Similarity-based Learning via Data Driven Embeddings
Supplementary Material

Purushottam Kar
Indian Institute of Technology
Kanpur, INDIA
purushot@cse.iitk.ac.in

Prateek Jain
Microsoft Research India
Bangalore, INDIA
prajain@microsoft.com

Abstract

This document contains detailed proofs of theorems stated in the main article entitled *Similarity-based Learning via Data Driven Embeddings*

1 Proof of Theorem 2

We first recall the definition of a good similarity function.

Definition 1 (Good Similarity Function)**.** *A similarity function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be an (ϵ, γ, B)-good similarity for a learning problem where $\epsilon, \gamma, B > 0$ if for some transfer function $f : \mathbb{R} \to \mathbb{R}$ and some weighing function $w : \mathcal{X} \times \mathcal{X} \to [-B, B]$, at least a $(1 - \epsilon)$ probability mass of examples $x \sim \mathcal{D}$ satisfies*

$$\mathop{\mathbb{E}}_{x',x'' \sim \mathcal{D} \times \mathcal{D}} \left[w\left(x', x''\right) f\left(K(x, x') - K(x, x'')\right) | \ell(x') = \ell(x), \ell(x'') \neq \ell(x) \right] \geq C_f \gamma \tag{1}$$

where $C_f = \sup_{x,x' \in \mathcal{X}} f(K(x, x')) - \inf_{x,x' \in \mathcal{X}} f(K(x, x'))$

Theorem 2 (Theorem 2 restated)**.** *If K is an (ϵ, γ, B)-good similarity with respect to transfer function f and weight function w then for any $\epsilon_1 > 0$, with probability at least $1 - \delta$ over the choice of $d = (8/\gamma^2) \ln(2/\delta\epsilon_1)$ positive and negative samples, $\left\{x_i^+\right\}_{i=1}^d \subset \mathcal{D}^+$ and $\left\{x_i^-\right\}_{i=1}^d \subset \mathcal{D}^-$ respectively, the following classifier has error no more than $\epsilon + \epsilon_1$ at margin $\frac{\gamma}{2}$*

$$h(x) = sgn[g(x)], g(x) = \frac{1}{d} \sum_{i=1}^{d} w(x_i^+, x_i^-) f\left(K(x, x_i^+) - K(x, x_i^-)\right).$$

Proof. We shall prove that with probability at least $1 - \delta$, at least a $1 - \epsilon_1$ fraction of points x that satisfy Equation 1 are classified correctly by the classifier $h(x)$. Overestimating the error by treating the points that do not satisfy Equation 1 as always being misclassified will give us the desired result.

For any fixed $x \in \mathcal{X}^+$ that satisfies Equation 1, we have

$$\mathop{\mathbb{E}}_{x',x'' \sim \mathcal{D} \times \mathcal{D}} \left[w(x', x'') f\left(K(x, x') - K(x, x'')\right) | \ell(x') = 1, \ell(x'') = -1 \right] \geq C_f \gamma$$

hence the Hoeffding Bounds give us

$$\Pr\left[g(x) < \frac{\gamma}{2}\right] = \Pr\left[\frac{1}{d} \sum_{i=1}^{d} w(x_i^+, x_i^-) f\left(K(x, x_i^+) - K(x, x_i^-)\right) < \frac{\gamma}{2}\right] \leq 2 \exp\left(-\frac{\gamma^2 d}{8}\right)$$

Similarly, for any fixed $x \in \mathcal{X}^-$ that satisfies Equation 1, we have

$$\mathop{\mathbb{E}}_{x',x'' \sim \mathcal{D} \times \mathcal{D}} \left[w(x',x'') f\left(K(x,x') - K(x,x'')\right) | \ell(x') = -1, \ell(x'') = 1 \right] \geq C_f \gamma$$

hence the Hoeffding Bounds give us

$$\begin{aligned} \Pr\left[g(x) > \frac{\gamma}{2}\right] &= \Pr\left[\frac{1}{d}\sum_{i=1}^{d} w(x_i^+, x_i^-) f\left(K(x,x_i^+) - K(x,x_i^-)\right) > \frac{\gamma}{2}\right] \\ &= \Pr\left[\frac{1}{d}\sum_{i=1}^{d} w(x_i^+, x_i^-) f\left(K(x,x_i^-) - K(x,x_i^+)\right) < \frac{\gamma}{2}\right] \leq 2\exp\left(-\frac{\gamma^2 d}{8}\right) \end{aligned}$$

where in the second step we have used antisymmetry of f.

