Generalization Guarantees for Multi-Task ~~and Meta-~~ Learning

Christoph Lampert (work done with Hossein Zakerinia)

Institute of Science and Technology Austria

Sep 22, 2025

- public research institute, opened in 2009
- located in outskirts of Vienna

Focus on curiosity-driven basic research

- avoiding boundaries between disciplines
- current 95 research groups
  - Computer Science, Mathematics, Physics, Astronomy, Chemistry, Biology, Neuroscience, Earth and Climate Sciences
- ELLIS Unit since 2019

We're hiring!

- interns, PhD students, **postdocs**
- faculty (tenure-track or tenured), sabbaticals, ...

More information: `chl@ist.ac.at` or `https://cvml.ist.ac.at`

A (Single) Learning Task

Setting:

- input set: $\mathcal{X}$,   e.g., text documents
- output set: $\mathcal{Y}$ e.g., labels $\mathcal{Y} = \{$"spam", "not spam"$\}$
- data distribution: $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ (fixed, but unknown)

Goal:

- find a good predictor/hypothesis/model: $f : \mathcal{X} \to \mathcal{Y}$ e.g. deep network

What do we mean by "good"?

- loss function: $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ e.g. $\ell(y, \bar{y}) = [\![y \neq \bar{y}]\!]$

- aim for model with small risk

$$\mathcal{R}(f) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \ell(\, y, f(x)\,)$$

A (Single) Learning Task

How to find a model, $f : \mathcal{X} \to \mathcal{Y}$, with small risk, $\mathcal{R}(f) = \mathop{\mathbb{E}}_{(x,y)} \ell(\, y, f(x)\,)$?

- training set: $S = \{(x_1, y_1) \ldots, (x_m, y_m)\} \overset{i.i.d.}{\sim} \mathcal{D}$,
- model class: $\mathcal{F} \subset \{f : \mathcal{X} \to \mathcal{Y}\}$
- learning algorithm ("learner"): $\mathcal{A} : \mathbb{P}(\mathcal{X} \times \mathcal{Y}) \to \mathcal{F}$
  - e.g., minimize the empirical risk

$$\widehat{\mathcal{R}}(f) = \frac{1}{m} \sum_{(x,y) \in S} \ell(\, y, f(x)\,)$$

Grand challenge:

- computable guarantees on true risk, $\mathcal{R}(f)$, e.g. based on empirical risk, $\widehat{\mathcal{R}}(f)$
  $\to$ generalization bound

Theorem (Theorem 7.7 in (Shalev-Shwartz, Ben-David. 2014))

Let $\mathcal{F}$ be a countable model class and let $E : \mathcal{F} \to \{0,1\}^*$ be a prefix-free encoding of the elements in $\mathcal{F}$. Then, for any data distribution, $\mathcal{D}$, any sample size, m, and any confidence value, $\delta > 0$, it holds with probability at least $1-\delta$ over the sampling of $S \sim \mathcal{D}^m$ that:

$$\forall f \in \mathcal{F}: \qquad \mathcal{R}(f) \leq \widehat{\mathcal{R}}(f) + \sqrt{\frac{|E(f)| + \log(2/\delta)}{2m}},$$

for $|E(f)| = \log 2 \cdot \text{length}(E(f))$, where length$(\cdot)$ denotes the length of a string.

Theorem (Theorem 7.7 in (Shalev-Shwartz, Ben-David. 2014))

Let $\mathcal{F}$ be a countable model class and let $E : \mathcal{F} \to \{0,1\}^*$ be a prefix-free encoding of the elements in $\mathcal{F}$. Then, for any data distribution, $\mathcal{D}$, any sample size, m, and any confidence value, $\delta > 0$, it holds with probability at least $1 - \delta$ over the sampling of $S \sim \mathcal{D}^m$ that:

$$\forall f \in \mathcal{F} : \qquad \mathcal{R}(f) \leq \widehat{\mathcal{R}}(f) + \sqrt{\frac{|E(f)| + \log(2/\delta)}{2m}},$$

for $|E(f)| = \log 2 \cdot \mathit{length}(E(f))$, where $\mathit{length}(\cdot)$ denotes the length of a string.

How to "encode"? For example, model with parameter vector $\theta \in \mathbb{R}^D$:

- store entries $(\theta_1, \ldots, \theta_D)$ as 32bit floats: length$(E(f)) = 32D$,
- if θ is sparse with s non-zeros: store positions+values: length$(E(f)) = (\lceil \log_2 D \rceil + 32)s$,
- if many entries of θ repeat: create a codebook, and store ids instead of values,
- many other: Huffman coding, arithmetic coding, run-level coding, ...

