Appendix: Proofs of Theorems

Proof of Theorem 2

We fix an RKHS H on the input space $X \subset \mathbb{R}^d$ with an RBF kernel k. Let $\mathbf{x} = \{x_1, \ldots, x_n\}$ be a set of objects to be ranked in $\mathbb{R}^d$ with labels $\mathbf{r} = \{r_1, \ldots, r_n\}$. Here r_i denotes the label of x_i, and $r_i \in \mathbb{R}$. We assume $\mathbf{x}$ to be a random variable distributed according to P, and $\mathbf{r}$ deterministic. Throughout L denotes the hinge loss.

The following notation will be useful in the proof of Theorem 2. Take T to be the set of pairs derived from $\mathbf{x}$ and define the *L-risk* of $f \in H$ as

$$\mathcal{R}_{L,P}(f) := E_{\mathbf{x}}[\mathcal{R}_{L,T}(f)]$$

where

$$\mathcal{R}_{L,T}(f) = \sum_{i,j:r_i > r_j} D(r_i, r_j) L(f(x_i) - f(x_j))$$

and $D(r_i, r_j)$ is some positive weight function, which we take for simplicity to be $1/|\mathcal{P}|$, $\mathcal{P} = \{(i,j) : r_i > r_j\}$. $\mathcal{R}_{L,T}(f)$ is the *empirical* L-risk of f, with respect to the empirical distribution over the pairs of samples, which we denote by T. This uniform weight is the setting we have taken in the main body of the paper. The smallest possible L-risk in H is denoted

$$\mathcal{R}_{L,P} := \inf_{f \in H} \mathcal{R}_{L,P}(f).$$

The *regularized L-risk* is

$$\mathcal{R}^{\text{reg}}_{L,P,\lambda}(f) := \lambda \|f\|^2 + \mathcal{R}_{L,P}(f), \tag{1}$$

$\lambda > 0$.

For simplicity we assume the preference pair set $\mathcal{P}$ contains all pairs over these n samples. Let $g_{\mathbf{x},\lambda}$ be the optimal solution to the rank-AD minimization step. We have,

$$g_{\mathbf{x},\lambda} = \arg\min_{f \in H} \mathcal{R}_{L,T}(f) + \lambda ||f||^2 \tag{2}$$

Let $\mathcal{H}_n$ denote a ball of radius $O(1/\sqrt{\lambda_n})$ in H. Let $C_k := \sup_{x,t} |k(x,t)|$ with k the rbf kernel associated to H. Given $\epsilon > 0$, we let $N(\mathcal{H}, \epsilon/4C_k)$ be the covering number of $\mathcal{H}$ by disks of radius $\epsilon/4C_k$. We first show that with appropriately chosen λ, as $n \to \infty$, $g_{\mathbf{x},\lambda}$ is consistent in the following sense.

Proposition 1. *Let λ_n be appropriately chosen such that $\lambda_n \to 0$ and $\frac{\log N(\mathcal{H}_n, \epsilon/4C_k)}{n\lambda_n} \to 0$, as $n \to \infty$. Then we have*

$$E_{\mathbf{x}}[\mathcal{R}_{L,T}(g_{\mathbf{x},\lambda_n})] \to \mathcal{R}_{L,P} = \min_{f \in H} \mathcal{R}_{L,P}(f), \quad n \to \infty.$$

Proof Let us outline the argument. In [Steinwart, 2001], the author shows that there exists a $f_{P,\lambda} \in H$ minimizing (1):

- For all Borel probability measures P on $X \times X$ and all $\lambda > 0$, there is an $f_{P,\lambda} \in H$ with
$$\mathcal{R}^{\text{reg}}_{L,P,\lambda}(f_{P,\lambda}) = \inf_{f \in H} \mathcal{R}^{\text{reg}}_{L,P,\lambda}(f)$$
such that $\|f_{P,\lambda}\| = O(1/\sqrt{\lambda})$. (If P is the empirical distribution over data T, then we denote this minimizer by $f_{T,\lambda}$.)

