Minimax Multi-Task Learning and a Generalized Loss-Compositional Paradigm for MTL

Nishant A. Mehta[†]**, Dongryeol Lee**[*]**, Alexander G. Gray**[†]
niche@cc.gatech.edu, drselee@gmail.com, agray@cc.gatech.edu
[†] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA
[*] GE Global Research, Niskayuna, NY 12309, USA

Abstract

Since its inception, the modus operandi of multi-task learning (MTL) has been to minimize the task-wise mean of the empirical risks. We introduce a generalized loss-compositional paradigm for MTL that includes a spectrum of formulations as a subfamily. One endpoint of this spectrum is minimax MTL: a new MTL formulation that minimizes the maximum of the tasks' empirical risks. Via a certain relaxation of minimax MTL, we obtain a continuum of MTL formulations spanning minimax MTL and classical MTL. The full paradigm itself is loss-compositional, operating on the vector of empirical risks. It incorporates minimax MTL, its relaxations, and many new MTL formulations as special cases. We show theoretically that minimax MTL tends to avoid worst case outcomes on newly drawn test tasks in the learning to learn (LTL) test setting. The results of several MTL formulations on synthetic and real problems in the MTL and LTL test settings are encouraging.

1 Introduction

The essence of machine learning is to exploit what we observe in order to form accurate predictors of what we cannot. 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 [15]; user preference prediction (including collaborative filtering) can be framed as MTL [16]; 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 [4, 12]. But if we see examples from a random set of tasks today, which of these tasks will matter tomorrow? Not knowing in the present what challenges nature has in store for the future, a sensible strategy is to mitigate the worst case by ensuring some 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 draws 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.

Classically, multi-task learning has sought to minimize the (regularized) sum of the empirical risks over a set of tasks. In this way, classical MTL implicitly assumes that once the learner has been trained, it will be tested on test tasks drawn uniformly at random from the empirical task distribution of the training tasks. Notably, there are several reasons why classical MTL may not be ideal:

---

[*]Work completed while at Georgia Institute of Technology

- While at training time the usual flavor of MTL commits to a fixed distribution over users (typically either uniform or proportional to the number of ratings available for each user), at test time there is no guarantee what user distribution we will encounter. In fact, there may not exist any fixed user distribution: the sequence of users for which ratings are elicited could be adversarial.

- Even in the case when the distribution over tasks is not adversarial, it may be in the interest of the music preference prediction company to guarantee some minimum level of accuracy per user in order to minimize negative feedback and a potential loss of business, rather than maximizing the mean level of accuracy over all users.

- Whereas minimizing the average prediction error is very much a teleological endeavor, typically at the expense of some locally egregious outcomes, minimizing the worst-case prediction error respects a notion of fairness to all tasks (or people).

This work introduces *minimax multi-task learning* as a response to the above scenario.[1] In addition, we cast a spectrum of multi-task learning. At one end of the spectrum lies minimax MTL, and 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 the problem of *learning to learn* (LTL), in which the goal is to learn a hypothesis space from a set of training tasks such that this representation admits good hypotheses on future tasks. In truth, MTL and LTL typically are handled equivalently at training time — this work will be no exception — and they diverge only in their test settings and hence the learning theoretic inquiries they inspire.

Contributions. The first contribution of this work is to introduce minimax MTL and a continuum of relaxations. Second, we introduce a generalized loss-compositional paradigm for MTL which admits a number of new MTL formulations and also includes classical MTL as a special case. Third, we empirically evaluate the performance of several MTL formulations from this paradigm in the multi-task learning and learning to learn settings, under the task-wise maximum test risk and task-wise mean test risk criteria, on four datasets (one synthetic, three real). Finally, Theorem 1 is the core theoretical contribution of this work and shows 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, the theory suggests minimizing the maximum of the empirical risks across the training tasks rather than their mean.

