Correction to "Stability and Hypothesis Transfer Learning"

Ilja Kuzborskij ILJA.KUZBORSKIJ@IDIAP.CH
Idiap Research Institute, Switzerland
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Francesco Orabona FRANCESCO@ORABONA.COM
Toyota Technological Institute at Chicago, USA

Abstract

There is an error in "Stability and Hypothesis Transfer Learning" (Kuzborskij & Orabona, 2013) which appeared in proceedings of ICML 2013. The Leave-One-Out generalization bound for Hypothesis Transfer Learning algorithm through Regularized Least Squares with biased regularization does not have the right convergence rate with respect to the regularization parameter λ and the source risk on the target domain, $R_\mu(f')$. This erratum describes the error and proves the correct generalization guarantees. The correct rate is in $\mathcal{O}\left(\frac{1}{m\lambda^{1.5}}\right)$, instead of incorrectly claimed $\mathcal{O}\left(\frac{1}{m\lambda}\right)$. However, the correct rate is still better than the usual one for Regularized Least Squares obtained via algorithmic stability analysis. Finally, corrected analysis still preserves the main contribution, that is, the relatedness of the source and target domains accelerates the convergence of the Leave-One-Out error to the generalization error.

1. Description of Error

The error was committed in the proof of Theorem 3 (Kuzborskij & Orabona, 2013), where Lemma 2 was applied incorrectly. This rendered Theorem 2, proving generalization guarantees for Hypothesis Transfer Learning (HTL) algorithm analyzed, invalid. Theorem 3 proves an upper-bound on hypothesis stability (Bousquet & Elisseeff, 2002) with respect to the square loss,

$$\forall i \in \{1,\dots,m\},$$
$$\mathbb{E}_{S,(\boldsymbol{x},y)}\left[\left|(f_S(\boldsymbol{x})-y)^2-(f_{S^{\setminus i}}(\boldsymbol{x})-y)^2\right|\right]\leq\gamma\ .$$

Here $(\boldsymbol{x},y)$ assumed to be any example drawn i.i.d. from p.d.f. μ. However, Lemma 2 proves an upper bound on the quantity $(f_{S^{\setminus i}}(\boldsymbol{x}_i)-y_i)^2$, where $(\boldsymbol{x}_i,y_i)\in S$, in other words, belongs to the training set. That is, in Theorem 3, $(f_{S^{\setminus i}}(\boldsymbol{x}_i)-y_i)^2$ appears instead of $(f_{S^{\setminus i}}(\boldsymbol{x})-y)^2$.

To use the correct quantity $(f_{S^{\setminus i}}(\boldsymbol{x})-y)^2$, we first obtain a closed form solution to $f_{S^{\setminus i}}(\boldsymbol{x})$, considering linear hypotheses $f(\boldsymbol{x}):=\boldsymbol{x}^\top\boldsymbol{w}$ and $f_{S^{\setminus i}}(\boldsymbol{x}):=\boldsymbol{x}^\top\boldsymbol{w}_{S^{\setminus i}}$. The derivation is given in Lemma 1.

A very similar error also appears in the proof of Lemma 4 (Kuzborskij & Orabona, 2013), first result. The nature of error is the same, and we fix it in the proof of the Theorem 1.

Lemma 1. *Let $\boldsymbol{w}_S$ be the hypothesis produced by the Regularized Least Squares (RLS) algorithm given training set S. For any sample $(\boldsymbol{x},y)\overset{i.i.d.}{\sim}\mu$ and $(\boldsymbol{x}_i,y_i)\in S$, such that $\|\boldsymbol{x}\|,\|\boldsymbol{x}_i\|\leq 1$, we have that the hypothesis $\boldsymbol{w}_{S^{\setminus i}}$ produced by the same RLS algorithm on a training set $S^{\setminus i}$ $\forall i\in\{1,\dots,m\}$, satisfies*

$$|\boldsymbol{x}^\top\boldsymbol{w}_S-\boldsymbol{x}^\top\boldsymbol{w}_{S^{\setminus i}}|\leq\frac{1}{m\lambda}|\boldsymbol{x}_i^\top\boldsymbol{w}_{S^{\setminus i}}-y_i|\ .$$