Next, a simple argument shows that

- $\lim_{\lambda \to 0} \mathcal{R}^{\text{reg}}_{L,P,\lambda}(f_{P,\lambda}) = \mathcal{R}_{L,P}$.

Finally, we will need a concentration inequality to relate the L-risk of $f_{P,\lambda}$ with the empirical L-risk of $f_{T,\lambda}$. We then derive consistency using the following argument:

$$\begin{aligned}
\mathcal{R}_{L,P}(f_{T,\lambda_n}) &\leq \lambda_n \|f_{T,\lambda_n}\|^2 + \mathcal{R}_{L,P}(f_{T,\lambda_n}) \\
&\leq \lambda_n \|f_{T,\lambda_n}\|^2 + \mathcal{R}_{L,T}(f_{T,\lambda_n}) + \delta/3 \\
&\leq \lambda_n \|f_{P,\lambda_n}\|^2 + \mathcal{R}_{L,T}(f_{P,\lambda_n}) + \delta/3 \\
&\leq \lambda_n \|f_{P,\lambda_n}\|^2 + \mathcal{R}_{L,P}(f_{P,\lambda_n}) + 2\delta/3 \\
&\leq \mathcal{R}_{L,P} + \delta
\end{aligned}$$

where λ_n is an appropriately chosen sequence $\to 0$, and n is large enough. The second and fourth inequality hold due to Concentration Inequalities, and the last one holds since $\lim_{\lambda \to 0} \mathcal{R}^{\text{reg}}_{L,P,\lambda}(f_{P,\lambda}) = \mathcal{R}_{L,P}$.

We now prove the appropriate concentration inequality [Cucker and Smale, 2001]. Recall H is an RKHS with smooth kernel k; thus the inclusion $I_k : H \to C(X)$ is compact, where $C(X)$ is given the $\|\cdot\|_\infty$-topology. That is, the "hypothesis space" $\mathcal{H} := \overline{I_k(B_R)}$ is compact in $C(X)$, where B_R denotes the ball of radius R in H. We denote by $N(\mathcal{H}, \epsilon)$ the covering number of $\mathcal{H}$ with disks of radius ϵ. We prove the following inequality:

Lemma 2. *For any probability distribution P on $X \times X$,*

$$P^{\epsilon_n}\{T \in (X \times X)^{\epsilon_n} : \sup_{f \in \mathcal{H}} |\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f)| \geq \epsilon\} \leq 2N(\mathcal{H}, \epsilon/4C_k) \exp\left(\frac{-\epsilon^2 n}{2(1 + 2\sqrt{C_k}R)^2}\right),$$

where $C_k := \sup_{x,t} |k(x,t)|$.

Proof Since $\mathcal{H}$ is compact, it has a finite covering number. Now suppose $\mathcal{H} = D_1 \cup \cdots \cup D_\ell$ is any finite covering of $\mathcal{H}$. Then

$$\text{Prob}\{\sup_{f \in \mathcal{H}} |\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f)| \geq \epsilon\} \leq \sum_{j=1}^{\ell} \text{Prob}\{\sup_{f \in D_j} |\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f)| \geq \epsilon\}$$

so we restrict attention to a disk D in $\mathcal{H}$ of appropriate radius ϵ.

Suppose $\|f-g\|_\infty \leq \epsilon$. We want to show that the difference

$$|(\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f)) - (\mathcal{R}_{L,T}(g) - \mathcal{R}_{L,P}(g))|$$

is also small. Rewrite this quantity as

$$|(\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,T}(g)) - E_{\mathbf{x}}[\mathcal{R}_{L,T}(g) - \mathcal{R}_{L,T}(f)]|.$$

Since $\|f-g\|_\infty \leq \epsilon$, for ϵ small enough we have

$$\begin{aligned} \max\{0, 1-(f(x_i)-f(x_j))\} - \max\{0, 1-(g(x_i)-g(x_j))\} &= \max\{0, (g(x_i)-g(x_j)-f(x_i)+f(x_j))\} \\ &= \max\{0, \langle g-f, \phi(x_i)-\phi(x_j)\rangle\}. \end{aligned}$$

