A Distribution-Dependent Analysis of Meta-Learning

Mikhail Konobeev[1] Ilja Kuzborskij[2] Csaba Szepesvári[1,2]

[1]University of Alberta

[2]DeepMind

Prior Work on Theoretical Analysis of Meta-Learning

- In learning theory, the most often used lower bounds are *distribution-free* or *problem independent*
- If the class of meta-distributions is sufficiently rich, the bounds simply tell us that the best meta-learner is competitive with the best "standard learner"
- For example, Lucas et al. (2020) gave a worst-case lower bound $\Omega(d/((2r)^{-d}M + m))$ for parameter identification which reduces to the standard bound on linear regression as $r \to \infty$
  - $r \geq 1$ is the radius of the ball that contains the parameters
  - M is the total number of data points in the training tasks
  - m is the number of data points in the training set of the target task

Prior Work on Theoretical Analysis of Meta-Learning

- In learning theory, the most often used lower bounds are *distribution-free* or *problem independent*
- If the class of meta-distributions is sufficiently rich, the bounds simply tell us that the best meta-learner is competitive with the best "standard learner"
- For example, Lucas et al. (2020) gave a worst-case lower bound $\Omega(d/((2r)^{-d}M + m))$ for parameter identification which reduces to the standard bound on linear regression as $r \to \infty$
    - $r \geq 1$ is the radius of the ball that contains the parameters
    - M is the total number of data points in the training tasks
    - m is the number of data points in the training set of the target task

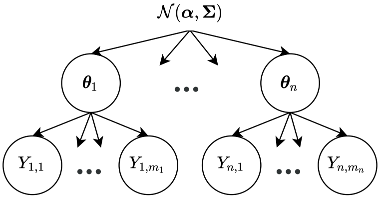
This work: *bounds that take into account* **task-relatedness** *via dependence on the parameters of the meta-distribution.*

Problem Setting: Mixed Linear Regression

Let the i-th task be parameterized by $\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{X}_i \boldsymbol{\theta}_i, \sigma^2 \boldsymbol{I}), \qquad (1)$$

and inputs $\boldsymbol{X}_i \in \mathbb{R}^{m_i \times d}$ be deterministic.



We can derive the marginal distribution over $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1^\top & \dots & \boldsymbol{Y}_n^\top \end{bmatrix}^\top$,

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\Psi} \boldsymbol{\alpha}, \boldsymbol{K}), \qquad (2)$$

where $\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{X}_1^\top & \dots & \boldsymbol{X}_n^\top \end{bmatrix}^\top$, $\boldsymbol{X} = \texttt{block\_diag}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$, and $\boldsymbol{K} = \boldsymbol{X}(\boldsymbol{I}_n \otimes \boldsymbol{\Sigma})\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I}$.

Bounding Squared Error

- We will study learning algorithms with performance measured by quadratic loss of adapting to the last task:

$$\mathcal{L}(\mathcal{A}, \boldsymbol{x}) = \mathbb{E}[(Y - \mathcal{A}(\mathcal{D}, \boldsymbol{x})))^2]. \tag{3}$$

  where $Y = \boldsymbol{x}^T \boldsymbol{\theta}_n + \varepsilon \sim \mathcal{N}(\boldsymbol{x}^T \boldsymbol{\theta}_n, \sigma^2)$.
- The risk decomposes into posterior mean estimation and posterior variance:

$$\mathcal{L}(\mathcal{A}, \boldsymbol{x}) = \mathbb{E}\left[(\mathbb{E}[Y|\mathcal{D}] - \mathcal{A}(\mathcal{D}, \boldsymbol{x}))^2\right] + \mathbb{E}[\mathbb{V}[Y|\mathcal{D}]] \tag{4}$$

Bounding Squared Error

- We will study learning algorithms with performance measured by quadratic loss of adapting to the last task:

$$\mathcal{L}(\mathcal{A}, \boldsymbol{x}) = \mathbb{E}[(Y - \mathcal{A}(\mathcal{D}, \boldsymbol{x})))^2]. \tag{3}$$

where $Y = \boldsymbol{x}^T \boldsymbol{\theta}_n + \varepsilon \sim \mathcal{N}(\boldsymbol{x}^T \boldsymbol{\theta}_n, \sigma^2)$.