Theorem (Theorem 7.7 in (Shalev-Shwartz, Ben-David. 2014))

Let $\mathcal{F}$ be a countable model class and let $E : \mathcal{F} \to \{0,1\}^*$ be a prefix-free encoding of the elements in $\mathcal{F}$. Then, for any data distribution, $\mathcal{D}$, any sample size, m, and any confidence value, $\delta > 0$, it holds with probability at least $1-\delta$ over the sampling of $S \sim \mathcal{D}^m$ that:

$$\forall f \in \mathcal{F}: \qquad \mathcal{R}(f) \leq \widehat{\mathcal{R}}(f) + \sqrt{\frac{|E(f)| + \log(2/\delta)}{2m}},$$

for $|E(f)| = \log 2 \cdot$ length$(E(f))$, where length$(\cdot)$ denotes the length of a string.

- principled learning algorithm: minimize the right hand size

$$\mathcal{A} : S \mapsto \operatorname*{\mathbf{argmin}}_{f \in \mathcal{F}} \left[\widehat{\mathcal{R}}(f) + \sqrt{\frac{|E(f)|}{2m}} \right]$$

- numeric values of (**??**) might or might not be informative (non-vacuous)

Theorem (Theorem 7.7 in (Shalev-Shwartz, Ben-David. 2014))

Let $\mathcal{F}$ be a countable model class and let $E : \mathcal{F} \to \{0,1\}^*$ be a prefix-free encoding of the elements in $\mathcal{F}$. Then, for any data distribution, $\mathcal{D}$, any sample size, m, it holds with high probability that:

$$\forall f \in \mathcal{F}: \quad \mathcal{R}(f) \lesssim \widehat{\mathcal{R}}(f) + \sqrt{\frac{|E(f)|}{2m}}, \quad \textit{(dropping } \log\textit{-terms)},$$

for $|E(f)| = \log 2 \cdot \mathit{length}(E(f))$, where length$(\cdot)$ denotes the length of a string.

- r.h.s. suggest a principled learning algorithm: minimize the right hand size

$$\mathcal{A} : S \mapsto \operatorname*{\mathbf{argmin}}_{f \in \mathcal{F}} \left[\widehat{\mathcal{R}}(f) + \sqrt{\frac{|E(f)|}{2m}}\right]$$

- numeric values of r.h.s. might or might not be informative (non-vacuous)

A (Single) Learning Task

Alternative analysis yields "fast-rate" bounds (for $m \geq 8$):

Theorem (Corollary of Theorem 5 in (Maurer, 2024))

Under the same assumption as above, it holds with high probability that

$$\forall f \in \mathcal{F}: \quad \mathrm{kl}\left(\widehat{\mathcal{R}}(f) \,\|\, \mathcal{R}(f)\right) \lesssim \frac{|E(f)|}{m},$$

with $\mathrm{kl}(q\|p) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}$.

Less interpretable left hand side, but:

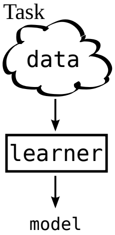
- recovers the classical $\sqrt{1/2m}$-rate using: $2(q-p)^2 \leq \mathrm{kl}(q\|p)$ (Pinsker's ineq)
- yields tighter guarantees on $\mathcal{R}(f)$ if $\widehat{\mathcal{R}}(f)$ is small. In particular (because $p \leq \mathrm{kl}(0\|p)$):

$$\forall f \in \mathcal{F} \text{ with } \widehat{\mathcal{R}}(f) = 0: \qquad \mathcal{R}(f) \lesssim \frac{|E(f)|}{m}.$$

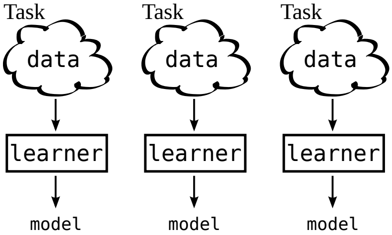
No closed form expression to invert kl, but numerically easy.

From Single-Task to Multi-Task Learning

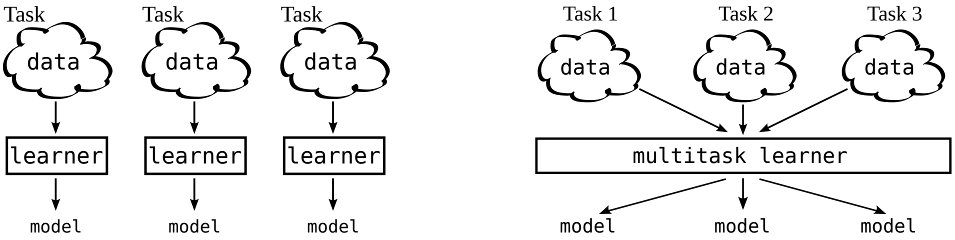
Single-Task Learning



Multi-Task Learning (MTL)



Multi-Task Learning (MTL)



Learning multiple tasks jointly,

- e.g. spam filters, recommender systems, next-word prediction:
  - many users, each has little annotated data, each has different preferences
- e.g. medical image analysis: different cancer types, different hospitals
- e.g. self-driving cars: different image analysis tasks

Sharing information between tasks might improve all models.

