Minimax Multi-Task Learning

Nishant A. Mehta, Dongryeol Lee, Alexander G. Gray
niche@cc.gatech.edu, drselee@gmail.com, agray@cc.gatech.edu
College of Computing, Georgia Insitute of Technology, Atlanta, GA 30332, USA

1 Introduction

A multi-task learning (MTL) algorithm learns an inductive bias to learn several tasks together. MTL is incredibly pervasive in machine learning: it has natural connections to random effects models [5]; user preference prediction (including collaborative filtering) can be framed as MTL [6]; multi-class classification admits the popular *one-vs-all* and *all-pairs* MTL reductions; and MTL admits provably good learning in settings where single-task learning is hopeless [3, 4]. But if we see a random set of tasks today, which of the tasks will matter tomorrow? Not knowing the challenges nature will pose in the future, it is wise to mitigate the worst case by ensuring a minimum proficiency on each task.

Consider a simple learning scenario: A music preference prediction company is in the business of predicting what 5-star ratings different users would assign to songs. At training time, the company learns a shared representation for predicting the users' song ratings by pooling together the company's limited data on each user's preferences. Given this learned representation, a separate predictor for each user can be trained very quickly. At test time, the environment selects a user according to some (possibly randomized) rule and solicits from the company a prediction of that user's preference for a particular song. The environment may also ask for predictions about new users, described by a few ratings each, and so the company must leverage its existing representation to rapidly learn new predictors and produce ratings for these new users.

Performing MTL well requires simultaneous navigation of several trade-offs. Classically, MTL minimizes a regularized sum of the empirical risks of a set of tasks, thereby implicitly assuming that the learner will be tested on test tasks drawn uniformly at random from the empirical task distribution of the training tasks. By considering the various trade-offs in MTL, it becomes apparent that classical MTL may not be ideal:

- There is a model order selection trade-off to motivate whether MTL should be done at all: Simpler, single-task learning models are ideal when the tasks are unrelated, while more complex MTL models are preferable when the tasks are related.
- Since MTL couples each task's model via a shared parameter, there is a trade-off between minimizing the risks of different tasks. One goal is to minimize the task-wise mean of the true risks. Another is to minimize the task-wise maximum of the true risks. A company might want the most accurate predictions on average, but it also may want to avoid doing very badly on any single task to minimize negative feedback and a potential loss of business. While at training time classical MTL commits to a fixed distribution over users, at test time the user distribution could change or the sequence of users for which ratings are elicited could be adversarial.
- *Teleology vs Deontology:* Whereas utilitarianism would seek to minimizing the average prediction error, typically at the expense of some locally egregious outcomes, minimizing the task-wise maximum of the prediction errors is the most fair to all tasks (or people).
- There is a second model order selection trade-off: Some tasks may overfit more than others, depending on the choice of the shared and model-specific parameters. Minimizing ℓ_p norms of the empirical risk, for $p > 1$, places more emphasis on tasks for which the model currently has high empirical risk. This can be a form of early stopping; in the extreme case of $p = \infty$, learning stops for tasks whose empirical risk is below the task-wise maximum of the empirical risks.

This work introduces *minimax multi-task learning* as a response to the above. We also cast a spectrum of multi-task learning. At one end of the spectrum lies minimax MTL; departing from this point progressively relaxes the "hardness" of the maximum until full relaxation reaches the second endpoint and recovers classical MTL. We further sculpt a generalized loss-compositional paradigm for MTL which includes this spectrum and several other new MTL formulations. This paradigm equally applies to *learning to learn* (LTL), where the goal is to learn a hypothesis space from a set of training tasks such that this representation admits good hypotheses on future tasks.

Theoretically, we show (via Theorem 1) the following: If it is possible to obtain maximum empirical risk across a set of training tasks below some level γ, then it is likely that the maximum true risk obtained by the learner on a new task is bounded by roughly γ. Hence, if the goal is to minimize the worst case outcome over new tasks, theory suggests minimizing the maximum of the empirical risks across the training tasks rather than their mean. Empirical evaluations of several MTL formulations from the new paradigm are promising, but for space reasons we cannot present these results here.