Since we have shown that this result holds true individually for any point x that satisfies Equation 1, the expected error (where the expectation is both over the choice of domain points as well as choice of the landmark points) itself turns out to be less than $2\exp\left(-\frac{\gamma^2 d}{8}\right) \leq \epsilon_1 \delta$. Applying Markov's inequality gives us that the probability of obtaining a set of landmarks such that the error on points satisfying Equation 1 is greater than ϵ_1 is at most δ.

Assuming, as mentioned earlier, that the points not satisfying Equation 1 can always be misclassified proves our desired result. □

2 Comparison with the models of Balcan-Blum and Wang *et al*

In [1], Wang *et al* consider a model of learning with distance functions. Their model is similar to our but for the difference that they restrict themselves to the use of a single transfer function namely the sign function $f = \text{sgn}()$. More formally they have the following notion of a *good distance function*.

Definition 3 ([2] Definition 4)**.** *A distance function $\mathcal{X}$, $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be an (ϵ, γ, B)-good distance for a learning problem where $\epsilon, \gamma, B > 0$ if there exist two class conditional probability distributions $\tilde{\mathcal{D}}(x|\ell(x) = 1)$ and $\tilde{\mathcal{D}}(x|\ell(x) = -1)$ such that for all $x \in \mathcal{X}$, $\frac{\tilde{\mathcal{D}}(x|\ell(x)=1)}{\mathcal{D}(x|\ell(x)=1)} < \sqrt{B}$ and $\frac{\tilde{\mathcal{D}}(x|\ell(x)=-1)}{\mathcal{D}(x|\ell(x)=-1)} < \sqrt{B}$ where $\mathcal{D}(x|\ell(x) = 1)$ and $\mathcal{D}(x|\ell(x) = -1)$ are the class conditional probability distributions of the problem, such that at least a $1 - \epsilon$ probability mass of examples $x \sim \mathcal{D}$ satisfies*

$$\mathop{\tilde{\mathcal{D}}}_{x',x'' \sim \tilde{\mathcal{D}} \times \tilde{\mathcal{D}}} \left[d(x,x') < d(x,x'') | \ell(x') = \ell(x), \ell(x'') \neq \ell(x) \right] \geq \frac{1}{2} + \gamma \tag{2}$$

It can be shown (and is implicit in the proof of Theorem 5 in [1]) that the above condition is equivalent to

$$\mathop{\mathbb{E}}_{x',x'' \sim \mathcal{D} \times \mathcal{D}} \left[w_{\ell(x)}(x') w_{-\ell(x)}(x'') \,\text{sgn}\left(d(x,x'') - d(x,x')\right) | \ell(x') = \ell(x), \ell(x'') \neq \ell(x) \right] \geq 2\gamma$$

where $w_1(x) := \frac{\tilde{\mathcal{D}}(x|\ell(x)=1)}{\mathcal{D}(x|\ell(x)=1)}$ and $w_{-1}(x) := \frac{\tilde{\mathcal{D}}(x|\ell(x)=-1)}{\mathcal{D}(x|\ell(x)=-1)}$. Now define $\varpi(x',x'') := w_{\ell(x')}(x') w_{\ell(x'')}(x'')$ and take $f = \text{sgn}()$ as the transfer function in our model. We have, for a $1 - \epsilon$ fraction of points,

$$\mathop{\mathbb{E}}_{x',x'' \sim \mathcal{D} \times \mathcal{D}} \left[\varpi\left(x',x''\right) f\left(K(x,x') - K(x,x'')\right) | \ell(x') = \ell(x), \ell(x'') \neq \ell(x) \right] \geq C_f \gamma$$

which is clearly seen to be equivalent to

$$\mathop{\mathbb{E}}_{x',x'' \sim \mathcal{D} \times \mathcal{D}} \left[w_{\ell(x)}(x') w_{-\ell(x)}(x'') \,\text{sgn}\left(K(x,x') - K(x,x'')\right) | \ell(x') = \ell(x), \ell(x'') \neq \ell(x) \right] \geq \gamma$$

since $C_f = 1$ for the sgn() function. Thus the Wang *et al* model of learning is an instantiation of our proposed model.

