arXiv:2403.19480v1 [cs.LG] 28 Mar 2024

# $\mathcal{H}$-Consistency Guarantees for Regression

**Anqi Mao** AQMAO@CIMS.NYU.EDU
*Courant Institute of Mathematical Sciences, New York*

**Mehryar Mohri** MOHRI@GOOGLE.COM
*Google Research and Courant Institute of Mathematical Sciences, New York*

**Yutao Zhong** YUTAO@CIMS.NYU.EDU
*Courant Institute of Mathematical Sciences, New York*

## Abstract

We present a detailed study of $\mathcal{H}$-consistency bounds for regression. We first present new theorems that generalize the tools previously given to establish $\mathcal{H}$-consistency bounds. This generalization proves essential for analyzing $\mathcal{H}$-consistency bounds specific to regression. Next, we prove a series of novel $\mathcal{H}$-consistency bounds for surrogate loss functions of the squared loss, under the assumption of a symmetric distribution and a bounded hypothesis set. This includes positive results for the Huber loss, all $\ell_p$ losses, $p \geq 1$, the squared $\epsilon$-insensitive loss, as well as a negative result for the $\epsilon$-insensitive loss used in squared Support Vector Regression (SVR). We further leverage our analysis of $\mathcal{H}$-consistency for regression and derive principled surrogate losses for adversarial regression (Section 5). This readily establishes novel algorithms for adversarial regression, for which we report favorable experimental results in Section 6.

## 1. Introduction

Learning algorithms often optimize loss functions that differ from the originally specified task. In classification, this divergence typically arises due to the computational intractability of optimizing the original loss or because it lacks certain desirable properties like differentiability or smoothness. In regression, the shift may occur because the surrogate loss used exhibits more favorable characteristics, such as handling outliers or ensuring sparser solutions. For instance, the Huber loss and $\ell_1$ loss are used to mitigate the impact of outliers since the squared loss is known to be sensitive to the presence of outliers, while $\epsilon$-insensitive losses promote sparsity. But, what guarantees do we have when training with a loss function distinct from the target squared loss?

Addressing this question can have significant implications in the design of regression algorithms. It can also strongly benefit the design of useful surrogate losses for other related problems, such as adversarial regression, as we shall see.

The statistical properties of surrogate losses have been extensively studied in the past. In particular, the Bayes-consistency of various convex loss functions, including margin-based surrogate losses in binary classification (Zhang, 2004a; Bartlett, Jordan, and McAuliffe, 2006), and other loss function families for multi-classification Zhang (2004b); Tewari and Bartlett (2007); Steinwart (2007), has been examined in detail.

However, prior work by Long and Servedio (2013) has highlighted the limitations of Bayes-consistency, since it does not account for the hypothesis set adopted. They established that for some hypothesis sets and distributions, algorithms minimizing Bayes-consistent losses may retain a constant expected error, while others minimizing inconsistent losses tend to have an expected error approaching zero. This indicates the significant role of the chosen hypothesis set in consistency.

© A. Mao, M. Mohri & Y. Zhong.

Recent seminal work by Awasthi, Mao, Mohri, and Zhong (2022a,b) and Mao, Mohri, and Zhong (2023f,c,e,b) has analyzed $\mathcal{H}$*-consistency bounds* for broad families of surrogate losses in binary classication, multi-class classification, structured prediction, and abstention (Mao et al., 2023a). These bounds are more informative than Bayes-consistency since they are hypothesis set-specific and do not require the entire family of measurable functions. Moreover, they offer finite sample, non-asymptotic guarantees. In light of these recent guarantees, the following questions naturally arise: Can we derive a non-asymptotic analysis of regression taking into account the hypothesis set? How can we benefit from that analysis?

While there is some previous work exploring Bayes-consistency in regression (Caponnetto, 2005; Christmann and Steinwart, 2007; Steinwart, 2007), we are not aware of any prior $\mathcal{H}$-consistency bounds or similar finite sample guarantees for surrogate losses in regression, such as, for example, the Huber loss or the squared $\epsilon$-insensitive loss.

This paper presents the first in-depth study of $\mathcal{H}$-consistency bounds in the context of regression. We first present new theorems that generalize the tools previously given by Awasthi et al. (2022a,b) and Mao et al. (2023f,c,e,b) to establish $\mathcal{H}$-consistency bounds (Section 3). This generalization proves essential in regression for analyzing $\mathcal{H}$-consistency bounds for surrogate losses such as Huber loss and the squared $\epsilon$-insensitive loss. It also provides finer bounds for the $\ell_1$ loss.

Next, we prove a series of $\mathcal{H}$-consistency bounds for surrogate loss functions of the squared loss, under the assumption of a symmetric distribution and a bounded hypothesis set (Section 4). We prove the first $\mathcal{H}$-consistency bound for the Huber loss, which is a commonly used surrogate loss used to handle outliers, contingent upon a specific condition concerning the Huber loss parameter $\delta$ and the distribution mass around the mean. We further prove that this condition is necessary when $\mathcal{H}$ is realizable.

We then extend our analysis to cover $\mathcal{H}$-consistency bounds for $\ell_p$ losses, for all values of $p \geq 1$. In particular, remarkably, we give guarantees for the $\ell_1$ loss and $\ell_p$ losses with $p \in (1, 2)$. We further analyze the $\epsilon$-insensitive and the squared $\epsilon$-insensitive losses integral to the definition of the SVR (Support Vector Regression) and quadratic SVR algorithms (Vapnik, 2000). These loss functions and SVR algorithms admit the benefit of yielding sparser solutions. We give the first $\mathcal{H}$-consistency bound for the quadratic $\epsilon$-insensitive loss. We also prove a negative result for the $\epsilon$-insensitive loss: this loss function used in the definition of SVR does not admit $\mathcal{H}$-consistency bounds with respect to the squared loss, even under some additional assumptions on the parameter $\epsilon$ and the distribution.

Subsequently, leveraging our analysis of $\mathcal{H}$-consistency for regression, we derive principled surrogate losses for adversarial regression (Section 5). This readily establishes a novel algorithm for adversarial regression, for which we report favorable experimental results in Section 6.

**Previous work.** Bayes-consistency has been extensively studied in various learning problems. These include binary classification (Zhang, 2004a; Bartlett et al., 2006), multi-class classification (Zhang, 2004b; Tewari and Bartlett, 2007; Narasimhan et al., 2015; Finocchiaro et al., 2019; Wang and Scott, 2020; Frongillo and Waggoner, 2021; Wang and Scott, 2023), ranking (Menon and Williamson, 2014; Gao and Zhou, 2015; Uematsu and Lee, 2017), multi-label classification (Gao and Zhou, 2011; Koyejo et al., 2015; Zhang et al., 2020), structured prediction (Ciliberto et al., 2016; Osokin et al., 2017; Blondel, 2019), and ordinal regression (Pedregosa et al., 2017). The concept of $\mathcal{H}$-consistency has been studied under the realizable assumption in (Long and Servedio, 2013; Zhang and Agarwal, 2020). The notion of $\mathcal{H}$-consistency bounds in classification is due to Awasthi et al. (2022a,b). $\mathcal{H}$-consistency bounds have been further analyzed in scenarios such as multi-class classification (Mao et al., 2023f; Zheng et al., 2023; Mao et al., 2023b), ranking (Mao et al., 2023c,d), structured prediction (Mao et al., 2023e), and abstention (Mao et al., 2024a,c,b; Mohri et al., 2024).

However, in the context of regression, there is limited work on the consistency properties of surrogate losses. The main related work we are aware are (Caponnetto, 2005; Christmann and Steinwart, 2007; Steinwart, 2007). In particular, Steinwart (2007) studied Bayes-consistency for a family of regression surrogate losses including $\ell_p$, but without presenting any non-asymptotic bound. Nevertheless, we partly benefit from this previous work. In particular, we adopt the same symmetric and bounded distribution assumption.

## 2. Preliminaries

**Bounded regression.** We first introduce the learning scenario of bounded regression. We denote by $\mathcal{X}$ the input space, $\mathcal{Y}$ a measurable subset of $\mathbb{R}$, and $\mathcal{D}$ a distribution over $\mathcal{X} \times \mathcal{Y}$. As for other supervised learning problems, the learner receives a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ drawn i.i.d. according to $\mathcal{D}$.

The measure of error is based on the magnitude of the difference between the predicted real-valued label and the true label. The function used to measure the error is denoted as $\mathsf{L}\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. Let $L\colon (h, x, y) \mapsto \mathsf{L}(h(x), y)$ be the associated loss function. Some common examples of loss functions used in regression are the squared loss $\ell_2$, defined by $\mathsf{L}(y', y) = |y' - y|^2$ for all $y, y' \in \mathcal{Y}$; or more generally the $\ell_p$ loss defined by $\mathsf{L}(y', y) = |y' - y|^p$, for $p \geq 1$. The squared loss is known to be quite sensitive to outliers. An alternative more robust surrogate loss is the Huber loss $\ell_\delta$ (Huber, 1964), which is defined for a parameter $\delta > 0$ as the following combination of the $\ell_2$ and $\ell_1$ loss functions: $\mathsf{L}(y', y) = \frac{1}{2}(y' - y)^2$ if $|y' - y| \leq \delta$, $\left(\delta |y' - y| - \frac{1}{2}\delta^2\right)$ otherwise. The $\epsilon$-insensitive loss $\ell_\epsilon$ and the squared $\epsilon$-insensitive loss $\ell_{\text{sq}-\epsilon}$ (Vapnik, 2000) are defined by $\mathsf{L}(y', y) = \max\{|y' - y| - \epsilon, 0\}$ and $\mathsf{L}(y', y) = \max\left\{|y' - y|^2 - \epsilon^2, 0\right\}$, for some $\epsilon > 0$.

**Bayes-Consistency.** Given a loss function $L$, we denote by $\mathcal{E}_L(h)$ the generalization error of a hypothesis $h \in \mathcal{H}$, and by $\mathcal{E}_L^*(\mathcal{H})$ the best-in-class error for a hypothesis set $\mathcal{H}$:

$$\mathcal{E}_L(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[L(h, x, y)] \quad \mathcal{E}_L^*(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathcal{E}_L(h).$$

A desirable property of surrogate losses in regression is *Bayes-consistency* (Zhang, 2004a; Bartlett et al., 2006; Steinwart, 2007), that is, minimizing the surrogate losses $L$ over the family of all measurable functions $\mathcal{H}_{\text{all}}$ leads to the minimization of the squared loss $\ell_2$ over $\mathcal{H}_{\text{all}}$. We say that $L$ is *Bayes-consistent* with respect to $\ell_2$, if, for all distributions and sequences of $\{h_n\}_{n\in\mathbb{N}} \subset \mathcal{H}_{\text{all}}$, $\lim_{n\to+\infty} \mathcal{E}_L(h_n) - \mathcal{E}_L^*(\mathcal{H}_{\text{all}}) = 0$ implies $\lim_{n\to+\infty} \mathcal{E}_{\ell_2}(h_n) - \mathcal{E}_{\ell_2}^*(\mathcal{H}_{\text{all}}) = 0$. Bayes-consistency stands as an essential prerequisite for a surrogate loss. Nonetheless, it has some shortcomings: it is only an asymptotic property and it fails to account for the hypothesis set $\mathcal{H}$ (Awasthi et al., 2022a,b).

**$\mathcal{H}$-Consistency bounds.** In contrast with Bayes-consistency, $\mathcal{H}$-Consistency bounds take into account the specific hypothesis set $\mathcal{H}$ and are non-asymptotic. Given a hypothesis set $\mathcal{H}$, we say that a regression loss function $L$ admits an *$\mathcal{H}$-consistency bound with respect to $\ell_2$* (Awasthi et al., 2022a,b), if for some non-decreasing function $f\colon \mathbb{R}_+ \to \mathbb{R}_+$, for all distributions and all $h \in \mathcal{H}$, the following inequality holds:

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}_{\ell_2}^*(\mathcal{H}) \leq f(\mathcal{E}_L(h) - \mathcal{E}_L^*(\mathcal{H})).$$

Thus, when the $L$-estimation error can be reduced to some $\eta > 0$, the squared loss estimation error is upper bounded by $f(\eta)$. An $\mathcal{H}$-Consistency bound is a stronger and more informative property than Bayes-consistency, which is implied by taking the limit.

In the next section, we will prove $\mathcal{H}$-consistency bounds for several common surrogate regression losses with respect to the squared loss $\ell_2$. A by-product of these guarantees is the Bayes-consistency of these losses.

For a regression loss function $L$ and a hypothesis $h$, the generalization error can be expressed as follows:

$$\mathcal{E}_L(h) = \mathbb{E}_x\left[\mathbb{E}_y[\mathsf{L}(h(x), y) \mid x]\right] = \mathbb{E}_x[\mathcal{C}_L(h, x)],$$

where $\mathcal{C}_L(h,x)$ is the *conditional error* $\mathbb{E}_y[\mathsf{L}(h(x), y) \mid x]$. We also write $\mathcal{C}_L^*(\mathcal{H}, x)$ to denote the best-in-class conditional error defined by $\mathcal{C}_L^*(\mathcal{H}, x) = \inf_{h \in \mathcal{H}} \mathcal{C}_L(h, x)$. The conditional regret or calibration gap, $\Delta\mathcal{C}_{L,\mathcal{H}}(h,x)$, measures the difference between the conditional error of $h$ and the best-in-class conditional error: $\Delta\mathcal{C}_{L,\mathcal{H}}(h,x) = \mathcal{C}_L(h,x) - \mathcal{C}_L^*(\mathcal{H}, x)$. A generalization of conditional regret is the conditional $\epsilon$-regret, defined as: $\left[\Delta\mathcal{C}_{L,\mathcal{H}}(h,x)\right]_\epsilon = \Delta\mathcal{C}_{L,\mathcal{H}}(h,x) 1_{\Delta\mathcal{C}_{L,\mathcal{H}}(h,x) > \epsilon}$.

A key term appearing in our bounds is the minimizability gap, defined for a loss function $L$ and a hypothesis set $\mathcal{H}$ as $\mathcal{M}_L(\mathcal{H}) = \mathcal{E}_L^*(\mathcal{H}) - \mathbb{E}_x[\mathcal{C}_L^*(\mathcal{H}, x)]$. It quantifies the discrepancy between the best-in-class generalization error and the expected best-in-class conditional error. An alternative expression for the minimizability gap is: $\mathcal{M}_L(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathbb{E}_x[\mathcal{C}_L(\mathcal{H}, x)] - \mathbb{E}_x[\inf_{h \in \mathcal{H}} \mathcal{C}_L(\mathcal{H}, x)]$. Due to the super-additivity of the infimum, the minimizability gap is always non-negative. As shown by Steinwart (2007, Lemma 2.5, Theorem 3.2), for the family of all measurable functions, the equality $\mathcal{E}_L^*(\mathcal{H}_{\text{all}}) = \mathbb{E}_x[\mathcal{C}_L^*(\mathcal{H}_{\text{all}}, x)]$ holds. Thus, the minimizability gap can be bounded above by the approximation error $\mathcal{E}_L^*(\mathcal{H}) - \mathcal{E}_L^*(\mathcal{H}_{\text{all}})$. The minimizability gap becomes zero when when $\mathcal{H} = \mathcal{H}_{\text{all}}$ or, more broadly, when $\mathcal{E}_L^*(\mathcal{H}) = \mathcal{E}_L^*(\mathcal{H}_{\text{all}})$.

