On the Rates of Information Transfer in Multi-Task Learning using Random Matrix Theory

24/02/2020 at 5:58pm

1 Problem Setup

In the multi-task learning (MTL) problem, we are given the input of k tasks $(X_1, Y_1), (X_2, Y_2), \ldots, (X_k, Y_k)$. We use a shared module B for all tasks and a separate module $\{A_i\}_{i=1}^k$ for each task. This corresponds to minimizing the following optimization objective.

$$f(B; A_1, \ldots, A_k) = \sum_{i=1}^{k} \|X_i B A_i - Y_i\|_F^2. \tag{1.1}$$

When the models are the same for all tasks, we use a simplified objective as follows.

$$f(w) = \sum_{i=1}^{k} \|X_i w - Y_i\|_F^2. \tag{1.2}$$

Equation (1.1) is a slightly more complicated parametrization of equation (1.2), in that each A_i can select a model from the subspace of B to fit (X_i, Y_i).

1.1 Hypothesis

Our hypothesis is that the heterogeneity among the multiple tasks can be categorized into two classes, *covariate shift* and *model shift*.

1.1.1 Covariate Shift

A natural setting is when the covariance matrices $X_i^\top X_i$ are different, i.e. having different spectrum. This is also known as covariate shift in the literature. Our hypothesis is that the covariate shift can slow down the convergence of learning the true θ as a function of the number of data points.

In this setting, we assume that the covariates of task i are drawn from $\Sigma_i \in \mathbb{R}^{d \times d}$, but the models are the same across all the tasks, i.e.

$$y_i = X_i \beta + \varepsilon_i, \text{ with } \frac{1}{n_i} X_i^\top X_i \sim \Sigma_i \tag{1.3}$$

1.1.2 Model Shift

The most general version is when the covariates are different between different tasks and the single-task models are also different. Here the hypothesis is that the optimal B is captured by a low-rank approximation of the single-task models.

We remove the assumption that the models are the same. The i-th task data can be viewed as generated by a separate model $\beta_i \in \mathbb{R}^d$.

$$y_i = X_i\beta_i + \varepsilon_i. \tag{1.4}$$

1.2 Objectives

We will designate the k-th task as the target. Our goal is to come up with an estimator $\hat{\beta}$ to provide accurate predictions for the target task. Concretely, we focus on two objectives.

- Estimation error for the target model β_t: we consider their distance
$$\mathop{\mathbb{E}}_{\varepsilon}\left[\left\|\hat{\beta} - \beta_t\right\|^2\right].$$
- Test error for the target task:
$$\mathop{\mathbb{E}}_{x\sim\Sigma_k}\left[(x^\top\hat{\beta} - x^\top\beta_t)^2\right].$$

2 Results for Two Tasks

We would like to get insight on how covariate and model shifts result in slower rates of transfer. For the case of two tasks, we can get precise rates in the low-dimensional setting using random matrix theory. Since there are only two tasks, we call task 1 the source task and task 2 the target task, i.e. $\beta_1 = \beta_s$ and $\beta_2 = \beta_t$.

2.1 Covariate Shift

First, we show that if β_s and β_t are equal, then combining the source and target task always helps. The estimation error using the source and target together is

$$e(\hat{\beta}_{s,t}) = \sigma^2 \cdot \mathrm{Tr}[(X_1^\top X_1 + X_2^\top X_2)^{-1}]. \tag{2.1}$$

The estimation error using the target alone is

$$e(\hat{\beta}_t) = \sigma^2 \cdot \mathrm{Tr}[(X_2^\top X_2)^{-1}]. \tag{2.2}$$

Proposition 2.1. *When there is no model shift, adding the source task data always reduces the estimation error, i.e.*

$$e(\hat{\beta}_{s,t}) \leqslant e(\hat{\beta}_t).$$

Proof. This is simply because $\mathrm{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right] \leqslant \mathrm{Tr}\left[(X_1^\top X_1)^{-1}\right]$. □

Next, we calculate the amount of improvement by comparing equation (2.1) to equation (2.2). We study the spectrum of the random matrix model:

$$Q = \Sigma_1^{1/2} X_1^T X_1 \Sigma_1^{1/2} + \Sigma_2^{1/2} X_2^T X_2 \Sigma_2^{1/2},$$

where $\Sigma_{1,2}$ are $p\times p$ deterministic covariance matrices, and $X_1 = (x_{ij})_{1\leqslant i\leqslant n_1, 1\leqslant j\leqslant p}$ and $X_2 = (x_{ij})_{n_1+1\leqslant i\leqslant n_1+n_2, 1\leqslant j\leqslant p}$ are $n_1 \times p$ and $n_2 \times p$ random matrices, respectively, where the entries x_{ij}, $1 \leqslant i \leqslant n_1 + n_2 \equiv n$, $1 \leqslant j \leqslant p$, are real independent random variables satisfying

$$\mathbb{E}x_{ij} = 0, \qquad \mathbb{E}|x_{ij}|^2 = n^{-1}. \tag{2.3}$$

For now, we assume that the random variables x_{ij} are i.i.d. Gaussian, but we know that universality holds for generally distributed entries. We shall consider the high-dimensional setting such that

$$\gamma_n := \frac{p}{n} \to \gamma, \quad c_n := \frac{n_1}{n} \to c, \quad \text{as} \ \ n \to \infty,$$

for some constants $\gamma \in (0, \infty)$ and $c \in (0, 1)$.