Proof. Define $\boldsymbol{X}=[\boldsymbol{x}_1,\dots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\dots,\boldsymbol{x}_m]$, $\boldsymbol{M}=\boldsymbol{X}^\top\boldsymbol{X}+m\lambda\boldsymbol{I}$. It is straightforward to see that $\boldsymbol{x}^\top\boldsymbol{w}_S$ is equal to

$$\begin{bmatrix}\boldsymbol{x}^\top\boldsymbol{X} & \boldsymbol{x}^\top\boldsymbol{x}_i\end{bmatrix}\begin{bmatrix}\boldsymbol{M} & \boldsymbol{X}^\top\boldsymbol{x}_i\\ \boldsymbol{x}_i^\top\boldsymbol{X} & \|\boldsymbol{x}_i\|^2+m\lambda\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{y}\\ y_i\end{bmatrix}. \tag{1}$$

Expanding the middle term and using the block-wise matrix inversion property (Petersen & Pedersen, 2008) we get

$$\begin{bmatrix}\boldsymbol{M} & \boldsymbol{X}^\top\boldsymbol{x}_i\\ \boldsymbol{x}_i^\top\boldsymbol{X} & \|\boldsymbol{x}_i\|^2+m\lambda\end{bmatrix}^{-1}=\begin{bmatrix}\boldsymbol{M}^{-1} & \boldsymbol{0}\\ \boldsymbol{0}^\top & 0\end{bmatrix}+\frac{1}{a}\begin{bmatrix}\boldsymbol{M}^{-1}\boldsymbol{X}^\top\boldsymbol{x}_i\\ -1\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_i^\top\boldsymbol{X}\boldsymbol{M}^{-1} & -1\end{bmatrix},$$

where $a:=\|\boldsymbol{x}_i\|^2+m\lambda-\boldsymbol{x}_i^\top\boldsymbol{X}\boldsymbol{M}^{-1}\boldsymbol{X}^\top\boldsymbol{x}_i$. Plugging

this result into (1) yields

$$\boldsymbol{x}^\top \boldsymbol{w}_S = \boldsymbol{x}^\top \boldsymbol{w}_{S^{\setminus i}} + \frac{\boldsymbol{x}^\top \left(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{M}^{-1}\boldsymbol{X}^\top\right)\boldsymbol{x}_i}{a}(y_i - \boldsymbol{x}_i^\top \boldsymbol{w}_{S^{\setminus i}}) \ .$$

Using result of Lemma 2, we have that $m\lambda \leq a$ and in addition by Cauchy-Schwarz inequality we have that $\boldsymbol{x}\left(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{M}^{-1}\boldsymbol{X}^\top\right)\boldsymbol{x}_i \leq 1$, since $\|\boldsymbol{x}\|, \|\boldsymbol{x}_i\| \leq 1$. □

Lemma 2. *For all $\boldsymbol{X} \in \mathbb{R}^{m\times d}$, $m, \lambda \geq 0$, we have that the matrix*

$$I - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X} + m\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top$$

is PSD and its maximum eigenvalue is less than 1.

Proof.

$$\begin{aligned} & I - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X} + m\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top \\ =& \boldsymbol{I} - \left(\boldsymbol{X}\boldsymbol{X}^\top + m\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}\boldsymbol{X}^\top \qquad (2) \\ =& \boldsymbol{I} - \boldsymbol{U}\left(\boldsymbol{\Sigma} + m\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top \\ =& \boldsymbol{U}\left(\boldsymbol{I} - \left(\boldsymbol{\Sigma} + m\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}\right)\boldsymbol{U}^\top, \end{aligned}$$

where we used identity $(\boldsymbol{X}\boldsymbol{X}^\top + m\lambda\boldsymbol{I})^{-1}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + m\lambda\boldsymbol{I})^{-1}$ to obtain (2). Subsequently we use decomposition $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$. □

2. Weaker HTL Guarantees through Correct Stability

The following HTL algorithm is analyzed by Kuzborskij & Orabona (2013).

Algorithm 1. *RLS transfer algorithm by altering training set as $\{(\boldsymbol{x}_i, y_i - f'(\boldsymbol{x}_i)) : 1 \leq i \leq m\}$ produces a hypothesis*

$$f_S^{htl'}(\boldsymbol{x}) = T_C(\boldsymbol{x}^\top\hat{\boldsymbol{w}}_S) + f'(\boldsymbol{x}),$$

where

$$\hat{\boldsymbol{w}}_S := \operatorname*{argmin}_{\boldsymbol{u}} \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{u}^\top\boldsymbol{x}_i - y_i + f'(\boldsymbol{x}_i))^2 + \lambda\|\boldsymbol{u}\|^2,$$

and the truncation function $T_C(\hat{y})$ is defined as $T_C(\hat{y}) = \min(\max(\hat{y}, -C), C)$.

