# Multi-source domain adaptation with guarantees

**Anastasia Pentina**
SDSC, ETH Zurich
anastasia.pentina@sdsc.ethz.ch

**Christoph H. Lampert**
IST Austria
chl@ist.ac.at

## Abstract

This paper addresses the problem of unsupervised multi-source domain adaptation and in particular the question of how to combine the source data in an optimal way. We prove a generalisation bound that characterises how the amount of transfer from different source tasks influences the final performance of a trained predictor. In contrast to most previous works, which provide guarantees for fixed transfer weights but do not allow choosing them in a data-dependent way, our bound yields a principled algorithm for weighting the source tasks effectively.

## 1 Introduction

Machine learning algorithms are most successful when given sufficient amount of training data coming from the same problem, as the one they will be tested on. However, there are scenarios, like adaptation of a personalised speech recognition system or recommender system to a new user, in which one wishes to make accurate predictions on a task for which there is only limited or even no annotated data available. The question of how to successfully utilise data available from one or more *source* tasks for solving a related, but different *target* problem is studied in the filed of domain adaptation. Different approached to domain adaptation have been studied in the literature, ranging from finding a weighted combination of source predictors [5] to optimal transport-based methods [2] to using variational autoencoders [4].

In this work we study unsupervised multi-source domain adaptation and focus on the question of how to weight the data from source tasks such that a hypothesis trained on this combination would perform well on the target domain. This question was first studied in [1] and recently revisited in [9]. Intuitively, one would want to use the most related source task for training a target hypothesis. However, including data from less related tasks in the learning procedure could be beneficial from the perspective of increasing the total amount of data used for training, and thereby reducing the danger of overfitting. The interplay between these two aspects - bias and variance - of the trained hypothesis as the result of using various weighted combinations of the source tasks has been theoretically characterised both in [1] and in [9]. However, in both these works the parameters of the weighted combination are assumed to be fixed in advance. The main result of this work is a high probability generalisation bound on the target error as a function of the weights that holds for all possible convex combinations. Thus, it extends the previous results in that it allows to derive an algorithm for selecting a good weighted combination in a principled data-dependent way.

## 2 Main result

We assume that there are $k$ source tasks $\langle D_1, f_1\rangle, \dots, \langle D_k, f_k\rangle$ and a target task $\langle D_T, f_T\rangle$, where for every $i = 1, \dots, k, T$ $D_i$ is a marginal distribution over a common input space $\mathcal{X}$ and $f_i : \mathcal{X} \to \{-1, 1\}$ is a deterministic labeling function. For every source task $i$ the learner is given a set $S_i^u$ of $n$ training examples, sampled i.i.d. from the corresponding data distribution $D_i$, and for a subset $S_i \subset S_i^u$ of size $m_i$ the labels according to $f_i$ are provided. For the target task the learner is given only a set $S_T^u$ of $n$ unlabeled examples sampled i.i.d. from $D_T$.

NeurIPS 2019 Workshop on Machine Learning with Guarantees, Vancouver, Canada.

The goal of the learner is to find a hypothesis $h$ in a hypotheses class $\mathcal{H}$ that would lead to low *expected error* on the target task:

$$\mathrm{er}_T(h) = \mathop{\mathbf{E}}_{x\sim D_T} [\![h(x) \neq f_T(x)]\!]. \tag{1}$$

We assume that to find such hypothesis the learner minimizes a convex combination of *training errors* on the source tasks:

$$\widehat{\mathrm{er}}_\alpha(h) = \sum_{i=1}^{k} \alpha_i \widehat{\mathrm{er}}_i(h) = \sum_{i=1}^{k} \frac{\alpha_i}{m_i} \sum_{(x_j^i, y_j^i)\in S_i} [\![h(x_j^i) \neq y_j^i]\!], \tag{2}$$

where $\alpha \in \Lambda = \{\alpha_i \geq 0,\ \sum_{i=1}^k \alpha_i = 1\}$. The success of any domain adaptation algorithm depends on how similar the target task is to the source ones. In the seminal work [1] the discrepancy measure was used for quantifying the dissimilarity between the marginal distributions of the tasks. In this work we use a very similar notion of *disparity discrepancy*:

$$d_{h,\mathcal{H}}(P, Q) = \sup_{h'\in\mathcal{H}} |\mathop{\mathbf{E}}_{x\sim P} [\![h(x) \neq h'(x)]\!] - \mathop{\mathbf{E}}_{x\sim Q} [\![h(x) \neq h'(x)]\!]|, \tag{3}$$

which was introduced in [8] and allows obtaining tighter generalization bounds.

