Supplementary material: A Theory of Multiple-Source Adaptation with Limited Target Labeled Data

A Related on domain adaptation

As stated in the introduction, various scenarios of adaptation can be distinguished depending on parameters such as the number of source domains available, the presence or absence of target labeled data, and access to labeled source data or only to predictors trained on each source domain. Single source domain adaptation has been studied in several papers including (Kifer et al., 2004; Ben-David et al., 2010; Mansour et al., 2009c).

Several algorithms have been proposed for multiple-source adaptation. Khosla et al. (2012); Blanchard et al. (2011) proposed to combine all the source data and train a single model. Duan et al. (2009, 2012) used unlabeled target data to obtain a regularizer. Domain adaptation via adversarial learning was studied by Pei et al. (2018); Zhao et al. (2018). Crammer et al. (2008) considered learning models for each source domain, using close-by data of other domains. Gong et al. (2012) ranked multiple source domains by how well they can adapt to a target domain. Other solutions to multiple-source domain adaptation include, clustering (Liu et al., 2016), learning domain-invariant features (Gong et al., 2013a), learning intermediate representations (Jhuo et al., 2012), subspace alignment techniques (Fernando et al., 2013), attributes detection (Gan et al., 2016), using a linear combination of pre-trained classifiers (Yang et al., 2007), using multitask auto-encoders (Ghifary et al., 2015), causal approaches (Sun et al., 2011), two-state weighting approaches (Sun et al., 2011), moments alignment techniques (Peng et al., 2019) and domain-invariant component analysis (Muandet et al., 2013).

B Proof of equation (2)

By the definition of discrepancy,

$$\mathcal{L}_{\mathcal{D}_0}(h_{\overline{\mathcal{D}}_\lambda}) \leq \mathcal{L}_{\mathcal{D}_\lambda}(h_{\overline{\mathcal{D}}_\lambda}) + \mathrm{disc}_{\mathcal{H}}(\mathcal{D}_\lambda, \mathcal{D}_0).$$

Similarly,

$$\begin{aligned}\mathcal{L}_{\mathcal{D}_0}(h_{\mathcal{D}_0}) &= \mathcal{L}_{\mathcal{D}_0}(h_{\mathcal{D}_0}) - \mathcal{L}_{\mathcal{D}_\lambda}(h_{\mathcal{D}_0}) + \mathcal{L}_{\mathcal{D}_\lambda}(h_{\mathcal{D}_0}) \\ &\geq \mathcal{L}_{\mathcal{D}_\lambda}(h_{\mathcal{D}_0}) - \mathrm{disc}_{\mathcal{H}}(\mathcal{D}_\lambda, \mathcal{D}_0) \\ &\geq \mathcal{L}_{\mathcal{D}_\lambda}(h_{\mathcal{D}_\lambda}) - \mathrm{disc}_{\mathcal{H}}(\mathcal{D}_\lambda, \mathcal{D}_0)\end{aligned}$$

Combining the above two equations yields

$$\mathcal{L}_{\mathcal{D}_0}(h_{\overline{\mathcal{D}}_\lambda}) - \mathcal{L}_{\mathcal{D}_0}(h_{\mathcal{D}_0}) \leq \mathcal{L}_{\mathcal{D}_\lambda}(h_{\overline{\mathcal{D}}_\lambda}) - \mathcal{L}_{\mathcal{D}_\lambda}(h_{\mathcal{D}_\lambda}) + 2\mathrm{disc}_{\mathcal{H}}(\mathcal{D}_\lambda, \mathcal{D}_0).$$

Next observe that, by rearranging terms,

$$\begin{aligned}\mathcal{L}_{\mathcal{D}_\lambda}(h_{\overline{\mathcal{D}}_\lambda}) - \mathcal{L}_{\mathcal{D}_\lambda}(h_{\mathcal{D}_\lambda}) &= \mathcal{L}_{\mathcal{D}_\lambda}(h_{\overline{\mathcal{D}}_\lambda}) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\overline{\mathcal{D}}_\lambda}) + \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\mathcal{D}_\lambda}) - \mathcal{L}_{\mathcal{D}_\lambda}(h_{\mathcal{D}_\lambda}) \\ &\quad + \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\overline{\mathcal{D}}_\lambda}) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\mathcal{D}_\lambda})\end{aligned}$$

