Supplement for the paper titled "Co-regularization Based Semi-supervised Domain Adaptation"

Hal Daumé III
Department of Computer Science
University of Maryland CP, MD, USA
hal@umiacs.umd.edu

Abhishek Kumar
Department of Computer Science
University of Maryland CP, MD, USA
abhishek@umiacs.umd.edu

Avishek Saha
School Of Computing
University of Utah, UT, USA
avishek@cs.utah.edu

In the following, we provide proofs for Theorem 4.2, Theorem 4.4 and Theorem 4.5. Note that the derivations and proofs make use of the kernel sub-matrices A, B, C, D, E, F (as defined in Eq. 4.6 of the original paper).

1 Proof of Theorem 4.2

Let h_s^* and h_t^* be the optimal source and target hypotheses in $\mathcal{H}_s$ and $\mathcal{H}_t$ respectively. Using triangle inequality for the loss function, we have

$$\epsilon_t(h_t, f_t) \leq \epsilon_t(h_t, h_t^*) + \epsilon_t(h_t^*, f_t).$$

We use the notion of $d_{\mathcal{H}\Delta\mathcal{H}}$-distance in the next step, which is defined as $\sup_{h_1,h_2\in\mathcal{H}} 2|\epsilon_s(h_1,h_2) - \epsilon_t(h_1,h_2)|$ [1]. This gives us

$$\epsilon_t(h_t, f_t) \leq \epsilon_s(h_t, h_t^*) + \frac{1}{2} d_{\mathcal{H}_t\Delta\mathcal{H}_t}(D_s, D_t) + \epsilon_t(h_t^*, f_t).$$

We make use of triangle inequality again to get

$$\epsilon_t(h_t, f_t) \leq \epsilon_s(h_t, f_s) + \epsilon_s(f_s, f_t) + \epsilon_s(h_t^*, f_t) + \frac{1}{2} d_{\mathcal{H}_t\Delta\mathcal{H}_t}(D_s, D_t) + \epsilon_t(h_t^*, f_t).$$

We denote $\eta_s := \epsilon_s(f_s, f_t)$, $\nu_s := \epsilon_s(h_t^*, f_t)$, and $\nu_t := \epsilon_t(h_t^*, f_t)$. Subtracting $\epsilon_s(h_s, f_s)$ from both sides, we get

$$\begin{aligned}
\epsilon_t(h_t, f_t) - \epsilon_s(h_s, f_s) &\leq (\epsilon_s(h_t, f_s) - \epsilon_s(h_s, f_s)) + \frac{1}{2} d_{\mathcal{H}_t\Delta\mathcal{H}_t}(D_s, D_t) + \eta_s + \nu_s + \nu_t \\
&\leq M E_s[h_t(x) - h_s(x)] + \frac{1}{2} d_{\mathcal{H}_t\Delta\mathcal{H}_t}(D_s, D_t) + \eta_s + \nu_s + \nu_t \\
&\qquad \text{(using M-Lipschitz property of loss function)} \\
&= M E_s[\langle h_t, k(x,\cdot)\rangle - \langle h_s, k(x,\cdot)\rangle] + \frac{1}{2} d_{\mathcal{H}_t\Delta\mathcal{H}_t}(D_s, D_t) + \eta_s + \nu_s + \nu_t \\
&\qquad \text{(using the reproducing kernel property)} \\
&= M E_s[\langle h_t - h_s, k(x,\cdot)\rangle] + \frac{1}{2} d_{\mathcal{H}_t\Delta\mathcal{H}_t}(D_s, D_t) + \eta_s + \nu_s + \nu_t \\
&\leq M ||h_t - h_s|| E_s[||k(x,\cdot)||] + \frac{1}{2} d_{\mathcal{H}_t\Delta\mathcal{H}_t}(D_s, D_t) + \eta_s + \nu_s + \nu_t \\
&= M ||h_t - h_s|| E_s[\sqrt{k(x,x)}] + \frac{1}{2} d_{\mathcal{H}_t\Delta\mathcal{H}_t}(D_s, D_t) + \eta_s + \nu_s + \nu_t.
\end{aligned}$$

(Note: Some of the steps involving reduction to the term $E_s\left[\sqrt{k(x,x)}\right]$ are similar to [2].)