- The risk decomposes into posterior mean estimation and posterior variance:

$$\mathcal{L}(\mathcal{A}, \boldsymbol{x}) = \mathbb{E}\left[(\mathbb{E}[Y|\mathcal{D}] - \mathcal{A}(\mathcal{D}, \boldsymbol{x}))^2\right] + \mathbb{E}[\mathbb{V}[Y|\mathcal{D}]] \tag{4}$$

- Letting $\boldsymbol{\mathcal{T}} = \mathbb{V}[\theta_n|\mathcal{D}] = \left(\boldsymbol{\Sigma}^{-1} + \sigma^{-2}\boldsymbol{X}_n^\top \boldsymbol{X}_n\right)^{-1}$ we have

$$\mathbb{E}[Y|\mathcal{D}] = \boldsymbol{x}^\top \boldsymbol{\mathcal{T}} \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} + \sigma^{-2}\boldsymbol{X}_n^\top \boldsymbol{Y}_n\right) \tag{5}$$

$$\mathbb{V}[Y|\mathcal{D}] = \boldsymbol{x}^\top \boldsymbol{\mathcal{T}} \boldsymbol{x} + \sigma^2 \tag{6}$$

Matching Lower and Upper Bounds

- Assume known covariance structure $(\sigma^2, \boldsymbol{\Sigma})$
- For any estimator $\mathcal{A}(\mathcal{D}, \boldsymbol{x})$ we have the following lower bound which depends on the parameters of the statistical model

$$\mathcal{L}(\mathcal{A}, \boldsymbol{x}) \geq \frac{1}{16\sqrt{e}} \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{\mathcal{T}} \boldsymbol{x} + \sigma^2, \tag{7}$$

  where $\boldsymbol{M} = \boldsymbol{\mathcal{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} \boldsymbol{\Psi})^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mathcal{T}}$
- We also provide special cases of this lower bound in the paper and compare them with prior work

Matching Lower and Upper Bounds

- Assume known covariance structure $(\sigma^2, \boldsymbol{\Sigma})$
- For any estimator $\mathcal{A}(\mathcal{D}, \boldsymbol{x})$ we have the following lower bound which depends on the parameters of the statistical model

$$\mathcal{L}(\mathcal{A}, \boldsymbol{x}) \geq \frac{1}{16\sqrt{e}} \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{\mathcal{T}} \boldsymbol{x} + \sigma^2, \tag{7}$$

  where $\boldsymbol{M} = \boldsymbol{\mathcal{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} \boldsymbol{\Psi})^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mathcal{T}}$
- We also provide special cases of this lower bound in the paper and compare them with prior work
- For $\mathcal{A}(\mathcal{D}, \boldsymbol{x})$ matching the form of $\mathbb{E}[Y|\mathcal{D}]$ with $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_{MLE} = (\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} \boldsymbol{Y}$ we have

$$\mathcal{L}(\mathcal{A}, \boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{\mathcal{T}} \boldsymbol{x} + \sigma^2 \tag{8}$$

- Optimal $\mathcal{A}(\mathcal{D}, \boldsymbol{x})$ matches the solution of a *weighted* version of biased regression

Special Cases of Our Lower Bounds

- If the input covariance for the i-th task is $\frac{m_i}{d}\boldsymbol{I}$ and $\boldsymbol{\Sigma} = \tau^2 \boldsymbol{I}$ we get

$$\frac{\mathcal{L}(\mathcal{A}, \boldsymbol{x}) - \sigma^2}{\sigma^2} \geq \frac{H_{\tau^2}}{16\sqrt{e}} \cdot \frac{d^2\sigma^2}{n(\tau^2 m_n + d\sigma^2)^2} + \frac{d\tau^2}{\tau^2 m_n + d\sigma^2} \tag{9}$$

$$\to \left(\frac{m_n}{d} + \frac{\sigma^2}{\tau^2}\right)^{-1} \text{ as } n \to \infty, \tag{10}$$

where H_z is the harmonic mean of the sequence $(z + d\sigma^2/m_i)_{i=1}^n$.