We prove Theorem 1 which is corrected version of originally presented Theorem 2 in (Kuzborskij & Orabona, 2013).

Theorem 1. *Set $\lambda \geq \frac{1}{m}$. If $C \geq B + \|f'\|_\infty$, then for Algorithm 1 we have*

$$\mathbb{E}_S[(R_\mu(f_S^{htl'}) - \hat{R}^{\,\mathrm{loo}}(f_S^{htl'}))^2] = \mathcal{O}\left(C^2\frac{\sqrt{R_\mu(f')T_{C^2}\left(\frac{R_\mu(f')}{\lambda}\right) + R_\mu(f')^2}}{m\lambda^{1.5}}\right) .$$

If $C = \infty$, then for Algorithm 1 we have

$$\mathbb{E}_S[(R_\mu(f_S^{htl'}) - \hat{R}^{\,\mathrm{loo}}(f_S^{htl'}))^2] = \mathcal{O}\left(\frac{R_\mu(f')(\|f'\|_\infty + B)^2}{m\lambda^3}\right) .$$

2.1. Implications of Weaker Result

The first and the main difference comes in weaker bound on the second order moment of the Leave-One-Out (LOO) error when one applies the truncation on the predictions in the range $[-B; B]$. In (Kuzborskij & Orabona, 2013) we have claimed it to be in $\mathcal{O}\left(\frac{B^2}{m\lambda}\right)$. However in the following we obtain weaker corrected rate in $\mathcal{O}\left(\frac{B^2}{m\lambda^{1.5}}\right)$. Note that this rate is still better than the one that can be obtained for RLS through stability analysis, $\mathcal{O}\left(\frac{1}{m\lambda^3}\right)$ (De Vito et al., 2005).

The second difference comes in dependence on the risk of the source on the target domain. The correct bound is in $\mathcal{O}\left(\frac{\sqrt{R_\mu(f')^2 + R_\mu(f')}}{m\lambda^{1.5}}\right)$, rather than $\mathcal{O}\left(\frac{R_\mu(f')}{m\lambda}\right)$, incorrectly suggested earlier. Nevertheless, the important behavior of HTL is still present, that is whenever the source hypothesis f' performs well on the target domain, in other words, for $\frac{\sqrt{R_\mu(f')^2 + R_\mu(f')}}{\lambda^{1.5}} \to 0$, the LOO error approaches expected risk with probability 1.

2.2. Proof of Theorem 1

In this section we give the proof of the main theorem. First we prove utility Lemma 3.

Lemma 3. $\forall a, b, \hat{y} \in \mathbb{R}$,

$$|(a - \hat{y})^2 - (b - \hat{y})^2| \leq (a - b)^2 + 2|(b - \hat{y})(a - b)| \ .$$

Proof.

$$\begin{aligned} & |(a - \hat{y})^2 - (b - \hat{y})^2| \\ =& |a^2 - b^2 - 2\hat{y}(a - b)| \\ =& |(a - b)^2 - 2b^2 + 2ab - 2\hat{y}(a - b)| \\ =& |(a - b)^2 + 2(b - \hat{y})(a - b)| \\ \leq& (a - b)^2 + 2|(b - \hat{y})(a - b)| \ . \end{aligned}$$

□

The following Theorem upper-bounds the hypothesis stability of the Algorithm 1.