Learning multiple tasks jointly:

- multiple data distributions $\mathcal{D}_1, \dots, \mathcal{D}_n$
- multiple training sets $S_1, \dots, S_n$ of sizes $m_1, \dots, m_n$
- for simplicity: same input/output sets, same model class, same loss function

Define analog quantities to single task learning:

- each task, i, has an expected risk and an empirical risk

$$\mathcal{R}_i(f) = \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}_i} \ell(\,y, f(x)\,), \qquad \widehat{\mathcal{R}}_i(f) = \frac{1}{m_i}\sum_{(x,y)\in S_i} \ell(\,y, f(x)\,).$$

- Goal: learn one model per task, $f_1, \dots, f_n$, with small multi-task risk

$$\mathcal{R}^{\mathsf{MT}}(f_1, \dots, f_n) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{R}_i(f_i), \qquad \widehat{\mathcal{R}}^{\mathsf{MT}}(f_1, \dots, f_n) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathcal{R}}_i(f_i).$$

What guarantees can we provide on $\mathcal{R}^{\mathsf{MT}}$? What are principled learning algorithms?

From Single-Task to Multi-Task Learning

Naive solution: control each task separately and combine the bounds

- for each task: $\mathcal{R}_i(f_i) \leq \widehat{\mathcal{R}}_i(f_i) + \mathcal{C}(f_i, m_i)$
- combine:

$$\mathcal{R}^{\mathsf{MT}}(f_1, \dots, f_n) = \widehat{\mathcal{R}}^{\mathsf{MT}}(f_1, \dots, f_n) + \frac{1}{n}\sum_i \mathcal{C}(f_i, m_i)$$

  no benefit from observing more tasks, regardless if tasks are related or not

Naive solution: control each task separately and combine the bounds

- for each task: $\mathcal{R}_i(f_i) \leq \widehat{\mathcal{R}}_i(f_i) + \mathcal{C}(f_i, m_i)$
- combine:

$$\mathcal{R}^{\mathsf{MT}}(f_1, \dots, f_n) = \widehat{\mathcal{R}}^{\mathsf{MT}}(f_1, \dots, f_n) + \frac{1}{n}\sum_i \mathcal{C}(f_i, m_i)$$

  no benefit from observing more tasks, regardless if tasks are related or not

Classic and ongoing research: exploiting that information can be shared between tasks

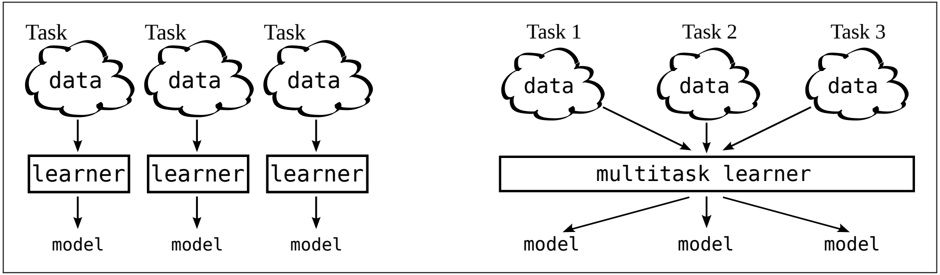
- architectures (what to share and how)
- task relatedness (which tasks should share or not)
- optimization (algorithms, convergence)
- trustworthiness (privacy, fairness, federated learning)
- applications (NLP, Computer Vision, Robotics)
- theory, e.g. **generalization guarantees**

Non-Vacuous Generalization Bounds in Deep Multi-Task Learning

Hossein Zakerinia
(ISTA)

Dorsa Ghobadi
(Sharif U)

[Hossein Zakerinia, Dorsa Ghobadi, Christoph H. Lampert. "From Low Intrinsic Dimensionality to Non-Vacuous Generalization Bounds in Deep Multi-Task Learning". arXiv arXiv:2501.19067 (under review)]



Observation: a multitask learning sees all data at once, it can exploit shared structure, e.g.

- learn one shared feature space and individual "classification heads" inside that space
- learn one prototype model, from which individual models are just minor modifications
- learn a small number of models, for each tasks select a suitable one

Common pattern: some parts are "shared", some parts of "individual"

Theorem (Reminder: Single-Task Generalization Bound)

Let $\mathcal{F}$ be a countable model class and let $E : \mathcal{F} \to \{0,1\}^*$ be a prefix-free encoding of the elements in $\mathcal{F}$. Then, [...] it holds with high probability:

$$\forall f \in \mathcal{F} : \quad \mathcal{R}(f) \lesssim \widehat{\mathcal{R}}(f) + \sqrt{\frac{|E(f)|}{2m}},$$

where $|E(f)| = \log 2 \cdot \mathit{length}(E(f))$.

How to derive a similar result for multi-task learning with information sharing?