However, by the definition of $h_{\overline{\mathcal{D}}_\lambda}$,

$$\mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\overline{\mathcal{D}}_\lambda}) \leq \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\mathcal{D}_\lambda}).$$

Hence,

$$\begin{aligned}\mathcal{L}_{\mathcal{D}_\lambda}(h_{\overline{\mathcal{D}}_\lambda}) - \mathcal{L}_{\mathcal{D}_\lambda}(h_{\mathcal{D}_\lambda}) &\leq \mathcal{L}_{\mathcal{D}_\lambda}(h_{\overline{\mathcal{D}}_\lambda}) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\overline{\mathcal{D}}_\lambda}) + \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\mathcal{D}_\lambda}) - \mathcal{L}_{\mathcal{D}_\lambda}(h_{\mathcal{D}_\lambda}) \\ &\leq 2\sup_h |\mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h) - \mathcal{L}_{\mathcal{D}_\lambda}(h)|,\end{aligned}$$

where the last inequality follows by taking the supremum. Combining the above equations yields the proof.

C Previous work

C.1 Bregman divergence based non-negative matrix factorization

A natural algorithm is a two step process, where we first identify the optimal λ by minimizing

$$\min_{\lambda \in \Delta_p} \mathsf{B}(\widehat{\mathcal{D}}_0 \| \overline{\mathcal{D}}_\lambda),$$

where B is a suitable Bregman divergence. We can then use λ to minimize the weighted loss. However, this approach has both practical and theoretical issues. On the practical side, if $\mathcal{X}$ is a continuous space, then the empirical distribution $\overline{\mathcal{D}}_\lambda$ would be a point mass distribution over observed points and would never converge to the true distribution $\mathcal{D}_\lambda$. To overcome this, we need to first use $\overline{\mathcal{D}}_\lambda$ to estimate the distribution $\mathcal{D}_\lambda$ via kernel density estimation or other methods and then use the estimate instead of $\overline{\mathcal{D}}_\lambda$. Even if we use these methods and find λ, it is likely that we would overfit as the generalization of the algorithm depends on the covering number of $\{\mathcal{D}_\lambda : \lambda \in \Delta_p\}$, which in general can be much larger than that of the class of hypotheses $\mathcal{H}$. Hence such an algorithm would not incur generalization loss of $\mathcal{E}(\lambda^*)$. One can try to reduce the generalization error by using a discrepancy based approach, which we discuss in the next section.

C.2 A convex combination discrepancy-based algorithm

Since pairwise discrepancies would result in identifying a sub-optimal λ, instead of just considering the pairwise discrepancies, one can consider the discrepancy between $\mathcal{D}_0$ and any $\mathcal{D}_\lambda$. Since

$$\mathcal{L}_{\mathcal{D}_0}(h) \le \min_{\lambda \in \Delta_p} \mathcal{L}_{\mathcal{D}_\lambda}(h) + \mathrm{disc}_{\mathcal{H}}(\mathcal{D}_0, \mathcal{D}_\lambda),$$

and the learner has more data from $\mathcal{D}_\lambda$ than from $\mathcal{D}_0$, a natural algorithm is to minimize $\mathcal{L}_{\mathcal{D}_\lambda}(h) + \mathrm{disc}_{\mathcal{H}}(\mathcal{D}_0, \mathcal{D}_\lambda)$. However, note that this requires estimating both the discrepancy $\mathrm{disc}_{\mathcal{H}}(\mathcal{D}_0, \mathcal{D}_\lambda)$ and the expected loss over $\mathcal{L}_{\mathcal{D}_\lambda}(h)$. In order to account for both terms, we propose to minimize the upper bound on $\min_{\lambda \in \Delta_p} \mathcal{L}_{\mathcal{D}_\lambda}(h) + \mathrm{disc}_{\mathcal{H}}(\mathcal{D}_0, \mathcal{D}_\lambda)$,

$$\min_{\lambda} \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h) + C_\epsilon(\lambda), \tag{12}$$

where $C_\epsilon(\lambda)$ is given by,

$$\mathrm{disc}_{\mathcal{H}}(\widehat{\mathcal{D}}_0, \overline{\mathcal{D}}_\lambda) + \frac{c\sqrt{d + \log \frac{1}{\delta}}}{\sqrt{m_0}} + \epsilon M + \frac{cM\sqrt{\mathfrak{s}(\lambda \| \mathbf{m})}}{\sqrt{m}} \cdot \left(\sqrt{d \log \frac{em}{d} + p \log \frac{p}{\epsilon\delta}} \right),$$

for some constant c. We first show that $\mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h) + C_\epsilon(\lambda)$ is an upper bound on $\mathcal{L}_{\mathcal{D}_0}(h)$.