2 Proof of Theorem 4.4: Complexity for EA

In this section, we bound the complexity of target hypothesis class $\mathcal{J}^t_{EA}$ for EA. The base hypothesis class $\mathcal{H}$ in Eq. 4.3 (of the original paper) is symmetric in source and target hypotheses. So the complexity of source class $\mathcal{J}^s_{EA}$ can be obtained by replacing adequate terms. We are interested in the complexity of the target hypothesis class $\mathcal{J}^t_{EA}$ which is defined as $\mathcal{J}^t_{EA} := \{h_2 : \mathcal{X} \mapsto \mathbb{R},\, (h_1, h_2) \in \mathcal{H}\}$, where h_1 is not fixed a priori.

The Rademacher complexity of $\mathcal{J}^t_{EA}$ is defined as

$$\hat{R}_n(\mathcal{J}^t_{EA}) = E_\sigma \left[\sup_{(h_1,h_2)\in\mathcal{H}} \left| \frac{2}{l_t} \sum_{i=1}^{l_t} \sigma_i h_2(x_i) \right| \right] \tag{2.1}$$

The basic framework of proof is similar to the proof of the main theorem of [3]. The hypothesis class considered in their work is different than ours. They find the complexity of average hypothesis class (i.e., $x \mapsto (h_1(x) + h_2(x))/2$), while we are interested in class $\mathcal{J}^t_{EA}$, as defined above. We also note that $h_2 \in \mathcal{J}^t_{EA} \implies -h_2 \in \mathcal{J}^t_{EA}$ since $(h_1, h_2) \in \mathcal{H} \implies (-h_1, -h_2) \in \mathcal{H}$. This means that we can remove the absolute value sign from Eq. 2.1. Since, $\forall i, h_2(x_i) = \langle k(x_i, \cdot), h_2 \rangle$, we can restrict the supremum to h_1 and h_2 that are in the span of all samples and also in $\mathcal{H}$. The restricted condition on (h_1, h_2) then becomes

$$\{(h_\alpha, h_\beta) : \lambda_1 \alpha' K \alpha + \lambda_2 \beta' K \beta + \lambda(\alpha - \beta)' K (\alpha - \beta) \le 1\} = \{(h_\alpha, h_\beta) : (\alpha'\ \beta') M (\alpha'\ \beta')' \le 1\}$$

where

$$M = \begin{pmatrix} (\lambda_1 + \lambda)K & -\lambda K \\ -\lambda K & (\lambda_2 + \lambda)K \end{pmatrix},$$

and K is the kernel matrix for source labeled and target labeled samples. Using the reproducing kernel property, we get

$$\hat{R}_n(\mathcal{J}^t_{EA}) = \frac{2}{l_t} E_\sigma \sup_{\alpha,\beta\in\mathbb{R}^{l_s+l_t}} \{\sigma'(C'B)\beta : (\alpha'\ \beta') M (\alpha'\ \beta')' \le 1\}.$$

For a symmetric positive definite matrix M, it can be shown that

$$\sup_{(\alpha,\beta):(\alpha'\ \ \beta')M(\alpha'\ \ \beta')'\le 1} x'\beta = ||(M/M_{11})^{-1/2} x|| = ||(M^{-1})^{1/2}_{22} x||, \tag{2.2}$$

and the maxima occurs at $\alpha = -M_{11}^{-1} M_{12} \beta$. M/M_{11} is the Schur complement of block M_{11} of matrix M (i.e. $M/M_{11} = M_{22} - M_{21} M_{11}^{-1} M_{12}$).