Multi-Task Generalization Bound with Explicit Sharing

Theorem (Zakerinia, Ghobadi, Lampert. arXiv:2501.19067)

Let $\mathcal{G}$ be a set of global parameters, and let $E : \mathcal{G} \to \{0,1\}^*$ be an encoder of its elements. For any $G \in \mathcal{G}$, let E_G be an encoder of potentially multiple models. For any $m \in \mathbb{N}$, it holds with high probability over the sampling of the training sets $S_i \sim \mathcal{D}_i^m$ that for all $G \in \mathcal{G}$ and all $f_1, \dots, f_n \in \mathcal{F}$:

$$\mathcal{R}^{MT}(f_1, \dots, f_n) \lesssim \widehat{\mathcal{R}}^{MT}(f_1, \dots, f_n) + \sqrt{\frac{|E(G)| + |E_G(f_1, \dots, f_n)|}{2mn}}$$

Multi-Task Generalization Bound with Explicit Sharing

Theorem (Zakerinia, Ghobadi, Lampert. arXiv:2501.19067)

Let $\mathcal{G}$ be a set of global parameters, and let $E : \mathcal{G} \to \{0,1\}^*$ be an encoder of its elements. For any $G \in \mathcal{G}$, let E_G be an encoder of potentially multiple models. For any $m \in \mathbb{N}$, it holds with high probability over the sampling of the training sets $S_i \sim \mathcal{D}_i^m$ that for all $G \in \mathcal{G}$ and all $f_1, \dots, f_n \in \mathcal{F}$:

$$\mathcal{R}^{MT}(f_1, \dots, f_n) \lesssim \widehat{\mathcal{R}}^{MT}(f_1, \dots, f_n) + \sqrt{\frac{|E(G)| + |E_G(f_1, \dots, f_n)|}{2mn}}$$

- Numerator, $|E(G)| + |E_G(f_1, ..., f_n)|$, exploits shared/task-specific encoding:
  1. identify shared information, G (for "global"), and encode it only once, $E(G)$
     $\to$ could later also be used for future tasks ("meta-learning")
  2. encode task-specific parts, relying on G as side information, $E_G(f_1, \dots, f_n)$
     $\to$ joint encoding can exploit further redundancy, e.g. arithmetic coding
- Denominator, mn, reflects *all available data*

Multi-Task Generalization Bound with Explicit Sharing

$$\mathcal{R}^{\mathsf{MT}}(f_1,\dots,f_n) \lesssim \widehat{\mathcal{R}}^{\mathsf{MT}}(f_1,\dots,f_n) + \sqrt{\frac{|E(G)| + |E_G(f_1,\dots,f_n)|}{2mn}}$$

Multi-task encoder setup allows for a lot of flexibility, e.g.

- $\mathcal{G} = \{\emptyset\}$, $|E(\emptyset)| = 0$, $|E_\emptyset(f_1,\dots,f_n)| = \sum_{i=1}^{n} |E(f_i)| \to$ recover independent learning
- G is a feature extractor, E_G encodes models with those features
- G is a prototype model, E_G encodes differences to prototype
- G is a set of base models, E_G encodes which tasks uses which base model
- G is a subspace of the parameter space, E_G encodes coordinates in subspace
- G is a codebook of values, E_G stores codebook id instead of parameter values

Multi-Task Generalization Bound with Explicit Sharing

$$\mathcal{R}^{\mathsf{MT}}(f_1,\dots,f_n) \lesssim \widehat{\mathcal{R}}^{\mathsf{MT}}(f_1,\dots,f_n) + \sqrt{\frac{|E(G)| + |E_G(f_1,\dots,f_n)|}{2mn}}$$

Multi-task encoder setup allows for a lot of flexibility, e.g.

- $\mathcal{G} = \{\emptyset\}$, $|E(\emptyset)| = 0$, $|E_\emptyset(f_1,\dots,f_n)| = \sum_{i=1}^n |E(f_i)| \to$ recover independent learning
- G is a feature extractor, E_G encodes models with those features
- G is a prototype model, E_G encodes differences to prototype
- G is a set of base models, E_G encodes which tasks uses which base model
- G is a subspace of the parameter space, E_G encodes coordinates in subspace
- G is a codebook of values, E_G stores codebook id instead of parameter values

Application: Non-Vacuous Generalization Bounds for MTL with Deep Network

Goal: learn n deep networks with parameter vectors $\theta_1, \dots, \theta_n \in \mathbb{R}^D$

Learnable Random Subspace Representation based on [Lotfi *et al*. 2022], [Li *et al*. 2018], [Baxter 2000]

- k-dimensional subspace of $\mathbb{R}^D$, parametrized by expansion matrix $Q \in \mathbb{R}^{D\times k}$
- task-specific: coordinates in subspace $\quad \theta_i = Q\alpha_i \quad$ for $\alpha_i \in \mathbb{R}^k$ for $i = 1, \dots, n$
- shared: learning Q itself via $\quad Q = [P_1 v_1, P_2 v_2, \cdots, P_k v_k] \in \mathbb{R}^{D\times k}$
  - $v_1, \dots, v_k \in \mathbb{R}^l$ are learnable vectors
  - $P_1, \dots, P_k \in \mathbb{R}^{D\times l}$ are fixed matrices (i.i.d. unit Gaussian entries)
- learnable parameters: $nk + kl$ total, i.e. $k + \frac{kl}{n}$ per task (instead of D).