Lemma 3. *With probability at least $1 - 2\delta$, for all $h \in \mathcal{H}$ and $\lambda \in \Delta_p$,*

$$|\mathcal{L}_{\mathcal{D}_0}(h) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)| \le C_\epsilon(\lambda).$$

Proof. By (2) and Proposition 1, with probability at least $1 - \delta$,

$$|\mathcal{L}_{\mathcal{D}_0}(h) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)| \le 2\mathrm{disc}_{\mathcal{H}}(\mathcal{D}_0, \mathcal{D}_\lambda) + \frac{4M\sqrt{s(\lambda \| \mathbf{m})}}{\sqrt{m}} \cdot \left(\sqrt{d \log \frac{em}{d} + \log \frac{1}{\delta}} \right).$$

Hence, by the union bound over an ϵ-ℓ_1 cover of Δ_p yields, with probability $\ge 1 - \delta$, for all $\lambda \in \Delta_p$,

$$|\mathcal{L}_{\mathcal{D}_0}(h) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)| \le 2\mathrm{disc}_{\mathcal{H}}(\mathcal{D}_0, \mathcal{D}_\lambda) + \epsilon M + \frac{4M\sqrt{\mathfrak{s}(\lambda \| \mathbf{m})}}{\sqrt{m}} \cdot \left(\sqrt{d \log \frac{em}{d} + p \log \frac{p}{\epsilon\delta}} \right).$$

With probability at least $1 - \delta$, discrepancy can be estimated as

$$|\mathrm{disc}_{\mathcal{H}}(\mathcal{D}_0, \mathcal{D}_\lambda) - \mathrm{disc}_{\mathcal{H}}(\widehat{\mathcal{D}}_0, \overline{\mathcal{D}}_\lambda)| \le \epsilon M + \frac{c\sqrt{d + \log \frac{1}{\delta}}}{\sqrt{m_0}} + \frac{cM\sqrt{\mathfrak{s}(\lambda \| \mathbf{m})}}{\sqrt{m}} \cdot \left(\sqrt{d \log \frac{em}{d} + p \log \frac{p}{\epsilon\delta}} \right),$$

for some constant $c > 0$. Combining the above equations yields, with probability at least $1 - 2\delta$,

$$\max_{\lambda} |\mathcal{L}_{\mathcal{D}_0}(h) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)| \le C_\epsilon(\lambda).$$

□

Let h_R be the solution to (12), we now give a generalization bound for the above algorithm.

Lemma 4. *With probability at least* $1-2\delta$*, the solution* h_R *for* (12) *satisfies*

$$\mathcal{L}_{\mathcal{D}_0}(h_R) \le \min_{h\in\mathcal{H}} \mathcal{L}_{\mathcal{D}_0}(h) + 2\min_{\lambda} C_\epsilon(\lambda).$$

Proof. By Lemma 3, with probability at least $1-2\delta$,

$$\min_{\lambda\in\Delta_p} |\mathcal{L}_{\mathcal{D}_0}(h) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)| \le \min_{\lambda\in\Delta_p} C_\epsilon(\lambda).$$

Let h_R be the output of the algorithm and $h_{\mathcal{D}_0}$ be the minimizer of $\mathcal{L}_{\mathcal{D}_0}(h)$.

$$\begin{aligned}
L_{\mathcal{D}_0}(h_R) - \mathcal{L}_{\mathcal{D}_0}(h_{\mathcal{D}_0}) &\le \min_\lambda\big(\mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_R) + C_\epsilon(\lambda)\big) - \max_\lambda\big(\mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\mathcal{D}_0}) - C_\epsilon(\lambda)\big)\\
&\le \min_\lambda\big(\mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_R) + C_\epsilon(\lambda)\big) + \min_\lambda\big(-\mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\mathcal{D}_0}) + C_\epsilon(\lambda)\big)\\
&\le \min_\lambda\big(\mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_R) + C_\epsilon(\lambda) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\mathcal{D}_0}) + C_\epsilon(\lambda)\big)\\
&\le 2\min_\lambda C_\epsilon(\lambda),
\end{aligned}$$

where the last inequality follows from the fact that h_R is the minimizer of (12).