## 3. General $\mathcal{H}$-consistency theorems

To derive $\mathcal{H}$-consistency bounds for regression, we first give two key theorems establishing that if a convex or concave function provides an inequality between the conditional regrets of regression loss functions $L_1$ and $L_2$, then this inequality translates into an $\mathcal{H}$-consistency bound involving the minimizability gaps of $L_1$ and $L_2$.

**Theorem 1 (General $\mathcal{H}$-consistency bound – convex function)** *Let $\mathcal{D}$ denote a distribution over $\mathcal{X} \times \mathcal{Y}$. Assume that there exists a convex function $\Psi\colon \mathbb{R}_+ \to \mathbb{R}$ with $\Psi(0) \geq 0$, a positive function $\alpha\colon \mathcal{H} \times \mathcal{X} \to \mathbb{R}_+^*$ with $\sup_{x \in \mathcal{X}} \alpha(h,x) < +\infty$ for all $h \in \mathcal{H}$, and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \mathcal{X}$: $\Psi\left(\left[\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)\right]_\epsilon\right) \leq \alpha(h,x)\, \Delta\mathcal{C}_{L_1,\mathcal{H}}(h,x)$. Then, for any hypothesis $h \in \mathcal{H}$, the following inequality holds:*

$$\Psi\left(\mathcal{E}_{L_2}(h) - \mathcal{E}_{L_2}^*(\mathcal{H}) + \mathcal{M}_{L_2}(\mathcal{H})\right) \leq \left[\sup_{x \in \mathcal{X}} \alpha(h,x)\right]\left(\mathcal{E}_{L_1}(h) - \mathcal{E}_{L_1}^*(\mathcal{H}) + \mathcal{M}_{L_1}(\mathcal{H})\right) + \max\{\Psi(0), \Psi(\epsilon)\}.$$

**Theorem 2 (General $\mathcal{H}$-consistency bound – concave function)** *Let $\mathcal{D}$ denote a distribution over $\mathcal{X} \times \mathcal{Y}$. Assume that there exists a concave function $\Gamma\colon \mathbb{R}_+ \to \mathbb{R}$, a positive function $\alpha\colon \mathcal{H} \times \mathcal{X} \to \mathbb{R}_+^*$ with $\sup_{x \in \mathcal{X}} \alpha(h,x) < +\infty$ for all $h \in \mathcal{H}$, and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$,*

*$x \in \mathcal{X}$: $\left[\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)\right]_{\epsilon} \leq \Gamma\big(\alpha(h,x)\Delta\mathcal{C}_{L_1,\mathcal{H}}(h,x)\big)$. Then, for any hypothesis $h \in \mathcal{H}$, the following inequality holds*

$$\mathcal{E}_{L_2}(h) - \mathcal{E}^*_{L_2}(\mathcal{H}) + \mathcal{M}_{L_2}(\mathcal{H}) \leq \Gamma\left(\left[\sup_{x\in\mathcal{X}} \alpha(h,x)\right]\left(\mathcal{E}_{L_1}(h) - \mathcal{E}^*_{L_1}(\mathcal{H}) + \mathcal{M}_{L_1}(\mathcal{H})\right)\right) + \epsilon.$$

*In the special case where $\Gamma(x) = x^{\frac{1}{q}}$ for some $q \geq 1$ with conjugate $p \geq 1$, that is $\frac{1}{p} + \frac{1}{q} = 1$, for any $h \in \mathcal{H}$, the following inequality holds, assuming $\mathbb{E}_X\left[\alpha^{\frac{p}{q}}(h,x)\right]^{\frac{1}{p}} < +\infty$ for all $h \in \mathcal{H}$:*

$$\mathcal{E}_{L_2}(h) - \mathcal{E}^*_{L_2}(\mathcal{H}) + \mathcal{M}_{L_2}(\mathcal{H}) \leq \mathop{\mathbb{E}}_{X}\left[\alpha^{\frac{p}{q}}(h,x)\right]^{\frac{1}{p}} \mathop{\mathbb{E}}_{X}\left[\mathcal{E}_{L_1}(h) - \mathcal{E}^*_{L_1}(\mathcal{H}) + \mathcal{M}_{L_1}(\mathcal{H})\right]^{\frac{1}{q}} + \epsilon.$$

Theorems 1 and 2 provide significantly more general tools for establishing $\mathcal{H}$-consistency bounds than previous results from (Awasthi et al., 2022a, Theorems 1 and 2) and (Awasthi et al., 2022b, Theorems 1 and 2) for binary and multi-class classification. They offer a more general framework for establishing consistency bounds by allowing for non-constant functions $\alpha$. This generalization is crucial for analyzing consistency bounds in regression, where $\alpha$ may not be constant for certain surrogate losses (e.g., Huber loss, squared $\epsilon$-insensitive loss). Our generalized theorems also enable finer consistency bounds, as demonstrated later in the case of the $\ell_1$ loss. The proofs of Theorems 1 and 2 are included in Appendix A.

To leverage these general theorems, we will characterize the best-in-class conditional error and the conditional regret of the squared loss. We first introduce some definitions we will need. We say that the conditional distribution is *bounded by* $B > 0$ if, for all $x \in \mathcal{X}, \mathbb{P}(|Y| \leq B \mid X = x) = 1$. We say that a hypothesis set $\mathcal{H}$ is bounded by $B > 0$ if, $|h(x)| \leq B$ for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$, and all values in $[-B, +B]$ are attainable by $h(x)$, $h \in \mathcal{H}$. The conditional mean of the distribution at $x$ is denoted as: $\mu(x) = \mathbb{E}[y \mid x]$.

**Theorem 3** *Assume that the conditional distribution and the hypothesis set $\mathcal{H}$ are bounded by $B > 0$. Then, the best-in-class conditional error and the conditional regret of the squared loss can be characterized as: for all $h \in \mathcal{H}, x \in \mathcal{X}$,*

$$\mathcal{C}^*_{\ell_2}(\mathcal{H},x) = \mathcal{C}_{\ell_2}(\mu(x),x) = \mathbb{E}[y^2 \mid x] - (\mu(x))^2$$
$$\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x) = (h(x) - \mu(x))^2.$$

Refer to Appendix A for the proof. As in (Steinwart, 2007), for our analysis, we will focus specifically on symmetric distributions, where the conditional mean and the conditional median coincide. This is because, otherwise, as shown by Steinwart (2007, Proposition 4.14) the squared loss is essentially the only distance-based and locally Lipschitz continuous loss function that is Bayes-consistent with respect to itself for all bounded conditional distributions.

A distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ is said to be *symmetric* if and only if for all $x \in \mathcal{X}$, there exits $y_0 \in \mathbb{R}$ such that $\mathcal{D}_{y|x}(y_0 - A) = \mathcal{D}_{y|x}(y_0 + A)$ for all measurable $A \subset [0, +\infty)$. The next result characterizes the best-in-class predictor for any symmetric regression loss functions for such distributions.

**Theorem 4** *Let $\psi: \mathbb{R} \to \mathbb{R}$ be a symmetric function such that $\psi(x) = \psi(-x)$ for all $x \in \mathbb{R}$. Furthermore, $\psi(x) \geq 0$ for all $x$ in its domain and it holds that $\psi(0) = 0$. Assume that the conditional distribution and the hypothesis set $\mathcal{H}$ is bounded by $B > 0$. Assume that the distribution is symmetric and the regression loss function is given by $\mathsf{L}(y', y) = \psi(y' - y)$. Then, we have $\mathcal{C}^*_L(\mathcal{H},x) = \mathcal{C}_L(\mu(x),x)$.*

The proof is included in Appendix A. It is straightforward to see that all the previously mentioned regression loss functions satisfy the assumptions in Theorem 4. Therefore, for these loss functions, the best-in-class conditional error is directly characterized by Theorem 4. Furthermore, if we have $x \mapsto \mu(x) \in \mathcal{H}$, then under the same assumption, we have $\mathcal{E}^*_L(\mathcal{H}) = \mathbb{E}_x[\mathcal{C}^*_L(\mathcal{H}, x)] = \mathbb{E}_x[\mathcal{C}_L(\mu(x), x)]$ and thus the minimizability gap vanishes: $\mathcal{M}_L(\mathcal{H}) = 0$.

**Definition 5** *A hypothesis set $\mathcal{H}$ is said to be* realizable *if the function that maps $x$ to the conditional mean $\mu(x)$ is included in $\mathcal{H}$: $x \mapsto \mu(x) \in \mathcal{H}$.*

**Corollary 6** *Under the same assumption as in Theorem 4, for realizable hypothesis sets, we have $\mathcal{M}_L(\mathcal{H}) = 0$.*

## 4. $\mathcal{H}$-Consistency bounds for regression

In this section, we will analyze the $\mathcal{H}$-consistency of several regression loss functions with respect to the squared loss.

### 4.1. Huber Loss

The Huber loss $\ell_\delta \colon (h, x, y) \mapsto \frac{1}{2}(h(x) - y)^2 1_{|h(x)-y| \leq \delta} + \big(\delta |h(x) - y| - \frac{1}{2}\delta^2\big) 1_{|h(x)-y| > \delta}$ is a frequently used loss function in regression for dealing with outliers. It imposes quadratic penalties on small errors and linear penalties on larger ones. The next result provides $\mathcal{H}$-consistency bounds for the Huber loss with respect to the squared loss.

**Theorem 7** *Assume that the distribution is symmetric, the conditional distribution and the hypothesis set $\mathcal{H}$ are bounded by $B > 0$. Assume that $p_{\min}(\delta) = \inf_{x \in \mathcal{X}} \mathbb{P}(0 \leq \mu(x) - y \leq \delta \mid x)$ is positive. Then, for all $h \in \mathcal{H}$, the following $\mathcal{H}$-consistency bound holds:*

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \mathcal{M}_{\ell_2}(\mathcal{H}) \leq \frac{\max\{\frac{2B}{\delta}, 2\}}{p_{\min}(\delta)} \big(\mathcal{E}_{\ell_\delta}(h) - \mathcal{E}^*_{\ell_\delta}(\mathcal{H}) + \mathcal{M}_{\ell_\delta}(\mathcal{H})\big).$$

The proof is presented in Appendix B.1. It leverages the general Theorem 1 with $\alpha(h, x) = \mathbb{P}(0 \leq \mu(x) - y \leq \delta \mid x)$. Note that the previous established general tools for $\mathcal{H}$-consistency bounds (Awasthi et al., 2022a,b) require $\alpha$ to be constant, which is not applicable in this context. This underscores the necessity of generalizing previous tools to accommodate any positive function $\alpha$.

As shown by Corollary 6, when $\mathcal{H}$ is realizable, the minimizability gap vanishes. Thus, by Theorem 7, we obtain the following corollary.

**Corollary 8** *Assume that the distribution is symmetric, the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Assume that $p_{\min}(\delta) = \inf_{x \in \mathcal{X}} \mathbb{P}(0 \leq \mu(x) - y \leq \delta \mid x)$ is positive. Then, for all $h \in \mathcal{H}$, the following $\mathcal{H}$-consistency bound holds:*

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \leq \frac{\max\{\frac{2B}{\delta}, 2\}}{p_{\min}(\delta)} \big(\mathcal{E}_{\ell_\delta}(h) - \mathcal{E}^*_{\ell_\delta}(\mathcal{H})\big).$$

Corollary 8 implies the Bayes-consistency of the Huber loss when $p_{\min}(\delta) > 0$, by taking the limit on both sides of the bound. Note that, as the value of $\delta$ increases, $\frac{2B}{\delta}$ decreases and $p_{\min}(\delta)$ increases, which improves the linear dependency on the Huber loss estimation error in this bound. However, this comes at the price of an Huber loss more similar to the squared loss and thus a higher sensitivity to outliers. Thus, selecting an appropriate value for $\delta$ involves considering these trade-offs.

The bound is uninformative when the probability mass $p_{\min}(\delta)$ is zero. However, the following theorem shows that the condition $p_{\min}(\delta) > 0$ is necessary and that otherwise, in general, the Huber loss is not $\mathcal{H}$-consistent with respect to the squared loss.

**Theorem 9** *Assume that the distribution is symmetric, the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Then, the Huber loss $\ell_\delta$ is not $\mathcal{H}$-consistent with respect to the squared loss.*

Refer to Appendix B.1 for the proof, which consists of considering a distribution that concentrates on an input $x$ with $\mathbb{P}(Y = y \mid x) = \frac{1}{2} = \mathbb{P}(Y = 2\mu(x) - y \mid x)$, where $-B \le y < \mu(x) \le B$ and $\mu(x) - y > \delta$. Then, we show that both $\overline{h}\colon x \mapsto y + \delta$ and $h^*\colon x \mapsto \mu(x)$ are best-in-class predictors of the Huber loss, while the best-in-class-predictor of the squared loss is uniquely $h^*\colon x \mapsto \mu(x)$.

### 4.2. $\ell_p$ Loss

Here, we analyze $\ell_p$ loss functions for any $p \ge 1$: $\ell_p\colon (h, x, y) \mapsto |h(x) - y|^p$. We show that this family of loss functions benefits from $\mathcal{H}$-consistency bounds with respect to the squared loss assuming, when adopting the same symmetry and boundedness assumptions as in the previous section.

**Theorem 10** *Assume that the distribution is symmetric, and that the conditional distribution and the hypothesis set $\mathcal{H}$ are bounded by $B > 0$. Then, for all $h \in \mathcal{H}$ and $p \ge 1$, the following $\mathcal{H}$-consistency bound holds:*

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \mathcal{M}_{\ell_2}(\mathcal{H}) \le \Gamma\Big(\mathcal{E}_{\ell_p}(h) - \mathcal{E}^*_{\ell_p}(\mathcal{H}) + \mathcal{M}_{\ell_p}(\mathcal{H})\Big),$$

*where $\Gamma(t) = \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} \{|h(x) - y| + |\mu(x) - y|\}\, t$ for $p = 1$, $\Gamma(t) = \frac{2}{(8B)^{p-2} p(p-1)}\, t$ for $p \in (1, 2]$, and $\Gamma(t) = t^{\frac{2}{p}}$ for $p \ge 2$.*

The proof is included in Appendix B.2. Note that for $p = 1$, $\Gamma$ can be further upper bounded as follows: $\Gamma(t) = \sup_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \{|h(x) - y| + |\mu(x) - y|\}\, t \le 4B\, t$ since the conditional distribution and the hypothesis set $\mathcal{H}$ are bounded by $B > 0$. This upper bound can also be obtained by using general theorems in Section 3 with $\alpha \equiv 1$. However, our generalized theorems, which apply to any positive function $\alpha$, yield a finer bound for the $\ell_1$ loss. This further shows that our generalized theorems are not only useful but can also yield finer bounds.