The success of this approach mainly depends on the choice of the weights $\alpha$ and the following theorem captures this dependence.

**Theorem 1.** *Let $d$ be the VC-dimension of the hypothesis set $\mathcal{H}$ and $M = \sum_{i\in I} m_i$ and assume that $M > d$. Then for any $\delta > 0$ with probability at least $1 - \delta$ over $S_1^u, \dots, S_k^u, S_T^u$ and $S_1, \dots, S_k$, the following inequality holds uniformly for all choices of weights $\alpha \in \Lambda$ and all possible choices of the predictor $h \in \mathcal{H}$:*

$$\mathrm{er}_T(h) \leq \widehat{\mathrm{er}}_\alpha(h) + \sum_{i=1}^{k} \alpha_i d_{h,\mathcal{H}}(S_T^u, S_i^u) + \sum_{i=1}^{k} \alpha_i \lambda_{Ti} + A + B, \tag{4}$$

*where*

$$A = \sqrt{\frac{\log(k) + 2\log(8/\delta)}{n}} + \sqrt{\frac{8d\log(en/4d)}{n}} \tag{5}$$

$$B = C\sqrt{\frac{\ln(32km_0\|\alpha\|_*^2)}{m_0}} + 4\|\alpha\|_*\sqrt{\pi d \ln\left(\frac{eM}{d}\right)} + \|\alpha\|_*\sqrt{8\ln\left(\frac{\log M}{\delta}\right)} \tag{6}$$

$$d_{h,\mathcal{H}}(S_T^u, S_i^u) = \max_{h'\in\mathcal{H}} |\widehat{\mathrm{er}}_T(h,h') - \widehat{\mathrm{er}}_i(h,h')| \qquad \widehat{\mathrm{er}}_t(h,h') = \frac{1}{n}\sum_{j=1}^{n} [\![h(x_j^t) \neq h'(x_j^t)]\!] \tag{7}$$

$$\lambda_{Ti} = \inf_h(\mathrm{er}_T(h) + \mathrm{er}_i(h)) \qquad \|\alpha\|_* = \sqrt{\sum_{i=1}^{k} \frac{(\alpha_i)^2}{m_i}} \qquad m_0 = \min_{i=1,\dots,k} m_i. \tag{8}$$

*and $C$ is a universal constant.*

**Discussion** Theorem 1 provides an upper bound on the expected error on the target task – the quantity of interest – in terms of averaged $\lambda$-s, a data-independent complexity term $A$ and three data-dependent components – the weighted empirical errors on the source tasks, weighted dissimilarities between unlabeled sample sets (measured by $d_{h,\mathcal{H}}$) and $\alpha$-dependent complexity term $B$.

The complexity term $A$ comes from the estimation of the disparity discrepancies from the finite sets of unlabeled samples and behaves as $\tilde{O}(\sqrt{(\log k + d)/n})$, where $n$ is the number of unlabeled examples per task. Since collecting unlabeled examples is typically much cheaper than annotated one, we can assume that $n \gg m_i$ for $i = 1, \dots, k$ and thus this term has a negligible effect on the whole bound.