□

The above bound is comparable to the model trained on only target data as $C_\epsilon(\lambda)$ contains $\mathcal{O}\left(\sqrt{\frac{d}{m_0}}\right)$, which can be large for a small values of m_0. This bound can be improved on certain favorable cases when $\mathcal{D}_0 = \mathcal{D}_k$ for some known k. In this case if we use the same set of samples for $\widehat{\mathcal{D}}_0$ and $\widehat{\mathcal{D}}_k$, then the bound can be improved to $\mathcal{O}\left(\frac{\sqrt{d(1-\lambda_k)}}{\sqrt{m_0}}\right)$, which in favorable cases such that λ_k is large, yields a better bound than the target-only model.

D Proof of Lemma 1

By the strong convexity of ℓ,

$$\begin{aligned}
\mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\lambda'}) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_\lambda) &\ge \nabla\mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_\lambda)\cdot(h_{\lambda'} - h_\lambda) + \frac{\mu}{2}\|h_{\lambda'} - h_\lambda\|^2\\
&= \frac{\mu}{2}\|h_{\lambda'} - h_\lambda\|^2,
\end{aligned}$$

where the equality follows from the definition of h_λ. Similarly, since the function ℓ is bounded by M

$$\begin{aligned}
\mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_{\lambda'}) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h_\lambda) &\le \mathcal{L}_{\overline{\mathcal{D}}_{\lambda'}}(h_{\lambda'}) - \mathcal{L}_{\overline{\mathcal{D}}_{\lambda'}}(h_\lambda) + \|\lambda-\lambda'\|_1 M\\
&\le -\nabla\mathcal{L}_{\overline{\mathcal{D}}_{\lambda'}}(h_{\lambda'})\cdot(h_{\lambda'} - h_\lambda) - \frac{\mu}{2}\|h_{\lambda'} - h_\lambda\|^2 + \|\lambda-\lambda'\|_1 M\\
&= -\frac{\mu}{2}\|h_{\lambda'} - h_\lambda\|^2 + \|\lambda-\lambda'\|_1 M.
\end{aligned}$$

Combining the above equations,

$$\mu\|h_{\lambda'} - h_\lambda\|^2 \le M\|\lambda-\lambda'\|_1.$$

Hence for any distribution $\mathcal{D}_0$,

$$\begin{aligned}
|\mathcal{L}_{\mathcal{D}_0}(h_{\lambda'}) - \mathcal{L}_{\mathcal{D}_0}(h_\lambda)| &\le |\nabla\mathcal{L}_{\mathcal{D}_0}(h_\lambda)\cdot(h_{\lambda'} - h_\lambda)|\\
&\le |\nabla\mathcal{L}_{\mathcal{D}_0}(h_\mathcal{D})\cdot(h_{\lambda'} - h_\lambda)|\\
&= G\|h_{\lambda'} - h_\lambda\|\\
&= \frac{G\sqrt{M}}{\sqrt{\mu}}\cdot\|\lambda-\lambda'\|_1^{1/2}.
\end{aligned}$$

E Proof of Theorem 3

Let p be a multiple of four. Let $\mathcal{X} = \{1, 2, \ldots, p/2\}$ and $\mathcal{Y} = \{0, 1\}$. For all $k \le p/2$, and $x \in \mathcal{X}$, let $\mathcal{D}_k(x) = \frac{2}{p}$. For every even k, let $\mathcal{D}_k(1|\lceil k/2 \rceil) = 1$ and for every odd k, $\mathcal{D}_k(1|\lceil k/2 \rceil) = 0$. For remaining x and k, let $\mathcal{D}_k(1|x) = \frac{1}{2}$.