Here $\phi : X \to H$ is the feature map, $\phi(x) := k(x,\cdot)$. Combining this with the Cauchy-Schwarz inequality, we have

$$|(\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,T}(g)) - E_{\mathbf{x}}[\mathcal{R}_{L,T}(g) - \mathcal{R}_{L,T}(f)]| \;\; \leq \tfrac{2}{n^2}(2n^2\|f-g\|_\infty C_k) \;\; \leq 4C_k\epsilon,$$

where $C_k := \sup_{x,t}|k(x,t)|$. From this inequality it follows that

$$|\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f)| \geq (4C_k+1)\epsilon \implies |(\mathcal{R}_{L,T}(g) - \mathcal{R}_{L,P}(g))| \geq \epsilon.$$

We thus choose to cover $\mathcal{H}$ with disks of radius $\epsilon/4C_k$, centered at $f_1, \ldots, f_\ell$. Here $\ell = N(\mathcal{H}, \epsilon/4C_k)$ is the covering number for this particular radius. We then have

$$\sup_{f\in D_j} |(\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f))| \geq 2\epsilon \implies |(\mathcal{R}_{L,T}(f_j) - \mathcal{R}_{L,P}(f_j))| \geq \epsilon.$$

Therefore,

$$\text{Prob}\{\sup_{f\in\mathcal{H}} |\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f)| \geq 2\epsilon\} \leq \sum_{j=1}^{n} \text{Prob}\{|\mathcal{R}_{L,T}(f_j) - \mathcal{R}_{L,P}(f_j)| \geq \epsilon\}$$

The probabilities on the RHS can be bounded using McDiarmid's inequality.

Define the random variable $g(x_1, \ldots, x_n) := \mathcal{R}_{L,T}(f)$, for fixed $f \in H$. We need to verify that g has bounded differences. If we change one of the variables, x_i, in g to x_i', then at most n summands will change:

$$\begin{aligned} |g(x_1,\ldots,x_i,\ldots,x_n) - g(x_1,\ldots,x_i',\ldots,x_n)| &\leq \frac{1}{n^2} 2n \sup_{x,y} |1-(f(x)-f(y))| \\ &\leq \frac{2}{n} + \frac{2}{n}\sup_{x,y}|f(x)-f(y)| \\ &\leq \frac{2}{n} + \frac{4}{n}\sqrt{C_k}\|f\|. \end{aligned}$$

Using that $\sup_{f\in\mathcal{H}}\|f\| \leq R$, McDiarmid's inequality thus gives

$$\text{Prob}\{\sup_{f\in\mathcal{H}} |\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f)| \geq \epsilon\} \leq 2N(\mathcal{H}, \epsilon/4C_k) \exp\left(\frac{-\epsilon^2 n}{2(1+2\sqrt{C_k}R)^2}\right).$$

■

We are now ready to prove Theorem 2. Take $R = \|f_{P,\lambda}\|$ and apply this result to $f_{P,\lambda}$:

$$\text{Prob}\{|\mathcal{R}_{L,T}(f_{P,\lambda}) - \mathcal{R}_{L,P}(f_{P,\lambda})| \geq \epsilon\} \leq 2N(\mathcal{H}, \epsilon/4C_k) \exp\left(\frac{-\epsilon^2 n}{2(1+2\sqrt{C_k}\|f_{P,\lambda}\|)^2}\right).$$

Since $\|f_{P,\lambda_n}\| = O(1/\sqrt{\lambda_n})$, the RHS converges to 0 so long as $\dfrac{n\lambda_n}{\log N(\mathcal{H}, \epsilon/4C_k)} \to \infty$ as $n \to \infty$. This completes the proof of Theorem 2.

■

We now establish that under mild conditions on the surrogate loss function, the solution minimizing the expected surrogate loss will asymptotically recover the correct preference relationships given by the density f.