The key term appearing in the bounds is the minimizability gap $\mathcal{M}_{\ell_p}(\mathcal{H}) = \mathcal{E}^*_{\ell_p}(\mathcal{H}) - \mathbb{E}_x\Big[\mathcal{C}^*_{\ell_p}(\mathcal{H}, x)\Big]$, which is helpful for comparing the bounds between $\ell_p$ losses for different $p \ge 1$. For example, for the $\ell_1$ and $\ell_2$ loss, by Theorem 4, we have $\mathcal{M}_{\ell_1}(\mathcal{H}) = \mathcal{E}^*_{\ell_1}(\mathcal{H}) - \mathbb{E}_x[\mathbb{E}_y[|\mu(x) - y|]]$ and $\mathcal{M}_{\ell_2}(\mathcal{H}) = \mathcal{E}^*_{\ell_2}(\mathcal{H}) - \mathbb{E}_x\big[\mathbb{E}_y\big[|\mu(x) - y|^2\big]\big]$. Thus, in the deterministic case, both $\mathbb{E}_y[|\mu(x) - y|]$ and $\mathbb{E}_y\big[|\mu(x) - y|^2\big]$ vanish, and $\mathcal{M}_{\ell_2} = \mathcal{E}^*_{\ell_2}(\mathcal{H}) \ge \big(\mathcal{E}^*_{\ell_1}(\mathcal{H})\big)^2 = (\mathcal{M}_{\ell_1})^2$.

In particular, when $\mathcal{H}$ is realizable, we have $\mathcal{M}_{\ell_p}(\mathcal{H}) = \mathcal{M}_{\ell_2}(\mathcal{H}) = 0$. This yields the following result.

**Corollary 11** *Assume that the distribution is symmetric, the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Then, for all $h \in \mathcal{H}$ and $p \geq 1$, the following $\mathcal{H}$-consistency bound holds:*

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \leq \Gamma\Big(\mathcal{E}_{\ell_p}(h) - \mathcal{E}^*_{\ell_p}(\mathcal{H})\Big),$$

*where $\Gamma(t) = \sup_{x \in \mathcal{X}, y \in \mathcal{Y}}\{|h(x) - y| + |\mu(x) - y|\}\, t$ for $p = 1$, $\Gamma(t) = \frac{2}{(8B)^{p-2}p(p-1)}\, t$ for $p \in (1, 2]$, and $\Gamma(t) = t^{\frac{2}{p}}$ for $p \geq 2$.*

Corollary 11 shows that when the estimation error of $\ell_p$ is reduced to $\epsilon$, the estimation error of the squared loss $\big(\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H})\big)$ is upper bounded by $\epsilon^{\frac{2}{p}}$ for $p > 2$, and by $\epsilon$ for $1 \leq p \leq 2$, which is more favorable, modulo a multiplicative constant.

### 4.3. Squared $\epsilon$-insensitive Loss

The $\epsilon$-insensitive loss and the squared $\epsilon$-insensitive loss functions are used in the support vector regression (SVR) algorithms (Vapnik, 2000). The use of these loss functions results in sparser solutions, characterized by fewer support vectors for the SVR algorithms. Moreover, the selection of the parameter $\epsilon$ determines a trade-off between accuracy and sparsity: larger $\epsilon$ values yield increasingly sparser solutions. We first provide a positive result for the squared $\epsilon$-insensitive loss $\ell_{\text{sq}-\epsilon}\colon (h, x, y) \mapsto \max\big\{|h(x) - y|^2 - \epsilon^2, 0\big\}$, by showing that it admits an $\mathcal{H}$-consistency bound with respect to $\ell_2$.

**Theorem 12** *Assume that the distribution is symmetric, and that the conditional distribution and the hypothesis set $\mathcal{H}$ are bounded by $B > 0$. Assume that $p_{\min}(\epsilon) = \inf_{x \in \mathcal{X}} \mathbb{P}(\mu(x) - y \geq \epsilon \mid x)$ is positive. Then, for all $h \in \mathcal{H}$, the following $\mathcal{H}$-consistency bound holds:*

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \mathcal{M}_{\ell_2}(\mathcal{H}) \leq \frac{\mathcal{E}_{\ell_{\text{sq}-\epsilon}}(h) - \mathcal{E}^*_{\ell_{\text{sq}-\epsilon}}(\mathcal{H}) + \mathcal{M}_{\ell_{\text{sq}-\epsilon}}(\mathcal{H})}{2p_{\min}(\epsilon)}.$$

The proof is presented in Appendix B.3. It requires the use of Theorem 1 with $\alpha(h, x) = \mathbb{P}(\mu(x) - y \geq \epsilon \mid x)$. As in the case of the Huber loss, the previous established general tools for $\mathcal{H}$-consistency bounds (Awasthi et al., 2022a,b) do not apply here. Our generalization of previous tools proves essential for analyzing $\mathcal{H}$-consistency bounds in regression. By Corollary 6, for realizable hypothesis sets, the minimizability gap vanishes. Thus, by Theorem 12, we obtain the following corollary.

**Corollary 13** *Assume that the distribution is symmetric, the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Assume that $p_{\min}(\epsilon) = \inf_{x \in \mathcal{X}} \mathbb{P}(\mu(x) - y \geq \epsilon \mid x)$ is positive. Then, for all $h \in \mathcal{H}$, the following $\mathcal{H}$-consistency bound holds:*

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \leq \frac{\mathcal{E}_{\ell_{\text{sq}-\epsilon}}(h) - \mathcal{E}^*_{\ell_{\text{sq}-\epsilon}}(\mathcal{H})}{2p_{\min}(\epsilon)}.$$

By taking the limit on both sides of the bound of Corollary 13, we can infer the $\mathcal{H}$-consistency of the squared $\epsilon$-insensitive loss under the assumption $p_{\min}(\epsilon) > 0$. Note that increasing $\epsilon$ diminishes

$p_{\min}(\epsilon)$, making the bound less favorable. Conversely, smaller $\epsilon$ values enhance the linear dependency bound but may hinder solution sparsity. Therefore, selecting the optimal $\epsilon$ involves balancing the trade-off between linear dependency and sparsity. When $p_{\min}(\epsilon)$ approaches zero, the bound derived from Corollary 13 becomes less informative. However, as demonstrated in the subsequent theorem, the squared $\epsilon$-insensitive loss fails to exhibit $\mathcal{H}$-consistency with the squared loss if the condition $p_{\min}(\epsilon) > 0$ is not satisfied.

**Theorem 14** *Assume that the distribution is symmetric, the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Then, the squared $\epsilon$-insensitive loss $\ell_{\mathrm{sq}-\epsilon}$ is not $\mathcal{H}$-consistent.*

The proof is given in Appendix B.3. It consists of considering a distribution that concentrates on an input $x$ with $\mathbb{P}(Y = y \mid x) = \frac{1}{2} = \mathbb{P}(Y = 2\mu(x) - y \mid x)$, where $-B \leq y < \mu(x) \leq B$ and $\mu(x) - y < \epsilon$. Then, we show that both $\overline{h}: x \mapsto y + \epsilon$ and $h^*: x \mapsto \mu(x)$ are best-in-class predictors of the squared $\epsilon$-insensitive loss, while the best-in-class-predictor of the squared loss is uniquely $h^*: x \mapsto \mu(x)$.

### 4.4. $\epsilon$-Insensitive Loss

In Appendix B.4, we present negative results, Theorem 22 and Theorem 23, for the $\epsilon$-insensitive loss $\ell_\epsilon: (h, x, y) \mapsto \max\{|h(x) - y| - \epsilon, 0\}$ used in the SVR algorithm, by showing that even under the assumption $\inf_{x \in \mathcal{X}} \mathbb{P}(\mu(x) - y \geq \epsilon) > 0$ or $\inf_{x \in \mathcal{X}} \mathbb{P}(0 \leq \mu(x) - y \leq \epsilon) > 0$, it is not $\mathcal{H}$-consistent with respect to the squared loss.

### 4.5. Generalization bounds

We can use our $\mathcal{H}$-consistency bounds to derive bounds on the squared loss estimation error of a surrogate loss minimizer. For a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ drawn i.i.d. according to $\mathcal{D}$, let $\widehat{h}_S \in \mathcal{H}$ be the empirical minimizer of a regression loss function $L$ over $S$ and $\mathfrak{R}_m^L(\mathcal{H})$ the Rademacher complexity of the hypothesis set $\{(x, y) \mapsto \mathsf{L}(h(x), y) : h \in \mathcal{H}\}$. We denote by $B_L$ an upper bound of the regression loss function $L$. Then, the following generalization bound holds.

**Theorem 15** *Assume that the distribution is symmetric, the conditional distribution and the hypothesis set $\mathcal{H}$ are bounded by $B > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample $S$ of size $m$, the following squared loss estimation bound holds for $\widehat{h}_S$:*

$$\mathcal{E}_{\ell_2}(\widehat{h}_S) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \leq \Gamma\left(\mathcal{M}_L(\mathcal{H}) + 4\mathfrak{R}_m^L(\mathcal{H}) + 2B_L\sqrt{\frac{\log \frac{2}{\delta}}{2m}}\right) - \mathcal{M}_{\ell_2}(\mathcal{H}).$$

*where $\Gamma(t) = \sup_{x \in \mathcal{X}} \sup_y \left\{|\widehat{h}_S(x) - y| + |\mu(x) - y|\right\} t$ for $L = \ell_1$, $\Gamma(t) = \frac{2}{(8B)^{p-2} p(p-1)} t$ for $L = \ell_p$, $p \in (1, 2]$, $\Gamma(t) = t^{\frac{2}{p}}$ for $L = \ell_p$, $p \geq 2$, $\Gamma(t) = \frac{\max\{\frac{2B}{\delta}, 2\}}{p_{\min}(\delta)} t$ for $L = \ell_\delta$, and $\Gamma(t) = \frac{1}{2p_{\min}(\epsilon)} t$ for $L = \ell_{\mathrm{sq}-\epsilon}$.*

The proof is included in Appendix C. Theorem 15 provides the first finite-sample guarantees for the squared loss estimation error of the empirical minimizer of the Huber loss, squared $\epsilon$-insensitive loss, $\ell_1$ loss, and more generally $\ell_p$ loss. The proof leverages the $\mathcal{H}$-consistency bounds of Theorems 7, 10, 12, along with standard Rademacher complexity bounds (Mohri et al., 2018). Under the boundedness assumption, we have $|h(x) - y| \leq |h(x)| + |y| \leq 2B$. Thus, an upper bound $B_L$ for the regression loss function can be derived. For example, for the $\ell_p$ loss, we have $|h(x) - y|^p \leq (2B)^p$ and thus $B_L = (2B)^p$.

## 5. Application to adversarial regression

In this section, we show how the $\mathcal{H}$-consistency guarantees we presented in the previous section can be applied to the design of new algorithms for adversarial regression. Deep neural networks are known to be vulnerable to small adversarial perturbations around input data (Krizhevsky et al., 2012; Szegedy et al., 2013; Sutskever et al., 2014; Awasthi et al., 2023, 2024).

Despite extensive previous work aimed at improving the robustness of neural networks, this often comes with a reduction in standard (non-adversarial) accuracy, leading to a trade-off between adversarial and standard generalization errors in both the classification (Madry et al., 2017; Tsipras et al., 2018; Zhang et al., 2019; Raghunathan et al., 2019; Min et al., 2021; Javanmard and Soltanolkotabi, 2022; Ma et al., 2022; Taheri et al., 2022; Dobriban et al., 2023) and regression scenarios (Javanmard et al., 2020; Dan et al., 2020; Xing et al., 2021; Hassani and Javanmard, 2022; Liu et al., 2023; Ribeiro and Schön, 2023; Ribeiro et al., 2023).

In the context of adversarial classification, Zhang et al. (2019) proposed algorithms seeking a trade-off between these two types of errors, by using the theory of Bayes-consistent binary classification surrogate losses functions. More recently, Mao et al. (2023f) introduced enhanced algorithms by minimizing smooth adversarial comp-sum losses, leveraging the $\mathcal{H}$-consistency guarantee of comp-sum losses in multi-class classification.

Building on the insights from these previous studies, we aim to leverage our novel $\mathcal{H}$-consistency theory tailored for standard regression to introduce a family of new loss functions for adversarial regression, termed as *smooth adversarial regression losses*. Minimizing these loss functions readily leads to new algorithms for adversarial regression.

### 5.1. Adversarial Squared Loss

In adversarial regression, the target adversarial generalization error is measured by the worst squared loss under the bounded $\gamma$ perturbation of $x$. This is defined for any $(h, x, y) \in \mathcal{H} \times \mathcal{X} \times \mathcal{Y}$ as follows:

$$\widetilde{\ell}_2(h, x, y) = \sup_{x' : \|x' - x\| \leq \gamma} \left(h(x') - y\right)^2$$

where $\|\cdot\|$ denotes a norm on $\mathcal{X}$, typically an $\ell_p$ norm for $p \geq 1$. We refer to $\widetilde{\ell}_2$ as the *adversarial squared loss*. By adding and subtracting the standard squared loss $\ell_2$, for any $(h, x, y) \in \mathcal{H} \times \mathcal{X} \times \mathcal{Y}$, we can write $\widetilde{\ell}_2$ as follows: $\widetilde{\ell}_2(h, x, y) = (h(x) - y)^2 + \sup_{x' : \|x' - x\| \leq \gamma} (h(x') - y)^2 - (h(x) - y)^2$. Then, assuming that the conditional distribution and the hypothesis set $\mathcal{H}$ are bounded by $B > 0$, we can write

$$\begin{aligned}
&\sup_{x' : \|x' - x\| \leq \gamma} \left(h(x') - y\right)^2 - \left(h(x) - y\right)^2 \\
&= \sup_{x' : \|x' - x\| \leq \gamma} \big(h(x') - h(x)\big)\big(h(x') + h(x) + y\big) \\
&\leq \sup_{x' : \|x' - x\| \leq \gamma} 3B\left|h(x') - h(x)\right|. && (|h(x)| \leq B, |y| \leq B)
\end{aligned}$$

Thus, the $\widetilde{\ell}_2$ loss can be upper bounded as follows for all $(x, y)$:

$$\widetilde{\ell}_2(h, x, y) \leq (h(x) - y)^2 + \nu \sup_{x' : \|x' - x\| \leq \gamma} \left| h(x') - h(x) \right|, \tag{1}$$

where $\nu \geq 3B$ is a positive constant.

### 5.2. Smooth Adversarial Regression Losses

Let $L$ be a standard regression loss function that admits an $\mathcal{H}$-consistency bound with respect to the squared loss:

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \leq \Gamma(\mathcal{E}_L(h) - \mathcal{E}^*_L(\mathcal{H})).$$

To trade-off the adversarial and standard generalization errors in regression, by using (1), we can upper bound the difference between the adversarial generalization and the best-in-class standard generalization error as follows:

$$\begin{aligned}
&\mathcal{E}_{\widetilde{\ell}_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \\
&\leq \mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \nu \sup_{x' : \|x' - x\| \leq \gamma} \left| h(x') - h(x) \right| \\
&\leq \Gamma(\mathcal{E}_L(h) - \mathcal{E}^*_L(\mathcal{H})) + \nu \sup_{x' : \|x' - x\| \leq \gamma} \left| h(x') - h(x) \right|.
\end{aligned}$$

Thus, by Corollaries 8, 11 and 13, we obtain the following guarantees with respect to the adversarial squared loss. The proofs are presented in Appendix D.