In the next section, we recall the settings of multi-task learning and learning to learn, formally introduce minimax MTL, and motivate it theoretically. In Section 3, we introduce a continuously parametrized family of minimax MTL relaxations and the new generalized loss-compositional paradigm. Section 4 presents an empirical evaluation of various MTL/LTL formulations with different models on four datasets. Finally, we close with a discussion.

2   Minimax multi-task learning

We begin with a promenade through the basic MTL and LTL setups, with an effort to abide by the notation introduced by Baxter [4]. Throughout the rest of the paper, each labeled example (x, y) will live in $\mathcal{X} \times \mathcal{Y}$ for input instance x and label y. Typical choices of $\mathcal{X}$ include $\mathbb{R}^n$ or a compact subset thereof, while $\mathcal{Y}$ typically is a compact subset of $\mathbb{R}$ or the binary $\{-1, 1\}$. In addition, 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 $h : \mathcal{X} \to \mathbb{R}$.

---

[1]Note that minimax MTL does not refer to the *minimax estimators* of statistical decision theory.

Learning to learn. In learning to learn, the goal is to achieve inductive transfer to learn the best $\mathcal{H}$ from $\mathbb{H}$. Unlike in MTL, there is a notion of 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.

2.1 Minimax MTL

A natural generalization of classical MTL results by introducing a prior distribution π over the index 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:

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

where the supremum is taken over the T-dimensional simplex. As the supremum (assuming it is attained) is attained at an extreme point of the simplex, this objective is equivalent to

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

In practice, we approximate the true objective via a regularized form of the empirical objective

$$\inf_{\mathcal{H} \in \mathbb{H}} \max_{t \in [T]} \inf_{h \in \mathcal{H}} \sum_{i=1}^{m} \ell(y_{t,i}, h(x_{t,i})).$$

In the next section, we 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. In this short paper, we restrict attention to the case of finite $\mathbb{H}$.

2.2 A learning to learn bound for the maximum risk

In this subsection, we use the following notation. Let $P^{(1)}, \ldots, P^{(T)}$ 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 $P_m^{(t)}$. Also, 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 would like to

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 learning to learn, 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 [4, 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:

$$\Pr\left\{\inf_{h\in\mathcal{H}} P\ell\circ h \geq \gamma\right\} \leq \frac{\frac{1}{T}\sum_{t\in[T]} P_m^{(t)}\ell\circ h_t + \varepsilon}{\gamma} \tag{4}$$

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 $P^{(1)}, \ldots, P^{(T)}$ 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:*

$$\Pr\left\{P\ell\circ\hat{h} > \gamma + \frac{1}{T} + 2L\max_{\mathcal{H}\in\mathbb{H}}\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{8\log\frac{4}{\delta}}{m}}\right\} \leq \frac{\log\frac{2\mathcal{C}}{\delta} + \log\lceil B\rceil + \log(T+1)}{T}. \tag{5}$$

In the above, $\mathcal{R}_m(\mathcal{H})$ is the Rademacher complexity of $\mathcal{H}$ (cf. [3]). Critically, in (5) the probability of observing a task with high true risk decays with T, whereas in (4) 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.

For the proof of Theorem 1, first consider the singleton case $\mathbb{H} = \{\mathcal{H}_1\}$. Suppose that for γ fixed a priori, the maximum of the empirical risks is bounded by γ, i.e. $\max_{t\in[T]}\min_{h\in\mathcal{H}_1} P_m^{(t)}\ell\circ h \leq \gamma$.

Let a new probability measure P drawn from Q correspond to a new test task. Suppose the probability of the event $[\min_{h\in\mathcal{H}_1} P_m\ell\circ h > \gamma]$ is at least ε. Then the probability that γ bounds all T empirical risks is at most $(1-\varepsilon)^T \leq e^{-T\varepsilon}$. Hence, with probability at least $1 - e^{-T\varepsilon}$:

$$\Pr\left\{\min_{h\in\mathcal{H}_1} P_m\ell\circ h > \gamma\right\} \leq \varepsilon.$$

A simple application of the union bound extends this result for finite $\mathbb{H}$:

Lemma 1. *Under the same conditions as Theorem 1, with probability at least $1 - \delta/2$ with respect to the random draw of the T tasks and their T corresponding m-samples:*

$$\Pr\left\{\min_{h\in\mathcal{H}} P_m\ell\circ h > \gamma\right\} \leq \frac{\log\frac{2\mathcal{C}}{\delta}}{T}.$$

The bound in the lemma states a $1/T$ rate of decay for the probability that the empirical risk obtained by $\mathcal{H}$ on a new task exceeds γ. Next, we relate this empirical risk to the true risk obtained by the empirical risk minimizer. Note that at test time $\mathcal{H}$ is fixed and hence independent of any test m-sample. Then, from by now standard learning theory results of Bartlett and Mendelson [3]:

Lemma 2. *Take loss ℓ as in Theorem 1. With probability at least $1 - \delta/2$, for all $h \in \mathcal{H}$ uniformly:*

$$P\ell\circ h \leq P_m\ell\circ h + 2L\mathcal{R}_m(\mathcal{H}) + \sqrt{(8\log(4/\delta))/m}.$$

In particular, with high probability the true risk of the empirical risk minimizer is not much larger than its empirical risk. Theorem 1 now follows from Lemmas 1 and 2 and a union bound over $\gamma \in \Gamma := \{0, 1/T, 2/T, \ldots, \lceil B\rceil\}$; note that mapping the observed maximum empirical risk γ to $\min\{\gamma' \in \Gamma \mid \gamma \leq \gamma'\}$ picks up the additional $\frac{1}{T}$ term in (5).

In the next section, we introduce a loss-compositional paradigm for multi-task learning which includes as special cases minimax MTL and classical MTL.

3 A generalized loss-compositional paradigm for MTL

The paradigm can benefit from a bit of notation. 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 $\mathbf{h} := (h_1, \ldots, h_T) \in \mathcal{H}^T$ and the vector of empirical risks $\hat{\boldsymbol{\ell}}(\mathbf{h}) := (\hat{\ell}_1(h_1), \ldots, \hat{\ell}_T(h_T))$.

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:

$$\inf_{\mathcal{H} \in \mathbb{H}} \inf_{\mathbf{h} \in \mathcal{H}^T} \phi\big(\hat{\boldsymbol{\ell}}(\mathbf{h})\big) + \Omega\big((\mathcal{H}, \mathbf{h})\big), \tag{6}$$

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

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

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

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

A natural question is why one might consider minimizing ℓ_p-norms of the empirical risks vector for $1 < p < \infty$, as in ℓ_2 MTL. The contour of the ℓ_1-norm of the empirical risks evenly trades off empirical risks between different tasks; however, it has been observed that overfitting often happens near the end of learning, rather than the beginning [14]. More precisely, when the empirical risk is high, the gradient of the empirical risk (taken with respect to the parameter $(\mathcal{H}, \mathbf{h})$) is likely to have positive inner product with the gradient of the true risk. Therefore, given a candidate solution with a corresponding vector of empirical risks, a sensible strategy is to take a step in solution space which places more emphasis on tasks with higher empirical risk. This strategy is particularly appropriate when the class of learners has high capacity relative to the amount of available data. This observation sets the foundation for an approach that minimizes norms of the empirical risks.

In this work, we also discuss an interesting subset of the loss-compositional paradigm which does not fit into ℓ_p MTL; this subfamily embodies a continuum of relaxations of minimax MTL.

α-minimax MTL. In some cases, minimizing the maximum loss can exhibit certain disadvantages because the maximum loss is not robust to situations when a small fraction of the tasks are fundamentally harder than the remaining tasks. Consider the case when the empirical risk for each task in this small fraction can not be reduced below a level u. Rather than rigidly minimizing the maximum loss, a more robust alternative is to minimize the maximize loss in a soft way. Intuitively, the idea is to ensure that most tasks have low empirical risk, but a small fraction of tasks are permitted to have higher loss. We formalize this as α-*minimax MTL*, via the relaxed objective:

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

In the above, ϕ from the loss-compositional paradigm (6) is a variational function of the empirical risks vector. The above optimization problem is equivalent to the perhaps more intuitive problem:

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

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.