2 Minimax multi-task learning

We begin with the MTL and LTL setups. In this work, each example (x, y) will live in $\mathcal{X} \times \mathcal{Y}$ for input instance x and label y. Typically, $\mathcal{X}$ is a subset of $\mathbb{R}^n$ while $\mathcal{Y}$ is $\{-1, 1\}$ or a compact subset of $\mathbb{R}$. Define a loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$. For simplicity, this work considers ℓ_2 loss (squared loss) $\ell(y', y) = (y' - y)^2$ for regression and hinge loss $\ell(y', y) = \max\{0, 1 - y'y\}$ for classification.

MTL and LTL often are framed as applying an inductive bias to learn a common hypothesis space, selected from a fixed family of hypothesis spaces, and thereafter learning from this hypothesis space a hypothesis for each task observed at training time. It will be useful to formalize the various sets and elements present in the preceding statement. Let $\mathbb{H}$ be a family of hypothesis spaces. Any hypothesis space $\mathcal{H} \in \mathbb{H}$ itself is a set of hypotheses; each hypothesis $h \in \mathcal{H}$ is a map $h : \mathcal{X} \to \mathbb{R}$.

Learning to learn In LTL, the goal is to achieve inductive transfer to learn the best $\mathcal{H}$ from $\mathbb{H}$. Unlike in MTL, there is an *environment* of tasks: an unknown probability measure Q over a space of task probability measures $\mathcal{P}$. The goal is to find the optimal representation via the objective

$$\inf_{\mathcal{H} \in \mathbb{H}} \mathsf{E}_{P \sim Q} \inf_{h \in \mathcal{H}} \mathsf{E}_{(x,y) \sim P} \ell(y, h(x)). \tag{1}$$

In practice, T (unobservable) training task probability measures $P_1, \ldots, P_T \in \mathcal{P}$ are drawn iid from Q, and from each task t a set of m examples are drawn iid from P_t.

Multi-task learning Whereas in learning to learn there is a distribution over tasks, in multi-task learning there is a fixed, finite set of tasks indexed by $[T] := \{1, \ldots, T\}$. Each task $t \in [T]$ is coupled with a fixed but unknown probability measure P_t. Classically, the goal of MTL is to minimize the expected loss at test time under the uniform distribution on $[T]$:

$$\inf_{\mathcal{H} \in \mathbb{H}} \frac{1}{T} \sum_{t \in [T]} \inf_{h \in \mathcal{H}} \mathsf{E}_{(x,y) \sim P_t} \ell(y, h(x)). \tag{2}$$

Notably, this objective is equivalent to (1) when Q is the uniform distribution on $\{P_1, \ldots, P_T\}$. In terms of the data generation model, MTL differs from LTL since the tasks are fixed; however, just as in LTL, from each task t a set of m examples are drawn iid from P_t .

Minimax MTL A natural generalization of classical MTL arises by introducing a prior distribution π over the set of tasks $[T]$. Given π, the (idealized) objective of this generalized MTL is

$$\inf_{\mathcal{H} \in \mathbb{H}} \mathsf{E}_{t \sim \pi} \inf_{h \in \mathcal{H}} \mathsf{E}_{(x,y) \sim P_t} \ell(y, h(x)), \tag{3}$$

given only the training data $\{(x_{t,1}, y_{t,1}), \ldots, (x_{t,m}, y_{t,m})\}_{t \in [T]}$. The classical MTL objective (2) equals (3) when π is taken to be the uniform prior over $[T]$. We argue that in many instances, that which is most relevant to minimize is not the expected error under a uniform distribution over tasks, or even any pre-specified π, but rather the expected error for the worst π. We propose to minimize the maximum error over tasks under an adversarial choice of π, yielding the objective

$$\inf_{\mathcal{H} \in \mathbb{H}} \sup_{\pi \in (T-1)\text{-simplex}} \mathsf{E}_{t \sim \pi} \inf_{h \in \mathcal{H}} \mathsf{E}_{(x,y) \sim P_t} \ell(y, h(x))$$

As the supremum is attained at an extreme point of the simplex, this objective is equivalent to

$$\inf_{\mathcal{H}\in\mathbb{H}} \max_{t\in[T]} \inf_{h\in\mathcal{H}} \mathsf{E}_{(x,y)\sim P_t}\ell(y,h(x)). \tag{4}$$

In practice, we approximate the true objective by using the (regularized) empirical objective: $\inf_{\mathcal{H}\in\mathbb{H}} \max_{t\in[T]} \inf_{h\in\mathcal{H}} \sum_{i=1}^m \ell(y_{t,i},h(x_{t,i}))$.