Observation:

- in practice, low training error possible even for small value of k, l
- few parameters, compressed with a learnable codebook $\to$ non-vacuous MTL bounds

Table: Necessary representation dimensions to achieve a pre-specified target accuracy for different datasets and model architectures. STL = single task learning, MTL = multitask learning.

Dataset	MNIST SP	MNIST PL	Folktables	Products	split-CIFAR10		split-CIFAR100	
Model	ConvNet	ConvNet	MLP	MLP	ConvNet	ViT	ConvNet	ViT
n / m	30 / 2000	30 / 2000	60 / 900	60 / 2000	100 / 453	30 / 1248	100 / 450	30 / 1250
model dim	21840	21840	11810	13730	121182	5526346	128832	5543716
necessary dim (STL)	400	300	50	50	200	200	1500	550
necessary dim (MTL)	31.6	166.6	10	10	12	26.7	36	100

Table: Generalization guarantees (upper bound on 0/1-test error) for STL and MTL

Dataset	MNIST SP	MNIST PL	Folktables	Products	split-CIFAR10		split-CIFAR100	
Model	ConvNet	ConvNet	MLP	MLP	ConvNet	ViT	ConvNet	ViT
STL	0.61	0.58	0.57	0.33	0.87	0.66	0.99	0.91
MTL (standard)	0.23	0.40	0.39	0.22	0.53	0.32	0.87	0.67
MTL (fast-rate)	0.20	0.35	0.39	0.20	0.53	0.28	0.83	0.66

Fast-Rate Bounds for Multi-Task Learning with Different Sample Sizes

Hossein Zakerinia
(ISTA)

[Hossein Zakerinia, Christoph H. Lampert. "Fast-Rate Bounds for Multi-Task and Meta-Learning with Different Sample Sizes". arXiv:2505.15496 (NeurIPS 2025)]

Remember, how we introduced the multi-task learning setting:

- multiple data distributions $\mathcal{D}_1, \dots, \mathcal{D}_n$
- multiple training sets $S_1, \dots, S_n$ of sizes $m_1, \dots, m_n$
- for simplicity: same input/output sets, same model class, same loss function

For the previous result, we had assumed $m_1 = m_2 = \cdots = m_n$ (balanced MTL).

But: arbitrary $m_1, \dots, m_n$ (unbalanced MTL) is much more relevant in practice.

Theorem (Balanced MTL)

For any $m \in \mathbb{N}$, it holds with high probability over the training sets, $S_i \sim \mathcal{D}_i^m$, that

$$\forall f_1, \ldots, f_n \in \mathcal{F}: \qquad \mathcal{R}^{\mathsf{MT}}(f_1, \ldots, f_n) \lesssim \widehat{\mathcal{R}}^{\mathsf{MT}}(f_1, \ldots, f_n) + \sqrt{\frac{|E(f_1, \ldots, f_n)|}{2mn}}.$$

Theorem (Balanced MTL)

For any $m \in \mathbb{N}$, it holds with high probability over the training sets, $S_i \sim \mathcal{D}_i^m$, that

$$\forall f_1, \dots, f_n \in \mathcal{F}: \qquad \mathcal{R}^{MT}(f_1, \dots, f_n) \lesssim \widehat{\mathcal{R}}^{MT}(f_1, \dots, f_n) + \sqrt{\frac{|E(f_1, \dots, f_n)|}{2mn}}.$$

Deriving an unbalanced analog is straight-forward:

Theorem (Unbalanced MTL)

For any $m_1, \dots, m_n \in \mathbb{N}$, it holds with high probability over the training sets, $S_i \sim \mathcal{D}_i^{m_i}$, that

$$\forall f_1, \dots, f_n \in \mathcal{F}: \qquad \mathcal{R}^{MT}(f_1, \dots, f_n) \lesssim \widehat{\mathcal{R}}^{MT}(f_1, \dots, f_n) + \sqrt{\frac{|E(f_1, \dots, f_n)|}{2\bar{m}n}},$$

where $\bar{m} = (\frac{1}{n}\sum_i \frac{1}{m_i})^{-1}$ is the harmonic mean of the training set sizes, $m_i = |S_i|$.

Harmonic mean makes sense here, e.g.,

- if $m_1 = \dots = m_n = m$, then $\bar{m} = m$, so we recover balanced MTL result,
- if $m_j \to \infty$ for all $j \neq i$, then $\bar{m} \to n m_i$, so $\sqrt{\frac{|E|}{\bar{m}n}} \to \frac{1}{n}\sqrt{\frac{|E|}{m_i}}$, like in single-task learning.