This work focuses on α-minimax MTL with $\alpha = 2/(\lceil 0.1T + 0.5 \rceil^{-1} + \lceil 0.1T + 1.5 \rceil^{-1})$ 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. We also try the 20% level (i.e. $0.2T$ replacing $0.1T$ in the above).

Models. We now provide examples of how specific models fit into this framework. We consider two convex multi-task learning formulations: Evgeniou and Pontil's regularized multi-task learning (the *EP model*) [5] and Argyriou, Evgeniou, and Pontil's convex multi-task feature learning (the *AEP model*) [1]. The EP model is a linear model with a shared parameter $v_0 \in \mathbb{R}^d$ and task-specific parameters $v_t \in \mathbb{R}^d$ (for $t \in [T]$). Evgeniou and Pontil presented this model as

$$\min_{v_0, \{v_t\}_{t \in [T]}} \sum_{t \in [T]} \sum_{i=1}^m \ell(y_{t,i}, \langle v_0 + v_t, x_{t,i} \rangle) + \lambda_0 \|v_0\|^2 + \frac{\lambda_1}{T} \sum_{t \in [T]} \|v_t\|^2,$$

for ℓ the hinge loss or squared loss. This can be set in the new paradigm via $\mathbb{H} = \{\mathcal{H}_{v_0} \mid v_0 \in \mathbb{R}^d\}$, $\mathcal{H}_{v_0} = \{h : x \mapsto \langle v_0 + v_t, x \rangle \mid v_t \in \mathbb{R}^d\}$, and $\hat{\ell}_t(h_t) = \frac{1}{m} \sum_{i=1}^m \ell(y_{t,i}, \langle v_0 + v_t, x_{t,i} \rangle)$.

The AEP model minimizes the task-wise average loss with the trace norm (nuclear norm) penalty:

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

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, $\hat{\ell}_t(h_t) = \frac{1}{m} \sum_{i=1}^m \ell(y_{t,i}, \langle h_t, x_{t,i} \rangle)$.

For easy empirical comparison between the various MTL formulations from the paradigm, at times it will be convenient to use constrained formulations of the EP and AEP model. If the regularized forms are used, a fair comparison of the methods warrants plotting results according to the size of the optimal parameter found (i.e. $\|W\|_{\mathrm{tr}}$ for AEP). For EP, the constrained form is:

$$\min_{v_0, \{v_t\}_{t \in [T]}} \sum_{t \in [T]} \sum_{i=1}^m \ell(y_{t,i}, \langle v_0 + v_t, x_{t,i} \rangle) \quad \text{subject to } \|v_0\| \le \tau_0, \ \|v_t\| \le \tau_1 \text{ for } t \in [T].$$

For AEP, the constrained form is: $\quad \min_W \sum_t \sum_{i=1}^m \ell(y_{t,i}, \langle W_t, x_{t,i} \rangle) \quad \text{subject to } \|W\|_{\mathrm{tr}} \le r.$

4 Empirical evaluation

We consider four learning problems; the first three involve regression (MTL model in parentheses):

- A synthetic dataset composed from *two modes* of tasks (EP model),
- The *school* dataset from the Inner London Education Authority (EP model),
- The conjoint analysis *personal computer* ratings dataset [2] [11] (AEP model).

The fourth problem is multi-class classification from the *MNIST* digits dataset [10] with a reduction to multi-task learning using a tournament of pairwise (binary) classifiers. We use the AEP model. Given data, each problem involved a choice of MTL formulation (e.g. minimax MTL), model (EP or AEP), and choice of regularized versus constrained. All the problems were solved with just a few lines of code using CVX [9, 8]. In this work, we considered convex multi-task learning formulations in order to make clear statements about the optimal solutions attained for various learning problems.