Theorem 2. *The hypothesis stability of Algorithm 1 is upper bounded as*

$$\gamma \leq T_{4C^2}\left(\frac{2R_\mu(f')}{m^2\lambda^2}\left(1+\frac{1}{\lambda}\right)\right) + 2T_{2C}\left(\frac{\sqrt{2R_\mu(f')}}{m\lambda}\sqrt{1+\frac{1}{\lambda}}\right) \cdot \sqrt{2T_{C^2}\left(\frac{R_\mu(f')}{\lambda}\right)+2R_\mu(f')}\ .$$

Proof. From Lemma 3 with $a = T_C(\Delta+\epsilon)$, $b = T_C(\Delta)$, $\hat{y} = y - f'(\boldsymbol{x})$ and also using the fact that $|T_C(\Delta+\epsilon) - T_C(\Delta)| \leq \min(|\epsilon|, 2C)$, we have

$$\left|(T_C(\Delta+\epsilon)-y+f'(\boldsymbol{x}))^2 - (T_C(\Delta)-y+f'(\boldsymbol{x}))^2\right| \leq \min(\epsilon^2, 4C^2) + 2\min(|\epsilon|, 2C)\left|T_C(\Delta)-y+f'(\boldsymbol{x})\right|\ .$$

Set $\Delta := \boldsymbol{x}^\top \boldsymbol{w}_{S^{\setminus i}}$, and $\Delta+\epsilon := \boldsymbol{x}^\top \boldsymbol{w}_S$. Taking the expectation $\mathbb{E}[\cdot] = \mathop{\mathbb{E}}_{S,(\boldsymbol{x},y)}[\cdot]$, and using Jensen's and Cauchy-Schwarz's inequalities, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|(T_C(\Delta+\epsilon)-y+f'(\boldsymbol{x}))^2 - (T_C(\Delta)-y+f'(\boldsymbol{x}))^2\right|\right] \\
&\leq \min\left(\mathbb{E}\left[\epsilon^2\right], 4C^2\right) + 2\min\left(\sqrt{\mathbb{E}\left[\epsilon^2\right]}, 2C\right) \cdot \\
&\quad \cdot \sqrt{\mathbb{E}\left[(T_C(\Delta)-y+f'(\boldsymbol{x}))^2\right]} \\
&\leq \min\left(\mathbb{E}\left[\epsilon^2\right], 4C^2\right) + 2\min\left(\sqrt{\mathbb{E}\left[\epsilon^2\right]}, 2C\right) \cdot \\
&\quad \cdot \sqrt{\mathbb{E}\left[2T_{C^2}(\|\hat{\boldsymbol{w}}_S\|^2) + 2(f'(\boldsymbol{x})-y)^2\right]} \qquad (3) \\
&\leq \min\left(\mathbb{E}\left[\epsilon^2\right], 4C^2\right) + 2\min\left(\sqrt{\mathbb{E}\left[\epsilon^2\right]}, 2C\right) \cdot \\
&\quad \cdot \sqrt{2T_{C^2}\left(\frac{R_\mu(f')}{\lambda}\right)+2R_\mu(f')}\ . \qquad (4)
\end{aligned}$$

In (3) we apply Cauchy-Schwarz inequality and elementary inequality $(a+b)^2 \leq 2a^2+2b^2$, while (4) comes from the first result of Lemma 3 in (Kuzborskij & Orabona, 2013).

We now use the fact that

$$\begin{aligned}
\mathbb{E}\left[\epsilon^2\right] &\leq \frac{1}{m^2\lambda^2}\mathbb{E}\left[(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_{S^{\setminus i}} - y_i + f'(\boldsymbol{x}))^2\right] \\
&\leq \frac{2}{m^2\lambda^2}\mathbb{E}\left[\|\hat{\boldsymbol{w}}_{S^{\setminus i}}\|^2 + (y-f'(\boldsymbol{x}))^2\right] \\
&\leq \frac{2R_\mu(f')}{m^2\lambda^2}\left(\frac{m-1}{m}\frac{1}{\lambda}+1\right) \\
&\leq \frac{2R_\mu(f')}{m^2\lambda^2}\left(\frac{1}{\lambda}+1\right)\ .
\end{aligned}$$

Putting all together we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S,(\boldsymbol{x},y)}\left[\left|(T_C(\Delta)-y+f'(\boldsymbol{x}))^2 - (T_C(\Delta+\epsilon)-y+f'(\boldsymbol{x}))^2\right|\right] \\
&\leq T_{4C^2}\left(\frac{2R_\mu(f')}{m^2\lambda^2}\left(1+\frac{1}{\lambda}\right)\right) \\
&+ 2T_{2C}\left(\frac{\sqrt{2R_\mu(f')}}{m\lambda}\sqrt{1+\frac{1}{\lambda}}\right) \\
&\cdot \sqrt{2T_{C^2}\left(\frac{R_\mu(f')}{\lambda}\right)+2R_\mu(f')}\ .
\end{aligned}$$