Results of [1] and [9] have shown that $\|\alpha\|_*$ affects the convergence rates, however, only when $\alpha$ is selected before observing the data. This left an undesirable gap in the theoretical understanding of multi-source domain adaptation, since for any practical algorithm, the mixture coefficients need to be chosen after the data had been observed, e.g. based on an estimate of the task relatedness. The complexity term $B$ in our results shows that the same behavior holds even in the uniform-in-$\alpha$ case, where $\alpha$ can be selected based on the observed data. Its last two terms are proportional to $\|\alpha\|_*$ and thus range from $\tilde{O}(\sqrt{d/m_0})$, when all the weight is put on the source task with the smallest training set, to the best possible rate of $\tilde{O}(\sqrt{d/M})$ when the weights are distributed proportionally to the sizes of the training sets. In contrast, the first term in $B$ does not exhibit this behaviour and converges only as $\tilde{O}(\sqrt{1/m_0})$. It is the consequence of the bound holding uniformly in $\alpha$. However, this component also does not depend on the VC-dimension of $\mathcal{H}$ and thus, if $d$ is sufficiently large, the last two terms in $B$ will be dominating.

The data-dependent part of the right-hand-side of (4) can be seen as a quality measure for the hypothesis $h$ and weights $\alpha$ and by minimizing it one could expect to find a beneficial combination of these parameters. This results in the following optimisation problem:

$$\min_{h\in\mathcal{H},\alpha\in\Lambda} \widehat{\mathrm{er}}_\alpha(h) + \sum_{i=1}^{k} \alpha_i d_{h,\mathcal{H}}(S_T^u, S_i^u) + \gamma\|\alpha\|_* \tag{9}$$

for some regularisation constant $\gamma \geq 0$. Optimisation (9) has an intuitive interpretation: we would like to give more weight to the source tasks that are easy (their empirical error is low) and similar to the target (small $d_{h,\mathcal{H}}$), while also sharing the weights among multiple source tasks to decrease the variance of the estimate (low $\|\alpha\|_*$), and select a hypothesis that works well on the source data and performs similarly on the source and on the target (low $d_{h,\mathcal{H}}$).

## 3 Experiments

**Data.** We use Amazon reviews dataset, containing reviews for four product categories - Books, DVDs, Electronics, and Kitchen appliances. Each review is encoded by a 5000-dimensional vector of unigrams and bigrams and has a binary label indicating the sentiment. We define four domain adaptation tasks, one for each target category with the remaining three categories serving as source tasks. For each domain we use 500 samples for training and the remaining samples from the target domain for testing.

**Methods.** To highlight the effects that the weighting of the source data has on the final performance, we omit the component of learning the representation that is often contained in domain adaptation methods (with MDAN [9] being no exception) and focus on the case of linear classifiers. As a result, MDAN is re-formulated as the following optimisation problem:

$$\min_h \frac{1}{\gamma} \log \sum_{i=1}^{k} \exp\left(\gamma\left(\widehat{\mathrm{er}}_i(h) + \mathrm{disc}(S_i^u, S_T^u)\right)\right), \tag{10}$$

where

$$\mathrm{disc}(S_i^u, S_T^u) = \sup_{h,h'\in\mathcal{H}} \left| \frac{1}{n}\sum_{j=1}^{n} [\![ h(x_j^i) \neq h'(x_j^i) ]\!] - \frac{1}{n}\sum_{j=1}^{n} [\![ h(x_j^T) \neq h'(x_j^T) ]\!] \right| \tag{11}$$

is the discrepancy measure [1]. To make the comparison to [9] more direct we also use it in (9) instead of the disparity discrepancy. We estimate it from the unlabeled data by training a linear classifier to separate $S_T^u$ from $S_i^u$ using logistic loss for every source task $i$. In both (9) and (10) we use logistic loss. Following [9] we set $\gamma$ in (10) to 10. In (9) we select $\gamma$ from the set $\{0., 0.0001, 0.001, 0.01, 0.1, 1., 2., 5., 10., 15., 20\}$ using leave-one-task-out cross-validation: we iterate over the source tasks by setting one of them aside, as if it was the target, and using the remaining source tasks for training.

**Results.** We report test accuracy averaged over 10 random data splits in Table 1. These results show that the algorithm derived from Theorem 1 is competitive with MDAN, and at the same time it enjoys stronger theoretical guarantees.

|  | D+E+K→ B | B+E+K→ D | B+D+K→ E | B+D+E→ K |
|---|---|---|---|---|
| MDAN | 74.26 | 76.44 | 80.56 | 82.38 |
| Ours | 74.29 | 76.52 | 81.07 | 82.84 |

Table 1: Test accuracy on the target task on Amazon reviews dataset.