---

[2]This data, collected at the University of Michigan MBA program, generously was provided by Peter Lenk.
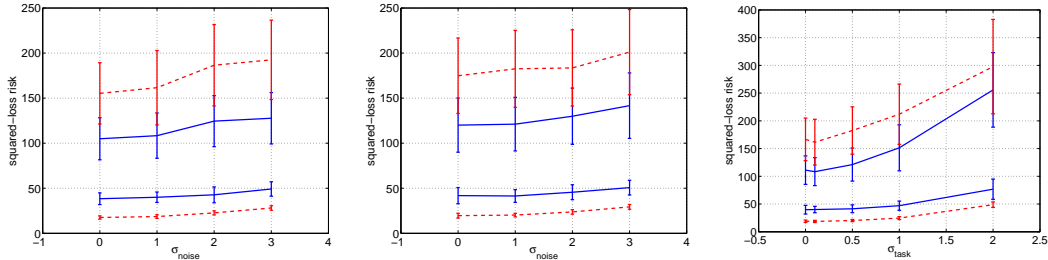
Figure 1: Max ℓ_2-risk (Top two lines) and mean ℓ_2-risk (Bottom two lines). At Left and Center: ℓ_2-risk vs noise level, for $\sigma_{\text{task}} = 0.1$ and $\sigma_{\text{task}} = 0.5$ respectively. At Right: ℓ_2-risk vs task variation, for $\sigma_{\text{noise}} = 0.1$. Dashed red is ℓ_1, dashed blue is minimax. Error bars indicate one standard deviation. MTL results (not shown) were similar to LTL results (shown), with MTL-LTL relative difference below 6.8% for all points plotted.

Two modes. The two modes regression problem consists of 50 linear prediction tasks for the first type of task and 5 linear prediction tasks for the second task type. The true parameter for the first task type is a vector μ drawn uniformly from the sphere of radius 5; the true parameter for the second task type is -2μ. Each task is drawn from an isotropic Gaussian with mean taken from the task type and the standard deviation of all dimensions set to σ_{task}. Each data point for each task is drawn from a product of 10 standard normals (so $x_{t,i} \in \mathbb{R}^{10}$). The targets are generated according to $\langle W_t, x_{t,i} \rangle + \varepsilon_t$, where the ε_t's are iid univariate centered normals with standard deviation σ_{noise}. We fixed τ_0 to a large value (in this case, $\tau_0 = 10$ is sufficient since the mean for the largest task fits into a ball of radius 10) and τ_1 to a small value ($\tau_1 = 2$). We compute the average mean and maximum test error over 100 instances of the 55-task multi-task problem. Each task's training set and test set are 5 and 15 points respectively. The average maximum (mean) test error is the 100-experiment-average of the task-wise maximum (mean) of the ℓ_2 risks. For each LTL experiment, 55 new test tasks were drawn using the same μ as from the training tasks.

Figure 1 shows a tradeoff: when each task group is fairly homogeneous (left and center plots), minimax is better at minimizing the maximum of the test risks while ℓ_1 is better at minimizing the mean of the test risks. As task homogeneity decreases (right plot), the gap in performance closes with respect to the maximum of the test risks and remains roughly the same with respect to the mean.
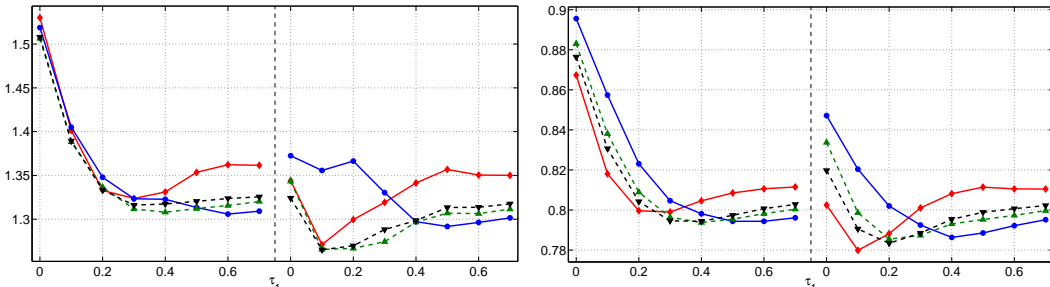