Proposition 3. *Let L be a non-negative, non-increasing convex surrogate loss function that is differentiable at zero and satisfies $L'(0) < 0$. If*

$$g^* = \arg\min_{g \in H} E_{\mathbf{x}}\left[\mathcal{R}_{L,T}(g)\right],$$

then g^* will correctly rank the samples according to their density, i.e. $\forall x_i \neq x_j, f(x_i) > f(x_j) \implies g^*(x_i) > g^*(x_j)$. Assume the input preference pairs satisfy: $\mathcal{P} = \{(x_i, x_j) : f(x_i) > f(x_j)\}$, where $\mathbf{x} = \{x_1, \ldots, x_n\}$ is drawn i.i.d. from distribution f. Let ℓ be some convex surrogate loss function that satisfies: (1) ℓ is non-negative and non-increasing; (2) ℓ is differentiable and $\ell'(0) < 0$. Then the optimal solution: g^*, will correctly rank the samples according to f, i.e. $g^*(x_i) > g^*(x_j)$, $\forall x_i \neq x_j, f(x_i) > f(x_j)$, .

The hinge-loss satisfies the conditions in the above theorem. Combining Theorem 1 and 3, we establish that asymptotically, the rankAD step yields a ranker that preserves the preference relationship on nominal samples given by the nominal density f.

Proof Our proof follows similar lines of Theorem 4 in [Lan et al., 2012]. Assume that $g(x_i) < g(x_j)$, and define a function g' such that $g'(x_i) = g(x_j)$, $g'(x_j) = g(x_i)$, and $g'(x_k) = g(x_k)$ for all $k \neq i, j$. We have $\mathcal{R}_{L,P}(g') - \mathcal{R}_{L,P}(g) = E_{\mathbf{x}}(A(\mathbf{x}))$, where

$$
\begin{aligned}
A(\mathbf{x}) = \sum_{k:r_j<r_i<r_k} & [D(r_k,r_j) - D(r_k,r_i)][L(g(x_k)-g(x_i)) - L(g(x_k)-g(x_j))] \\
& + \sum_{k:r_j<r_k<r_i} D(r_i,r_k)[L(g(x_j)-g(x_k)) - L(g(x_i)-g(x_k))] \\
& + \sum_{k:r_j<r_k<r_i} D(r_k,r_j)[L(g(x_k)-g(x_i)) - L(g(x_k)-g(x_j))] \\
+ \sum_{k:r_j<r_i<r_k} & [D(r_k,r_j) - D(r_k,r_i)][L(g(x_k)-g(x_i)) - L(g(x_k)-g(x_j))] \\
+ \sum_{k:r_j<r_i<r_k} & [D(r_i,r_k) - D(r_j,r_k)][L(g(x_j)-g(x_k)) - L(g(x_i)-g(x_k))] \\
& + (L(g(x_j)-g(x_i)) - L(g(x_i)-g(x_j)))D(r_i,r_j).
\end{aligned}
$$

Using the requirements of the weight function D and the assumption that L is non-increasing and non-negative, we see that all six sums in the above equation for $A(\mathbf{x})$ are negative. Thus $A(\mathbf{x}) < 0$, so $\mathcal{R}_{L,P}(g') - \mathcal{R}_{L,P}(g) = E_{\mathbf{x}}(A(\mathbf{x})) < 0$, contradicting the minimality of g. Therefore $g(x_i) \geq g(x_j)$.

Now we assume that $g(x_i) = g(x_j) = g_0$. Since $\mathcal{R}_{L,P}(g) = \inf_{h\in H} \mathcal{R}_{L,P}(h)$, we have $\left.\dfrac{\partial \ell_L(g;x)}{\partial g(x_i)}\right|_{g_0} = A = 0$, and $\left.\dfrac{\partial \ell_L(g;x)}{\partial g(x_j)}\right|_{g_0} = B = 0$, where