In [2], Balcan-Blum present a model of learning with similarity functions. Their model does not consider landmark pairs, just singletons. Accordingly, instead of assigning a weight to each landmark pair, one simply assigns a weight to each element of the domain. Consequently one arrives at the following notion of a *good similarity*.

Definition 4 ([1], Definition 3)**.** *A similarity measure $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be an (ϵ, γ)-good similarity for a learning problem where $\epsilon, \gamma > 0$ if for some weighing function $w : \mathcal{X} \to [-1, 1]$, at least a $1 - \epsilon$ probability mass of examples $x \sim \mathcal{D}$ satisfies*

$$\mathop{\mathbb{E}}_{x' \sim \mathcal{D}} [w(x') K(x, x') | \ell(x') = \ell(x)] \geq \mathop{\mathbb{E}}_{x' \sim \mathcal{D}} [w(x') K(x, x') | \ell(x') \neq \ell(x)] + \gamma \tag{3}$$

Now define $w_+ := \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [w(x) | \ell(x) = 1]$ and $w_- := \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [w(x) | \ell(x) = -1]$. Furthermore, take $\varpi(x', x'') = w(x')w(x'')$ as the weight function and $f = \text{id}()$ as the transfer function in our model. Then we have, for a $1 - \epsilon$ fraction of the points,

$$\begin{aligned}
& \mathop{\mathbb{E}}_{x', x'' \sim \mathcal{D} \times \mathcal{D}} [\varpi(x', x'') f(K(x, x') - K(x, x'')) | \ell(x') = \ell(x), \ell(x'') \neq \ell(x)] \geq C_f \gamma \\
\equiv & \mathop{\mathbb{E}}_{x', x'' \sim \mathcal{D} \times \mathcal{D}} [\varpi(x', x'') (K(x, x') - K(x, x'')) | \ell(x') = \ell(x), \ell(x'') \neq \ell(x)] \geq \gamma \\
\equiv & \mathop{\mathbb{E}}_{x', x'' \sim \mathcal{D} \times \mathcal{D}} [\varpi(x', x'') K(x, x') | \ell(x') = \ell(x), \ell(x'') \neq \ell(x)] \geq \\
& \mathop{\mathbb{E}}_{x', x'' \sim \mathcal{D} \times \mathcal{D}} [\varpi(x', x'') K(x, x'') | \ell(x') = \ell(x), \ell(x'') \neq \ell(x)] + \gamma \\
\equiv & w_{-\ell(x)} \mathop{\mathbb{E}}_{x' \sim \mathcal{D}} [w(x') K(x, x') | \ell(x') = \ell(x)] \geq w_{\ell(x)} \mathop{\mathbb{E}}_{x' \sim \mathcal{D}} [w(x') K(x, x') | \ell(x') \neq \ell(x)] + \gamma \\
\equiv & \mathop{\mathbb{E}}_{x' \sim \mathcal{D}} [w'(x') K(x, x') | \ell(x') = \ell(x)] \geq \mathop{\mathbb{E}}_{x' \sim \mathcal{D}} [w'(x') K(x, x') | \ell(x') \neq \ell(x)] + \gamma
\end{aligned}$$

where $C_f = 1$ for the id() function and $w'(x) = w(x) w_{-\ell(x)}$. Note that this again guarantees a classifier with margin γ in the landmarked space. Thus the Balcan-Blum model can also be derived in our model.