Figure 2: Maximum RMSE (Left) and normalized mean RMSE (Right) versus task-specific parameter bound τ_1, for shared parameter bound τ_0 fixed. In each figure, Left section is τ_0 is 0.2 and Right section is $\tau_0 = 0.6$. Solid red ♦ is ℓ_1, solid blue ● is minimax, dashed green ▲ is $(0.1T)$-minimax, dashed black ▼ is $(0.2T)$-minimax. The results for ℓ_2 MTL were visually identical to ℓ_1 MTL and hence were not plotted.

School. The school dataset has appeared in many previous works [7, 2, 6]. For brevity we just say the goal is to predict student test scores using certain student-level features. Each school is treated as a separate task. We report both the task-wise maximum of the root mean square error (RMSE) and the taskwise-mean of the RMSE (normalized by number of points per task, as in previous works).

The results (see Figure 2) demonstrate that when the learner has moderate shared capacity τ_0 and high task-specific capacity τ_1, minimax MTL outperforms ℓ_1 MTL for the max objective; additionally, for the max objective in almost all parameter settings $(0.1T)$-minimax and $(0.2T)$-minimax MTL outperform ℓ_1 MTL, and they also outperform minimax MTL when the task-specific capacity τ_1 is not too large. We hypothesize that minimax MTL performs the best in the high$-\tau_1$ regime because stopping learning once the maximum of the empirical risks cannot be improved invokes early stopping and its built-in regularization properties (see e.g. [13]). Interestingly, for the normalized mean RMSE objective, both minimax relaxations are competitive with ℓ_1 MTL; however, when the

shared capacity τ_0 is high (right section, right plot), ℓ_1 MTL performs the best. For high task-specific capacity τ_1, minimax MTL and its relaxations again seem to resist overfitting compared to ℓ_1 MTL.

Personal computer. The personal computer dataset is composed of 189 human subjects each of which rated on a 0-10 scale the same 20 computers (16 training, 4 test). Each computer has 13 binary features (amount of memory, screen size, price, etc.).

The results are shown in Figure 3. In the MTL setting, for both the maximum RMSE objective and the mean RMSE objective, ℓ_1 MTL appears to perform the best. When the trace norm of W is high, minimax MTL displays resistance to overfitting and obtains the lowest mean RMSE. In the LTL setting for the maximum RMSE objective, ℓ_2, minimax, and $(0.1T)$-minimax MTL all outperform ℓ_1 MTL. For the mean RMSE, ℓ_1 MTL obtains the lowest risk for almost all parameter settings.
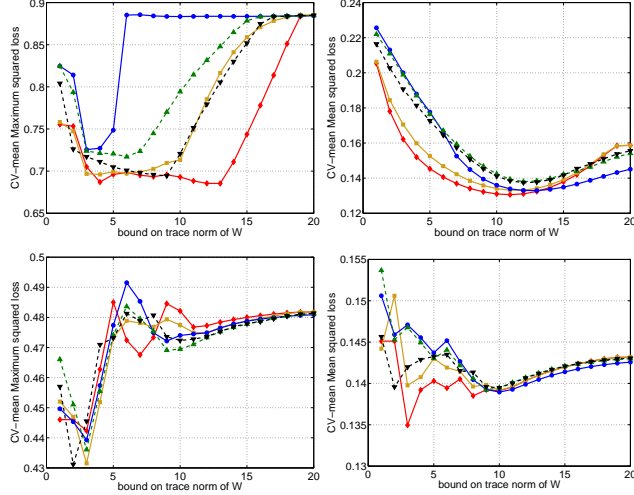

Figure 3: MTL (Top) and LTL (Bottom). Maximum ℓ_2 risk (Left) and Mean ℓ_2 risk (Right) vs bound on $\|W\|_{\mathrm{tr}}$. LTL used 10-fold cross-validation (10% of tasks left out in each fold). Solid red ◆ is ℓ_1, solid blue ● is minimax, dashed green ▲ is $(0.1T)$-minimax, dashed black ▼ is $(0.2T)$-minimax, solid gold ■ is ℓ_2.