**Theorem 16** *Assume that the distribution is symmetric, the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Assume that $p_{\min}(\delta) = \inf_{x \in \mathcal{X}} \mathbb{P}(0 \leq \mu(x) - y \leq \delta \mid x)$ is positive. Then, for any $\nu \geq 3B$ and all $h \in \mathcal{H}$, the following bound holds:*

$$\mathcal{E}_{\widetilde{\ell}_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \leq \frac{\max\left\{\frac{2B}{\delta}, 2\right\}}{p_{\min}(\delta)} \left( \mathcal{E}_{\ell_\delta}(h) - \mathcal{E}^*_{\ell_\delta}(\mathcal{H}) \right) + \nu \sup_{x' : \|x' - x\| \leq \gamma} \left| h(x') - h(x) \right|$$

**Theorem 17** *Assume that the distribution is symmetric, the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Then, for any $\nu \geq 3B$ and all $h \in \mathcal{H}$, the following bound holds:*

$$\mathcal{E}_{\widetilde{\ell}_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \leq \Gamma\left( \mathcal{E}_{\ell_p}(h) - \mathcal{E}^*_{\ell_p}(\mathcal{H}) \right) + \nu \sup_{x' : \|x' - x\| \leq \gamma} \left| h(x') - h(x) \right|,$$

*where $\Gamma(t) = t^{\frac{2}{p}}$ if $p \geq 2$, $\frac{2}{(8B)^{p-2} p(p-1)}\, t$ for $p \in (1, 2)$ and $4B\, t$, if $p = 1$.*

**Theorem 18** *Assume that the distribution is symmetric, the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Assume that $p_{\min}(\epsilon) = \inf_{x \in \mathcal{X}} \mathbb{P}(\mu(x) - y \geq \epsilon \mid x)$ is positive. Then, for any $\nu \geq 3B$ and all $h \in \mathcal{H}$, the following bound holds:*

$$\mathcal{E}_{\widetilde{\ell}_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \leq \frac{\mathcal{E}_{\ell_{\mathrm{sq}-\epsilon}}(h) - \mathcal{E}^*_{\ell_{\mathrm{sq}-\epsilon}}(\mathcal{H})}{2 p_{\min}(\epsilon)} + \nu \sup_{x' : \|x' - x\| \leq \gamma} \left| h(x') - h(x) \right|.$$

Table 1: Comparison of the performance of the ADV-SQ algorithm and our smooth adversarial regression algorithms for $L = \ell_2$ and $L = \ell_\delta$ for $\ell_\infty$ adversarial training with perturbation size $\gamma \in \{0.001, 0.005, 0.01\}$ on the Diverse MAGIC wheat dataset.

| METHOD | SIZE | CLEAN | ROBUST |
|---|---|---|---|
| ADV-SQ | 0.001 | **1.28 ± 0.10** | **1.32 ± 0.11** |
| OURS ($L = \ell_2$) | | **1.28 ± 0.09** | **1.32 ± 0.09** |
| OURS ($L = \ell_\delta$) | | 1.30 ± 0.08 | 1.34 ± 0.09 |
| ADV-SQ | 0.005 | 1.30 ± 0.09 | 1.53 ± 0.10 |
| OURS ($L = \ell_2$) | | 1.26 ± 0.09 | 1.46 ± 0.10 |
| OURS ($L = \ell_\delta$) | | **1.03 ± 0.09** | **1.12 ± 0.10** |
| ADV-SQ | 0.01 | 1.30 ± 0.08 | 1.78 ± 0.11 |
| OURS ($L = \ell_2$) | | 1.22 ± 0.11 | 1.62 ± 0.14 |
| OURS ($L = \ell_\delta$) | | **0.97 ± 0.02** | **1.01 ± 0.02** |

Theorems 16, 17 and 18 suggest minimizing

$$L(h, x, y) + \tau \sup_{x' : \|x' - x\| \le \gamma} \left| h(x') - h(x) \right| \tag{2}$$

where $L$ can be chosen as $\ell_\delta$, $\ell_p$ and $\ell_{\text{sq}-\epsilon}$ and $\tau > 0$ is a parameter. For simplicity, we use the parameter $\tau$ to approximate the effect of the functional form $\Gamma$ in these bounds, as with the approach adopted in (Zhang et al., 2019). Given that $L$ is a convex function of $h$, the minimization of (2) can be achieved equivalently and more efficiently by applying the standard Lagrange method. This allows for the replacement of the $\ell_1$ norm with its square, since the regularization term can be moved to a constraint, where it can then be squared.

We refer to (2) as *smooth adversarial regression loss functions*. They can be obtained by augmenting the standard regression loss function such as the Huber loss, the $\ell_p$ loss and the $\epsilon$-insensitive loss with a natural smoothness term. Minimizing the regularized empirical smooth adversarial regression loss functions leads to a new family of algorithms for adversarial regression, *smooth adversarial regression algorithms*. In the next section, we report experimental results illustrating the effectiveness of these new algorithms, in particular in terms of the trade-off between the adversarial accuracy and standard accuracy, as guaranteed by Theorems 16, 17 and 18. We will show that these algorithms outperform the direct minimization of the adversarial squared loss.

It is important to note that the regularizer $\sup_{x' : \|x' - x\| \le \gamma} |h(x') - h(x)|$ in the smooth adversarial regression loss closely relates to the local Lipschitz constant and the gradient norm, which are established methods in adversarially robust training (Hein and Andriushchenko, 2017; Finlay and Oberman, 2019; Yang et al., 2020; Gouk et al., 2021). Furthermore, by building upon the derivation in Section 5.1, we can develop new surrogate losses for adversarial regression scenarios beyond the adversarial squared loss, such as the adversarial $\ell_1$ loss. In this context, the formulation of the smooth adversarial regression loss would replace the absolute value with the local Lipschitz constant of the target loss. To establish guarantees for these new surrogate losses, the $\mathcal{H}$-consistency bounds shown in Section 4 can be extended to other target losses in regression, such as the $\ell_1$ loss. An intriguing direction for future exploration is investigating how our surrogate losses relate to Moreau envelope theory (see, for example, (Zhou et al., 2022)).

Table 2: Comparison of the performance of the ADV-SQ algorithm and our smooth adversarial regression algorithms for $L = \ell_2$ and $L = \ell_\delta$ for $\ell_\infty$ adversarial training with perturbation size $\gamma \in \{0.001, 0.005, 0.01\}$ on the Diabetes dataset.

| METHOD | SIZE | CLEAN | ROBUST |
|---|---|---|---|
| ADV-SQ | 0.001 | $2.53 \pm 0.48$ | $2.57 \pm 0.49$ |
| OURS ($L = \ell_2$) | | $\mathbf{1.24 \pm 0.21}$ | $\mathbf{1.26 \pm 0.21}$ |
| OURS ($L = \ell_\delta$) | | $1.31 \pm 0.15$ | $1.32 \pm 0.15$ |
| ADV-SQ | 0.005 | $1.12 \pm 0.12$ | $1.18 \pm 0.13$ |
| OURS ($L = \ell_2$) | | $0.80 \pm 0.04$ | $0.82 \pm 0.04$ |
| OURS ($L = \ell_\delta$) | | $\mathbf{0.78 \pm 0.06}$ | $\mathbf{0.79 \pm 0.06}$ |
| ADV-SQ | 0.01 | $0.83 \pm 0.05$ | $0.87 \pm 0.05$ |
| OURS ($L = \ell_2$) | | $\mathbf{0.74 \pm 0.05}$ | $\mathbf{0.76 \pm 0.05}$ |
| OURS ($L = \ell_\delta$) | | $0.81 \pm 0.05$ | $0.82 \pm 0.05$ |

## 6. Experiments

In this section, we demonstrate empirically the effectiveness of the smooth adversarial regression algorithms introduced in the previous section.

**Experimental settings.** We studied two real-world datasets: the Diabetes dataset (Efron et al., 2004) and the Diverse MAGIC wheat dataset (Scott et al., 2021), and adopted the same exact settings for feature engineering as (Ribeiro et al., 2023, Example 3 and Example 5 in Appendix D). For the sake of a fair comparison, we used a linear hypothesis set. We considered an $\ell_\infty$ perturbation with perturbation size $\gamma \in \{0.001, 0.005, 0.01\}$ for adversarial training. For our smooth adversarial regression losses (2), we chose $L = \ell_2$, the squared loss, and $L = \ell_\delta$ with $\delta = 0.2$, the Huber loss, setting $\tau = 1$ as the default. Other choices for the regression loss functions and the value of $\tau$ may yield better performance, which can typically be selected by cross-validation in practice. Both our smooth adversarial regression losses and the adversarial squared loss were optimized using the CVXPY library (Diamond and Boyd, 2016).

**Evaluation.** We report the standard error, measured by the squared loss (or MSE) on the test data, and the robust error, measured by the adversarial squared loss with $\ell_\infty$ perturbation and the corresponding perturbation size used for training. We averaged both errors over five runs and report the standard deviation for both our smooth adversarial regression losses and the adversarial squared loss.

**Results.** Tables 1 and 2 present the experimental results of our adversarial regression algorithms with both the squared ($L = \ell_2$) and Huber ($L = \ell_\delta$) losses. The results suggest that these algorithms consistently surpass the adversarial squared loss in clean and robust error metrics across various settings. In particular, on the Diabetes dataset with a perturbation size of $\gamma = 0.01$, our method ($L = \ell_2$) outperforms the adversarial squared loss by more than $0.5\%$ in both robust error and clean error. Similarly, on the Diverse MAGIC wheat dataset with a perturbation size of $\gamma = 0.01$, our method ($L = \ell_\delta$) surpasses the adversarial squared loss by more than $0.3\%$ in terms of robust and clean errors.

Remarkably, the surrogate loss using the Huber loss occasionally outperforms the squared loss variant. This highlights the importance of using surrogate losses, even within the adversarial training framework, to enhance performance.

## 7. Conclusion

We presented the first study of $\mathcal{H}$-consistency bounds for regression.This involved generalizing existing tools that were previously used to prove $\mathcal{H}$-consistency bounds. Leveraging our generalized tools, we proved a series of novel $\mathcal{H}$-consistency bounds for surrogate losses of the squared loss. Our $\mathcal{H}$-consistency guarantees can be beneficial in designing new algorithms for adversarial regression. This study can be useful for the later studies of other surrogate losses for other target losses in regression.

## References

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. $H$-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, pages 1117–1174, 2022a.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-class $H$-consistency bounds. In *Advances in neural information processing systems*, pages 782–795, 2022b.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for adversarial robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 10077–10094, 2023.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. DC-programming for neural network optimizations. *Journal of Global Optimization*, pages 1–17, 2024.

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

Mathieu Blondel. Structured prediction with projection oracles. In *Advances in neural information processing systems*, 2019.

A. Caponnetto. A note on the role of squared loss in regression. Technical report, Massachusetts Institute of Technology, 2005.

Andreas Christmann and Ingo Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3), 2007.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Advances in neural information processing systems*, 2016.

James A Clarkson. Uniformly convex spaces. *Transactions of the American Mathematical Society*, 40(3):396–414, 1936.

Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pages 2345–2355, 2020.

Steven Diamond and Stephen Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, 17(1):2909–2913, 2016.

Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. *IEEE Transactions on Information Theory*, 2023.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–451, 2004.

Chris Finlay and Adam M Oberman. Scaleable input gradient regularization for adversarial robustness. *arXiv preprint arXiv:1905.11468*, 2019.

Jessica Finocchiaro, Rafael Frongillo, and Bo Waggoner. An embedding framework for consistent polyhedral surrogates. In *Advances in neural information processing systems*, 2019.

Rafael Frongillo and Bo Waggoner. Surrogate regret bounds for polyhedral losses. In *Advances in Neural Information Processing Systems*, pages 21569–21580, 2021.

Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In *Conference on learning theory*, pages 341–358, 2011.

Wei Gao and Zhi-Hua Zhou. On the consistency of AUC pairwise optimization. In *International Joint Conference on Artificial Intelligence*, 2015.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110:393–416, 2021.

Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. *arXiv preprint arXiv:2201.05149*, 2022.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in neural information processing systems*, 2017.

P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist*, 35:73—-101, 1964.

Adel Javanmard and Mahdi Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics*, 50(4):2127–2156, 2022.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078, 2020.

Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon. Consistent multilabel classification. In *Advances in Neural Information Processing Systems*, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Changyu Liu, Yuling Jiao, Junhui Wang, and Jian Huang. Non-asymptotic bounds for adversarial excess risk under misspecified models. *arXiv preprint arXiv:2309.00771*, 2023.

Phil Long and Rocco Servedio. Consistency versus realizable H-consistency for multiclass classification. In *International Conference on Machine Learning*, pages 801–809, 2013.

Xinsong Ma, Zekai Wang, and Weiwei Liu. On the tradeoff between robustness and fairness. In *Advances in Neural Information Processing Systems*, pages 26230–26241, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *Advances in neural information processing systems*, 2023a.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds: Characterization and extensions. In *Advances in Neural Information Processing Systems*, 2023b.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds for pairwise misranking loss surrogates. In *International conference on Machine learning*, 2023c.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Ranking with abstention. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023d.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Structured prediction with stronger consistency guarantees. In *Advances in Neural Information Processing Systems*, 2023e.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International Conference on Machine Learning*, 2023f.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with multiple experts. In *International Symposium on Artificial Intelligence and Mathematics*, 2024a.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *Algorithmic Learning Theory*, 2024b.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In *International Conference on Artificial Intelligence and Statistics*, 2024c.

Aditya Krishna Menon and Robert C Williamson. Bayes-optimal scorers for bipartite ranking. In *Conference on Learning Theory*, pages 68–106, 2014.

Yifei Min, Lin Chen, and Amin Karbasi. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. In *Uncertainty in Artificial Intelligence*, pages 129–139, 2021.

Christopher Mohri, Daniel Andor, Eunsol Choi, Michael Collins, Anqi Mao, and Yutao Zhong. Learning to reject with a fixed predictor: Application to decontextualization. In *International Conference on Learning Representations*, 2024.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Harikrishna Narasimhan, Harish Ramaswamy, Aadirupa Saha, and Shivani Agarwal. Consistent multiclass algorithms for complex performance measures. In *International Conference on Machine Learning*, pages 2398–2407, 2015.

Anton Osokin, Francis Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*, 2017.

Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18:1–35, 2017.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

Antônio H Ribeiro and Thomas B Schön. Overparameterized linear regression under adversarial attacks. *IEEE Transactions on Signal Processing*, 71:601–614, 2023.

Antonio H Ribeiro, Dave Zachariah, Francis Bach, and Thomas B Schön. Regularization properties of adversarially-trained linear regression. In *Advances in Neural Information Processing Systems*, 2023.

Michael F Scott, Nick Fradgley, Alison R Bentley, Thomas Brabbs, Fiona Corke, Keith A Gardner, Richard Horsnell, Phil Howell, Olufunmilayo Ladejobi, Ian J Mackay, et al. Limited haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding. *Genome Biology*, 22(1):1–30, 2021.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Asymptotic behavior of adversarial training in binary linear classification. In *IEEE International Symposium on Information Theory (ISIT)*, pages 127–132, 2022.

Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025, 2007.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Kazuki Uematsu and Yoonkyung Lee. On theoretically optimal ranking functions in bipartite ranking. *Journal of the American Statistical Association*, 112(519):1311–1322, 2017.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 2000.

Yutong Wang and Clayton Scott. Weston-Watkins hinge loss and ordered partitions. In *Advances in neural information processing systems*, pages 19873–19883, 2020.

Yutong Wang and Clayton D Scott. On classification-calibration of gamma-phi losses. *arXiv preprint arXiv:2302.07321*, 2023.

Yue Xing, Ruizhi Zhang, and Guang Cheng. Adversarially robust estimate and risk analysis in linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 514–522, 2021.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. In *Advances in neural information processing systems*, pages 8588–8601, 2020.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.