Lemma 2.2. *Put in the statement We have with high probability* $1 - \mathrm{o}(1)$,

$$\frac{1}{p}\operatorname{Tr}(Q^{-1}) = \frac{1}{p}\operatorname{Tr}\left[\frac{1}{a_3\Sigma_1 + a_4\Sigma_2}\right] + \mathrm{O}(n^{-1+\varepsilon})$$

for any constant $\varepsilon > 0$, *where* $a_{3,4}$ *are found using equations in* (2.7).

We assume that $\Sigma_1^{-1/2}\Sigma_2$ has eigendecomposition

$$\Sigma_1^{-1/2}\Sigma_2^{1/2} = ODO^T, \quad D = \operatorname{diag}(d_1, \cdots, d_p). \tag{2.4}$$

Then by the rotational invariance of Gaussian matrices, we have

$$\widetilde{Q} \overset{d}{=} \Sigma_1^{1/2} O \widetilde{Q} O^T \Sigma_1^{1/2}, \quad \widetilde{Q} := X_1^T X_1 + D X_2^T X_2 D.$$

Thus we study the spectrum of $\widetilde{Q}$ instead. We define $\mathcal{G}(z) := (\widetilde{Q} - z)^{-1}$ for $z \in \mathbb{C}_+$. With some random matrix tools, we have that

$$\mathcal{G}(z) \approx \operatorname{Diag}\left(\frac{1}{-z\left(1 + m_3(z) + d_i^2 m_4(z)\right)}\right)_{1 \leqslant i \leqslant p} = \frac{1}{-z\left(1 + m_3(z) + D^2 m_4(z)\right)}$$

in certain sense. Here $m_{3,4}(z)$ satisfy the following self-consistent equations

$$\frac{n_1}{n}\frac{1}{m_3} = -z + \frac{1}{n}\sum_{i=1}^{p}\frac{1}{1 + m_3 + d_i^2 m_4}, \quad \frac{n_2}{n}\frac{1}{m_4} = -z + \frac{1}{n}\sum_{i=1}^{p}\frac{d_i^2}{1 + m_3 + d_i^2 m_4} \tag{2.5}$$

When $z \to 0$, we shall have

$$m_3(z) = -\frac{a_3}{z} + \mathrm{O}(1), \quad m_4(z) = -\frac{a_4}{z} + \mathrm{O}(1), \quad a_3, a_4 > 0.$$

Then for $z \to 0$, the equations in (2.5) are reduced to

$$\frac{n_1}{n}\frac{1}{a_3} = 1 + \frac{1}{n}\sum_{i=1}^{p}\frac{1}{a_3 + d_i^2 a_4}, \quad \frac{n_2}{n}\frac{1}{a_4} = 1 + \frac{1}{n}\sum_{i=1}^{p}\frac{d_i^2}{a_3 + d_i^2 a_4}. \tag{2.6}$$

First, it is easy to see that these equations are equivalent to

$$a_3 + a_4 = 1 - \gamma_n, \quad a_3 + \frac{1}{n}\sum_{i=1}^{p}\frac{a_3}{a_3 + d_i^2[(1-\gamma_n) - a_3]} = c_n. \tag{2.7}$$

Furthermore, we have

$$\begin{aligned}\operatorname{Tr}(Q^{-1}) &= \lim_{z\to 0}\operatorname{Tr}\left[\Sigma_1^{-1/2} O \mathcal{G}(z) O^T \Sigma_1^{-1/2}\right] = \operatorname{Tr}\left[\Sigma_1^{-1/2} O\left(\frac{1}{a_3 + D^2 a_4}\right) O^T \Sigma_1^{-1/2}\right] \\ &= \operatorname{Tr}\left[\Sigma_1^{-1/2}\frac{1}{a_3 + \Sigma_1^{-1}\Sigma_2 a_4}\Sigma_1^{-1/2}\right] = \operatorname{Tr}\left[\frac{1}{a_3\Sigma_1 + a_4\Sigma_2}\right].\end{aligned}$$

2.2 Model Shift

In this case, β_s and β_t are different. If we put the two tasks together, we get

$$\hat{\beta}_{s,t} = (X_1^\top X_1 + X_2^\top X_2)^{-1}(X_1^\top Y_1 + X_2^\top Y_2) \tag{2.8}$$

$$= (X_1^\top X_1 + X_2^\top X_2)^{-1}\left((X_1^\top X_1\beta_s + X_2^\top X_2\beta_t) + (X_1^\top \varepsilon_1 + X_2^\top \varepsilon_2)\right) \tag{2.9}$$