3 Proof of Theorem 3

Theorem 5 (Theorem 3 restated)**.** *Let $\mathcal{F}$ be a compact class of transfer functions with respect to the infinity norm and $\epsilon_1, \delta > 0$. Let $\mathcal{N}(\mathcal{F}, r)$ be the size of the smallest ϵ-net over $\mathcal{F}$ with respect to the infinity norm at scale $r = \frac{\epsilon_1}{4 C_L B}$. Then if one chooses $d = \frac{64 B^2 C_L^2}{\epsilon_1^2} \ln\left(\frac{16 B \cdot \mathcal{N}(\mathcal{F}, r)}{\delta \epsilon_1}\right)$ random landmark pairs then we have the following with probability greater than $(1 - \delta)$*

$$\sup_{f \in \mathcal{F}} \left[\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[L\left(g_{(f, w_{(g,f)})}(x) \right) \right] - \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[L\left(G_{(f, w_{(G,f)})}(x) \right) \right] \right| \right] \leq \epsilon_1$$

We shall prove the theorem in two parts. As we shall see, one of the parts is fairly simple to prove. To prove the other part, we shall exploit the Lipschitz properties of the loss function as well as the fact that the class of transfer functions chosen form a compact set. Let us call a given set of landmark pairs to be *good* with respect to a fixed transfer function $f \in \mathcal{F}$ if for the corresponding g, $\mathop{\mathbb{E}}_{x}[L(g(x))] \leq \mathop{\mathbb{E}}_{x}[L(G(x))] + \epsilon_1$ for some small fixed $\epsilon_1 > 0$.

We will first prove, using Lipschitz properties of the loss function that if a given set of landmarks is good with respect to a given transfer function, then it is also good with respect to all transfer functions in its neighborhood. Having proved this, we will apply a standard covering number argument in which we will ensure that a large enough set of landmarks is good with respect to a set of transfer functions that form an ϵ-net over $\mathcal{F}$ and use the previous result to complete the proof.

We first prove a series of simple results which will be used in the first part of the proof. In the following f and f' are two transfer functions such that $f' \in \mathcal{B}_\infty(f, r) \cap \mathcal{F}$.

Lemma 6. *The following results are true*

1. *For any fixed $f \in F$, $\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[L\left(G_{(f, w_{(G,f)})}(x) \right) \right] \leq \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[L\left(G_{(f,w)}(x) \right) \right]$ for all $w \in \mathcal{W}$.*
2. *For any fixed $f \in F$, any fixed g obtained by an arbitrary choice of landmark pairs, $\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[L\left(g_{(f, w_{(g,f)})}(x) \right) \right] \leq \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[L\left(g_{(f,w)}(x) \right) \right]$ for all $w \in \mathcal{W}$.*
3. *For any $f' \in \mathcal{B}_\infty(f, r) \cap \mathcal{F}$, $\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[L\left(G_{(f, w_{(G,f)})}(x) \right) \right] - \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[L\left(G_{(f', w_{(G,f')})}(x) \right) \right] \right| \leq C_L r B$.*

4. *For any fixed g obtained by an arbitrary choice of landmark pairs, $f' \in \mathcal{B}_\infty(f,r) \cap \mathcal{F}$,* $\left| \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(g_{(f,w_{(g,f)})}(x) \right) \right] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(g_{(f',w_{(g,f')})}(x) \right) \right] \right| \leq C_L r B$.

Proof. We prove the results in order,

1. Immediate from the definition of $w_{(G,f)}$.
2. Immediate from the definition of $w_{(g,f)}$.
3. We have $\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(G_{(f',w_{(G,f')})}(x) \right) \right] \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(G_{(f',w_{(G,f)})}(x) \right) \right]$ by an application of Lemma 6.1 proven above. For sake of simplicity let us denote $w_{(G,f)} = w$ for the next set of calculations. Now we have

$$\begin{aligned}
G_{(f',w)}(x) &= \mathop{\mathbb{E}}_{x',x''\sim\mathcal{D}\times\mathcal{D}} \left[w\left(x',x''\right) f'\left(K(x,x') - K(x,x'')\right) | \ell(x') = \ell(x), \ell(x'') \neq \ell(x) \right] \\
&\leq \mathop{\mathbb{E}}_{x',x''\sim\mathcal{D}\times\mathcal{D}} \left[w\left(x',x''\right) \left(f\left(K(x,x') - K(x,x'')\right) + r\right) | \ell(x') = \ell(x), \ell(x'') \neq \ell(x) \right] \\
&= \mathop{\mathbb{E}}_{x',x''\sim\mathcal{D}\times\mathcal{D}} \left[w\left(x',x''\right) f\left(K(x,x') - K(x,x'')\right) | \ell(x') = \ell(x), \ell(x'') \neq \ell(x) \right] \\
&\quad + \quad r \cdot \mathop{\mathbb{E}}_{x',x''\sim\mathcal{D}\times\mathcal{D}} \left[w\left(x',x''\right) | \ell(x') = \ell(x), \ell(x'') \neq \ell(x) \right] \\
&\leq G_{(f,w)}(x) + rB
\end{aligned}$$

