rho}_\lambda$:

$$\left|\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right| \leq \sup_{|y|,|y'| \leq \hat{\rho}_\lambda^i} \frac{\alpha\left|\phi'(h_S(X_i)Y_i + y)\phi'(h_S(X)Y + y')\right|}{n}, \tag{B.2}$$

with $\alpha = \frac{\kappa}{\lambda}$. Now, by taking the expectation and using the fact that ϕ verifies assumption 3, Inequality (B.2) becomes:

$$\mathbb{E}\left[\left|\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right|\right] \leq \alpha \frac{\Psi_1(\mathcal{R}\left[h_S\right])}{n}.$$

Besides, notice that by Equation (B.1),

$$\forall \mathcal{D}_T \in \mathcal{Z}_T^n,\ \forall Z \in \mathcal{Z}_T\ ,\ \left|\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right| \leq \alpha \frac{\left\|\phi'\right\|_\infty^2}{n}.$$

It remains to take the expectation to complete the proof.

B.3. Proof of Proposition 3.2

The proof is similar to the previous one thus we will only give the key step: replace $Z = (X, Y)$ by $Z_i = (X_i, Y_i)$ in Equation (B.2) to obtain:

$$\left|\ell\left(\mathcal{A}\left(\mathcal{D}\right), Z_i\right) - \ell\left(\mathcal{A}\left(\mathcal{D}^{\setminus i}\right), Z_i\right)\right| \leq \sup_{|y|,|y'| \leq \hat{\rho}_\lambda^i} \frac{\alpha\left|\phi'(h_S(X_i)Y_i + y)\phi'(h_S(X_i)Y_i + y')\right|}{n}.$$

To conclude the proof, take the expectation of both sides of the last inequality and use the Assumption 4.

B.4. proof of Proposition 3.3

First, let $i \in [n]$ and $|y|, |y'| \leq \hat{\rho}_\lambda^i$. Furthermore let $Z = (X, Y)$ and $Z = (X', Y')$ be two observations independent of $\mathcal{D}^{\setminus i}$. We start by showing that the MSE and squared hinge verify Assumptions 3, 4 and explicit their corresponding function

Ψ_1, Ψ_2 . To do so, remind that:

$$\begin{aligned}\left(\hat{\rho}_\lambda^i\right)^2 &= \max(\hat{r}_\lambda^i, \hat{r}_\lambda)^2 \\ &\leq (\hat{r}_\lambda^i + \hat{r}_\lambda)^2 \\ &\leq 2\left(\hat{r}_\lambda^i\right)^2 + 2\left(\hat{r}_\lambda\right)^2 \\ &= 2\alpha\left(\widehat{\mathcal{R}}[h_S] + \widehat{\mathcal{R}}^{\setminus i}[h_S]\right). \end{aligned} \tag{B.3}$$

B.4.1. MSE

Recall the MSE loss $\phi(x) = (1-x)^2$. For all $x \in \mathbb{R}$, one has:

$$\begin{aligned}|\phi'(x+y)| &= 2\,|1-x-y| \\ &\leq 2\,|1-x| + 2\,|y| \\ &\leq 2\sqrt{\phi(x)} + 2\hat{\rho}_\lambda^i. \end{aligned} \tag{B.4}$$

Thus,

$$\sup_{|y'|,|y|\leq\hat{\rho}_\lambda^i} |\phi'(h_S(X')Y' + y')\phi'(h_S(X)Y + y)| \leq 4\left(\sqrt{\phi\left(h_S(X')Y'\right)} + \hat{\rho}_\lambda^i\right)\left(\sqrt{\phi\left(h_S(X)Y\right)} + \hat{\rho}_\lambda^i\right).$$