Mingyuan Zhang and Shivani Agarwal. Bayes consistency vs. H-consistency: The interplay between surrogate loss functions and the scoring function class. In *Advances in Neural Information Processing Systems*, pages 16927–16936, 2020.

Mingyuan Zhang, Harish Guruprasad Ramaswamy, and Shivani Agarwal. Convex calibrated surrogates for the multi-label f-measure. In *International Conference on Machine Learning*, pages 11246–11255, 2020.

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

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004b.

Chenyu Zheng, Guoqiang Wu, Fan Bao, Yue Cao, Chongxuan Li, and Jun Zhu. Revisiting discriminative vs. generative classifiers: Theory and implications. *arXiv preprint arXiv:2302.02334*, 2023.

Lijia Zhou, Frederic Koehler, Pragya Sur, Danica J Sutherland, and Nati Srebro. A non-asymptotic moreau envelope theory for high-dimensional generalized linear models. In *Advances in Neural Information Processing Systems*, pages 21286–21299, 2022.

## Contents of Appendix

**A Proofs of general $\mathcal{H}$-consistency theorems** **20**

A.1 Proof of Theorem 1 . . . . . . 20

A.2 Proof of Theorem 2 . . . . . . 20

A.3 Proof of Theorem 3 . . . . . . 21

A.4 Proof of Theorem 4 . . . . . . 22

**B Proofs of $\mathcal{H}$-consistency bounds for common surrogate losses** **22**

B.1 $\mathcal{H}$-consistency of $\ell_\delta$ with respect to $\ell_2$ . . . . . . 22

B.2 $\mathcal{H}$-consistency of $\ell_p$ with respect to $\ell_2$ . . . . . . 27

B.3 $\mathcal{H}$-consistency of $\ell_{\mathrm{sq}-\epsilon}$ with respect to $\ell_2$ . . . . . . 29

B.4 $\mathcal{H}$-consistency of $\ell_\epsilon$ with respect to $\ell_2$ . . . . . . 31

**C Proofs of generalization bound** **33**

**D Proofs of adversarial regression** **33**

D.1 Proof of Theorem 16 . . . . . . 33

D.2 Proof of Theorem 17 . . . . . . 34

D.3 Proof of Theorem 18 . . . . . . 34

# Appendix A. Proofs of general $\mathcal{H}$-consistency theorems

## A.1. Proof of Theorem 1

**Theorem 1 (General $\mathcal{H}$-consistency bound – convex function)** *Let $\mathcal{D}$ denote a distribution over $\mathcal{X} \times \mathcal{Y}$. Assume that there exists a convex function $\Psi \colon \mathbb{R}_+ \to \mathbb{R}$ with $\Psi(0) \geq 0$, a positive function $\alpha \colon \mathcal{H} \times \mathcal{X} \to \mathbb{R}_+^*$ with $\sup_{x \in \mathcal{X}} \alpha(h, x) < +\infty$ for all $h \in \mathcal{H}$, and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \mathcal{X}$: $\Psi\big(\big[\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)\big]_\epsilon\big) \leq \alpha(h,x)\, \Delta\mathcal{C}_{L_1,\mathcal{H}}(h,x)$. Then, for any hypothesis $h \in \mathcal{H}$, the following inequality holds:*

$$\Psi\big(\mathcal{E}_{L_2}(h) - \mathcal{E}^*_{L_2}(\mathcal{H}) + \mathcal{M}_{L_2}(\mathcal{H})\big) \leq \Big[\sup_{x \in \mathcal{X}} \alpha(h,x)\Big]\big(\mathcal{E}_{L_1}(h) - \mathcal{E}^*_{L_1}(\mathcal{H}) + \mathcal{M}_{L_1}(\mathcal{H})\big) + \max\{\Psi(0), \Psi(\epsilon)\}.$$

**Proof** For any $h \in \mathcal{H}$, we can write

$$\begin{aligned}
&\Psi\big(\mathcal{E}_{L_2}(h) - \mathcal{E}^*_{L_2,\mathcal{H}} + \mathcal{M}_{L_2,\mathcal{H}}\big)\\
&= \Psi\Big(\mathbb{E}_X\big[\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)\big]\Big)\\
&\leq \mathbb{E}_X\big[\Psi\big(\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)\big)\big] && \text{(Jensen's ineq.)}\\
&= \mathbb{E}_X\Big[\Psi\Big(\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)1_{\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x) > \epsilon} + \Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)1_{\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x) \leq \epsilon}\Big)\Big]\\
&\leq \mathbb{E}_X\Big[\Psi\Big(\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)1_{\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x) > \epsilon}\Big) + \Psi\Big(\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)1_{\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x) \leq \epsilon}\Big)\Big] && (\Psi(0) \geq 0)\\
&\leq \mathbb{E}_X\big[\alpha(h,x)\Delta\mathcal{C}_{L_1,\mathcal{H}}(h,x)\big] + \sup_{t \in [0,\epsilon]} \Psi(t) && \text{(assumption)}\\
&\leq \Big[\sup_{x \in \mathcal{X}} \alpha(h,x)\Big] \mathbb{E}_x\big[\Delta\mathcal{C}_{L_1,\mathcal{H}}(h,x)\big] + \sup_{t \in [0,\epsilon]} \Psi(t) && \text{(Hölder's ineq.)}\\
&= \Big[\sup_{x \in \mathcal{X}} \alpha(h,x)\Big]\big(\mathcal{E}_{L_1}(h) - \mathcal{E}^*_{L_1,\mathcal{H}} + \mathcal{M}_{L_1,\mathcal{H}}\big) + \max\{\Psi(0), \Psi(\epsilon)\}, && \text{(convexity of } \Psi)
\end{aligned}$$

which completes the proof. ■

## A.2. Proof of Theorem 2

**Theorem 2 (General $\mathcal{H}$-consistency bound – concave function)** *Let $\mathcal{D}$ denote a distribution over $\mathcal{X} \times \mathcal{Y}$. Assume that there exists a concave function $\Gamma \colon \mathbb{R}_+ \to \mathbb{R}$, a positive function $\alpha \colon \mathcal{H} \times \mathcal{X} \to \mathbb{R}_+^*$ with $\sup_{x \in \mathcal{X}} \alpha(h,x) < +\infty$ for all $h \in \mathcal{H}$, and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \mathcal{X}$: $\big[\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)\big]_\epsilon \leq \Gamma\big(\alpha(h,x)\Delta\mathcal{C}_{L_1,\mathcal{H}}(h,x)\big)$. Then, for any hypothesis $h \in \mathcal{H}$, the following inequality holds*

$$\mathcal{E}_{L_2}(h) - \mathcal{E}^*_{L_2}(\mathcal{H}) + \mathcal{M}_{L_2}(\mathcal{H}) \leq \Gamma\Bigg(\Big[\sup_{x \in \mathcal{X}} \alpha(h,x)\Big]\big(\mathcal{E}_{L_1}(h) - \mathcal{E}^*_{L_1}(\mathcal{H}) + \mathcal{M}_{L_1}(\mathcal{H})\big)\Bigg) + \epsilon.$$

*In the special case where $\Gamma(x) = x^{\frac{1}{q}}$ for some $q \geq 1$ with conjugate $p \geq 1$, that is $\frac{1}{p} + \frac{1}{q} = 1$, for any $h \in \mathcal{H}$, the following inequality holds, assuming $\mathbb{E}_X\big[\alpha^{\frac{p}{q}}(h,x)\big]^{\frac{1}{p}} < +\infty$ for all $h \in \mathcal{H}$:*

$$\mathcal{E}_{L_2}(h) - \mathcal{E}^*_{L_2}(\mathcal{H}) + \mathcal{M}_{L_2}(\mathcal{H}) \leq \mathbb{E}_X\Big[\alpha^{\frac{p}{q}}(h,x)\Big]^{\frac{1}{p}} \mathbb{E}_X\big[\mathcal{E}_{L_1}(h) - \mathcal{E}^*_{L_1}(\mathcal{H}) + \mathcal{M}_{L_1}(\mathcal{H})\big]^{\frac{1}{q}} + \epsilon.$$

**Proof** For any $h \in \mathcal{H}$, we can write

$$\begin{aligned}
&\mathcal{E}_{L_2}(h) - \mathcal{E}^*_{L_2,\mathcal{H}} + \mathcal{M}_{L_2,\mathcal{H}} \\
&= \mathbb{E}_X\big[\mathcal{C}_{L_2}(h,x) - \mathcal{C}^*_{L_2,\mathcal{H}}(x)\big] \\
&= \mathbb{E}_X\big[\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)\big] \\
&= \mathbb{E}_X\Big[\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)1_{\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)>\epsilon} + \Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)1_{\Delta\mathcal{C}_{L_2,\mathcal{H}}(h,x)\le\epsilon}\Big] \\
&\le \mathbb{E}_X\big[\Gamma\big(\alpha(h,x)\Delta\mathcal{C}_{L_1,\mathcal{H}}(h,x)\big)\big] + \epsilon && \text{(assumption)} \\
&\le \Gamma\Big(\mathbb{E}_X\big[\alpha(h,x)\Delta\mathcal{C}_{L_1,\mathcal{H}}(h,x)\big]\Big) + \epsilon && \text{(Jensen's ineq.)} \\
&\le \Gamma\bigg(\Big[\sup_{x\in\mathcal{X}}\alpha(h,x)\Big]\mathbb{E}_X\big[\Delta\mathcal{C}_{L_1,\mathcal{H}}(h,x)\big]\bigg) + \epsilon && \text{(Hölder's ineq.)} \\
&= \Gamma\bigg(\Big[\sup_{x\in\mathcal{X}}\alpha(h,x)\Big]\big(\mathcal{E}_{L_1}(h) - \mathcal{E}^*_{L_1,\mathcal{H}} + \mathcal{M}_{L_1,\mathcal{H}}\big)\bigg) + \epsilon.
\end{aligned}$$

When $\Gamma(x) = x^{\frac{1}{q}}$ for some $q \ge 1$ with conjugate number $p$, starting from the fourth inequality above, we can write

$$\begin{aligned}
\mathcal{E}_{L_2}(h) - \mathcal{E}^*_{L_2,\mathcal{H}} + \mathcal{M}_{L_2,\mathcal{H}} &\le \mathbb{E}_X\Big[\alpha^{\frac{1}{q}}(h,x)\Delta\mathcal{C}^{\frac{1}{q}}_{L_2,\mathcal{H}}(h,x)\Big] + \epsilon \\
&\le \mathbb{E}_X\Big[\alpha^{\frac{p}{q}}(h,x)\Big]^{\frac{1}{p}}\mathbb{E}_X\big[\Delta\mathcal{C}_{L_1,\mathcal{H}}(h,x)\big]^{\frac{1}{q}} + \epsilon && \text{(Hölder's ineq)} \\
&= \mathbb{E}_X\Big[\alpha^{\frac{p}{q}}(h,x)\Big]^{\frac{1}{p}}\mathbb{E}_X\big[\mathcal{E}_{L_1}(h) - \mathcal{E}^*_{L_1,\mathcal{H}} + \mathcal{M}_{L_1,\mathcal{H}}\big]^{\frac{1}{q}} + \epsilon.
\end{aligned}$$

This completes the proof. ■

### A.3. Proof of Theorem 3

**Theorem 3** *Assume that the conditional distribution and the hypothesis set $\mathcal{H}$ are bounded by $B > 0$. Then, the best-in-class conditional error and the conditional regret of the squared loss can be characterized as: for all $h \in \mathcal{H}, x \in \mathcal{X}$,*

$$\begin{aligned}
\mathcal{C}^*_{\ell_2}(\mathcal{H},x) &= \mathcal{C}_{\ell_2}(\mu(x),x) = \mathbb{E}[y^2 \mid x] - (\mu(x))^2 \\
\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x) &= (h(x) - \mu(x))^2.
\end{aligned}$$

**Proof** By definition,

$$\begin{aligned}
\mathcal{C}^*_{\ell_2}(\mathcal{H}, x) &= \inf_{h \in \mathcal{H}} \mathbb{E}\big[(h(x) - y)^2 \mid x\big] \\
&= \inf_{h \in \mathcal{H}} \big[(h(x) - \mathbb{E}[y \mid x])^2 + \mathbb{E}[y^2 \mid x] - (\mathbb{E}[y \mid x])^2\big] \\
&= \mathbb{E}[y^2 \mid x] - (\mathbb{E}[y \mid x])^2 \\
\Delta\mathcal{C}_{\ell_2, \mathcal{H}}(h, x) &= \mathbb{E}\big[(h(x) - y)^2 \mid x\big] - \inf_{h \in \mathcal{H}} \mathbb{E}\big[(h(x) - y)^2 \mid x\big] \\
&= (h(x) - \mathbb{E}[y \mid x])^2 + \mathbb{E}[y^2 \mid x] - (\mathbb{E}[y \mid x])^2 - \big(\mathbb{E}[y^2 \mid x] - (\mathbb{E}[y \mid x])^2\big) \\
&= (h(x) - \mathbb{E}[y \mid x])^2.
\end{aligned}$$

This completes the proof. ■

### A.4. Proof of Theorem 4

**Theorem 4** *Let $\psi: \mathbb{R} \to \mathbb{R}$ be a symmetric function such that $\psi(x) = \psi(-x)$ for all $x \in \mathbb{R}$. Furthermore, $\psi(x) \geq 0$ for all $x$ in its domain and it holds that $\psi(0) = 0$. Assume that the conditional distribution and the hypothesis set $\mathcal{H}$ is bounded by $B > 0$. Assume that the distribution is symmetric and the regression loss function is given by $\mathsf{L}(y', y) = \psi(y' - y)$. Then, we have $\mathcal{C}^*_L(\mathcal{H}, x) = \mathcal{C}_L(\mu(x), x)$.*

**Proof** By the symmetry of the distribution, we can write

$$\begin{aligned}
\mathop{\mathbb{E}}_y[\psi(h(x) - y) \mid x] &= \frac{\mathbb{E}_y[\psi(h(x) - y) \mid x] + \mathbb{E}_y[\psi(h(x) - 2\mu(x) + y) \mid x]}{2} \\
&= \frac{\mathbb{E}_y[\psi(h(x) - y) \mid x] + \mathbb{E}_y[\psi(-h(x) + 2\mu(x) - y) \mid x]}{2} && (\psi \text{ is symmetric}) \\
&= \frac{\mathbb{E}_y[\psi(h(x) - y) + \psi(-h(x) + 2\mu(x) - y) \mid x]}{2} \\
&\geq \mathop{\mathbb{E}}_y[\psi(\mu(x) - y)] && (\text{Jensen's inequality})
\end{aligned}$$

where the equality is achieved when $h(x) = \mu(x) \in \mathcal{H}$. This completes the proof. ■

## Appendix B. Proofs of $\mathcal{H}$-consistency bounds for common surrogate losses

### B.1. $\mathcal{H}$-consistency of $\ell_\delta$ with respect to $\ell_2$

Define the function $g$ as $g: t \mapsto \frac{1}{2} t^2 1_{|t| \leq \delta} + \big(\delta |t| - \frac{1}{2}\delta^2\big) 1_{|t| > \delta}$. Consider the function $F$ defined over $[-B, B]^2$ by $F(x, y) = \frac{g(x+y) + g(x-y)}{2} - g(y)$. We prove a useful lemma as follows.