Fast-Rate Bounds for Unbalanced MTL

Theorem (Fast-Rate Bound – Balanced MTL)

For any $m \in \mathbb{N}$, it holds with high probability over the training sets, $S_i \sim \mathcal{D}_i^m$, that

$$\forall f_1, \dots, f_n \in \mathcal{F} : \qquad \mathrm{kl}\left(\widehat{\mathcal{R}}^{MT}(f_1, \dots, f_n) \,\|\, \mathcal{R}^{MT}(f_1, \dots, f_n)\right) \lesssim \frac{|E(f_1, \dots, f_n)|}{mn}.$$

What's an unbalanced analog?

Fast-Rate Bounds for Unbalanced MTL

Theorem (Fast-Rate Bound – Balanced MTL)

For any $m \in \mathbb{N}$, it holds with high probability over the training sets, $S_i \sim \mathcal{D}_i^m$, that

$$\forall f_1, \ldots, f_n \in \mathcal{F}: \qquad \mathrm{kl}\left(\widehat{\mathcal{R}}^{MT}(f_1, \ldots, f_n) \,\|\, \mathcal{R}^{MT}(f_1, \ldots, f_n)\right) \lesssim \frac{|E(f_1, \ldots, f_n)|}{mn}.$$

What's an unbalanced analog?

Theorem (Fast-Rate Bound – Unbalanced MTL)

For any $m_1, \ldots, m_n \in \mathbb{N}$, it holds with high probability over the training sets, $S_i \sim \mathcal{D}_i^{m_i}$, that

$$\forall f_1, \ldots, f_n \in \mathcal{F}: \qquad \mathrm{kl}\left(\widehat{\mathcal{R}}^{MT}(f_1, \ldots, f_n) \,\|\, \mathcal{R}^{MT}(f_1, \ldots, f_n)\right) \lesssim \frac{|E(f_1, \ldots, f_n)|}{\underline{m}n},$$

where $\underline{m} = \mathbf{min}_i\, m_i$ is the smallest training set size.

Fast-Rate Bounds for Unbalanced MTL

Theorem (Fast-Rate Bound – Balanced MTL)

For any $m \in \mathbb{N}$, it holds with high probability over the training sets, $S_i \sim \mathcal{D}_i^m$, that

$$\forall f_1, \ldots, f_n \in \mathcal{F}: \qquad \mathrm{kl}\left(\widehat{\mathcal{R}}^{MT}(f_1, \ldots, f_n) \,\|\, \mathcal{R}^{MT}(f_1, \ldots, f_n)\right) \lesssim \frac{|E(f_1, \ldots, f_n)|}{mn}.$$

What's an unbalanced analog?

Theorem (Fast-Rate Bound – Unbalanced MTL)

For any $m_1, \ldots, m_n \in \mathbb{N}$, it holds with high probability over the training sets, $S_i \sim \mathcal{D}_i^{m_i}$, that

$$\forall f_1, \ldots, f_n \in \mathcal{F}: \qquad \mathrm{kl}\left(\widehat{\mathcal{R}}^{MT}(f_1, \ldots, f_n) \,\|\, \mathcal{R}^{MT}(f_1, \ldots, f_n)\right) \lesssim \frac{|E(f_1, \ldots, f_n)|}{\underline{m}n},$$

where $\underline{m} = \mathbf{min}_i\, m_i$ is the smallest training set size. ← *that can't be right?!?*

Fast-Rate Bounds for Unbalanced MTL

Theorem (Fast-Rate Bound – Balanced MTL)

For any $m \in \mathbb{N}$, it holds with high probability over the training sets, $S_i \sim \mathcal{D}_i^m$, that

$$\forall f_1, \dots, f_n \in \mathcal{F}: \qquad \mathrm{kl}\left(\widehat{\mathcal{R}}^{MT}(f_1, \dots, f_n) \,\|\, \mathcal{R}^{MT}(f_1, \dots, f_n)\right) \lesssim \frac{|E(f_1, \dots, f_n)|}{mn}.$$

What's an unbalanced analog?

Theorem (Fast-Rate Bound – Unbalanced MTL)

For any $m_1, \dots, m_n \in \mathbb{N}$, it holds with high probability over the training sets, $S_i \sim \mathcal{D}_i^{m_i}$, that

$$\forall f_1, \dots, f_n \in \mathcal{F}: \qquad \mathrm{kl}\left(\widehat{\mathcal{R}}^{MT}(f_1, \dots, f_n) \,\|\, \mathcal{R}^{MT}(f_1, \dots, f_n)\right) \lesssim \frac{|E(f_1, \dots, f_n)|}{\underline{m}n},$$

where $\underline{m} = \mathbf{min}_i\, m_i$ is the smallest training set size. ← *that can't be right?!?*

- if $m_1 = \cdots = m_n = m$, then $\underline{m} = m$, so we recover balanced MTL result,
- if $m_j \to \infty$ for all $j \neq i$, then $\underline{m} = m_i$, so no gain at all from other tasks.