$$
\begin{aligned}
A = \sum_{k:r_j<r_i<r_k} D(r_k,r_i)[-L'(g(x_k)-g_0)] + \sum_{k:r_j<r_k<r_i} D(r_i,r_k)L'(g_0-g(x_k)) + \\
\sum_{k:r_k<r_j<r_i} D(r_i,r_k)L'(g_0-g(x_k)) + D(r_i,r_j)[-L'(0)].
\end{aligned}
$$

$$
\begin{aligned}
B = \sum_{k:r_j<r_i<r_k} D(r_k,r_j)[-L'(g(x_k)-g_0)] + \sum_{k:r_j<r_k<r_i} D(r_k,r_j)L'(g_0-g(x_k)) + \\
\sum_{k:r_k<r_j<r_i} D(r_j,r_k)L'(g_0-g(x_k)) + D(r_i,r_j)[-L'(0)].
\end{aligned}
$$

However, using $L'(0) < 0$ and the requirements of D we have

$$
A - B \leq 2L'(0)D(r_i,r_j) < 0,
$$

contradicting $A = B = 0$.

■

The following lemma completes the proof of Theorem 2:

Lemma 4. *Assume G is any function that gives the same order relationship as the density: $G(x_i) > G(x_j)$, $\forall x_i \neq x_j$ such that $f(x_i) > f(x_j)$. Then*

$$
\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{G(x_i)\leq G(\eta)\}} \to p(\eta). \tag{3}
$$

Proof of Theorem 3

To prove Theorem 3 we need the following lemma [Vapnik, 1979]:

Lemma 5. *Let $\mathcal{X}$ be a set and S a system of sets in $\mathcal{X}$, and P a probability measure on S. For $\mathbf{X} \in \mathcal{X}^n$ and $A \in S$, define $\nu_{\mathbf{X}}(A) := |\mathbf{X} \cap A|/n$. If $n > 2/\epsilon$, then*

$$P^n \left\{ \mathbf{X} : \sup_{A \in S} |\nu_{\mathbf{X}}(A) - P(A)| > \epsilon \right\} \leq 2P^{2n} \left\{ \mathbf{X}\mathbf{X}' : \sup_{A \in S} |\nu_{\mathbf{X}}(A) - \nu_{\mathbf{X}'}(A)| > \epsilon/2 \right\}.$$

Now to the proof of the Theorem. Consider the event

$$J := \left\{ \mathbf{X} \in \mathcal{X}^n : \exists f \in \mathcal{F}, P\{x : f(x) < f^{(m)} - 2\gamma\} > \frac{m-1}{n} + \epsilon \right\}.$$

We must show that $P^n(J) \leq \delta$ for $\epsilon = \epsilon(n, k, \delta)$. Fix k and apply lemma 5 with

$$A = \{x : f(x) < f^{(m)} - 2\gamma\}$$

with γ small enough so that

$$\nu_{\mathbf{X}}(A) = |\{x_j \in \mathbf{X} : f(x_j) < f^{(m)} - 2\gamma\}|/n = \frac{m-1}{n}.$$

We obtain

$$P^n(J) \leq 2P^{2n} \left\{ \mathbf{X}\mathbf{X}' : \exists f \in \mathcal{F}, |\{x'_j \in \mathbf{X}' : f(x'_j) < f^{(m)} - 2\gamma\}| > \epsilon n/2 \right\}.$$

The remaining portion of the proof follows as Theorem 12 in [Schölkopf et al., 2001].

References

F. Cucker and S. Smale. On the mathematical foundations of learning. In *Bull. Amer. Math. Soc.*, pages 1–49, 2001.

Y. Lan, J. Guo, X. Cheng, and T. Liu. Statistical consistency of ranking methods in a rank-differentiable probability space. In *Advances in Neural Information Processing Systems*, pages 1241–1249, 2012.

B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7): 1443–1471, 2001.

I. Steinwart. Consistency of support vector machines and other regularized kernel machines. In *IEEE Trans. Inform. Theory*, pages 67–93, 2001.

V. Vapnik. *Estimation of Dependences Based on Empirical Data [in Russian]*. English translation: Springer Verlag, New York, 1982, 1979.