   where in the second inequality we have used the fact that $\|f - f'\|_\infty \leq r$ and in the fourth inequality we have used the fact that $w \in \mathcal{W}$. Thus we have $G_{(f',w)}(x) \leq G_{(f,w)}(x) + rB$. Using the Lipschitz properties of L we can now get $\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(G_{(f',w)}(x) \right) \right] \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(G_{(f,w)}(x) \right) \right] + C_L r B$. Thus we have $\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(G_{(f',w_{(G,f')})}(x) \right) \right] \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(G_{(f',w_{(G,f)})}(x) \right) \right] \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(G_{(f,w_{(G,f)})}(x) \right) \right] + C_L r B$.

   Similarly we can also prove $\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(G_{(f,w_{(G,f)})}(x) \right) \right] \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(G_{(f',w_{(G,f')})}(x) \right) \right] + C_L r B$. This gives us the desired result.
4. The proof follows in a manner similar to the one for Lemma 6.3 proven above. □

Using the above results we get a preliminary form of the first part of our proof as follows :

Lemma 7. *Suppose a set of landmarks is $(\epsilon_1/2)$-good for a particular landmark $f \in \mathcal{F}$ (i.e. $\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(g_{(f,w_{(G,f)})}(x) \right) \right] < \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(G_{(f,w_{(G,f)})}(x) \right) \right] + \epsilon_1/2$), then the same set of landmarks are also ϵ_1-good for any $f' \in \mathcal{B}_\infty(f,r) \cap \mathcal{F}$ (i.e. for all $f' \in \mathcal{B}_\infty(f,r) \cap \mathcal{F}$, $\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(g_{(f',w_{(g,f')})}(x) \right) \right] \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(G_{(f',w_{(G,f')})}(x) \right) \right] + \epsilon_1$) for some $r = r\left(\epsilon_1\right)$.*

Proof. Theorem 9 proven below guarantees that for any fixed $f \in \mathcal{F}$, with probability $1 - \delta$ that $\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(g_{(f,w_{(G,f)})}(x) \right) \right] < \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[L\left(G_{(f,w_{(G,f)})}(x) \right) \right] + \epsilon_1/2$. This can be achieved with $d = (64B^2C_L^2/\epsilon_1^2)\ln(8B/\delta\epsilon_1)$. Now assuming that the above holds, using the above results we can get

the following for any $f' \in \mathcal{B}_\infty(f,r) \cap \mathcal{F}$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(g_{(f',w_{(g,f')})}(x)\right)\right] &\leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(g_{(f,w_{(g,f)})}(x)\right)\right] + C_L r B \\
&\quad \text{(using Lemma 6.4)} \\
&\leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(g_{(f,w_{(G,f)})}(x)\right)\right] + C_L r B \\
&\quad \text{(using Lemma 6.2)} \\
&\leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(G_{(f,w_{(G,f)})}(x)\right)\right] + \epsilon_1/2 + C_L r B \\
&\quad \text{(using Theorem 9)} \\
&\leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(G_{(f',w_{(G,f')})}(x)\right)\right] + \epsilon_1/2 + 2C_L r B \\
&\quad \text{(using Lemma 6.3)}
\end{aligned}$$

Setting $r = \frac{\epsilon_1}{4C_L B}$ gives us the desired result. □

Proof. (of Theorem 3) As mentioned earlier we shall prove the theorem in two parts as follows :

1. (Part I) In this part we shall prove the following :

$$\sup_{f\in\mathcal{F}}\left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(g_{(f,w_{(g,f)})}(x)\right)\right] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(G_{(f,w_{(G,f)})}(x)\right)\right]\right] \leq \epsilon_1$$

   We first set up an ϵ-net over $\mathcal{F}$ at scale $r = \frac{\epsilon_1}{4C_L B}$. Let there be $\mathcal{N}(\mathcal{F},r)$ elements in this net. Taking $d = (64B^2C_L^2/\epsilon_1^2)\ln(8B\cdot\mathcal{N}(\mathcal{F},r)/\delta\epsilon_1)$ landmarks should ensure that the landmarks, with very high probability, are good for all functions in the net by an application of union bound. Since every function in $\mathcal{F}$ is at least r-close to some function in the net, Lemma 7 tells us that the same set of landmarks are, with very high probability, good for all the functions in $\mathcal{F}$. This proves the first part of our result.