Taking the expectation of the latter inequality and using Lemma A.1 with $X = \sqrt{\phi\left(h_S(X')Y'\right)}$, $Y = \sqrt{\phi\left(h_S(X)Y\right)}$ and $Z = \hat{\rho}_\lambda^i$ yields:

$$\begin{aligned}\mathbb{E}\left[\sup_{|y'|,|y|\leq\hat{\rho}_\lambda^i} |\phi'(h_S(X')Y' + y')\phi'(h_S(X)Y + y)|\right] &\leq 8\left(\mathbb{E}\left[\sqrt{\phi\left(h_S(X)Y\right)}\right]^2 + \mathbb{E}\left[\left(\hat{\rho}_\lambda^i\right)^2\right]\right) \\ \text{(by Jensen's Inequality)} &\leq 8\left(\mathbb{E}\left[\phi\left(h_S(X)Y\right)\right] + \mathbb{E}\left[\left(\hat{\rho}_\lambda^i\right)^2\right]\right) \\ \left(\phi\left(h_S(X)Y\right) = \ell(h_S, Z)\right) &\leq 8\left(\mathcal{R}[h_S] + \mathbb{E}\left[\left(\hat{\rho}_\lambda^i\right)^2\right]\right) \\ \text{(Inequality (B.3))} &\leq 8\left(\mathcal{R}[h_S] + 4\alpha\mathcal{R}[h_S]\right). \end{aligned}$$

This means that the MSE verifies Assumption 3 with $\Psi_1(x) = 8x(1+4\alpha)$. Now using Inequality (B.4) again yields:

$$\sup_{|y'|,|y|\leq\hat{\rho}_\lambda^i} |\phi'(h_S(X)Y + y')\phi'(h_S(X)Y + y)| \leq 4\left(\sqrt{\phi\left(h_S(X)Y\right)} + \hat{\rho}_\lambda^i\right)^2.$$

By taking the expectation and mimicking the previous step one can show that the MSE verifies Assumption 4 with $\Psi_2(x) = 8x(1+4\alpha)$.

B.4.2. SQUARED HINGE

First recall the loss function $\phi(x) = \max\left(0, 1-x\right)^2$. By simple calculation we obtain:

$$|\phi'(x+y)| = 2\max\left(0, 1-x-y\right).$$

On the other hand, one has:

$$\begin{cases} 0 \leq \max(0, 1-x) + |y|\,, \\ 1-x-y \leq \max(0, 1-x) + |y|\,. \end{cases}$$

Thus, it holds:

$$\begin{aligned}|\phi'(x+y)| &\leq 2\max(0, 1-x) + 2\,|y| \\ &\leq 2\sqrt{\phi(x)} + 2\hat{\rho}_\lambda^i. \end{aligned}$$

The result follows using the same steps as in the MSE case.

B.4.3. EXPONENTIAL

Recalling the loss function $\phi(x) = e^{-x}$, first notice that the exponential loss verifies:

$$|\phi'(x+y)| = e^{-x}e^{-y} = \phi(x)e^{-y} \leq \phi(x)e^{\hat{\rho}^i_\lambda} \leq \phi(x)e^{\hat{\tau}^i_\lambda}, \tag{B.5}$$

where $\hat{\tau}^i_\lambda$ is given by Equation (A.1). Thus, we get:

$$\begin{aligned}
\mathbb{E}\left[\sup_{|y'|,|y|\leq\hat{\rho}^i_\lambda} |\phi'(h_S(X')Y'+y')\phi'(h_S(X)Y+y)|\right] &\leq \mathbb{E}\left[\phi(h_S(X')Y')\phi(h_S(X)Y)e^{2\hat{\tau}^i_\lambda}\right] \\
(Z \perp\!\!\!\perp Z' \perp\!\!\!\perp \hat{\tau}^i_\lambda) &\leq \mathcal{R}[h_S]^2\mathbb{E}\left[e^{2\hat{\tau}^i_\lambda}\right]. \\
\text{(By Lemma A.3 with } p=2) &\leq \mathcal{R}[h_S]^2 e^{2+\frac{2\alpha M_S}{n}+\frac{4\alpha^2 M_S^2}{n-1}} e^{2\alpha\mathcal{R}[g]}
\end{aligned}$$