- If the input covariance for the i-th task is $\frac{m_i}{d}\boldsymbol{I}$ and $\boldsymbol{\Sigma}$ is an arbitrary rank $s \leq d$ positive semi-definite matrix

$$\frac{\mathcal{L}(\mathcal{A}, \boldsymbol{x}) - \sigma^2}{\sigma^2} \geq \frac{H_{\lambda_s}}{16\sqrt{e}} \cdot \frac{sd\sigma^2}{n(\lambda_1 m_n + d\sigma^2)^2} + \frac{s\lambda_s}{\lambda_s m_n + d\sigma^2}, \tag{11}$$

where $\lambda_1 > \cdots > \lambda_s > 0$ are the eigenvalues of $\boldsymbol{\Sigma}$.

Practical Adaptation via EM Algorithm

Algorithm 1 EM procedure to estimate $(\boldsymbol{\alpha}, \sigma^2, \boldsymbol{\Sigma})$

Require: Initial parameter estimates $\widehat{\mathcal{E}}_1 = (\hat{\boldsymbol{\alpha}}_1, \widehat{\sigma}_1^2, \hat{\boldsymbol{\Sigma}}_1)$
Ensure: Final parameter estimates $\widehat{\mathcal{E}}_t = (\hat{\boldsymbol{\alpha}}_t, \widehat{\sigma}_t^2, \hat{\boldsymbol{\Sigma}}_t)$

$\hat{\boldsymbol{\mathcal{T}}}_{1,i} \leftarrow \mathbf{0}, \ \hat{\boldsymbol{\mu}}_{1,i} \leftarrow \mathbf{0} \quad i \in \{1, \ldots, n\}$
repeat
  for $i = 1, \ldots, n$ **do** ▷ E-step
    $\hat{\boldsymbol{\mathcal{T}}}_{t,i} \leftarrow \left(\hat{\boldsymbol{\Sigma}}_t^{-1} + \widehat{\sigma}_t^{-2} \boldsymbol{X}_i^\top \boldsymbol{X}_i\right)^{-1}$
    $\hat{\boldsymbol{\mu}}_{t,i} \leftarrow \hat{\boldsymbol{\mathcal{T}}}_{t,i} \left(\hat{\boldsymbol{\Sigma}}_t^{-1} \hat{\boldsymbol{\alpha}}_t + \widehat{\sigma}_t^{-2} \boldsymbol{X}_i^\top \boldsymbol{Y}_i\right)$
  end for
  $\hat{\boldsymbol{\alpha}}_t \leftarrow \frac{1}{n} \sum_{i=1}^n \hat{\boldsymbol{\mu}}_{t,i}$ ▷ M-step
  $\hat{\boldsymbol{\Sigma}}_t \leftarrow \frac{1}{n} \sum_{i=1}^n \left(\hat{\boldsymbol{\mathcal{T}}}_{t,i} + (\hat{\boldsymbol{\mu}}_{t,i} - \hat{\boldsymbol{\alpha}}_t)(\hat{\boldsymbol{\mu}}_{t,i} - \hat{\boldsymbol{\alpha}}_t)^\top\right)$
  $\widehat{\sigma}_t^2 \leftarrow \frac{1}{n} \sum_{i=1}^n \frac{1}{m_i} \left(\sum_{j=1}^{m_i} (Y_{i,j} - \hat{\boldsymbol{\mu}}_i^T \boldsymbol{x}_{i,j})^2 + \operatorname{tr}\left(\boldsymbol{X}_i \hat{\boldsymbol{\mathcal{T}}}_{t,i} \boldsymbol{X}_i^\top\right)\right)$
  $t \leftarrow t + 1$
until Convergence

At the end use plug-in estimate of $\hat{\boldsymbol{\theta}}_n$:

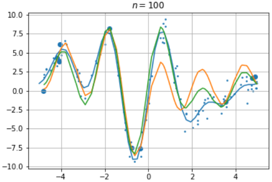
$$\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\mathcal{T}}} \left(\hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\alpha}} + \hat{\sigma}^{-2} \boldsymbol{X}_n^T \boldsymbol{Y}_n\right)$$

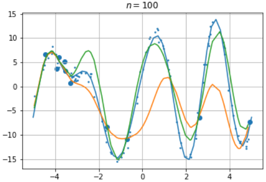
and predict $\mathcal{A}(\mathcal{D}, \boldsymbol{x}) = \hat{\boldsymbol{\theta}}_n^T \boldsymbol{x}$.