The matrix M may not always be full rank, however it can be noted that if β is in the null space of K, $(C'\ B)\beta$ will be zero. So, we can project β onto the column space of K (or row space due to K being a symmetric matrix) to get β_{pr} and the term $(C'\ B)\beta_{pr}$ is equal to $(C'\ B)\beta$. Specifically, β_{pr} can be thought as computed by the operation $UU_{pr}^T\beta$ where U is the full eigenvector matrix and U_{pr} is the eigenvector matrix consisting of only the vectors having nonzero eigenvalues. So, the sup is restricted to the projected α_{pr} and β_{pr}, and the expression for Rademacher complexity can be rewritten as

$$\hat{R}_n(\mathcal{J}^t_{EA}) = \frac{2}{l_t} E_\sigma \sup_{\alpha_{pr},\beta_{pr}\in ColSpace\{K\}} \Big\{ \sigma'(C'\ B)\beta_{pr} : (\alpha'_{pr}\beta'_{pr}) M (\alpha'_{pr}\beta'_{pr})' \le 1 \Big\}.$$

We proceed in a manner similar to that used in [3] and diagonalize the kernel matrix K to get orthonormal bases U corresponding the nonzero eigenvalues ($K = U'\Lambda U$). Λ is a diagonal matrix of size $r \times r$, containing just the nonzero eigenvalues and r is the rank of matrix K. Since α_{pr} and β_{pr} are in the span of column space of K, there exist a_s and b such that

$$\alpha_{pr} = Ua \qquad \text{and} \qquad \beta_{pr} = Ub$$

The expression for complexity now becomes, $\hat{R}_n(\mathcal{J}^t_{EA}) = \frac{2}{l_t} E_\sigma \sup \{\sigma' W b : (a'\ b') P (a'\ b')' \le 1\}$ where $W = (C'\ B)U$ and

$$P = \begin{pmatrix} (\lambda_1 + \lambda)\Lambda & -\lambda\Lambda \\ -\lambda\Lambda & (\lambda_2 + \lambda)\Lambda \end{pmatrix}$$

Using Eq. 2.2, the supremum can be evaluated as

$$\hat{R}_n(\mathcal{J}^t_{EA}) = \frac{2}{l_t} E_\sigma ||(P^{-1/2})_{22} W'\sigma||.$$

We now make use of Kahane-Khintchine inequality [4] which is stated in the following lemma.

Lemma 2.1. *For any vectors* $a_1, a_2, \ldots, a_n$ *and independent Rademacher random variables* $\sigma_1, \sigma_2, \ldots, \sigma_n$*, we have*

$$\frac{1}{\sqrt{2}} E \left\|\sigma^n_{i=1}\sigma_i a_i\right\|^2 \leq \left(E \left\|\sigma^n_{i=1}\sigma_i a_i\right\|\right)^2 \leq E \left\|\sigma^n_{i=1}\sigma_i a_i\right\|^2$$

Using the above inequality we get a lower and upper bound on the complexity as

$$\frac{2C^t_{EA}}{2^{1/4} l_t} \leq \hat{R}_n(\mathcal{J}^t_{EA}) \leq \frac{2C^t_{EA}}{l_t}, \tag{2.3}$$

where

$$\begin{aligned} \left(C^t_{EA}\right)^2 &= E_\sigma ||(P^{-1})^{1/2}_{22} W'\sigma||^2 \\ &= E_\sigma \left(\sigma' W (P^{-1})_{22} W'\sigma\right) \\ &= E_\sigma tr\{\sigma\sigma' W (P^{-1})_{22} W'\} \\ &= tr\{W (P^{-1})_{22} W'\}. \end{aligned} \tag{2.4}$$

The above expression can be written in terms of original kernel sub-matrices by doing algebraic manipulations on the eigenbases using similar steps as in [3]. We finally get the result

$$\left(C^t_{EA}\right)^2 = \frac{1}{\lambda_2}\left(\frac{1}{1 + \frac{1}{\frac{\lambda_2}{\lambda_1} + \frac{\lambda_2}{\lambda}}}.\right) tr(B).$$

Plugging it into Eq. 2.3 gives the desired bounds on the Rademacher complexity of the EA target hypothesis class.