Thus the exponential loss verifies Assumption 3 with $\Psi_1(x) = C_S x^2 e^{2\alpha x}$ and $C_S = e^{2+\frac{2\alpha M_S}{n}+\frac{4\alpha^2 M_S^2}{n-1}}$. Besides, using (B.5) again yields:

$$\begin{aligned}
\mathbb{E}\left[\sup_{|y'|,|y|\leq\hat{\rho}^i_\lambda} |\phi'(h_S(X)Y+y')\phi'(h_S(X)Y+y)|\right] &\leq \mathbb{E}\left[\phi(h_S(X)Y)^2 e^{2\hat{\tau}^i_\lambda}\right] \\
(M_S = \sup_{Z\in\mathcal{Z}_T} \ell(h_S, Z)) &\leq M_S\mathbb{E}\left[\phi(h_S(X)Y)e^{2\hat{\tau}^i_\lambda}\right] \\
(Z \perp\!\!\!\perp \hat{\tau}^i_\lambda) &\leq M_S\mathcal{R}[h_S]\mathbb{E}\left[e^{2\hat{\tau}^i_\lambda}\right] \\
\text{(By Lemma A.3 with } p=2) &\leq M_S\mathcal{R}[h_S] e^{2+\frac{2\alpha M_S}{n}+\frac{4\alpha^2 M_S^2}{n-1}} e^{2\alpha\mathcal{R}[g]}.
\end{aligned}$$

Therefore the exponential loss verifies Assumption 4 with $\Psi_2(x) = C_S M_S x e^{2\alpha x}$.

B.4.4. LOGISTIC

Recall the loss function $\phi(x) = \log(1+e^{-x})$ and its derivative:

$$|\phi'(x)| = \frac{e^{-x}}{e^{-x}+1}.$$

Thus, we have:

$$\begin{aligned}
|\phi'(x+y)| &= \frac{e^{-x-y}}{e^{-x-y}+1} \\
&\leq e^{-y}e^{-x} \\
&= e^{-y}\left(e^{\phi(x)}-1\right) \\
&\leq e^{\hat{\rho}^i_\lambda}\left(e^{\phi(x)}-1\right) \\
&\leq e^{\hat{\tau}^i_\lambda}\left(e^{\phi(x)}-1\right),
\end{aligned}$$

where the two last inequalities result from the facts that $y \leq \hat{\rho}^i_\lambda$ and $\hat{\rho}^i_\lambda \leq \hat{\tau}^i_\lambda$ respectively. Using the facts that $\|\phi'\|_\infty \leq 1$

and $e^{\hat{\tau}^i_\lambda} \leq 1$, one obtains:

$$\begin{aligned}
|\phi'(h_S(X)Y + y)| &\leq \min\left(e^{\hat{\tau}^i_\lambda}\left(e^{\phi(h_S(X)Y)} - 1\right), 1\right) \\
&= \min\left(e^{\hat{\tau}^i_\lambda}\left(e^{\ell(h_S,Z)} - 1\right), 1\right) \\
&\leq \min\left(e^{\hat{\tau}^i_\lambda}\left(e^{\ell(h_S,Z)} - 1\right), e^{\hat{\tau}^i_\lambda}\right) \\
&\leq e^{\hat{\tau}^i_\lambda} \min\left(\left(e^{\ell(h_S,Z)} - 1\right), 1\right). \qquad \text{(B.6)}
\end{aligned}$$

The latter inequality yields:

$$\sup_{|y'|,|y| \leq \hat{\rho}^i_\lambda} |\phi'(h_S(X')Y' + y')\phi'(h_S(X)Y + y)| \leq e^{2\hat{\tau}^i_\lambda} \min\left(e^{\ell(h_S,Z)} - 1, 1\right) \min\left(e^{\ell\left(h_S,Z'\right)} - 1, 1\right).$$

Thus, since Z, Z' are independent of $\mathcal{D}_T^{\setminus i}$, they are also independent of $\hat{\tau}^i_\lambda$. It follows:

$$\begin{aligned}
\mathbb{E}\left[\sup_{|y'|,|y| \leq \hat{\rho}^i_\lambda} |\phi'(h_S(X')Y' + y')\phi'(h_S(X)Y + y)|\right] &\leq \mathbb{E}\left[e^{2\hat{\tau}^i_\lambda}\right] \mathbb{E}\left[\min\left(e^{\ell(h_S,Z)} - 1, 1\right)\right]^2 \\
\text{(By Lemma A.3)} &\leq C_S e^{2\alpha\mathcal{R}[h_S]} \mathbb{E}\left[\min\left(e^{\ell(h_S,Z)} - 1, 1\right)\right]^2. \qquad \text{(B.7)}
\end{aligned}$$