2. (Part II) In this part we shall prove the following :

$$\sup_{f\in\mathcal{F}}\left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(G_{(f,w_{(G,f)})}(x)\right)\right] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(g_{(f,w_{(g,f)})}(x)\right)\right]\right] \leq \epsilon_1$$

   This part is actually fairly simple to prove. Intuitively, since one can imagine G as being the output of an algorithm that is allowed to take the entire domain as its landmark set, we should expect $\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(G_{(f,w_{(G,f)})}(x)\right)\right] \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(g_{(f,w_{(g,f)})}(x)\right)\right]$ to hold unconditionally for every f. For a formal argument, let us build up some more notation. As we have said before, for any transfer function f and arbitrary choice of d landmark pairs $\mathcal{P}$, we let $w_{(g,f)} \in [-B,B]^d$ be the best weighing function for this choice of transfer function and landmark pairs. Now let $\overline{w_{(g,f)}}$ be the best possible extension of $w_{(g,f)}$ to the entire domain. More formally, for any $w^* \in [-B,B]^d$ let $\overline{w^*} = \mathop{\arg\min}_{w\in\mathcal{W}, w_{|\mathcal{P}}=w^*} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(G_{(f,w)}(x)\right)\right]$.

   Now Lemma 6.1 tells us that for any $f \in \mathcal{F}$ and any choice of landmark pairs $\mathcal{P}$, $\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(G_{(f,w_{(G,f)})}(x)\right)\right] \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(G_{(f,\overline{w_{(g,f)}})}(x)\right)\right]$. Furthermore, since $\overline{w_{(g,f)}}$ is chosen to be the most beneficial extension of $w_{(g,f)}$, we also have $\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(G_{(f,\overline{w_{(g,f)}})}(x)\right)\right] \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L\left(g_{(f,w_{(g,f)})}(x)\right)\right]$. Together, these two inequalities give us the second part of the proof. □

4 Proof of Theorem 5

We first recall the definition of a good similarity under a given loss function.

Definition 8. *A similarity measure over a domain $\mathcal{X}$, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be an (ϵ, B)-good similarity for a learning problem with respect to a loss function $L : \mathbb{R} \to \mathbb{R}^+$ where $\epsilon > 0$ if for some transfer function $f : \mathbb{R} \to \mathbb{R}$ and some weighing function $w : \mathcal{X} \times \mathcal{X} \to [-B, B]$, the following holds true*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L(G(x))\right] \leq \epsilon \tag{4}$$

where $G(x) = \mathop{\mathbb{E}}_{x',x''\sim\mathcal{D}\times\mathcal{D}}\left[w\left(x',x''\right) f\left(K(x,x') - K(x,x'')\right) | \ell(x') = \ell(x), \ell(x'') \neq \ell(x)\right]$

Theorem 9 (Theorem 5 restated)**.** *If K is an (ϵ, B)-good similarity measure with respect to a C_L-Lipschitz loss function L then for any $\epsilon_1 > 0$, with probability at least $1 - \delta$ over the choice of $d = (16B^2C_L^2/\epsilon_1^2)\ln(4B/\delta\epsilon_1)$ positive and negative samples from $\mathcal{D}^+$ and $\mathcal{D}^-$ respectively, the following classifier has expected loss no more than $\epsilon + \epsilon_1$ with respect to the loss function L*

$$h(x) = sgn[g(x)], g(x) = \frac{1}{d}\sum_{i=1}^{d} w\left(x_i^+, x_i^-\right) f\left(K(x,x_i^+) - K(x,x_i^-)\right).$$

i.e. $\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[L(g(x))\right] \leq \epsilon + \epsilon_1$ *where* $\left\{x_i^+\right\}_{i=1}^d$ *are the positive samples and* $\left\{x_i^-\right\}_{i=1}^d$ *are the negative samples.*