3 Proof of Theorem 4.5: Complexity for EA++

In this section, we bound the complexity of the target hypothesis class $\mathcal{J}^s_{++}$ for EA++. The base hypothesis class $\mathcal{H}_{++}$ in Eq. 4.3 (of the original paper) in source and target hypotheses. So the complexity of source class $\mathcal{J}^s_{++}$ can be obtained by replacing adequate terms. We are interested in the complexity of the hypothesis class $\mathcal{J}^t_{++}$ which is defined as $\mathcal{J}^t_{++} := \{h_2 : \mathcal{X} \mapsto \mathbb{R},\, (h_1, h_2) \in \mathcal{H}_{++}\}$, where h_1 is not fixed a priori.

The Rademacher complexity of $\mathcal{J}^t_{++}$ is defined as

$$\hat{R}_n(\mathcal{J}^t_{++}) = E_\sigma \left[\sup_{(h_1,h_2)\in\mathcal{H}_{++}} \left|\frac{2}{l_t}\sum_{i=1}^{l_t} \sigma_i h_2(x_i)\right|\right] \tag{3.1}$$

We proceed similar to the complexity proof of EA given in previous section. Note that $h_2 \in \mathcal{J}^t_{++} \implies -h_2 \in \mathcal{J}^t_{++}$ since $(h_1, h_2) \in \mathcal{H}_{++} \implies (-h_1, -h_2) \in \mathcal{H}_{++}$. This means that we can remove the absolute value sign from Eq. 3.1. Since, $\forall i, h_2(x_i) = \langle k(x_i, \cdot), h_2\rangle$, we can restrict the supremum to h_1 and h_2 that are in the span of all samples and also in $\mathcal{H}_{++}$. The restricted condition on (h_1, h_2) then becomes

$$\begin{aligned} &\{(h_\alpha, h_\beta) : \lambda_1\alpha' K\alpha + \lambda_2\beta' K\beta + \lambda(\alpha-\beta)' K(\alpha-\beta) + \lambda_u(\alpha-\beta)' M(\alpha-\beta) \leq 1\} \\ &= \{(h_\alpha, h_\beta) : (\alpha'\ \beta') N (\alpha'\ \beta')' \leq 1\} \end{aligned}$$

where

$$M = \begin{pmatrix} D \\ E \\ F \end{pmatrix} \begin{pmatrix} D' & E' & F' \end{pmatrix},$$

$$N = \begin{pmatrix} (\lambda_1+\lambda)K & -\lambda K \\ -\lambda K & (\lambda_2+\lambda)K \end{pmatrix} + \lambda_u \begin{pmatrix} M & -M \\ -M & M \end{pmatrix},$$

and K is the kernel matrix for source labeled, target labeled and target unlabeled samples. Using the reproducing kernel property, we get

$$\hat{R}_n(\mathcal{J}^t_{++}) = \frac{2}{l_t} E_\sigma \sup_{(\alpha,\beta)\in\mathbb{R}^{l_s+l_t+l_u}} \{\sigma'(C'\ B\ E)\beta : (\alpha'\ \beta')N(\alpha'\ \beta')' \le 1\}.$$

Using Eq. 2.2, the supremum in the above equation becomes $||(N^{-1})^{1/2}_{22}(C'\ B\ E)'\sigma||$.