Now using the fact that:

$$e^{\ell(h_S,Z)} - 1 \leq 1 \implies \ell(h_S, Z) \leq 1 \implies \ell(h_S, Z) \leq \sqrt{\ell(h_S, Z)},$$

we have:

$$\mathbb{E}\left[\min\left(e^{\ell(h_S,Z)} - 1, 1\right)\right] \leq \mathbb{E}\left[\min\left(e^{\sqrt{\ell(h_S,Z)}} - 1, 1\right)\right].$$

In addition, notice that:

$$(e^{\sqrt{x}} - 1) \wedge 1 = \begin{cases} e^{\sqrt{x}} - 1 & \text{if } x \leq \ln(2)^2, \\ 1 & \text{otherwise,} \end{cases}$$

which is concave. Therefore, it holds:

$$\mathbb{E}\left[\min\left(e^{\ell(h_S,Z)} - 1, 1\right)\right] \leq \min\left(e^{\sqrt{\mathcal{R}[h_S]}} - 1, 1\right).$$

To show that the logistic loss verifies Assumption 3 with $\Psi_1(x) = C_S e^{2\alpha\mathcal{R}[h_S]}(e^{\sqrt{x}} - 1)^2$, it suffices to plug the latter inequality in (B.7). Now, using (B.6) again yields:

$$\begin{aligned}
\mathbb{E}\left[\sup_{|y'|,|y| \leq \hat{\rho}^i_\lambda} |\phi'(h_S(X)Y + y')\phi'(h_S(X)Y + y)|\right] &\leq \mathbb{E}\left[e^{2\hat{\tau}^i_\lambda} \min\left(e^{\ell(h_S,Z)} - 1, 1\right)^2\right] \\
&\leq \mathbb{E}\left[e^{2\hat{\tau}^i_\lambda}\right] \mathbb{E}\left[\min\left(e^{\ell(h_S,Z)} - 1, 1\right)\right].
\end{aligned}$$

Finally, using the same steps as before, we show that the logistic loss verifies Assumption 4 with $\Psi_1(x) = C_S e^{2\alpha\mathcal{R}[h_S]}(e^{\sqrt{x}} - 1)$.

B.4.5. SOFTPLUS

The proof is similar to that of the logistic loss and is left for the reader.

B.5. Proof of Theorem 4.1

First, notice that:

$$\mathcal{E}_{\text{gen}} = \left|\mathbb{E}\left[\widehat{\mathcal{R}}\left[\mathcal{A}(\mathcal{D}_T)\right] - \mathcal{R}\left[\mathcal{A}(\mathcal{D}_T)\right]\right]\right| = \left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z_i\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right]\right|$$
$$= \left|\mathbb{E}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z_1\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right]\right|.$$

Using triangle inequality and the fact that Z and Z_1 have the same distributions, we obtains:

$$\mathcal{E}_{\text{gen}} \leq \left|\mathbb{E}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z_1\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z_1\right)\right]\right| + \left|\mathbb{E}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z_1\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right]\right|$$
$$= \left|\mathbb{E}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z_1\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z_1\right)\right]\right| + \left|\mathbb{E}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right]\right|.$$

The desired result follows from Propositions 3.1 and 3.2.

B.6. Proof of Theorem 4.2

First introduce

$$h_\lambda = \underset{h\in\mathcal{H}}{\arg\min}\, \mathcal{R}\left[h_S + h\right] + \lambda\left\|h\right\|_k^2,$$

and write

$$\mathcal{R}\left[\mathcal{A}\right] - \mathcal{R}[h^* + h_S] = \mathcal{R}\left[\mathcal{A}\right] - \widehat{\mathcal{R}}\left[\mathcal{A}\right] + \widehat{\mathcal{R}}\left[\mathcal{A}\right] + \lambda\left\|f\right\|_k^2 - \mathcal{R}[h_\lambda + h_S] + \mathcal{R}[h_\lambda + g'] - \mathcal{R}[h^* + h_S].$$