Proof sketch for balanced case, $m_1 = \cdots = m_n = m$:

1) For any $(f_1, \ldots, f_n)$: control kl-term by moment-generating function:

$$\Pr\left\{ \mathrm{kl}(\widehat{\mathcal{R}}^{\mathsf{MT}} | \mathcal{R}^{\mathsf{MT}}) \geq t \right\} = \Pr\left\{ e^{mn\,\mathrm{kl}(\widehat{\mathcal{R}}^{\mathsf{MT}} | \mathcal{R}^{\mathsf{MT}})} \geq e^{mnt} \right\} \lesssim \frac{\mathbb{E}[e^{mn\,\mathrm{kl}(\widehat{\mathcal{R}}^{\mathsf{MT}} | \mathcal{R}^{\mathsf{MT}})}]}{e^{mnt}}.$$

2) derive that $\mathbb{E}[e^{mn\,\mathrm{kl}(\widehat{\mathcal{R}}^{\mathsf{MT}} | \mathcal{R}^{\mathsf{MT}})}] \leq 2\sqrt{mn}$ using

Theorem (Maurer, 2004)

For any $\mu \in (0,1)$, let $Z_{i,j} \overset{i.i.d.}{\sim} \mathrm{Bernoulli}(\mu)$ for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$. Set $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m}\sum_{j=1}^{m} Z_{i,j}$ as the average of their averages. Then it holds that

$$\mathbb{E}\left[e^{mn\,\mathrm{kl}(\hat{\mu}|\mu)}\right] \leq \sqrt{2mn}.$$

3) result follows by weighted union bound using Kraft-McMillan's inequality for prefix codes.

Fast-Rate Bounds for Unbalanced MTL – Why the bad rate?

Proof sketch for balanced case, $m_1 = \cdots = m_n = m$:

1) For any $(f_1, \ldots, f_n)$: control kl-term by moment-generating function:

$$\Pr\left\{ \mathrm{kl}(\widehat{\mathcal{R}}^{\mathsf{MT}} | \mathcal{R}^{\mathsf{MT}}) \geq t \right\} = \Pr\left\{ e^{mn\,\mathrm{kl}(\widehat{\mathcal{R}}^{\mathsf{MT}} | \mathcal{R}^{\mathsf{MT}})} \geq e^{mnt} \right\} \lesssim \frac{\mathbb{E}[e^{mn\,\mathrm{kl}(\widehat{\mathcal{R}}^{\mathsf{MT}} | \mathcal{R}^{\mathsf{MT}})}]}{e^{mnt}}.$$

2) derive that $\mathbb{E}[e^{mn\,\mathrm{kl}(\widehat{\mathcal{R}}^{\mathsf{MT}} | \mathcal{R}^{\mathsf{MT}})}] \leq 2\sqrt{mn}$ using

Theorem (Maurer, 2004)

For any $\mu \in (0,1)$, let $Z_{i,j} \overset{i.i.d.}{\sim} \mathrm{Bernoulli}(\mu)$ for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$. Set $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} Z_{i,j}$ as the average of their averages. Then it holds that

$$\mathbb{E}\left[e^{mn\,\mathrm{kl}(\hat{\mu}|\mu)}\right] \leq \sqrt{2mn}.$$

3) result follows by weighted union bound using Kraft-McMillan's inequality for prefix codes.

Unbalanced case: step 2) fails!

Lemma

For any $\mu \in (0,1)$, *let* $Z_{i,j} \overset{i.i.d.}{\sim} \mathrm{Bernoulli}(\mu)$ *for* $i \in \{1,\dots,n\}$ *and* $j \in \{1,\dots,m_i\}$. *Set* $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i} Z_{i,j}$ *as the average of their averages and write* $M_\mu(\lambda) = \mathbb{E}\big[e^{n\lambda\,\mathrm{kl}(\hat{\mu}|\mu)}\big]$.
Then, if $\lambda > \underline{m} = \mathbf{min}_i\, m_i$, *it holds that*

$$\sup_{0<\mu<1} M_\mu(\lambda) \;=\; +\infty.$$

In particular, no upper bound on $M_\mu(\lambda)$ *exists that depends only on* n *and the* m_i.

Lemma

For any $\mu \in (0,1)$, *let* $Z_{i,j} \overset{i.i.d.}{\sim} \mathrm{Bernoulli}(\mu)$ *for* $i \in \{1,\dots,n\}$ *and* $j \in \{1,\dots,m_i\}$. *Set* $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i} Z_{i,j}$ *as the average of their averages and write* $M_\mu(\lambda) = \mathbb{E}\big[e^{n\lambda\,\mathrm{kl}(\hat{\mu}|\mu)}\big]$.
Then, if $\lambda > \underline{m} = \mathbf{min}_i\, m_i$, *it holds that*

$$\sup_{0<\mu<1} M_\mu(\lambda) = +\infty.$$

In particular, no upper bound on $M_\mu(\lambda)$ *exists that depends only on* n *and the* m_i.