□

Proof of Theorem 1. We apply Theorem 1 from (Kuzborskij & Orabona, 2013). To apply this theorem, we need to upper-bound quantities $M, \mathop{\mathbb{E}}_S[\ell(\boldsymbol{w}_{S^{\setminus i}}, (\boldsymbol{x}_i, y_i))]$ and γ. As $\mathop{\mathbb{E}}_S[\ell(\boldsymbol{w}_{S^{\setminus i}}, (\boldsymbol{x}_i, y_i))]$ is already correctly bounded in (Kuzborskij & Orabona, 2013) as

$$\begin{aligned}
&\mathbb{E}_S[\ell(f_{S^{\setminus i}}, (\boldsymbol{x}_i, y_i))] \\
&\leq 2\left(1+\frac{1}{m\lambda}\right)^2\left(T_{C^2}\left(\frac{R_\mu(f')}{\lambda}\right)+R_\mu(f')\right)\ ,
\end{aligned}$$

we use bound on γ given by Theorem 2.

By definition of Theorem 1 in (Kuzborskij & Orabona, 2013),

$$\ell(\boldsymbol{w}_{S^{\setminus i}}, (\boldsymbol{x}, \hat{y})) \leq M,\ \forall \boldsymbol{x} \in \mathcal{X}, \hat{y} \in \mathcal{Y}\ .$$

So we have

$$\begin{aligned}
&\sup_{\boldsymbol{x},y}\ \left(T_C(\boldsymbol{x}^\top \boldsymbol{w}_{S^{\setminus i}}) - \hat{y} + f'(\boldsymbol{x})\right)^2 \\
&\leq \left(T_C\left(\frac{B+\|f'\|_\infty}{\sqrt{\lambda}}\right) + B + \|f'\|_\infty\right)^2\ .
\end{aligned}$$

We have this result, because term $T_C(\boldsymbol{x}^\top \boldsymbol{w}_{S^{\setminus i}})$ can be simultaneously upper-bounded by C and, using Cauchy-Schwarz inequality, $\|\boldsymbol{w}_{S^{\setminus i}}\|$. Consequently, $\|\boldsymbol{w}_{S^{\setminus i}}\|$ is bounded using second result of Lemma 3 in (Kuzborskij & Orabona, 2013).

Putting it all together and applying Theorem 1 (Kuzborskij & Orabona, 2013), we have the stated result. The dominant rates in $\mathcal{O}(\cdot)$ notation, in both truncated and untruncated cases, come from the bound on the component $M\gamma$ in Theorem 1 in (Kuzborskij & Orabona, 2013). □

3. Conclusions

For an Algorithm 1 we have obtained the generalization bound of the form,

$$R_\mu(f_S^{htl'}) \leq \hat{R}^{\,\mathrm{loo}}(f_S^{htl'}) + \mathcal{O}\left(\frac{\sqrt[4]{R_\mu(f')^2 + R_\mu(f')}}{\sqrt{m}\lambda^{0.75}}\right), \tag{5}$$

which is slightly weaker than originally claimed,

$$R_\mu(f_S^{htl'}) \leq \hat{R}^{\,\mathrm{loo}}(f_S^{htl'}) + \mathcal{O}\left(\sqrt{\frac{R_\mu(f')}{m\lambda}}\right). \tag{6}$$

The result is marginally worse in dependency on regularization parameter λ and performance of the source hypothesis on the target domain, $R_\mu(f')$. However, corrected bound preserves the important message for the scenario of transfer learning: for a stable algorithm, good performance of the source hypothesis on the target domain accelerates the convergence of the error measured on the training set to the expected risk.

Obtaining bound of a form (6) remains an open problem.

References

Bousquet, O. and Elisseeff, A. Stability and Generalization. *Journal of Machine Learning Research*, 2: 499–526, 2002.

De Vito, E., Caponnetto, A., and Rosasco, L. Model Selection for Regularized Least-Squares Algorithm in Learning Theory. *Found. Comput. Math.*, 5(1): 59–85, February 2005.

Kuzborskij, I. and Orabona, F. Stability and hypothesis transfer learning. In *Proceedings of the International Conference on Machine Learning*, 2013.

Petersen, K.B. and Pedersen, M.S. The matrix cookbook. *Technical University of Denmark*, 2008.