Now by rearranging and reminding that:

$$\widehat{\mathcal{R}}\left[\mathcal{A}\right] + \lambda\left\|f\right\|_k^2 \leq \widehat{\mathcal{R}}\left[h_\lambda + h_S\right] + \lambda\left\|h_\lambda\right\|_k^2,$$

we obtain:

$$\mathcal{R}\left[\mathcal{A}\right] - \mathcal{R}[h^* + h_S] \leq \mathcal{R}\left[\mathcal{A}\right] - \widehat{\mathcal{R}}\left[\mathcal{A}\right] + \widehat{\mathcal{R}}\left[h_\lambda + h_S\right] - \mathcal{R}[h_\lambda + h_S]$$
$$+ \mathcal{R}[h_\lambda + h_S] + \lambda\left\|h_\lambda\right\|_k^2 - \mathcal{R}[h^* + h_S].$$

For the first term notice that:

$$\mathbb{E}\left[\mathcal{R}\left[\mathcal{A}\right] - \widehat{\mathcal{R}}\left[\mathcal{A}\right]\right] \leq \mathcal{E}_{\text{gen}} \leq \beta(n) + \gamma(n).$$

Regarding the second term, since h_λ is independent of $\mathcal{D}_T$ we have:

$$\mathbb{E}\left[\widehat{\mathcal{R}}\left[h_\lambda + h_S\right] - \mathcal{R}[h_\lambda + h_S]\right] = 0.$$

Finally, notice that by definition of g_λ that:

$$\mathcal{R}[h_\lambda + h_S] + \lambda\left\|h_\lambda\right\|_k^2 - \mathcal{R}[h^* + h_S] \leq \lambda\left\|h^*\right\|_k^2.$$

Combining the latter four inequalities yields:

$$\mathcal{E}_{\text{ex}} = \mathcal{R}\left[\mathcal{A}\right] - \mathcal{R}[h^* + h_S] \leq \beta(n) + \gamma(n) + \lambda\left\|h^*\right\|_k, \tag{B.8}$$

which concludes the first part. For the second part we shall use Table 1 and the fact that

$$\gamma(n) + \beta(n) \leq \alpha\frac{\left(\Psi_1\left(\mathcal{R}\left[h_S\right]\right) + \Psi_2\left(\mathcal{R}\left[h_S\right]\right)\right) \wedge \left(2\left\|\phi'\right\|_\infty^2\right)}{n}. \tag{B.9}$$

B.6.1. MSE AND SQUARED HINGE

For these two losses, $\Psi_1(x) = \Psi_2(x) = 8x(4\alpha + 1)$, so that by inequality (B.9) we get:

$$\begin{aligned}\gamma(n) + \beta(n) &\leq \alpha \frac{16\mathcal{R}\left[h_S\right](4\alpha + 1)}{n} \\ &= \frac{16\kappa\mathcal{R}\left[h_S\right](4\frac{\kappa}{\lambda} + 1)}{\lambda n}.\end{aligned}$$

Thus for small λ one has:

$$\gamma(n) + \beta(n) = \mathcal{O}\left(\frac{\mathcal{R}[h_S]}{\lambda^2 n}\right).$$

To conclude, set $\lambda = \sqrt{\frac{\mathcal{R}[h_S]}{\sqrt{n}}}$ and use Inequality (B.8) to obtain:

$$\mathcal{E}_{\text{ex}} = \mathcal{O}\left(\sqrt{\frac{\mathcal{R}[h_S]}{\sqrt{n}}}\right).$$

B.6.2. EXPONENTIAL

Using Table 1, remind that the functions $\Psi_1(x)$ and Ψ_2 are given by: $\Psi_1(x) = C_S x^2 e^{2\alpha x}$, $\Psi_2(x) = C_S M_S x e^{2\alpha x}$ with $M_S = \sup_{z\in\mathcal{Z}_T} \ell(h_S, z)$ and

$$C_S = \exp\left\{2 + \frac{2\alpha M_S}{n} + \frac{4\alpha^2 M_S^2}{n-1}\right\} = \exp\left\{2 + \frac{2\kappa M_S}{\lambda n} + \frac{4\kappa^2 M_S^2}{\lambda^2(n-1)}\right\}.$$