**Lemma 19** *For any $x, y \in [-B, B]$ and $|y| \leq \delta$, the following inequality holds:*

$$F(x, y) \geq \min\left\{\frac{\delta}{2B}, \frac{1}{4}\right\} x^2.$$

**Proof** Given the definition of $g$ and the symmetry of $F$ with respect to $y = 0$, we can assume, without loss of generality, that $y \geq 0$. Next, we will analyze case by case.

**Case I:** $|x + y| \leq \delta$, $|x - y| \leq \delta$, $0 \leq y \leq \delta$. In this case, we have

$$F(x, y) = \frac{\frac{1}{2}(x + y)^2 + \frac{1}{2}(x - y)^2}{2} - \frac{1}{2}y^2 = \frac{1}{2}x^2 \geq \min\left\{\frac{\delta}{2B}, \frac{1}{4}\right\}x^2.$$

**Case II:** $|x + y| \leq \delta$, $|x - y| > \delta$, $0 \leq y \leq \delta$. In this case, we must have $-y - \delta \leq x < y - \delta$ and $\delta \geq y \geq \max\{-x - \delta, x + \delta\} \geq x + \delta$. Thus,

$$\begin{aligned}
F(x, y) &= \frac{\frac{1}{2}(x + y)^2 + \delta|x - y| - \frac{1}{2}\delta^2}{2} - \frac{1}{2}y^2 \\
&= \frac{\frac{1}{2}(x + y)^2 + \delta(y - x) - \frac{1}{2}\delta^2}{2} - \frac{1}{2}y^2 && (x - y < 0) \\
&= \frac{-\frac{1}{2}y^2 + (x + \delta)y + \frac{1}{2}x^2 - \delta x - \frac{1}{2}\delta^2}{2} \\
&\geq \frac{-\frac{1}{2}\delta^2 + (x + \delta)\delta + \frac{1}{2}x^2 - \delta x - \frac{1}{2}\delta^2}{2} \\
&\quad \text{(the minimum of the quadratic function is attained when } y = \delta) \\
&= \frac{x^2}{4} \\
&\geq \min\left\{\frac{\delta}{2B}, \frac{1}{4}\right\}x^2.
\end{aligned}$$

**Case III:** $|x + y| > \delta$, $|x - y| \leq \delta$, $0 \leq y \leq \delta$. In this case, we must have $-y + \delta \leq x \leq y + \delta$ and $\delta \geq y \geq \max\{-x + \delta, x - \delta\} \geq -x + \delta$. Thus,

$$\begin{aligned}
F(x, y) &= \frac{\delta|x + y| - \frac{1}{2}\delta^2 + \frac{1}{2}(x - y)^2}{2} - \frac{1}{2}y^2 \\
&= \frac{\delta(x + y) - \frac{1}{2}\delta^2 + \frac{1}{2}(x - y)^2}{2} - \frac{1}{2}y^2 && (x + y > 0) \\
&= \frac{-\frac{1}{2}y^2 + (-x + \delta)y + \frac{1}{2}x^2 + \delta x - \frac{1}{2}\delta^2}{2} \\
&\geq \frac{-\frac{1}{2}\delta^2 + (-x + \delta)\delta + \frac{1}{2}x^2 + \delta x - \frac{1}{2}\delta^2}{2} \\
&\quad \text{(the minimum of the quadratic function is attained when } y = \delta) \\
&= \frac{x^2}{4} \\
&\geq \min\left\{\frac{\delta}{2B}, \frac{1}{4}\right\}x^2.
\end{aligned}$$

**Case IV:** $x + y > \delta$, $|x - y| > \delta$, $0 \leq y \leq \delta$. In this case, we must have $x > y + \delta \geq \delta$ and $0 \leq y < \min\{x - \delta, \delta\}$. Thus, we have

$$\begin{aligned} F(x,y) &= \frac{\delta|x+y| - \frac{1}{2}\delta^2 + \delta|x-y| - \frac{1}{2}\delta^2}{2} - \frac{1}{2}y^2 \\ &= \frac{\delta(x+y) - \frac{1}{2}\delta^2 + \delta(x-y) - \frac{1}{2}\delta^2}{2} - \frac{1}{2}y^2 && (x+y>0 \text{ and } x-y>0) \\ &= \frac{-y^2 + 2\delta x - \delta^2}{2}. \end{aligned}$$

Then, if $\delta < x \leq 2\delta$ and $\min\{x - \delta, \delta\} = x - \delta$,

$$\begin{aligned} F(x,y) &= \frac{-y^2 + 2\delta x - \delta^2}{2} \\ &\geq \frac{-(x-\delta)^2 + 2\delta x - \delta^2}{2} \\ &\text{(the minimum of the quadratic function is attained when } y = x - \delta) \\ &= \frac{-x^2 + 4\delta x - 2\delta^2}{2} \\ &\geq \frac{x^2}{4} && (\delta < x \leq 2\delta) \\ &\geq \min\left\{\frac{\delta}{2B}, \frac{1}{4}\right\}x^2. \end{aligned}$$

If $2\delta < x \leq B$ and $\min\{x - \delta, \delta\} = \delta$,

$$\begin{aligned} F(x,y) &= \frac{-y^2 + 2\delta x - \delta^2}{2} \\ &\geq \frac{-\delta^2 + 2\delta x - \delta^2}{2} && \text{(the minimum of the quadratic function is attained when } y = \delta) \\ &= \delta x - \delta^2 \\ &\geq \frac{\delta}{2B}x^2 && (2\delta < x \leq B) \\ &\geq \min\left\{\frac{\delta}{2B}, \frac{1}{4}\right\}x^2. \end{aligned}$$

**Case V:** $x + y < -\delta$, $|x - y| > \delta$, $0 \leq y \leq \delta$. In this case, we must have $x < -y - \delta \leq -\delta$, and $0 \leq y < \min\{-x - \delta, \delta\}$. Thus, we have

$$\begin{aligned} F(x,y) &= \frac{\delta|x+y| - \frac{1}{2}\delta^2 + \delta|x-y| - \frac{1}{2}\delta^2}{2} - \frac{1}{2}y^2 \\ &= \frac{-\delta(x+y) - \frac{1}{2}\delta^2 - \delta(x-y) - \frac{1}{2}\delta^2}{2} - \frac{1}{2}y^2 && (x+y<0 \text{ and } x-y<0) \\ &= \frac{-y^2 - 2\delta x - \delta^2}{2}. \end{aligned}$$

Then, if $-2\delta \le x < -\delta$ and $\min\{-x-\delta, \delta\} = -x-\delta$,

$$\begin{aligned}
F(x,y) &= \frac{-y^2 - 2\delta x - \delta^2}{2}\\
&\geq \frac{-(-x-\delta)^2 - 2\delta x - \delta^2}{2}\\
&\qquad \text{(the minimum of the quadratic function is attained when } y = -x-\delta\text{)}\\
&= \frac{-x^2 - 4\delta x - 2\delta^2}{2}\\
&\geq \frac{x^2}{4} && (-2\delta \le x < -\delta)\\
&\geq \min\left\{\frac{\delta}{2B}, \frac{1}{4}\right\} x^2.
\end{aligned}$$

If $-B \le x < -2\delta$ and $\min\{-x-\delta, \delta\} = \delta$,

$$\begin{aligned}
F(x,y) &= \frac{-y^2 - 2\delta x - \delta^2}{2}\\
&\geq \frac{-\delta^2 - 2\delta x - \delta^2}{2} && \text{(the minimum of the quadratic function is attained when } y = \delta\text{)}\\
&= -\delta x - \delta^2\\
&\geq \frac{\delta}{2B} x^2 && (-B \le x < -2\delta)\\
&\geq \min\left\{\frac{\delta}{2B}, \frac{1}{4}\right\} x^2.
\end{aligned}$$

In summary, we complete the proof. ■

**Theorem 7** *Assume that the distribution is symmetric, the conditional distribution and the hypothesis set $\mathcal{H}$ are bounded by $B > 0$. Assume that $p_{\min}(\delta) = \inf_{x \in \mathcal{X}} \mathbb{P}(0 \le \mu(x) - y \le \delta \mid x)$ is positive. Then, for all $h \in \mathcal{H}$, the following $\mathcal{H}$-consistency bound holds:*

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \mathcal{M}_{\ell_2}(\mathcal{H}) \le \frac{\max\left\{\frac{2B}{\delta}, 2\right\}}{p_{\min}(\delta)} \left(\mathcal{E}_{\ell_\delta}(h) - \mathcal{E}^*_{\ell_\delta}(\mathcal{H}) + \mathcal{M}_{\ell_\delta}(\mathcal{H})\right).$$

**Proof** By Theorem 4, we can write $\forall h \in \mathcal{H}, x \in \mathcal{X}$,

$$\begin{aligned}
&\Delta\mathcal{C}_{\ell_\delta}(h,x)\\
&= \mathbb{E}_y\bigg[\frac{1}{2}(h(x)-y)^2 1_{|h(x)-y|\le\delta} + \Big(\delta|h(x)-y| - \frac{1}{2}\delta^2\Big)1_{|h(x)-y|>\delta} \mid x\bigg]\\
&\qquad - \mathbb{E}_y\bigg[\frac{1}{2}(\mu(x)-y)^2 1_{|\mu(x)-y|\le\delta} + \Big(\delta|\mu(x)-y| - \frac{1}{2}\delta^2\Big)1_{|\mu(x)-y|>\delta} \mid x\bigg]\\
&= \mathbb{E}_y\bigg[\frac{g(h(x)-\mu(x)+\mu(x)-y) + g(h(x)-\mu(x)-(\mu(x)-y))}{2} - g(\mu(x)-y) \mid x\bigg]\\
&\hspace{20em}\text{(distribution is symmetric with respect to } \mu(x))\\
&= \mathbb{E}_y[F(h(x)-\mu(x), \mu(x)-y) \mid x]\\
&\ge 2\,\mathbb{P}(0 \le \mu(x)-y \le \delta \mid x)\,\mathbb{E}_y[F(h(x)-\mu(x), \mu(x)-y) \mid 0 \le \mu(x)-y \le \delta]\\
&\ge \mathbb{P}(0 \le \mu(x)-y \le \delta \mid x)\min\Big\{\frac{\delta}{2B}, \frac{1}{2}\Big\}(h(x)-\mu(x))^2 \quad (|h(x)-\mu(x)| \le 2B, |\mu(x)-y| \le 2B)\\
&= \mathbb{P}(0 \le \mu(x)-y \le \delta \mid x)\min\Big\{\frac{\delta}{2B}, \frac{1}{2}\Big\}\Delta\mathcal{C}_{\ell_2}(h,x).
\end{aligned}$$

By Theorems 1 or 2 with $\alpha(h,x) = \frac{1}{\mathbb{P}(0 \le \mu(x)-y \le \delta | x)}$, we have

$$\mathcal{E}_{\ell_2}(r) - \mathcal{E}^*_{\ell_2}(\mathcal{R}) + \mathcal{M}_{\ell_2}(\mathcal{R}) \le \frac{\max\{\frac{2B}{\delta}, 2\}}{p_{\min}(\delta)}\big(\mathcal{E}_{\ell_\delta}(r) - \mathcal{E}^*_{\ell_\delta}(\mathcal{R}) + \mathcal{M}_{\ell_\delta}(\mathcal{R})\big).$$

■

**Theorem 9** *Assume that the distribution is symmetric, the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Then, the Huber loss $\ell_\delta$ is not $\mathcal{H}$-consistent with respect to the squared loss.*

**Proof** Consider a distribution that concentrates on an input $x$. Choose $y, \mu(x), \delta \in \mathbb{R}$ such that $-B \le y < \mu(x) \le B$ and $\mu(x) - y > \delta$. Consider the conditional distribution as $\mathbb{P}(Y = y \mid x) = \frac{1}{2} = \mathbb{P}(Y = 2\mu(x) - y \mid x)$. Thus, the distribution is symmetric with respect to $y = \mu(x)$. For such a distribution, the best-in-class predictor for the squared loss is $h^*(x) = \mu(x)$. However, for the Huber loss, we have

$$\begin{aligned}
&\mathcal{C}_{\ell_\delta}(h,x)\\
&= \mathbb{E}_y\bigg[\frac{1}{2}(h(x)-y)^2 1_{|h(x)-y|\le\delta} + \Big(\delta|h(x)-y| - \frac{1}{2}\delta^2\Big)1_{|h(x)-y|>\delta} \mid x\bigg]\\
&= \frac{1}{2}\bigg(\frac{1}{2}(h(x)-y)^2 1_{|h(x)-y|\le\delta} + \Big(\delta|h(x)-y| - \frac{1}{2}\delta^2\Big)1_{|h(x)-y|>\delta}\bigg)\\
&\quad + \frac{1}{2}\bigg(\frac{1}{2}(h(x)-2\mu(x)+y)^2 1_{|h(x)-2\mu(x)+y|\le\delta} + \Big(\delta|h(x)-2\mu(x)+y| - \frac{1}{2}\delta^2\Big)1_{|h(x)-2\mu(x)+y|>\delta}\bigg).
\end{aligned}$$

Thus, plugging $\overline{h}: x \mapsto y + \delta$ and $h^*: x \mapsto \mu(x)$, we obtain that

$$\begin{aligned}
\mathcal{C}_{\ell_\delta}(\overline{h}, x) &= \frac{1}{2}\left(\frac{1}{2}\delta^2\right) + \frac{1}{2}\left(\delta|2y + \delta - 2\mu(x)| - \frac{1}{2}\delta^2\right) \quad (\overline{h}(x) - y = \delta \text{ and } \overline{h}(x) - 2\mu(x) + y < -\delta)\\
&= \delta\left(\mu(x) - \frac{1}{2}\delta - y\right)\\
\mathcal{C}_{\ell_\delta}(h^*, x) &= \frac{1}{2}\left(\delta|\mu(x) - y| - \frac{1}{2}\delta^2 + \delta|-\mu(x) + y| - \frac{1}{2}\delta^2\right)\\
&\qquad\qquad (h^*(x) - y > \delta \text{ and } h^*(x) - 2\mu(x) + y < -\delta)\\
&= \delta\left(\mu(x) - \frac{1}{2}\delta - y\right).
\end{aligned}$$

Therefore, $\mathcal{C}_{\ell_\delta}(\overline{h}, x) = \mathcal{C}_{\ell_\delta}(h^*, x)$, and both $\overline{h}$ and $h^*$ are the best-in-class predictors for the Huber loss. This implies that the Huber loss is not $\mathcal{H}$-consistent with respect to the squared loss. ■

### B.2. $\mathcal{H}$-consistency of $\ell_p$ with respect to $\ell_2$

**Theorem 10** *Assume that the distribution is symmetric, and that the conditional distribution and the hypothesis set $\mathcal{H}$ are bounded by $B > 0$. Then, for all $h \in \mathcal{H}$ and $p \geq 1$, the following $\mathcal{H}$-consistency bound holds:*

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \mathcal{M}_{\ell_2}(\mathcal{H}) \leq \Gamma\left(\mathcal{E}_{\ell_p}(h) - \mathcal{E}^*_{\ell_p}(\mathcal{H}) + \mathcal{M}_{\ell_p}(\mathcal{H})\right),$$

*where $\Gamma(t) = \sup_{x \in \mathcal{X}, y \in \mathcal{Y}}\{|h(x) - y| + |\mu(x) - y|\}\, t$ for $p = 1$, $\Gamma(t) = \frac{2}{(8B)^{p-2}p(p-1)}\, t$ for $p \in (1, 2]$, and $\Gamma(t) = t^{\frac{2}{p}}$ for $p \geq 2$.*

**Proof** We will analyze case by case.

**Case I:** $p \geq 2$**.** By Theorem 4, we can write

$$\begin{aligned}
\forall h \in \mathcal{H}, x \in \mathcal{X}, \quad & \Delta\mathcal{C}_{\ell_p}(h, x)\\
&= \mathop{\mathbb{E}}_{y}[|h(x) - y|^p - |\mu(x) - y|^p \mid x]\\
&= \mathop{\mathbb{E}}_{y}\left[\frac{|h(x) - y|^p + |h(x) - 2\mu(x) + y|^p}{2} - |\mu(x) - y|^p \mid x\right]\\
&\qquad\qquad \text{(distribution is symmetric with respect to } \mu(x))\\
&= \mathop{\mathbb{E}}_{y}\left[\frac{|h(x) - \mu(x) + \mu(x) - y|^p + |h(x) - \mu(x) - (\mu(x) - y)|^p}{2} - |\mu(x) - y|^p \mid x\right]\\
&\geq |h(x) - \mu(x)|^p \qquad \text{(by Clarkson's inequality (Clarkson, 1936))}\\
&= \left((h(x) - \mu(x))^2\right)^{\frac{p}{2}}\\
&= (\Delta\mathcal{C}_{\ell_2}(h, x))^{\frac{p}{2}}.
\end{aligned}$$

By Theorem 1, we have

$$\mathcal{E}_{\ell_2}(r) - \mathcal{E}^*_{\ell_2}(\mathcal{R}) + \mathcal{M}_{\ell_2}(\mathcal{R}) \leq \left(\mathcal{E}_{\ell_p}(r) - \mathcal{E}^*_{\ell_p}(\mathcal{R}) + \mathcal{M}_{\ell_p}(\mathcal{R})\right)^{\frac{2}{p}}.$$

**Case II:** $1 < p \leq 2$**.** In this case, the Clarkson's inequality cannot be used directly. We first prove a useful lemma as follows.