## References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 2010.

[2] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Conference on Neural Information Processing Systems (NIPS)*, 2017.

[3] Y. Gordon, A. E. Litvak, S. Mendelson, and A. Pajor. Gaussian averages of interpolated bodies and applications to approximate reconstruction. *Journal of Approximation Theory*, 149(1):59 – 73, 2007.

[4] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair auto encoder. In *International Conference on Learning Representations (ICLR)*, 2015.

[5] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Conference on Neural Information Processing Systems (NIPS)*, 2009.

[6] A. Maurer, M. Pontil, and B. Romera-Paredes. An inequality with applications to structured sparsity and multitask dictionary learning. In *Conference on Learning Theory (COLT)*, 2014.

[7] Andreas Maurer. Concentration inequalities for functions of independent variables. *Random Structures and Algorithms*, 2006.

[8] Mingsheng Long Michael I. Jordan Yuchen Zhang, Tianle Liu. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learing (ICML)*, 2019.

[9] Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In *Conference on Neural Information Processing Systems (NIPS)*, 2018.

## A Proof

Our analysis heavily relies on the following two results from the literature:

**Lemma 1** (Theorem 1 in [7]). *Let $X_1, \ldots, X_n$ be independent random variables taking values in the set $\mathcal{X}$ and $f$ be a function $f : \mathcal{X}^n \to \mathbb{R}$. For any $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $y \in \mathcal{X}$ define:*

$$\begin{aligned} x_{y,k} &= (x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_n) \\ (\inf_k f)(x) &= \inf_{y \in \mathcal{X}} f(x_{y,k}) \\ \Delta_{+,f} &= \sum_{i=1}^{n} (f - \inf_k f)^2. \end{aligned}$$

*Then for $t > 0$:*

$$\Pr\{f - \mathbf{E} f \geq t\} \leq \exp\left( \frac{-t^2}{2\|\Delta_+\|_\infty} \right). \tag{12}$$

**Lemma 2** (Slightly modified Lemma 2 in [6]). *Let $M \geq 2$, $A_1, \ldots, A_M \subset \mathbb{R}^n$ and $A = \cup_m A_m$. Then:*

$$G(A) \leq \max_m G(A_m) + \sup_{z \in A} \|z\| \sqrt{2 \ln(M)}. \tag{13}$$

*Proof of Theorem 1.* We start with applying Theorem B.2 from [8] to every pair of tasks which results in the following bound on the average expected error over all tasks in terms of the error on the labeled tasks:

$$\mathrm{er}_T(h) \leq \mathrm{er}_\alpha(h) + \sum_{i=1}^{k} \alpha_i^t d_{h,\mathcal{H}}(D_T, D_i) + \sum_{i=1}^{k} \alpha_i \lambda_{Ti}. \tag{14}$$

According to Theorem B.4 from [8] for any pair of tasks $i, j$ and any $\delta > 0$ with probability at least $1-\delta$ the following holds for all $h$ in $\mathcal{H}$:

$$d_{h,\mathcal{H}}(D_i, D_T) \leq d_{h,\mathcal{H}}(S_i^u, S_T^u) + \sqrt{\frac{8d\log(en/4d)}{n}} + \sqrt{\frac{2\log(4/\delta)}{n}}. \tag{15}$$

Therefore, using the union bound argument, we obtain that for any $\delta > 0$ with probability at least $1-\delta/2$ for any choice of weights $\alpha$:

$$\sum_{i=1}^{k} \alpha_i d_{h,\mathcal{H}}(D_T, D_i) \leq \sum_{i=1}^{k} \alpha_i d_{h,\mathcal{H}}(S_T^u, S_i^u) + \sqrt{\frac{8d\log(en/4d)}{n}} + \sqrt{\frac{\log(k) + 2\log(8/\delta)}{n}}. \tag{16}$$

What remains is to upper-bound the difference between weighted expected errors and their empirical counter-parts uniformly in weights $\alpha \in \Lambda$ and predictor $h \in \mathcal{H}$.