Assume that $n \geq \max\left(\frac{M_S^2 \ln(n)^2}{\mathcal{R}[h_S]}, 2\right)$ and $\lambda = 4\frac{\sqrt{\mathcal{R}[h_S]}\wedge 1}{\ln(n)} = 4\frac{\sqrt{\mathcal{R}[h_S]}}{\ln(n)}$. The case where $\mathcal{R}[h_S] \geq 1$ is similar and thus omitted. Now, write

$$n \geq \frac{M_S^2 \ln(n)^2}{\mathcal{R}[h_S]} = \frac{M_S^2}{\lambda^2} \implies \frac{M_S^2}{\lambda^2(n-1)} \leq \frac{n}{n-1} \leq 2.$$

The latter condition also implies that $\frac{M_S}{\lambda n} \leq \frac{\lambda}{M_S} \leq \frac{\sqrt{\mathcal{R}[h_S]}}{M_S \ln(n)} \leq \frac{1}{2M_S}$. By these two facts, we deduce that C_S can be bounded independently of n. Thus, using (B.9) yields:

$$\begin{aligned}\gamma(n) + \beta(n) &\leq \alpha \frac{C_S\left(\mathcal{R}[h_S]^2 + M_S\mathcal{R}[h_S]\right)e^{2\alpha\mathcal{R}[h_S]}}{n} \\ &= \ln(n)\frac{C_S\left(\mathcal{R}[h_S]^{3/2} + M_S\mathcal{R}[h_S]^{1/2}\right)(\sqrt{n})^{\kappa\sqrt{\mathcal{R}[h_S]}}}{n} \\ &= \mathcal{O}\left(\frac{\sqrt{\mathcal{R}[h_S]}}{\sqrt{n}}\right),\end{aligned} \tag{B.10}$$

where the two last inequalities follow from the facts that $\alpha = \frac{\kappa}{\lambda} = \frac{\kappa\ln(n)}{\sqrt{\mathcal{R}[h_S]}}$ and $\kappa \leq 1$. It remains to use the (B.8) to conclude the first part. For the second part, set $\lambda = \frac{\ln(n)^2}{\sqrt{n}}$ and notice that, if $n \leq \frac{M_S^2\ln(n)^2}{\mathcal{R}[h_S]}$ then $\mathcal{R}[h_S] \leq \frac{M_S^2\ln(n)^2}{n} \leq M_S^2\lambda$ and $\alpha\mathcal{R}[h_s] \leq M_S^2\kappa \leq 1$. Furthermore, the constant C_S can be bounded independently of n with such a choice of λ. Inequality (B.10) becomes:

$$\gamma(n) + \beta(n) = \mathcal{O}\left(\frac{\mathcal{R}[h_S]}{\sqrt{n}\ln(n)^2}\right).$$

It remains to use Inequality (B.8) to complete the proof.

B.6.3. LOGISTIC

For this loss, we have $\|\phi'\|_\infty = 1$ and Inequality (B.9) becomes:

$$\beta(n) + \gamma(n) \leq \frac{2\alpha}{n} = \frac{2\kappa}{\lambda n} \leq \frac{2}{\lambda n}.$$

Thus, setting $\lambda = \frac{1}{\sqrt{n}}$ and using Inequality (B.8) yields:

$$\mathcal{E}_{\text{ex}} = \mathcal{O}\left(\sqrt{\frac{1}{n}}\right).$$

Furthermore, if $n \geq 9$ and $\mathcal{R}[h_S] \leq \frac{1}{\sqrt{n}} \leq \frac{1}{e}$, then with the choice $\lambda = \frac{8}{\sqrt{-n\ln(\mathcal{R}[h_S])}}$ one has:

$$\mathcal{E}_{\text{ex}} = \mathcal{O}\left(\frac{1}{\sqrt{-n\ln\left(\mathcal{R}[h_S]\right)}}\right).$$