We can motivate minimax MTL theoretically by showing that the worst-case performance on future tasks likely will not be much higher than the maximum of the empirical risks for the training tasks.

LTL bound for the maximum risk Let $P^{(1)},\dots,P^{(T)}$ be probability measures drawn iid from Q, and for $t\in[T]$ let $\mathbf{z}^{(t)}$ be an m-sample (a sample of m points) from $P^{(t)}$ with corresponding empirical measure $P_m^{(t)}$. If P is a probability measure then $P\ell\circ h := \mathsf{E}\ell(y,h(x))$; similarly, if P_m is an empirical measure, then $P_m\ell\circ h := \frac{1}{m}\sum_{i=1}^m \ell(y_i,h(x_i))$. Our focus is the learning to learn setting with a minimax lens: when one learns a representation $\mathcal{H}\in\mathbb{H}$ from multiple training tasks and observes maximum empirical risk γ, we seek a guarantee that $\mathcal{H}$'s true risk on a newly drawn test task will be bounded by roughly γ. Such a goal is in striking contrast to the classical emphasis of LTL, where the goal is to obtain bounds on $\mathcal{H}$'s expected true risk. Using $\mathcal{H}$'s expected true risk and Markov's inequality, Baxter [3, the display prior to (25)] showed that the probability that $\mathcal{H}$'s true risk on a newly drawn test task is above some level γ decays as the expected true risk over γ:

$$\mathsf{Pr}\left\{\inf_{h\in\mathcal{H}} P\ell\circ h \geq \gamma\right\} \leq \frac{1}{\gamma T}\left(\sum_{t\in[T]} P_m^{(t)}\ell\circ h_t + \varepsilon\right), \tag{5}$$

where the size of ε is controlled by T, m, and the complexities of certain spaces.

The expected true risk is not of primary interest for controlling the tail of the (random) true risk, and a more direct approach yields a much better bound. In this short paper we restrict the space of representations $\mathbb{H}$ to be finite with cardinality $\mathcal{C}$; in this case, the analysis is particularly simple and illuminates the idea for proving the general case. The next theorem is the main result of this section:

Theorem 1. *Let $|\mathbb{H}| = \mathcal{C}$, and let the loss ℓ be L-Lipschitz in its second argument and bounded by B. Suppose T tasks $P^{(1)},\dots,P^{(T)}$ are drawn iid from Q and from each task $P^{(t)}$ an iid m-sample $\mathbf{z}^{(t)}$ is drawn. Suppose there exists $\mathcal{H}\in\mathbb{H}$ such that all $t\in[T]$ satisfy $\min_{h\in\mathcal{H}} P_m^{(t)}\ell\circ h \leq \gamma$. Let P be newly drawn probability measure from Q. Let $\hat{h}$ be the empirical risk minimizer over the test m-sample. With probability at least $1-\delta$ with respect to the random draw of the T tasks and their T corresponding m-samples:*

$$\mathsf{Pr}\left\{P\ell\circ\hat{h} > \gamma + \frac{1}{T} + 2L\max_{\mathcal{H}\in\mathbb{H}}\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{8\log\frac{4}{\delta}}{m}}\right\} \leq \frac{\log\frac{2\mathcal{C}}{\delta} + \log\lceil B\rceil + \log(T+1)}{T}. \tag{6}$$

In the above, $\mathcal{R}_m(\mathcal{H})$ is the Rademacher complexity of $\mathcal{H}$ (cf. [2]). Critically, in (6) the probability of observing a task with high true risk decays with T, whereas in (5) the decay is independent of T. Hence, when the goal is to minimize the probability of bad performance on future tasks uniformly, this theorem motivates minimizing the *maximum* of the empirical risks as opposed to their mean.

3 A generalized loss-compositional paradigm for MTL

Given a set of T tasks, we represent the empirical risk for hypothesis $h_t\in\mathcal{H}$ ($\in\mathbb{H}$) on task $t\in[T]$ as $\hat{\ell}_t(h_t) := \sum_{i=1}^m \ell(y_{t,i},h_t(x_{t,i}))$. Additionally define a set of hypotheses for multiple tasks $\mathbf{h} := (h_1,\dots,h_T)\in\mathcal{H}^T$ and the vector of empirical risks $\hat{\boldsymbol{\ell}}(\mathbf{h}) := (\hat{\ell}_1(h_1),\dots,\hat{\ell}_T(h_T))$.

