Correction to "The Importance of Convexity in Learning with Squared Loss"

Wee Sun Lee, *Senior Member, IEEE,* Peter L. Bartlett, *Member, IEEE,* and Robert C. Williamson, *Member, IEEE*

***Abstract*—The paper *The Importance of Convexity in Learning with Squared Loss* gave a lower bound on the sample complexity of learning with quadratic loss using a nonconvex function class. The proof contains an error. We show that the lower bound is true under a stronger condition that holds for many cases of interest.**

***Index Terms*—Agnostic learning, sample complexity, lower bound.**

In Theorem 2 of [2], it was claimed that if the closure of a function class F under the metric induced by some probability distribution is not convex, then the sample complexity for agnostically learning F with squared loss (using only hypotheses in F) is $\Omega(\ln(1/\delta)/\epsilon^2)$ where $1-\delta$ is the probability of success and ϵ is the required accuracy. The proof of this result—in particular, the proof of Lemma 5—is incorrect. Thus, we only know that this theorem is true for cases where the function class F is finite dimensional. This weakens the result. However, the lower bound for the sample complexity for agnostic learning still holds for many cases of interest, including any case where the closure of the class of restrictions of functions to a finite subset of the input space $\mathcal{X}$ is not convex. This is the case, for instance, for all of the examples mentioned in [2], including the set of linear combinations of a fixed number of linear threshold functions. (A counting argument, exploiting the finite pseudodimension of such a class, demonstrates this.)

The following is a corrected version of Lemma 5 of [2]; it differs from that lemma by the addition of the words "finite dimensional."

Lemma 5'. *Suppose that $P_{\mathcal{X}}$ is a probability distribution on $\mathcal{X}$, H is the corresponding Hilbert space, and $\mathcal{Y}'$ is a bounded interval in $\mathbb{R}$. Let $H_{\mathcal{Y}'}$ denote the set of functions f in H with $f(x) \in \mathcal{Y}'$ for all $x \in \mathcal{X}$. Let F be a totally bounded finite dimensional subset of $H_{\mathcal{Y}'}$. If $\bar{F}$ is not convex, there is a bounded interval $\mathcal{Y}$ in $\mathbb{R}$ and functions $c \in H_{\mathcal{Y}}$, and $f_1, f_2 \in \bar{F}$ satisfying $||f_1 - f_2|| \neq 0$, $||c - f_1|| = ||c - f_2|| > 0$, and for all $f \in \bar{F}$, $||c-f|| \geq ||c - f_1||$.*

While it is true that for any closed, totally bounded nonconvex subset $\bar{F}$ of $H_{\mathcal{Y}'}$ there is a function c with two best approximations (see, e.g., Theorem 12.6 in [1]), we do not know if there is a uniformly bounded c with this property. However, if $\bar{F}$ is finite dimensional, we can project a function c with two best approximations to the subspace spanned by functions in $\bar{F}$. This projection, c', would still have two best approximations. As c' can be represented as a finite linear combination of functions in $\bar{F}$ and every function in $\bar{F}$ has bounded range, c' has bounded range as well, which proves Lemma 5'.

Acknowledgements

Thanks to Shahar Mendelson for pointing out the error; see [3].

References

[1] F. Deutsch. *Best Approximation in Inner Product Spaces*. Canadian Mathematical Society, 2001.
[2] W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory 44(5):1974-1980*, 1998.
[3] S. Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Transactions on Information Theory*, 2008 (To appear).

Wee Sun Lee is with the Department of Computer Science, National University of Singapore.

Peter L. Bartlett is with the Computer Science Division and Department of Statistics, University of California, Berkeley.

Robert C. Williamson is with NICTA and the Research School of Information Sciences and Engineering, Australian National University.