We start with applying Lemma 1 to:

$$f(S_1, \dots, S_k) = \sup_{\alpha,h} \mathrm{er}_\alpha(h) - \widehat{\mathrm{er}}_\alpha(h) = \sup \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{\alpha_i}{m_i}(\mathrm{er}_i(h) - \ell(h(x_j^i), y_j^i)) \tag{17}$$

Note that:

$$\Delta_{+,f} \leq \sum_{i=1}^{k} \sum_{j=1}^{m_i} \left(\frac{\alpha_i}{m_i}\right)^2 = \sum_{i=1}^{k} \frac{(\alpha_i)^2}{m_i} \tag{18}$$

Therefore, by applying Lemma 1 to:

$$\Lambda_\rho = \left\{ \alpha \in \mathbb{R}^k : \ \alpha_i \geq 0, \ \sum_{i=1}^{k} \alpha_i = 1, \ \sum_{i=1}^{k} \frac{(\alpha_i)^2}{m_i} \leq \rho^2 \right\} \tag{19}$$

we obtain that with probability at least $1-\delta/2$ for all $\alpha \in \Lambda_\rho$ and $h \in \mathcal{H}$:

$$\mathrm{er}_\alpha(h) - \widehat{\mathrm{er}}_\alpha(h) \leq \mathbf{E} \sup_{\alpha,h} \Big( \mathrm{er}_\alpha(h) - \widehat{\mathrm{er}}_\alpha(h) \Big) + \rho\sqrt{2\log(2/\delta)}. \tag{20}$$

To bound the first term on the right-hand-side, we use the symmetrization trick and switch to Gaussian random variables:

$$\begin{aligned}
\mathbf{E} \sup_{\alpha,h} \Big( \mathrm{er}_\alpha(h) - \widehat{\mathrm{er}}_\alpha(h) \Big) &= \mathbf{E} \sup_{\alpha,h} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{\alpha_i}{m_i}(\mathrm{er}_i(h) - \ell(h(x_j^i), y_j^i)) && (21)\\
&= \mathbf{E} \sup_{\alpha,h} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{\alpha_i}{m_i}(\mathop{\mathbf{E}}_{\bar{S}} \ell(h(\bar{x}_j^i), \bar{y}_j^i) - \ell(h(x_j^i), y_j^i)) && (22)\\
&\leq \mathop{\mathbf{E}}_{S} \mathop{\mathbf{E}}_{\bar{S}} \sup_{\alpha,h} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{\alpha_i}{m_i}(\ell(h(\bar{x}_j^i), \bar{y}_j^i) - \ell(h(x_j^i), y_j^i)) && (23)\\
&\leq 2 \mathop{\mathbf{E}}_{S} \mathop{\mathbf{E}}_{\sigma} \sup_{\alpha,h} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{\sigma_j^i \alpha_i}{m_i} \ell(h(x_j^i), y_j^i) && (24)\\
&\leq \sqrt{2\pi} \mathop{\mathbf{E}}_{S} \mathop{\mathbf{E}}_{\xi} \sup_{\alpha,h} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{\xi_j^i \alpha_i}{m_i} \ell(h(x_j^i), y_j^i). && (25)
\end{aligned}$$

where $\sigma_j^i$ are independent Rademacher random variables that take values $+1$ and $-1$ with probability $0.5$ and $\xi_j^i$ are independent standard Gaussian random variables. Now, fix $h \in \mathcal{H}$. Then:

$$\mathop{\mathbf{E}}_{S}\mathop{\mathbf{E}}_{\xi}\sup_{\alpha}\sum_{i=1}^{k}\sum_{j=1}^{m_i}\frac{\xi_j^i\alpha_i}{m_i}\ell(h(x_j^i),y_j^i) = \mathop{\mathbf{E}}_{S}\mathop{\mathbf{E}}_{\xi}\sup\sum_{i=1}^{k}\sum_{j=1}^{m_i}\frac{\xi_j^i\alpha_i}{m_i}\frac{1-h(x_j^i)y_j^i}{2} \tag{26}$$