With this notation set, the proposed loss-compositional paradigm encompasses any regularized minimization of a (typically convex) function $\phi:\mathbb{R}_+^T\to\mathbb{R}_+$ of the empirical risks:

$$\inf_{\mathcal{H}\in\mathbb{H}} \inf_{\mathbf{h}\in\mathcal{H}^T} \phi\big(\hat{\boldsymbol{\ell}}(\mathbf{h})\big) + \Omega\big((\mathcal{H},\mathbf{h})\big), \tag{7}$$

where $\Omega(\cdot):\mathbb{H}\times\cup_{\mathcal{H}\in\mathbb{H}}\mathcal{H}^T\to\mathbb{R}_+$ is a regularizer.

$\boldsymbol{\ell}_\mathbf{p}$ MTL One notable specialization that is still quite general is the case when ϕ is an ℓ_p-norm, yielding ℓ_p *MTL*. This subfamily encompasses classical MTL and many new MTL formulations:

Classical MTL as ℓ_1 *MTL*: $$\inf_{\mathcal{H}\in\mathbb{H}} \inf_{\mathbf{h}\in\mathcal{H}^T} \frac{1}{T}\sum_{t\in[T]} \hat{\ell}(h_t)+\Omega\big((\mathcal{H},\mathbf{h})\big) \equiv \inf_{\mathcal{H}\in\mathbb{H}} \inf_{\mathbf{h}\in\mathcal{H}^T} \frac{1}{T}\|\hat{\boldsymbol{\ell}}(\mathbf{h})\|_1+\Omega\big((\mathcal{H},\mathbf{h})\big).$$

Minimax MTL as ℓ_∞ *MTL*: $$\inf_{\mathcal{H}\in\mathbb{H}} \inf_{\mathbf{h}\in\mathcal{H}^T} \max_{t\in[T]} \hat{\ell}(h_t) + \Omega\big((\mathcal{H},\mathbf{h})\big) \equiv \inf_{\mathcal{H}\in\mathbb{H}} \inf_{\mathbf{h}\in\mathcal{H}^T} \|\hat{\boldsymbol{\ell}}(\mathbf{h})\|_\infty + \Omega\big((\mathcal{H},\mathbf{h})\big).$$

ℓ_2 *MTL* (New) : $$\inf_{\mathcal{H}\in\mathbb{H}} \inf_{\mathbf{h}\in\mathcal{H}^T} \Big(\frac{1}{T}\sum_{t\in[T]} \big(\hat{\ell}(h_t)\big)^2\Big)^{1/2}+\Omega\big((\mathcal{H},\mathbf{h})\big) \equiv \inf_{\mathcal{H}\in\mathbb{H}} \inf_{\mathbf{h}\in\mathcal{H}^T} \frac{1}{\sqrt{T}}\|\hat{\boldsymbol{\ell}}(\mathbf{h})\|_2+\Omega\big((\mathcal{H},\mathbf{h})\big).$$

We now see a continuum of relaxations of minimax MTL that are in the paradigm but not in ℓ_p MTL.

α-minimax MTL Minimizing the maximum loss can be problematic since the max is not robust to cases when a small fraction of the tasks are much harder than the other tasks. Consider the case when the empirical risk for each task in this small fraction cannot be reduced below some u. Rather than rigidly minimizing the maximum loss, a more robust alternative is to minimize the maximize loss in a soft way. The idea is to ensure that most tasks have low empirical risk, but a small fraction of tasks are permitted a higher loss. We formalize this as α*-minimax MTL*, via the relaxed objective:

$$\underset{\mathcal{H}\in\mathbb{H},\mathbf{h}\in\mathcal{H}^T}{\text{minimize}} \quad \min_{b\geq 0}\Big\{b + \frac{1}{\alpha}\sum_{t\in[T]} \max\{0, \hat{\ell}_t(h_t) - b\}\Big\} + \Omega\big((\mathcal{H},\mathbf{h})\big). \tag{8}$$