Hence

$$\mathop{\mathbb{E}}_{\varepsilon}\left[\left\|\hat{\beta}_{s,t}-\beta_t\right\|\right]=\left\|(X_1^\top X_1+X_2^\top X_2)^{-1}X_1^\top X_1(\beta_s-\beta_t)\right\|^2+\sigma^2\operatorname{Tr}\left[(X_1^\top X_1+X_2^\top X_2)^{-1}\right] \tag{2.10}$$

We are interested in the eigenvalues of

$$(X_1^\top X_1)^{-1}(X_2^\top X_2),$$

which is called a generalized Fisher matrix. As in the previous setting, we write out their covariance explicitly and consider

$$(\Sigma_1^{1/2}X_1^TX_1\Sigma_1^{1/2})^{-1}\Sigma_2^{1/2}X_2^TX_2\Sigma_2^{1/2},$$

where $\Sigma_{1,2}$ are $p\times p$ deterministic covariance matrices, and $X_1=(x_{ij})_{1\leqslant i\leqslant n_1,1\leqslant j\leqslant p}$ and $X_2=(x_{ij})_{n_1+1\leqslant i\leqslant n_1+n_2,1\leqslant j\leqslant p}$ are $n_1\times p$ and $n_2\times p$ random matrices, respectively, where the entries x_{ij}, $1\leqslant i\leqslant n_1+n_2\equiv n$, $1\leqslant j\leqslant p$, are real independent random variables satisfying (2.3).

For now, we assume that the random variables x_{ij} are i.i.d. Gaussian, and that $\Sigma_1^{-1/2}\Sigma_2$ has singular value decomposition

$$\Sigma_1^{-1/2}\Sigma_2^{1/2}=O_1DO_2,\quad D=\operatorname{diag}(d_1,\cdots,d_p).$$

Then it is equivalent to study

$$Q:=(X_1^TX_1)^{-1}DX_2^TX_2D,$$

which has the same nonzero eigenvalues of

$$\mathcal{Q}:=X_2D(X_1^TX_1)^{-1}DX_2^T,$$

i.e. $\mathcal{Q}$ has the same nonzero eigenvalues of Q, but has (n_2-p) more zero eigenvalues.

We can study it using the linearization matrix (for my own purpose right now)

$$H=\begin{pmatrix}-zI & X_2D & 0\\ DX_2^T & 0 & X_1^T\\ 0 & X_1 & I\end{pmatrix},\quad G(z):=H(z)^{-1}.$$

Then the $(1,1)$-th block is equal to $\mathcal{G}(z):=(\mathcal{Q}-z)^{-1}$. We denote

$$m_1(z):=\frac{1}{n}\operatorname{Tr}\mathcal{G}(z),\quad m(z)=\frac{1}{p}\left[nm_1(z)+\frac{n_2-p}{z}\right]. \tag{2.11}$$

We can show that it satisfies the following self-consistent equations together with another $m_3(z)$:

$$\frac{n_2}{n}\frac{1}{m_1}=-z+\frac{1}{n}\sum_{i=1}^p\frac{d_i^2}{m_3+d_i^2m_1},\quad \frac{n_1}{n}\frac{1}{m_3}=1+\frac{1}{n}\sum_{i=1}^p\frac{1}{m_3+d_i^2m_1}. \tag{2.12}$$

For these two equations, we can obtain one single equation for $m_1(z)$:

$$m_3=1-\frac{p}{n}+zm_1,\quad \frac{n_2}{n}\frac{1}{m_1}=-z+\frac{1}{n}\sum_{i=1}^p\frac{d_i^2}{1-\frac{p}{n}+(z+d_i^2)m_1}. \tag{2.13}$$

One can solve the above equation for m_1 with positive imaginary parts, and then calculate $m(z)$ using (2.11).

With $m(z)$, we can define

$$\rho_c(z):=\frac{1}{\pi}\lim_{\eta\downarrow 0}m(z).$$

It will be a compact supported probability density which gives the eigenvalue distribution of Q. Moreover, we have

$$\frac{1}{p}\operatorname{Tr}\frac{1}{(\Sigma_1^{1/2}X_1^TX_1\Sigma_1^{1/2})^{-1}\Sigma_2^{1/2}X_2^TX_2\Sigma_2^{1/2}+1}\left(\frac{1}{(\Sigma_1^{1/2}X_1^TX_1\Sigma_1^{1/2})^{-1}\Sigma_2^{1/2}X_2^TX_2\Sigma_2^{1/2}+1}\right)^T\approx\int\frac{\rho_c(x)}{(x+1)^2}\mathrm{d}x.$$

Moreover, the right edge λ_+ of ρ_c gives the location of the largest eigenvalue, while the left edge λ_- of ρ_c gives the location of the smallest eigenvalue (we have ways to determine the edges by solving some equations, state them later). They will provide the upper and lower bounds on the operator norm:

$$\frac{1}{1+\lambda_+} \leqslant \frac{1}{(\Sigma_1^{1/2} X_1^T X_1 \Sigma_1^{1/2})^{-1} \Sigma_2^{1/2} X_2^T X_2 \Sigma_2^{1/2} + 1} \leqslant \frac{1}{1+\lambda_-}.$$