MNIST. The MNIST task is a 10-class problem; we approach it via a reduction to a tournament of 45 binary classifiers trained via the AEP model. The dimensionality was reduced to 50 using principal component analysis (computed on the full training set), and only the first 2% of each class's training points were used for training.

Intuitively, the performance of the tournament tree of binary classifiers can only be as accurate as its paths, and the accuracy of each path depends on the accuracy of the nodes. Hence, our hypothesis is that minimax MTL should outperform ℓ_1 MTL. The results in Figure 4 confirm our hypothesis. Minimax MTL outperforms ℓ_1 MTL when the capacity $\|W\|_{\mathrm{tr}}$ is somewhat limited, with the gap widening as the capacity decreases. Furthermore, at every capacity minimax MTL is competitive with ℓ_1 MTL.
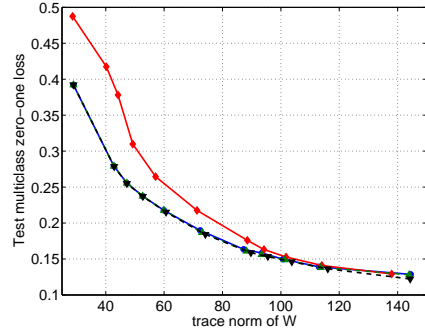

Figure 4: Test multiclass 0-1 loss vs $\|W\|_{\mathrm{tr}}$. Solid red is ℓ_1 MTL, solid blue is minimax, dashed green is $(0.1T)$-minimax, dashed black is $(0.2T)$-minimax. Regularized AEP used for speed and trace norm of W's computed, so samples differ per curve.

5   Discussion

We have established a continuum 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. The empirical evaluations indicate that α-minimax MTL at either the 10% or 20% level often outperforms ℓ_1 MTL in terms of the maximum test risk objective and sometimes even in the mean test risk objective. 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.

Although efficient algorithms may make the various new MTL learning formulations practical for large problems, a proper effort to develop fast algorithms in this setting would have detracted from the main point of this first study. A good direction for the future is to obtain efficient algorithms for minimax and α-minimax MTL. In fact, such algorithms might have applications beyond MTL and even machine learning. Another area ripe for exploration is to establish more general learning bounds for minimax MTL and to extend these bounds to α-minimax MTL.

References

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

[2] Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4:83–99, 2003.

[3] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[4] Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12(1):149–198, 2000.

[5] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004.

[6] Theodoros Evgeniou, Massimiliano Pontil, and Olivier Toubia. A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. *Marketing Science*, 26(6):805–818, 2007.

[7] Harvey Goldstein. Multilevel modelling of survey data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 40(2):235–244, 1991.

[8] Michael C. Grant and Stephen P. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008.

[9] Michael C. Grant and Stephen P. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21, April 2011.

[10] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[11] Peter J. Lenk, Wayne S. DeSarbo, Paul E. Green, and Martin R. Young. Hierarchical bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, pages 173–191, 1996.

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

[13] Noboru Murata and Shun-ichi Amari. Statistical analysis of learning dynamics. *Signal Processing*, 74(1):3–28, 1999.

[14] Nicolas Le Roux, Pierre-Antoine Manzagol, and Yoshua Bengio. Topmoumoute online natural gradient algorithm. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 849–856. MIT Press, Cambridge, MA, 2008.

[15] Kai Yu, John Lafferty, Shenghuo Zhu, and Yihong Gong. Large-scale collaborative prediction using a nonparametric random effects model. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1185–1192. ACM, 2009.

[16] Liang Zhang, Deepak Agarwal, and Bee-Chung Chen. Generalizing matrix factorization through flexible regression priors. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 13–20. ACM, 2011.