**Lemma 20** *For any $x, y \in [-B, B]$ and $1 < p \leq 2$, the following inequality holds:*

$$\frac{|x+y|^p + |x-y|^p}{2} - |y|^p \geq \frac{(2B)^{p-2}p(p-1)}{2}x^2.$$

**Proof** For any $y \in [-B, B]$, consider the function $f_y \colon x \mapsto \frac{|x+y|^p+|x-y|^p}{2} - |y|^p - \frac{(2B)^{p-2}p(p-1)}{2}x^2$. We compute the first derivative and second derivative of $f_y$ as follows:

$$\begin{aligned} f_y'(x) &= \frac{\frac{p|x+y|^p}{x+y} + \frac{p|x-y|^p}{x-y}}{2} - (2B)^{p-2}p(p-1)x \\ f_y''(x) &= \frac{\frac{p(p-1)}{|x+y|^{2-p}} + \frac{p(p-1)}{|x-y|^{2-p}}}{2} - (2B)^{p-2}p(p-1). \end{aligned}$$

Thus, using the fact that $1 < p \leq 2$ and $|x+y| \leq 2B$, $|x-y| \leq 2B$, we have

$$\forall x \in [-B, B], \quad f_y''(x) \geq \frac{\frac{p(p-1)}{(2B)^{2-p}} + \frac{p(p-1)}{(2B)^{2-p}}}{2} - (2B)^{p-2}p(p-1) = 0.$$

Therefore, $f_y(x)$ is convex. Since $f_y'(0) = 0$, $x = 0$ achieves the minimum:

$$\forall x, y \in [-B, B], \quad f_y(x) \geq f_y(0) = 0.$$

This completes the proof. ■

By Theorem 4, we can write

$$\begin{aligned} \forall h \in \mathcal{H}, x \in \mathcal{X}, \quad & \Delta\mathcal{C}_{\ell_p}(h, x) \\ &= \mathbb{E}_y\big[|h(x)-y|^p - |\mu(x)-y|^p \mid x\big] \\ &= \mathbb{E}_y\left[\frac{|h(x)-y|^p + |h(x)-2\mu(x)+y|^p}{2} - |\mu(x)-y|^p \mid x\right] \\ &\qquad\qquad \text{(distribution is symmetric with respect to } \mu(x)\text{)} \\ &= \mathbb{E}_y\left[\frac{|h(x)-\mu(x)+\mu(x)-y|^p + |h(x)-\mu(x)-(\mu(x)-y)|^p}{2} - |\mu(x)-y|^p \mid x\right] \\ &\geq \frac{(8B)^{p-2}p(p-1)}{2}(h(x)-\mu(x))^2 \\ &\qquad\qquad \text{(by Lemma 20 and } |h(x)-\mu(x)| \leq 4B, |\mu(x)-y| \leq 4B\text{)} \\ &= \frac{(8B)^{p-2}p(p-1)}{2}\Delta\mathcal{C}_{\ell_2}(h, x). \end{aligned}$$

By Theorem 1, we have

$$\mathcal{E}_{\ell_2}(r) - \mathcal{E}^*_{\ell_2}(\mathcal{R}) + \mathcal{M}_{\ell_2}(\mathcal{R}) \leq \frac{2}{(8B)^{p-2}p(p-1)}\big(\mathcal{E}_{\ell_p}(r) - \mathcal{E}^*_{\ell_p}(\mathcal{R}) + \mathcal{M}_{\ell_p}(\mathcal{R})\big).$$

**Case III:** $p = 1$**.** By Theorem 4, we can write

$$\begin{aligned}
\forall h \in \mathcal{H}, x \in \mathcal{X}, \quad \Delta\mathcal{C}_{\ell_2}(h,x) &= \mathop{\mathbb{E}}_{y}\big[(h(x)-y)^2 - (\mu(x)-y)^2 \mid x\big] \\
&= \mathop{\mathbb{E}}_{y}[(|h(x)-y| + |\mu(x)-y|)(|h(x)-y| - |\mu(x)-y|) \mid x] \\
&\leq \sup_{y\in\mathcal{Y}}\{|h(x)-y| + |\mu(x)-y|\} \mathop{\mathbb{E}}_{y}[|h(x)-y| - |\mu(x)-y| \mid x] \\
&= \sup_{y\in\mathcal{Y}}\{|h(x)-y| + |\mu(x)-y|\} \, \Delta\mathcal{C}_{\ell_1}(h,x).
\end{aligned}$$

By Theorems 1 or 2 with $\alpha(h,x) = \sup_{y\in\mathcal{Y}}\{|h(x)-y| + |\mu(x)-y|\}$, we have

$$\mathcal{E}_{\ell_2}(r) - \mathcal{E}^*_{\ell_2}(\mathcal{R}) + \mathcal{M}_{\ell_2}(\mathcal{R}) \leq \sup_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}}\{|h(x)-y| + |\mu(x)-y|\}\big(\mathcal{E}_{\ell_1}(r) - \mathcal{E}^*_{\ell_1}(\mathcal{R}) + \mathcal{M}_{\ell_1}(\mathcal{R})\big).$$

■

### B.3. $\mathcal{H}$-consistency of $\ell_{\text{sq}-\epsilon}$ with respect to $\ell_2$

Define the function $g$ as $g\colon t \mapsto \max\{t^2 - \epsilon^2, 0\}$. Consider the function $F$ defined over $\mathbb{R}^2$ by $F(x,y) = \frac{g(x+y)+g(x-y)}{2} - g(y)$. We first prove a useful lemma as follows.

**Lemma 21** *For any $x \in \mathbb{R}$ and $|y| \geq \epsilon$, the following inequality holds:*

$$F(x,y) \geq x^2.$$

**Proof** Given the definition of $g$ and the symmetry of $F$ with respect to $y = 0$, we can assume, without loss of generality, that $y \geq 0$. Next, we will analyze case by case.
**Case I:** $|x+y| > \epsilon$, $|x-y| > \epsilon$, $y \geq \epsilon$. In this case, we have

$$F(x,y) = \frac{(x+y)^2 - \epsilon^2 + (x-y)^2 - \epsilon^2}{2} - y^2 + \epsilon^2 = x^2.$$

**Case II:** $|x+y| > \epsilon$, $|x-y| \leq \epsilon$, $y \geq \epsilon$. In this case, we must have $y - \epsilon \leq x \leq y + \epsilon$ and $x + \epsilon \geq y \geq \max\{x - \epsilon, \epsilon\} \geq x - \epsilon$. Thus,

$$\begin{aligned}
F(x,y) &= \frac{(x+y)^2 - \epsilon^2 + 0}{2} - y^2 + \epsilon^2 \\
&= \frac{-y^2 + 2xy + x^2 + \epsilon^2}{2} \\
&\geq \frac{-(x+\epsilon)^2 + 2x(x+\epsilon) + x^2 + \epsilon^2}{2} \\
&\qquad \text{(the minimum of the quadratic function is attained when } y = x + \epsilon) \\
&= x^2.
\end{aligned}$$

**Case III:** $|x+y| \le \epsilon$, $|x-y| > \epsilon$, $y \ge \epsilon$. In this case, we must have $-y-\epsilon \le x \le -y+\epsilon$ and $-x+\epsilon \ge y \ge \max\{-x-\epsilon, \epsilon\} \ge -x-\epsilon$. Thus,

$$\begin{aligned} F(x,y) &= \frac{0 + (x-y)^2 - \epsilon^2}{2} - y^2 + \epsilon^2 \\ &= \frac{-y^2 - 2xy + x^2 + \epsilon^2}{2} \\ &\ge \frac{-(-x+\epsilon)^2 - 2x(-x+\epsilon) + x^2 + \epsilon^2}{2} \\ &\qquad \text{(the minimum of the quadratic function is attained when } y = -x+\epsilon) \\ &= x^2. \end{aligned}$$

**Case IV:** $|x+y| \le \epsilon$, $|x-y| \le \epsilon$, $y \ge \epsilon$. In this case, we must have $x = 0$ and $y = \epsilon$. Thus,

$$F(x,y) = \frac{0+0}{2} - 0 = 0 = x^2.$$

In summary, we complete the proof. ■

**Theorem 12** *Assume that the distribution is symmetric, and that the conditional distribution and the hypothesis set $\mathcal{H}$ are bounded by $B > 0$. Assume that $p_{\min}(\epsilon) = \inf_{x \in \mathcal{X}} \mathbb{P}(\mu(x) - y \ge \epsilon \mid x)$ is positive. Then, for all $h \in \mathcal{H}$, the following $\mathcal{H}$-consistency bound holds:*

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \mathcal{M}_{\ell_2}(\mathcal{H}) \le \frac{\mathcal{E}_{\ell_{\text{sq}-\epsilon}}(h) - \mathcal{E}^*_{\ell_{\text{sq}-\epsilon}}(\mathcal{H}) + \mathcal{M}_{\ell_{\text{sq}-\epsilon}}(\mathcal{H})}{2p_{\min}(\epsilon)}.$$

**Proof** By Theorem 4, we can write $\forall h \in \mathcal{H}, x \in \mathcal{X}$,

$$\begin{aligned} &\Delta\mathcal{C}_{\ell_{\text{sq}-\epsilon}}(h,x) \\ &= \mathop{\mathbb{E}}_y\big[\max\{(h(x)-y)^2, \epsilon^2\} \mid x\big] - \mathop{\mathbb{E}}_y\big[\max\{(\mu(x)-y)^2, \epsilon^2\} \mid x\big] \\ &= \mathop{\mathbb{E}}_y\left[\frac{g(h(x)-\mu(x)+\mu(x)-y) + g(h(x)-\mu(x)-(\mu(x)-y))}{2} - g(\mu(x)-y) \mid x\right] \\ &\qquad\qquad\qquad \text{(distribution is symmetric with respect to } \mu(x)) \\ &= \mathop{\mathbb{E}}_y\big[F(h(x)-\mu(x), \mu(x)-y) \mid x\big] \\ &\ge 2\,\mathbb{P}(\mu(x)-y \ge \epsilon \mid x) \mathop{\mathbb{E}}_y\big[F(h(x)-\mu(x), \mu(x)-y) \mid \mu(x)-y \ge \epsilon\big] \\ &\ge 2\,\mathbb{P}(\mu(x)-y \ge \epsilon \mid x)(h(x)-\mu(x))^2 \qquad \text{(by Lemma 21)} \\ &= 2\,\mathbb{P}(\mu(x)-y \ge \epsilon \mid x)\Delta\mathcal{C}_{\ell_2}(h,x). \end{aligned}$$

By Theorems 1 or 2 with $\alpha(h,x) = \frac{1}{2\,\mathbb{P}(\mu(x)-y \ge \epsilon \mid x)}$, we have

$$\mathcal{E}_{\ell_2}(r) - \mathcal{E}^*_{\ell_2}(\mathcal{R}) + \mathcal{M}_{\ell_2}(\mathcal{R}) \le \frac{\mathcal{E}_{\ell_{\text{sq}-\epsilon}}(r) - \mathcal{E}^*_{\ell_{\text{sq}-\epsilon}}(\mathcal{R}) + \mathcal{M}_{\ell_{\text{sq}-\epsilon}}(\mathcal{R})}{2p_{\min}(\epsilon)}.$$

■

**Theorem 14** *Assume that the distribution is symmetric, the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Then, the squared $\epsilon$-insensitive loss $\ell_{\mathrm{sq}-\epsilon}$ is not $\mathcal{H}$-consistent.*

**Proof** Consider a distribution that concentrates on an input $x$. Choose $y, \mu(x), \epsilon \in \mathbb{R}$ such that $-B \leq y < \mu(x) \leq B$ and $\mu(x) - y < \epsilon$. Consider the conditional distribution as $\mathbb{P}(Y = y \mid x) = \frac{1}{2} = \mathbb{P}(Y = 2\mu(x) - y \mid x)$. Thus, the distribution is symmetric with respect to $y = \mu(x)$. For such a distribution, the best-in-class predictor for the squared loss is $h^*(x) = \mu(x)$. However, for the $\epsilon$-insensitive loss, we have

$$\begin{aligned}
&\mathcal{C}_{\ell_{\mathrm{sq}-\epsilon}}(h, x)\\
&= \mathbb{E}_{y}\Big[\max\big\{(h(x) - y)^2 - \epsilon^2, 0\big\} \mid x\Big]\\
&= \frac{1}{2}\max\big\{(h(x) - y)^2 - \epsilon^2, 0\big\} + \frac{1}{2}\max\big\{(h(x) - 2\mu(x) + y)^2 - \epsilon^2, 0\big\}.
\end{aligned}$$

Thus, plugging $\overline{h}: x \mapsto y + \epsilon$ and $h^*: x \mapsto \mu(x)$, we obtain that

$$\begin{aligned}
\mathcal{C}_{\ell_{\mathrm{sq}-\epsilon}}(\overline{h}, x) &= \frac{1}{2}(0) + \frac{1}{2}(0) && (\overline{h}(x) - y = \epsilon \text{ and } \epsilon > \overline{h}(x) - 2\mu(x) + y > -\epsilon)\\
&= 0\\
\mathcal{C}_{\ell_{\mathrm{sq}-\epsilon}}(h^*, x) &= \frac{1}{2}(0) + \frac{1}{2}(0) && (0 < h^*(x) - y < \epsilon \text{ and } 0 > h^*(x) - 2\mu(x) + y > -\epsilon)\\
&= 0.
\end{aligned}$$

Therefore, $\mathcal{C}_{\ell_{\mathrm{sq}-\epsilon}}(\overline{h}, x) = \mathcal{C}_{\ell_{\mathrm{sq}-\epsilon}}(h^*, x)$, and both $\overline{h}$ and $h^*$ are the best-in-class predictors for the $\epsilon$-insensitive loss. This implies that the $\epsilon$-insensitive loss is not $\mathcal{H}$-consistent with respect to the squared loss. ■

### B.4. $\mathcal{H}$-consistency of $\ell_\epsilon$ with respect to $\ell_2$

Here, we present negative results for the $\epsilon$-insensitive loss $\ell_\epsilon: (h, x, y) \mapsto \max\{|h(x) - y| - \epsilon, 0\}$ used in the SVR algorithm, by showing that even under the assumption $\inf_{x \in \mathcal{X}} \mathbb{P}(\mu(x) - y \geq \epsilon) > 0$ or $\inf_{x \in \mathcal{X}} \mathbb{P}(0 \leq \mu(x) - y \leq \epsilon) > 0$, it is not $\mathcal{H}$-consistent with respect to the squared loss. In the proof, we consider distributions that concentrate on an input $x$, leading to both $\overline{h}: x \mapsto y + \epsilon$ and $h^*: x \mapsto \mu(x)$ being the best-in-class predictors for the $\epsilon$-insensitive loss.