Let $\mathcal{H}$ be the set of all mappings from $\mathcal{X} \to \mathcal{Y}$ and the loss function be zero-one loss. Let $\mathcal{D}_0 = \mathcal{D}_\lambda$ for some λ. Hence, the optimal estimator h^* is

$$h^*_\lambda(1|x) = 1_{\lambda_{2x} > \lambda_{2x-1}}.$$

Given infinitely number of samples from each $\mathcal{D}_k$, the learner knows the distributions $\mathcal{D}_k$. Hence, roughly speaking the algorithm has to find if $\lambda_{2x} > \lambda_{2x-1}$ for each x.

Let $\epsilon = \frac{1}{100} \cdot \sqrt{\frac{p}{m_0}}$. We restrict $\lambda \in \Lambda$, where Λ is defined as follows. Let Λ be the set of all distributions such that for each $\lambda \in \Lambda$ and x,

$$\lambda_{2x} + \lambda_{2x-1} = \frac{2}{p},$$

and $\lambda_{2x} \in \{\frac{1+\epsilon}{p}, \frac{1-\epsilon}{p}\}$. Note that $|\Lambda| = 2^{p/4}$. For $x \le p/4$, let $s_x = \{2x, 2x-1\}$. Let m_{s_x} be the number of occurrences of elements from s_x. Given m_{s_x}, m_{2x} and m_{2x-1} are random variables from Binomial distribution with parameters m_{s_x} and $\frac{\lambda_{2x}}{\lambda_{2x}+\lambda_{2x-1}}$. This reduces the problem of learning the best classifier into testing $p/2$ Bernoulli distributions and we can use standard tools from information theory such as Fano's inequality (Cover and Thomas, 2012) to provide a lower bound. We provide a proof sketch.

Since $\sum_{x=1}^{p/4} m_{s_x} = m_0$, there are at least $p/8$ values of s_x for which $m_{s_x} \le 8m_0/p$. Consider one such s_x, where $m_{s_x} \le 8m_0/p$. For that x, given m_{s_x} samples from s_x, by Fano's inequality, with probability at least $1/4$, any algorithm cannot differentiate between $\lambda_{2x} > \lambda_{2x-1}$ and $\lambda_{2x} < \lambda_{2x-1}$. Thus, with probability at least $1/4$, any algorithm incorrectly finds the wrong hypothesis for h, and hence,

$$\mathbb{E}[\mathcal{L}(h)|x \in s_x] \ge \mathbb{E}[\mathcal{L}(h^*_\lambda)|x \in s_x] + c\frac{|\lambda_{2x} - \lambda_{2x-1}|}{\lambda_{2x} + \lambda_{2x-1}} \ge \mathbb{E}[\mathcal{L}(h^*_\lambda)|x \in s_x] + c\epsilon,$$

for some constant c. Averaging over all symbols x, yields

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(h)] &= \sum_{s_x} \mathcal{D}_\lambda(s_x)\, \mathbb{E}[\mathcal{L}(h)|x \in s_x] \\
&= \sum_{s_x : m_{s_x} \le 8m_0/p} \frac{2}{p}\, \mathbb{E}[\mathcal{L}(h)|x \in s_x] + \sum_{s_x : m_{s_x} > 8m_0/p} \frac{2}{p}\, \mathbb{E}[\mathcal{L}(h)|x \in s_x] \\
&\ge \sum_{s_x : m_{s_x} \le 8m_0/p} \frac{2}{p}\, (\mathbb{E}[\mathcal{L}(h^*_\lambda)|x \in s_x] + c\epsilon) + \sum_{s_x : m_{s_x} > 8m_0/p} \frac{2}{p}\, \mathbb{E}[\mathcal{L}(h^*_\lambda)|x \in s_x] \\
&= \mathbb{E}[\mathcal{L}(h^*_\lambda)] + \sum_{s_x : m_{s_x} \le 8m_0/p} \frac{2}{p} c\epsilon \\
&\ge \mathbb{E}[\mathcal{L}(h^*_\lambda)] + \frac{p}{8} \cdot \frac{2}{p} c\epsilon \\
&= \mathbb{E}[\mathcal{L}(h^*_\lambda)] + \frac{c\epsilon}{4} = \mathbb{E}[\mathcal{L}(h^*_\lambda)] + \frac{c\epsilon}{400}\sqrt{\frac{p}{m_0}}.
\end{aligned}$$