If the matrix N is not full rank, we can project β and α onto the column space of K without changing the supremum (as it is done in the previous proof). So, the $\sup$ is restricted to the projected α_{pr} and β_{pr}, and the expression for Rademacher complexity can be rewritten as

$$\hat{R}_n(\mathcal{J}^t_{++}) = \frac{2}{l_t} E_\sigma \sup_{\alpha_{pr},\beta_{pr}\in ColSpace\{K\}} \Big\{\sigma'(C'\ B\ E)\beta_{pr} : (\alpha'_{pr}\beta'_{pr})N(\alpha'_{pr}\beta'_{pr})' \le 1\Big\}.$$

We proceed in a manner similar to the previous proof and diagonalize the kernel matrix K to get orthonormal bases U corresponding the nonzero eigenvalues ($K = U'\Lambda U$). Λ is a diagonal matrix of size $r \times r$, containing just the nonzero eigenvalues and r is the rank of matrix K. Since α_{pr} and β_{pr} are in the span of column space of K, there exist a_s and b such that $\alpha_{pr} = Ua,\ \beta_{pr} = Ub$.

The expression for complexity now becomes,

$$\hat{R}_n(\mathcal{J}^t_{++}) = \frac{2}{l_t} E_\sigma \sup\{\sigma' W b : (a'\ b')P(a'\ b')' \le 1\}$$

where $W = (C'\ B\ E)U$ and

$$P = \begin{pmatrix} (\lambda_1+\lambda)\Lambda & -\lambda\Lambda \\ -\lambda\Lambda & (\lambda_2+\lambda)\Lambda \end{pmatrix} + \lambda_u \begin{pmatrix} V' & 0 \\ 0 & V' \end{pmatrix} \begin{pmatrix} M & -M \\ -M & M \end{pmatrix} \begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix}$$

The solution to the above maximization problem is given by $||(P^{-1})^{1/2}_{22}W'\sigma||$. Using Kahane-Khintchine inequality and taking similar steps as in Eq. 2.4, we get the following result:

$$\frac{2C^t_{++}}{2^{1/4}l_t} \le \hat{R}_n(\mathcal{J}^t_{++}) \le \frac{2C^t_{++}}{l_t}, \tag{3.2}$$

where $\left(C^t_{++}\right)^2 = tr\{W(P^{-1})_{22}W'\}$.

Let T be the first term in the above expression for P. The second term can be written as RR' where

$$R = \begin{pmatrix} V' & 0 \\ 0 & V' \end{pmatrix} \begin{pmatrix} D \\ E \\ F \\ D \\ E \\ F \end{pmatrix}$$

Using the matrix inversion lemma, we have $(T + \lambda_u RR')^{-1} = T^{-1} - \lambda_u T^{-1}R(I + \lambda_u R'T^{-1}R)^{-1}R'T^{-1}$. The term $tr\{W(T^{-1})_{22}W'\}$ evaluates to the same expression as the complexity of EA in previous proof. The second term can also be reduced in terms of original kernel sub-matrices by performing algebraic manipulations on eigenbases using similar steps as used in [3]. We finally get the result

$$\left(C^t_{++}\right)^2 = \left(\frac{1}{\lambda_2 + \left(\frac{1}{\lambda_1} + \frac{1}{\lambda}\right)^{-1}}\right) tr(B) - \lambda_u \left(\frac{\lambda_1}{\lambda\lambda_1 + \lambda\lambda_2 + \lambda_1\lambda_2}\right)^2 tr\left(E(I + kF)^{-1}E'\right),$$

where $k = \frac{\lambda_u(\lambda_1+\lambda_2)}{\lambda\lambda_1+\lambda\lambda_2+\lambda_1\lambda_2}$. Plugging it into Eq. 3.2 gives the desired bounds on the Rademacher complexity of EA++ target hypothesis class.

References

[1] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *NIPS'07*, pages 129–136, Vancouver, B.C., December 2007.

[2] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT'09*, Montreal, Quebec, June 2009.

[3] D. S. Rosenberg and P. L. Bartlett. The Rademacher complexity of co-regularized kernel classes. In *AISTATS'07*, pages 396–403, San Juan, Puerto Rico, March 2007.

[4] Rafal Latala and Krzysztof Oleszkiewicz. On the best constant in the Khinchin-Kahane inequality. *Studia Mathematica*, 109:101–104, 1994.