**Theorem 22** *Assume that the distribution is symmetric and satisfies $\inf_{x \in \mathcal{X}} \mathbb{P}(\mu(x) - y \geq \epsilon \mid x) > 0$. Assume that the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Then, the $\epsilon$-insensitive loss $\ell_\epsilon$ is not $\mathcal{H}$-consistent with respect to the squared loss.*

**Proof** Consider a distribution that concentrates on an input $x$. Choose $y, \mu(x), \epsilon \in \mathbb{R}$ such that $-B \leq y < \mu(x) \leq B$ and $\mu(x) - y > \epsilon$. Consider the conditional distribution as $\mathbb{P}(Y = y \mid x) = \frac{1}{2} = \mathbb{P}(Y = 2\mu(x) - y \mid x)$. Thus, the distribution is symmetric with respect to $y = \mu(x)$. For such a distribution, the best-in-class predictor for the squared loss is $h^*(x) = \mu(x)$. However, for the $\epsilon$-insensitive loss, we have

$$\begin{aligned}
&\mathcal{C}_{\ell_{\text{sq}-\epsilon}}(h, x)\\
&= \mathop{\mathbb{E}}_{y}[\max\{|h(x) - y| - \epsilon, 0\} \mid x]\\
&= \frac{1}{2}\max\{|h(x) - y| - \epsilon, 0\} + \frac{1}{2}\max\{|h(x) - 2\mu(x) + y| - \epsilon, 0\}.
\end{aligned}$$

Thus, plugging $\overline{h}\colon x \mapsto y + \epsilon$ and $h^*\colon x \mapsto \mu(x)$, we obtain that

$$\begin{aligned}
\mathcal{C}_{\ell_{\text{sq}-\epsilon}}(\overline{h}, x) &= \frac{1}{2}(0) + \frac{1}{2}(2\mu(x) - 2y - 2\epsilon) && (\overline{h}(x) - y = \epsilon \text{ and } \overline{h}(x) - 2\mu(x) + y < -\epsilon)\\
&= \mu(x) - y - \epsilon.\\
\mathcal{C}_{\ell_{\text{sq}-\epsilon}}(h^*, x) &= \frac{1}{2}(\mu(x) - y - \epsilon) + \frac{1}{2}(\mu(x) - y - \epsilon) && (h^*(x) - y > \epsilon \text{ and } h^*(x) - 2\mu(x) + y < -\epsilon)\\
&= \mu(x) - y - \epsilon.
\end{aligned}$$

Therefore, $\mathcal{C}_{\ell_{\text{sq}-\epsilon}}(\overline{h}, x) = \mathcal{C}_{\ell_{\text{sq}-\epsilon}}(h^*, x)$, and both $\overline{h}$ and $h^*$ are the best-in-class predictors for the $\epsilon$-insensitive loss. This implies that the $\epsilon$-insensitive loss is not $\mathcal{H}$-consistent with respect to the squared loss. ■

**Theorem 23** *Assume that the distribution is symmetric and satisfies $p_{\min}(\epsilon) = \inf_{x \in \mathcal{X}} \mathbb{P}(0 \le \mu(x) - y \le \epsilon \mid x) > 0$. Assume further that the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Then, the $\epsilon$-insensitive loss $\ell_\epsilon$ is not $\mathcal{H}$-consistent with respect to the squared loss.*

**Proof** Consider a distribution that concentrates on an input $x$. Choose $y, \mu(x), \epsilon \in \mathbb{R}$ such that $-B \le y < \mu(x) \le B$ and $\mu(x) - y < \epsilon$. Consider the conditional distribution as $\mathbb{P}(Y = y \mid x) = \frac{1}{2} = \mathbb{P}(Y = 2\mu(x) - y \mid x)$. Thus, the distribution is symmetric with respect to $y = \mu(x)$. For such a distribution, the best-in-class predictor for the squared loss is $h^*(x) = \mu(x)$. However, for the $\epsilon$-insensitive loss, we have

$$\begin{aligned}
&\mathcal{C}_{\ell_{\text{sq}-\epsilon}}(h, x)\\
&= \mathop{\mathbb{E}}_{y}[\max\{|h(x) - y| - \epsilon, 0\} \mid x]\\
&= \frac{1}{2}\max\{|h(x) - y| - \epsilon, 0\} + \frac{1}{2}\max\{|h(x) - 2\mu(x) + y| - \epsilon, 0\}.
\end{aligned}$$

Thus, plugging $\overline{h}\colon x \mapsto y + \epsilon$ and $h^*\colon x \mapsto \mu(x)$, we obtain that

$$\begin{aligned}
\mathcal{C}_{\ell_{\text{sq}-\epsilon}}(\overline{h}, x) &= \frac{1}{2}(0) + \frac{1}{2}(0) && (\overline{h}(x) - y = \epsilon \text{ and } \epsilon > \overline{h}(x) - 2\mu(x) + y > -\epsilon)\\
&= 0.\\
\mathcal{C}_{\ell_{\text{sq}-\epsilon}}(h^*, x) &= \frac{1}{2}(0) + \frac{1}{2}(0) && (0 < h^*(x) - y < \epsilon \text{ and } 0 > h^*(x) - 2\mu(x) + y > -\epsilon)\\
&= 0.
\end{aligned}$$

Therefore, $\mathcal{C}_{\ell_{\text{sq}-\epsilon}}(\overline{h}, x) = \mathcal{C}_{\ell_{\text{sq}-\epsilon}}(h^*, x)$, and both $\overline{h}$ and $h^*$ are the best-in-class predictors for the $\epsilon$-insensitive loss. This implies that the $\epsilon$-insensitive loss is not $\mathcal{H}$-consistent with respect to the squared loss. ■

## Appendix C. Proofs of generalization bound

**Theorem 15** *Assume that the distribution is symmetric, the conditional distribution and the hypothesis set $\mathcal{H}$ are bounded by $B > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample $S$ of size $m$, the following squared loss estimation bound holds for $\widehat{h}_S$:*

$$\mathcal{E}_{\ell_2}(\widehat{h}_S) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \leq \Gamma\left(\mathcal{M}_L(\mathcal{H}) + 4\mathfrak{R}^L_m(\mathcal{H}) + 2B_L\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right) - \mathcal{M}_{\ell_2}(\mathcal{H}).$$

*where $\Gamma(t) = \sup_{x\in\mathcal{X}}\sup_y\left\{|\widehat{h}_S(x) - y| + |\mu(x) - y|\right\} t$ for $L = \ell_1$, $\Gamma(t) = \frac{2}{(8B)^{p-2}p(p-1)} t$ for $L = \ell_p$, $p \in (1, 2]$, $\Gamma(t) = t^{\frac{2}{p}}$ for $L = \ell_p$, $p \geq 2$, $\Gamma(t) = \frac{\max\{\frac{2B}{\delta}, 2\}}{p_{\min}(\delta)} t$ for $L = \ell_\delta$, and $\Gamma(t) = \frac{1}{2p_{\min}(\epsilon)} t$ for $L = \ell_{\text{sq}-\epsilon}$.*

**Proof** By using the standard Rademacher complexity bounds (Mohri et al., 2018), for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in \mathcal{H}$:

$$\left|\mathcal{E}_L(h) - \widehat{\mathcal{E}}_{L,S}(h)\right| \leq 2\mathfrak{R}^L_m(\mathcal{H}) + B_L\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Fix $\epsilon > 0$. By the definition of the infimum, there exists $h^* \in \mathcal{H}$ such that $\mathcal{E}_L(h^*) \leq \mathcal{E}^*_L(\mathcal{H}) + \epsilon$. By definition of $\widehat{h}_S$, we have

$$\begin{aligned}
&\mathcal{E}_L(\widehat{h}_S) - \mathcal{E}^*_L(\mathcal{H})\\
&= \mathcal{E}_L(\widehat{h}_S) - \widehat{\mathcal{E}}_{L,S}(\widehat{h}_S) + \widehat{\mathcal{E}}_{L,S}(\widehat{h}_S) - \mathcal{E}^*_L(\mathcal{H})\\
&\leq \mathcal{E}_L(\widehat{h}_S) - \widehat{\mathcal{E}}_{L,S}(\widehat{h}_S) + \widehat{\mathcal{E}}_{L,S}(h^*) - \mathcal{E}^*_L(\mathcal{H})\\
&\leq \mathcal{E}_L(\widehat{h}_S) - \widehat{\mathcal{E}}_{L,S}(\widehat{h}_S) + \widehat{\mathcal{E}}_{L,S}(h^*) - \mathcal{E}^*_L(h^*) + \epsilon\\
&\leq 2\left[2\mathfrak{R}^L_m(\mathcal{H}) + B_L\sqrt{\frac{\log(2/\delta)}{2m}}\right] + \epsilon.
\end{aligned}$$

Since the inequality holds for all $\epsilon > 0$, it implies:

$$\mathcal{E}_L(\widehat{h}_S) - \mathcal{E}^*_L(\mathcal{H}) \leq 4\mathfrak{R}^L_m(\mathcal{H}) + 2B_L\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Plugging in this inequality in the bound of Theorems 7, 10, 12 completes the proof. ■

## Appendix D. Proofs of adversarial regression

### D.1. Proof of Theorem 16

**Theorem 16** *Assume that the distribution is symmetric, the conditional distribution is bounded by $B > 0$, and the hypothesis set $\mathcal{H}$ is realizable and bounded by $B > 0$. Assume that $p_{\min}(\delta) =$*

$\inf_{x\in\mathcal{X}}\mathbb{P}(0 \leq \mu(x) - y \leq \delta \mid x)$ *is positive. Then, for any* $\nu \geq 3B$ *and all* $h \in \mathcal{H}$*, the following bound holds:*

$$\mathcal{E}_{\widetilde{\ell}_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \leq \frac{\max\{\frac{2B}{\delta}, 2\}}{p_{\min}(\delta)}\big(\mathcal{E}_{\ell_\delta}(h) - \mathcal{E}^*_{\ell_\delta}(\mathcal{H})\big) + \nu \sup_{x':\|x'-x\|\leq\gamma}\big|h(x') - h(x)\big|$$

**Proof** By (1), we have

$$\begin{aligned}\mathcal{E}_{\widetilde{\ell}_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) &\leq \mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \nu \sup_{x':\|x'-x\|\leq\gamma}\big|h(x') - h(x)\big|\\ &\leq \frac{\max\{\frac{2B}{\delta}, 2\}}{p_{\min}(\delta)}\big(\mathcal{E}_{\ell_\delta}(h) - \mathcal{E}^*_{\ell_\delta}(\mathcal{H})\big) + \nu \sup_{x':\|x'-x\|\leq\gamma}\big|h(x') - h(x)\big|. \quad \text{(Corollary 8)}\end{aligned}$$

This completes the proof. ■

## D.2. Proof of Theorem 17

**Theorem 17** *Assume that the distribution is symmetric, the conditional distribution is bounded by* $B > 0$*, and the hypothesis set* $\mathcal{H}$ *is realizable and bounded by* $B > 0$*. Then, for any* $\nu \geq 3B$ *and all* $h \in \mathcal{H}$*, the following bound holds:*

$$\mathcal{E}_{\widetilde{\ell}_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \leq \Gamma\big(\mathcal{E}_{\ell_p}(h) - \mathcal{E}^*_{\ell_p}(\mathcal{H})\big) + \nu \sup_{x':\|x'-x\|\leq\gamma}\big|h(x') - h(x)\big|,$$

*where* $\Gamma(t) = t^{\frac{2}{p}}$ *if* $p \geq 2$*,* $\frac{2}{(8B)^{p-2}p(p-1)}\,t$ *for* $p \in (1,2)$ *and* $4B\,t$*, if* $p = 1$*.*

**Proof** By (1), we have

$$\begin{aligned}\mathcal{E}_{\widetilde{\ell}_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) &\leq \mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \nu \sup_{x':\|x'-x\|\leq\gamma}\big|h(x') - h(x)\big|\\ &\leq \Gamma\big(\mathcal{E}_{\ell_p}(h) - \mathcal{E}^*_{\ell_p}(\mathcal{H})\big) + \nu \sup_{x':\|x'-x\|\leq\gamma}\big|h(x') - h(x)\big|. \quad \text{(Corollary 11)}\end{aligned}$$

where $\Gamma(t) = \begin{cases} t^{\frac{2}{p}} & p > 2\\ \frac{2}{(8B)^{p-2}p(p-1)}\,t & p \in (1,2]\\ 4B\,t & p = 1.\end{cases}$. This completes the proof. ■

## D.3. Proof of Theorem 18

**Theorem 18** *Assume that the distribution is symmetric, the conditional distribution is bounded by* $B > 0$*, and the hypothesis set* $\mathcal{H}$ *is realizable and bounded by* $B > 0$*. Assume that* $p_{\min}(\epsilon) = \inf_{x\in\mathcal{X}}\mathbb{P}(\mu(x) - y \geq \epsilon \mid x)$ *is positive. Then, for any* $\nu \geq 3B$ *and all* $h \in \mathcal{H}$*, the following bound holds:*

$$\mathcal{E}_{\widetilde{\ell}_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) \leq \frac{\mathcal{E}_{\ell_{\text{sq}-\epsilon}}(h) - \mathcal{E}^*_{\ell_{\text{sq}-\epsilon}}(\mathcal{H})}{2p_{\min}(\epsilon)} + \nu \sup_{x':\|x'-x\|\leq\gamma}\big|h(x') - h(x)\big|.$$

**Proof** By (1), we have

$$\begin{aligned}\mathcal{E}_{\widetilde{\ell}_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) &\leq \mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \nu \sup_{x' : \|x' - x\| \leq \gamma} \left| h(x') - h(x) \right| \\ &\leq \frac{\mathcal{E}_{\ell_{\text{sq}-\epsilon}}(h) - \mathcal{E}^*_{\ell_{\text{sq}-\epsilon}}(\mathcal{H})}{2 p_{\min}(\epsilon)} + \nu \sup_{x' : \|x' - x\| \leq \gamma} \left| h(x') - h(x) \right|. \qquad \text{(Corollary 13)}\end{aligned}$$

This completes the proof. ■