$$\leq \frac{1}{2}\left(\mathop{\mathbf{E}}_{\xi}\sup\sum_{i=1}^{k}\sum_{j=1}^{m_i}\xi_j^i\frac{\alpha_i}{m_i} + \mathop{\mathbf{E}}_{S,\xi}\sup\sum_{i=1}^{k}\sum_{j=1}^{m_i}\xi_j^i h(x_j^i)y_j^i\frac{\alpha_i}{m_i}\right) = \mathop{\mathbf{E}}_{\xi}\sup\sum_{i=1}^{k}\sum_{j=1}^{m_i}\xi_j^i\frac{\alpha_i}{m_i}. \tag{27}$$

Define:

$$B = \left\{\beta \in \mathbb{R}^k : \sum_{i=1}^{k}\beta_i = 1,\ \sum_{i=1}^{k}\frac{(\beta_i)^2}{m_i} \leq \rho^2\right\} \tag{28}$$

$$K = \left\{\gamma \in \mathbb{R}^k : \|\gamma\|_2 \leq 1,\ \|\gamma\|_{(1)} = \rho\sum_{i=1}^{k}\sqrt{m_i}\gamma_i \leq 1\right\} \tag{29}$$

Then $\Lambda_\rho \subset B$. Therefore:

$$\mathop{\mathbf{E}}_{\xi}\sup_{\alpha\in\Lambda_\rho}\sum_{i=1}^{k}\sum_{j=1}^{m_i}\xi_j^i\frac{\alpha_i}{m_i} \leq \mathop{\mathbf{E}}_{\xi}\sup_{\beta\in B}\sum_{i=1}^{k}\sum_{j=1}^{m_i}\xi_j^i\frac{\beta_i}{m_i} = \mathop{\mathbf{E}}_{\xi}\sup_{\beta\in B}\sum_{i=1}^{k}\frac{\beta_i}{\sqrt{m_i}}\cdot\frac{1}{\sqrt{m_i}}\sum_{j=1}^{m_i}\xi_j^i \tag{30}$$

$$= \mathop{\mathbf{E}}_{\xi}\sup_{\beta\in B}\sum_{i=1}^{k}\frac{\beta_i}{\sqrt{m_i}}\xi_i = \rho\mathop{\mathbf{E}}_{\xi}\sup_{\beta\in B}\sum_{i=1}^{k}\frac{\beta_i}{\rho\sqrt{m_i}}\xi_i \leq \rho\mathop{\mathbf{E}}_{\xi}\sup_{\gamma\in K}\sum_{i=1}^{k}\gamma_i\xi_i \tag{31}$$

$K$ is a unit ball with respect to the norm $\|\gamma\|_* = \max\{\|\gamma\|_2, \|\gamma\|_{(1)}\}$. Therefore, for a fixed $(\xi_1,\dots,\xi_k)$:

$$\sup\sum_{i=1}^{k}\xi_i\gamma_i = \|\xi\|_*^*,$$

where $\|\cdot\|_*^*$ is dual norm to $\|\cdot\|_*$. It can be expressed as:

$$\|\xi\|_*^* = \inf_{\xi_1+\xi_2=\xi}\|\xi_1\|_2^* + \|\xi_2\|_{(1)}^*, \tag{32}$$

where $\|\cdot\|_2^*$ is dual to $\|\cdot\|_2$, so just $\|\cdot\|_2$, and $\|\cdot\|_{(1)}^*$ is dual to $\|\cdot\|_{(1)}$:

$$\|\xi\|_{(1)}^* = \max_i\frac{|\xi_i|}{\rho\sqrt{m_i}}.$$

To upper-bound the infinum in (32), consider the following split: let $I_1$ be the set of indices of $\left\lceil\left(\frac{1}{\rho\sqrt{m_0}}\right)^2\right\rceil$ largest elements of $(|\xi_1|,\dots,|\xi_k|)$, where $m_0 = \min m_i$, and let $I_2$ be the set of remaining indices. Then:

$$\|\xi\|_*^* \leq \sqrt{\sum_{i\in I_1}\xi_i^2} + \max_{i\in I_2}\frac{|\xi_i|}{\rho\sqrt{m_i}}. \tag{33}$$