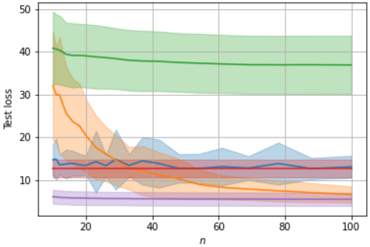
Fourier Experiments

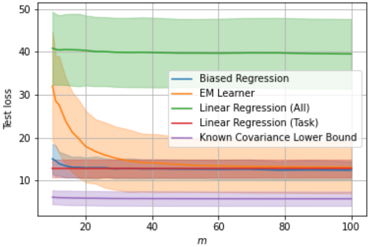
$$u \sim \mathcal{U}nif[-5, 5]$$

$$x_j = \begin{cases} \sin\left(5^{-1}\pi j u\right), & \text{if } 1 \le j \le 5 \\ \cos\left(5^{-1}\pi (j-5) u\right), & \text{if } 6 \le j \le 10 \\ 1, & \text{if } j = 11 \end{cases}$$









Spherical Synthetic Experiments

x is sampled from a unit sphere with $d = 42$

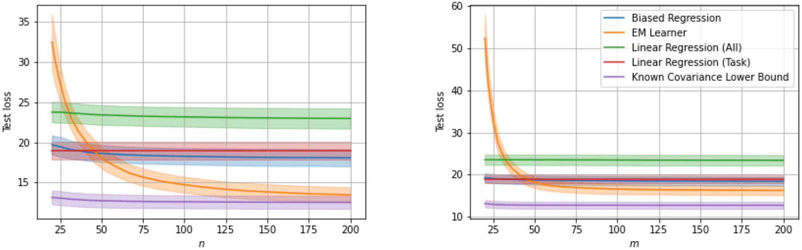


Figure: Spherical Synthetic Experiment Results

School Data Experiment

Predicting exam scores for students from different schools with $d = 27$. Each school could be thought of as a separate meta-learning task.

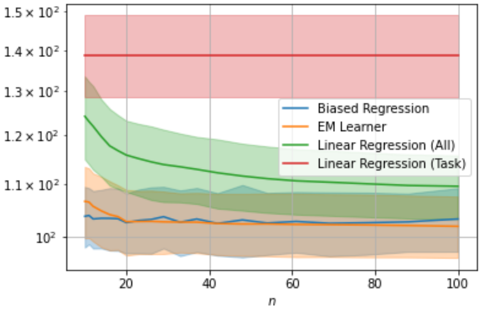


Figure: School Data Experiment Results

Subspace Estimation

EM Learner can estimate subspace matrix by zeroing out the smallest eigenvalues of $\hat{\Sigma}$.

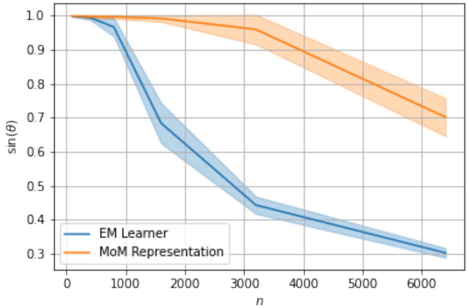


Figure: Comparison with the Method of Moments subspace estimation algorithm of Tripuraneni et al. (2020) in the same setting as theirs.

Summary of the Contributions

- Derived, up to a universal constant, matching lower and upper bounds for the studied problem
- Showed that the upper bound holds for the weighted version of biased regularized regression
- Proposed to use the EM algorithm for the case of unknown covariances and derived analytic expressions for the two steps of the algorithm
- Experimentally showed that EM attains the lower bound for sufficient number of tasks and that it is competitive as a representation learner.

J. Lucas, M. Ren, I. Kameni, T. Pitassi, and R. Zemel. Theoretical bounds on estimation error for meta-learning. arXiv:2010.07140, 2020.

N. Tripuraneni, C. Jin, and M. I. Jordan. Provable meta-learning of linear representations. arXiv:2002.11684, 2020.