In the above, ϕ from the loss-compositional paradigm (7) is a variational function of the empirical risks vector. The above optimization problem is equivalent to the perhaps more intuitive problem:

$$\underset{\mathcal{H}\in\mathbb{H},\mathbf{h}\in\mathcal{H}^T,b\geq 0,\boldsymbol{\xi}\geq 0}{\text{minimize}} \quad b + \frac{1}{\alpha}\sum_{t\in[T]} \xi_t + \Omega\big((\mathcal{H},\mathbf{h})\big) \qquad \text{subject to} \quad \hat{\ell}_t(h_t) \leq b + \xi_t,\ t\in[T]. \tag{9}$$

Here, b plays the role of the relaxed maximum, and each ξ_t's deviation from zero indicates the deviation from the (loosely enforced) maximum. We expect $\boldsymbol{\xi}$ to be sparse. To help understand how α affects the learning problem, let us consider a few cases:

(1) When $\alpha > T$, the optimal value of b is zero, and the problem is equivalent to classical MTL. To see this, note that for a given candidate solution with $b > 0$ the objective always can be reduced by reducing b by some ε and increasing each ξ_t by the same ε.

(2) Suppose one task is much harder than all the other tasks (e.g. an outlier task), and its empirical risk is separated from the maximum empirical risk of the other tasks by ρ. Let $1 < \alpha < 2$; now, at the optimal hard maximum solution (where $\boldsymbol{\xi} = \mathbf{0}$), the objective can be reduced by increasing one of the ξ_t's by ρ and decreasing b by ρ. Thus, the objective can focus on minimizing the maximum risk of the set of $T-1$ easier tasks. In this special setting, this argument can be extended to the more general case $k < \alpha < k+1$ and k outlier tasks, for $k\in[T]$.

(3) As α approaches 0, we recover the hard maximum case of minimax MTL.

An interesting choice of α is $\alpha = 2/(\lceil 0.1T + 0.5\rceil^{-1} + \lceil 0.1T + 1.5\rceil^{-1})$ i.e. the harmonic mean of $\lceil 0.1T + 0.5\rceil$ and $\lceil 0.1T + 1.5\rceil$. The reason for this choice is that in the idealized case (2) above, for large T this setting of α makes the relaxed maximum consider all but the hardest 10% of the tasks.

Models Let us see how a specific model, convex multi-task feature learning [1], fits into this framework. This model minimizes the task-wise average loss with the trace norm penalty:

$$\min_W \sum_t \sum_{i=1}^m \ell(y_{t,i}, \langle W_t, x_{t,i}\rangle) + \lambda\|W\|_{\mathrm{tr}}, \tag{10}$$

where $\|\cdot\|_{\mathrm{tr}} : W \mapsto \sum_i \sigma_i(W)$ is the trace norm. In the new paradigm, $\mathbb{H}$ is a set where each element is a k-dimensional subspace of linear estimators (for $k \ll d$). Each $h_t = W_t$ in some $\mathcal{H}\in\mathbb{H}$ lives in $\mathcal{H}$'s corresponding low-dimensional subspace. Also, $\hat{\ell}_t(h_t) = \frac{1}{m}\sum_{i=1}^m \ell\big(y_{t,i}, \langle h_t, x_{t,i}\rangle\big)$.

4 Discussion

We have established a spectrum of formulations for MTL which recovers as special cases classical MTL and the newly formulated minimax MTL. In between these extreme points lies a continuum of relaxed minimax MTL formulations. More generally, we introduced a loss-compositional paradigm that operates on the vector of empirical risks, inducing the additional ℓ_p MTL paradigms. All the minimax or α-minimax MTL formulations exhibit a built-in safeguard against overfitting in the case of learning with a model that is very complex relative to the available data.

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[2] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.

[3] J. Baxter. A model of inductive bias learning. *JAIR*, 12(1):149–198, 2000.

[4] A. Maurer. Transfer bounds for linear feature learning. *Machine learning*, 75(3):327–350, 2009.

[5] K. Yu, J. Lafferty, S. Zhu, and Y. Gong. Large-scale collaborative prediction using a nonparametric random effects model. In *ICML*, pages 1185–1192. ACM, 2009.

[6] L. Zhang, D. Agarwal, and B.C. Chen. Generalizing matrix factorization through flexible regression priors. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 13–20. ACM, 2011.