By the choice of $I_1$ and $I_2$, the largest element in $I_2$ is not bigger than any element in $I_1$, and thus not bigger than their average. Therefore:

$$\max_{i\in I_2}\frac{|\xi_i|}{\rho\sqrt{m_i}} \leq \max_{i\in I_2}\frac{|\xi_i|}{\rho\sqrt{m_0}} \leq \frac{\frac{1}{|I_1|}\sum_{i\in I_1}\xi_i^2}{\rho\sqrt{m_0}} \leq \frac{\sqrt{\frac{1}{|I_1|}\sum_{i\in I_1}\xi_i^2}}{\rho\sqrt{m_0}} = \sqrt{\sum_{i\in I_1}\xi_i^2}. \tag{34}$$

Thus:

$$\|\xi\|_*^* \leq 2\sqrt{\sum_{i\in I_1}\xi_i^2} \tag{35}$$

By Corollary 3.4 in [3]:

$$\mathbf{E}\sqrt{\sum_{i\in I_1}\xi_i^2} \sim \sqrt{|I_1|}\sqrt{2+\ln(2k/|I_1|)} \leq \frac{1}{\rho\sqrt{m_0}}\sqrt{\log(8km_0\rho^2)} \tag{36}$$

Thus, we obtain that:

$$\rho\,\mathbf{E}_{\xi}\sup_{\gamma\in K}\sum_{i=1}^{k}\gamma_i\xi_i = \rho\,\mathbf{E}\,\|\xi\|_*^* \leq \rho\frac{C}{\rho\sqrt{m_0}}\sqrt{\log(8km_0\rho)} = C\sqrt{\frac{\log(8km_0\rho^2)}{m_0}} \tag{37}$$

for some absolute constant $C$.

According to Lemma 2

$$\sqrt{2\pi}\,\mathbf{E}_{S}\,\mathbf{E}_{\xi}\sup_{\alpha,h}\sum_{i=1}^{k}\sum_{j=1}^{m_i}\frac{\xi_j^i\alpha_i}{m_i}\ell(h(x_j^i),y_j^i) \leq C\sqrt{\frac{\log(8km_0\rho^2)}{m_0}} + \rho\sqrt{2d\ln\left(\frac{eM}{d}\right)}. \tag{38}$$

Combining this with (20) and (27) we obtain that for any fixed $\rho > 0$ and $\delta > 0$ with probability at least $1-\delta/2$ the following inequality holds for all $\alpha \in \Lambda_\rho$ and $h \in \mathcal{H}$:

$$\mathrm{er}_\alpha(h) \leq \widehat{\mathrm{er}}_\alpha(h) + C\sqrt{\frac{\log(8km_0\rho^2)}{m_0}} + 2\rho\sqrt{\pi d\log\left(\frac{eM}{d}\right)} + \rho\sqrt{2\log\left(\frac{2}{\delta}\right)}. \tag{39}$$

By combining these bounds for $\rho = \frac{1}{2^s}$ for various $s$ using the union bound argument (and taking into account that $\frac{1}{\sqrt{M}} \leq \rho \leq \frac{1}{\sqrt{m_0}}$) we obtain that for any $\delta > 0$ with probability at least $1-\delta/2$ the following inequality holds for all $\alpha \in \Lambda$ and $h \in \mathcal{H}$:

$$\mathrm{er}_\alpha(h) \leq \widehat{\mathrm{er}}_\alpha(h) + C\sqrt{\frac{\log(32km_0\|\alpha\|_*^2)}{m_0}} + 4\|\alpha\|_*\sqrt{\pi d\log\left(\frac{eM}{d}\right)} + \|\alpha\|_*\sqrt{8\log\left(\frac{\log M}{\delta}\right)}. \tag{40}$$

Combination of (40) with (16) gives the statement of the theorem. □