Two suggested fixed:

- re-weight the kl-terms
- re-weight the sample contributions

Theorem (Fast-Rate Bound for Task-Centric MTL [Zakerinia, Lampert. arXiv 2505.15496])

In the setting above with task sizes $m_1, \dots, m_n$, set $M = \sum_i m_i$. Then, it holds with high probability over the training sets $S_i \sim \mathcal{D}_i^{m_i}$ that

$$\forall f_1, \dots, f_n \in \mathcal{F}: \qquad \sum_{i=1}^{n} \frac{m_i}{M} \operatorname{kl}(\widehat{\mathcal{R}}_i(f_i) \,\|\, \mathcal{R}_i(f_i)) \lesssim \frac{|E(f_1, \dots, f_n)|}{M}$$

Observation: we can recover (up to $\log$-terms)

- standard-rate bound with $\frac{1}{\bar{m}n}$ (Pinsker's ineq., Cauchy-Schwartz ineq.)
- the balanced fast-rate bound with $\frac{1}{mn}$, if actually $m_1 = \dots = m_n = m$ (Jensen's).

Observation: if we multiply both sides by M, r.h.s. is a constant.

- if any m_i increases, its $\operatorname{kl}(\widehat{\mathcal{R}}_i(f_i) \,\|\, \mathcal{R}_i(f_i))$ decreases at least proportionally
  $\to$ same rate as for single-tasks, but better constants possible by information sharing

Fast-Rate Bounds for Unbalanced Multi-Task Learning – Sample-Centric

For datasets $S_i = \{(x_{i,1}, y_{i,1}), \dots, (x_{i,m_i}, y_{i,m_i})\}$, let $M := \sum_i m_i$. Define the *sample-centric* expected and empirical risks as

$$\mathcal{R}^{\text{MT-S}}(f_1, \dots, f_n) = \sum_{i=1}^{n} \frac{m_i}{M} \mathcal{R}_i(f_i) = \sum_{i=1}^{n} \frac{m_i}{M} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i} \ell(\,y, f(x)\,),$$

$$\widehat{\mathcal{R}}^{\text{MT-S}}(f_1, \dots, f_n) = \sum_{i=1}^{n} \frac{m_i}{M} \widehat{\mathcal{R}}_i(f_i) = \frac{1}{M} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \ell(\,y_{i,j}, f_i(x_{i,j}\,)).$$

Theorem (Fast-Rate Bound for Sample-Centric MTL [Zakerinia, Lampert. arXiv 2505.15496])

In the setting above with task sizes $m_1, \dots, m_n$, set $M = \sum_i m_i$. Then, it holds with high probability over the training sets $S_i \sim \mathcal{D}_i^{m_i}$ that

$$\forall f_1, \dots, f_n \in \mathcal{F}: \qquad \mathrm{kl}\left(\widehat{\mathcal{R}}^{\textit{MT-S}}(f_1, \dots, f_n) \,\|\, \mathcal{R}^{\textit{MT-S}}(f_1, \dots, f_n)\right) \lesssim \frac{|E(f_1, \dots, f_n)|}{M}$$

Observation:

- for $m_1 = \dots = m_n$, identical to previous setting, same guarantees

Task-centric		
Dataset	CIFAR10	CIFAR100
Standard rate	0.31	0.59
Fast-rate with $m_{\min}$	0.35	0.62
Fast-rate (unbalanced)	0.27	0.59

Sample-centric		
Dataset	CIFAR10	CIFAR100
Standard rate	0.30	0.59
Fast-rate	0.26	0.59

Table: Generalization bounds for low-rank parametrized deep networks on split-CIFAR.

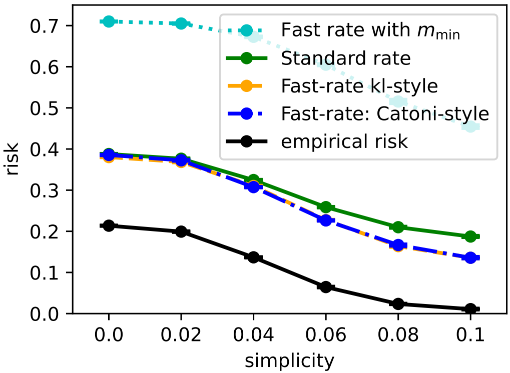


Figure: Generalization bounds of task-centric risk for linear classifiers on MDPR dataset ($n = 953$ tasks; $102 \leq m_i \leq 22530$).

Summary: Generalization Guaranteed for Multi-Task Learning

We presented compression-based generalization bounds for multi-task learning (really, in the background are PAC-Bayesian bounds)

- first non-vacuous guarantees for MTL with deep networks,
- first fast-rate bounds for unbalanced MTL.

Open Questions

Practice:

- How to model information sharing between tasks to simultaneously achieve high accuracy and strong generalization guarantees?

Theory:

- What's the best possible bound on $\mathrm{kl}(\widehat{\mathcal{R}}^{\mathsf{MT}} \| \mathcal{R}^{\mathsf{MT}})$ in the unbalanced setting?

Thank you!

We're hiring: `chl@ist.ac.at`