Proof. For any $x \in \mathcal{X}$, we have, by an application of Hoeffding bounds $\Pr_g\left[|G(x) - g(x)| > \epsilon_1\right] < 2\exp\left(-\frac{\epsilon_1^2 d}{2B^2}\right)$ since $|g(x)| \leq B$. Here the notation $\Pr_g$ signifies that the probability is over the choice of the landmark points. Thus for $d > \frac{4B^2}{\epsilon_1^2}\ln\left(\frac{2}{\delta}\right)$, we have $\Pr_g\left[|G(x) - g(x)| > \epsilon_1\right] < \delta^2$. For sake of simplicity let us denote by $\mathsf{BAD}\,(x)$ the event $|G(x) - g(x)| > \epsilon_1$. Thus we have, for every $x \in \mathcal{X}$, $\mathop{\mathbb{E}}_g\left[1_{\mathsf{BAD}(x)}\right] < \delta^2$. Since this is true for every $x \in \mathcal{X}$, this also holds in expectation i.e. $\mathop{\mathbb{E}}_x\mathop{\mathbb{E}}_g\left[1_{\mathsf{BAD}(x)}\right] < \delta^2$. The expectation over x is with respect to the problem distribution $\mathcal{D}$. Applying Fubini's Theorem gives us $\mathop{\mathbb{E}}_g\mathop{\mathbb{E}}_x\left[1_{\mathsf{BAD}(x)}\right] < \delta^2$ which upon application of Markov's inequality gives us $\Pr_g\left[\mathop{\mathbb{E}}_x\left[1_{\mathsf{BAD}(x)}\right] > \delta\right] < \delta$. Thus, with very high probability we would always choose landmarks such that $\Pr_x\left[BAD(x)\right] < \delta$. Thus we have, in such a situation, $\mathop{\mathbb{E}}_x\left[|G(x) - g(x)|\right] \leq (1-\delta)\epsilon_1 + \delta\cdot 2B$ since $\sup_{x\in\mathcal{X}}|G(x) - g(x)| \leq 2B$. For small enough δ we have $\mathop{\mathbb{E}}_x\left[|G(x) - g(x)|\right] \leq 2\epsilon_1$.

Thus we have $\mathop{\mathbb{E}}_x\left[L(g(x))\right] - \mathop{\mathbb{E}}_x\left[L(G(x))\right] = \mathop{\mathbb{E}}_x\left[L(g(x)) - L(G(x))\right] \leq \mathop{\mathbb{E}}_x\left[C_L\cdot|g(x) - G(x)|\right] = C_L\cdot\mathop{\mathbb{E}}_x\left[|g(x) - G(x)|\right] \leq 2C_L\epsilon_1$ where we used the Lipschitz properties of the loss function L to arrive at the second inequality. Putting $\epsilon_1 = \frac{\epsilon_1'}{2C_L}$ we have $\mathop{\mathbb{E}}_x\left[L(g(x))\right] \leq \mathop{\mathbb{E}}_x\left[L(G(x))\right] + \epsilon_1' \leq \epsilon + \epsilon_1'$ which gives us our desired result.

Actually we can prove something stronger since $\left|\mathop{\mathbb{E}}_x\left[L(g(x))\right] - \mathop{\mathbb{E}}_x\left[L(G(x))\right]\right| = \left|\mathop{\mathbb{E}}_x\left[L(g(x)) - L(G(x))\right]\right| \leq \mathop{\mathbb{E}}_x\left[|L(g(x)) - L(G(x))|\right] \leq \mathop{\mathbb{E}}_x\left[C_L\cdot|g(x) - G(x)|\right] \leq \epsilon_1'$. Thus we have $\epsilon - \epsilon_1' \leq \mathop{\mathbb{E}}_x\left[L(g(x))\right] \leq \epsilon + \epsilon_1'$. □

References

[1] Liwei Wang, Cheng Yang, and Jufu Feng. On Learning with Dissimilarity Functions. In *International Conference on Machine Learning*, pages 991–998, 2007.

[2] Maria-Florina Balcan and Avrim Blum. On a Theory of Learning with Similarity Functions. In *International Conference on Machine Learning*, pages 73–80, 2006.