Indeed, in the setting above, it leads to that:

$$\begin{aligned} e^{\alpha\mathcal{R}[h_S]} = e^{\frac{\kappa}{\lambda}\mathcal{R}[h_S]} &= e^{\frac{\kappa\sqrt{-\ln(\mathcal{R}[h_S])}}{8}} \\ (\kappa \leq 1) &\leq e^{\frac{\sqrt{-\ln(\mathcal{R}[h_S])}}{8}} \\ (-\ln(\mathcal{R}[h_S]) \geq 1) &\leq e^{\frac{-\ln(\mathcal{R}[h_S])}{8}} = \mathcal{R}[h_S]^{-1/8}, \end{aligned}$$

and

$$e^{\frac{2\kappa M_S}{\lambda n}} \leq e^{\frac{2}{\lambda n}} = e^{-\frac{\sqrt{\ln(\mathcal{R}[h_S])}}{4\sqrt{n}}} \leq e^{\sqrt{\frac{\ln(n)}{4n}}} \leq e^{1/4}.$$

Besides, since $\frac{n}{n-1} \leq 2$,

$$e^{\frac{4\kappa M_S}{\lambda^2 n}} \leq e^{\frac{4}{\lambda^2(n-1)}} = e^{-\frac{\ln(\mathcal{R}[h_S])(n)}{16(n-1)}} \leq \mathcal{R}[h_S]^{-1/8}.$$

Moreover, using Inequality (B.9) and Table 1 gives:

$$\begin{aligned} \gamma(n) + \beta(n) &\leq \alpha\frac{C_S e^{\sqrt{\mathcal{R}[h_S]}} e^{2\alpha\mathcal{R}[h_S]}\left(e^{\sqrt{\mathcal{R}[h_S]}} - 1\right)}{n} \\ &= \kappa\frac{\exp\left\{2 + \frac{2\alpha M_S}{n} + \frac{4\alpha^2 M_S^2}{n-1} + \sqrt{\mathcal{R}[h_S]} + 2\alpha\mathcal{R}[h_S]\right\}\left(e^{\sqrt{\mathcal{R}[h_S]}} - 1\right)}{\lambda n} \\ &= \mathcal{O}\left(\frac{\sqrt{-\ln\left(\mathcal{R}[h_S]\right)}\left(e^{\sqrt{\mathcal{R}[h_S]}} - 1\right)}{\mathcal{R}[h_S]^{1/4}\sqrt{n}}\right). \end{aligned}$$

Now, since the function $e^{\sqrt{x}} - 1 \leq 2\sqrt{x}$ for all $x \leq \ln(2)^2$ and $\mathcal{R}[h_S] \leq \frac{1}{n} \leq \frac{1}{3} \leq \ln(2)^2$ the latter inequality becomes:

$$\gamma(n) + \beta(n) = \mathcal{O}\left(\frac{\sqrt{-\ln\left(\mathcal{R}[h_S]\right)}\mathcal{R}[h_S]^{1/4}}{\sqrt{n}}\right).$$

To conclude the proof notice that, for all $x \leq 1$, we have $\ln(x^{-1/4}) \leq x^{-1/4}$ and thus $x^{1/4}(-\ln(x)) \leq 4$. This leads to :

$$x^{1/4}\sqrt{-\ln(x)} \leq \frac{4}{\sqrt{-\ln(x)}}.$$

Therefore,

$$\gamma(n) + \beta(n) = \mathcal{O}\left(\frac{1}{\sqrt{-n\ln\left(\mathcal{R}[h_S]\right)}}\right).$$

It remains to use Inequality (B.8) to complete the proof.

B.6.4. SOFTPLUS

For the softplus, the choice $\lambda = 1/\sqrt{n}$ yields:

$$\mathcal{E}_{\text{ex}} = \mathcal{O}\left(\sqrt{\frac{1}{n}}\right).$$

Furthermore, if $n \geq 9$ and $\mathcal{R}[h_S] \leq \frac{1}{\sqrt{n}}$ and $\frac{1}{s} \leq -\ln(\mathcal{R}[h_S])$, then with the choice $\lambda = \frac{8}{\sqrt{-n\ln(\mathcal{R}[h_S])}}$, one has:

$$\mathcal{E}_{\text{ex}} = \mathcal{O}\left(\frac{1}{\sqrt{-sn\ln\left(\mathcal{R}[h_S]\right)}}\right).$$

The proof is identical to the previous one and thus omitted.