Journal of Machine Learning Research 15 (2014) Submitted ; Published -

Revisiting Stein's Paradox: Multi-Task Averaging

Sergey Feldman SERGEY.FELDMAN@GMAIL.COM
Data Cowboys
9126 23rd Ave. NE
Seattle, WA 98115, USA
Maya R. Gupta MAYAGUPTA@GOOGLE.COM
Google
1225 Charleston Rd
Mountain View, CA 94301, USA
Bela A. Frigyik FRIGYIK@GMAIL.COM
Institute of Mathematics and Informatics
University of Pécs
H-7624 Pécs, Ifjúság St. 6, Hungary

Editor: Massimiliano Pontil

Abstract

We present a multi-task learning approach to jointly estimate the means of multiple independent distributions from samples. The proposed multi-task averaging (MTA) algorithm results in a convex combination of the individual task's sample averages. We derive the optimal amount of regularization for the two task case for the minimum risk estimator and a minimax estimator, and show that the optimal amount of regularization can be practically estimated without cross-validation. We extend the practical estimators to an arbitrary number of tasks. Simulations and real data experiments demonstrate the advantage of the proposed MTA estimators over standard averaging and James-Stein estimation.

Keywords: multi-task learning, James-Stein, Stein's paradox

1. Introduction

The mean is one of the most fundamental and useful tools in statistics (Salsburg, 2001). By the 16th century Tycho Brahe was using the mean to reduce measurement error in astrononimical investigations (Plackett, 1958). Legendre (1805) noted that the mean minimizes the sum of squared errors to a set of samples:

$$\bar{y} = \arg\min_{\tilde{y}} \sum_{i=1}^{N} (y_i - \tilde{y})^2. \tag{1}$$

More recently it has been shown that the mean minimizes the sum of any Bregman divergence to a set of samples (Banerjee et al., 2005; Frigyik et al., 2008). Gauss (1857) commented on the mean's popularity in his time:

> *"It has been customary certainly to regard as an axiom the hypothesis that if any quantity has been determined by several direct observations, made under the*

same circumstances and with equal care, the arithmetical mean of the observed values affords the most probable value, if not rigorously, yet very nearly at least, so that it is always most safe to adhere to it."

But the mean is a more subtle quantity than it first appears. In a surprising result popularly called *Stein's paradox* (Efron and Morris, 1977), Stein (1956) showed that it is better (in a summed squared error sense) to estimate *each* of the means of T Gaussian random variables using data sampled from *all* of them, even if the random variables are independent and have different means. That is, it is beneficial to consider samples from seemingly *unrelated* distributions to estimate a mean. Stein's result is an early example of the motivating hypothesis behind multi-task learning (MTL): that leveraging data from multiple tasks can yield superior performance over learning from each task independently. In this work we consider a multi-task regularization approach to the problem of estimating multiple means; we call this *multi-task averaging* (MTA).

Multi-task learning is most intuitive when the multiple tasks are *conceptually* similar. But we argue that it is really the *statistical* similarity of the multiple tasks that determines how well multi-task learning works. In fact, a key result of this paper is that proposed multi-task estimation achieves lower total squared error than the sample mean *if* the true means of the multiple tasks are close compared to the variance of the samples (see equation (12)). Of course, in practice cognitive notions of similarity can be a useful guide for multi-task learning, as tasks that seem similar to humans often do have similar statistics.

We begin the paper with the proposed MTA objective in Section 2, and review related work in Section 3. We show that MTA has provably nice theoretical properties in Section 4; in particular, we derive the optimal notion of task similarity for the two task case, which is also the optimal amount of regularization to be used in the MTA estimation. We generalize this analysis to form practical estimators for the general case of T tasks. Simulations in Section 5 verify the advantage of MTA over standard sample means and James-Stein estimation if the true means are close compared to the variance of the underlying distributions. In Section 6 we present four experiments on real data: (i) estimating Amazon customer reviews, (ii) estimating class grades, (iii) forecasting sales, and (iv) estimating the length of kings' reigns. These real-data experiments show that MTA is generally 10-20% better than the sample mean.

A short version of this paper was published in NIPS 2012 (Feldman et al., 2012). This paper substantially differs from that conference paper that it contains more analysis, proofs, and new and expanded experiments.

2. Multi-Task Averaging

Consider the problem of estimating the means of T random variables that have finite means $\{\mu_t\}$ and variances $\{\sigma_t^2\}$ for $t = 1, \ldots, T$. We treat this as a T-task multi-task learning problem, and estimate the T means jointly. We take as given N_t independent and identically distributed (iid) random samples $\{Y_{ti}\}_{i=1}^{N_t}$ for each task t. Key notation is summarized in Table 2.

T	number of tasks
N_t	number of samples for tth task
μ_t	true mean of task t
σ_t^2	variance of the tth task
$Y_{ti} \in \mathbb{R}$	ith random sample from tth task
$\bar{Y}_t \in \mathbb{R}$	sample average for tth task: $\frac{1}{N_t}\sum_i Y_{ti}$, also referred to as the *single-task* mean estimate
$\bar{Y} \in \mathbb{R}^T$	vector with tth component $\bar{Y}_t$
$Y_t^* \in \mathbb{R}$	MTA estimate of tth mean
$Y^* \in \mathbb{R}^T$	vector with tth component Y_t^*
$\hat{Y}_t \in \mathbb{R}$	an estimate of the tth mean
$\tilde{Y}_t \in \mathbb{R}$	dummy variable
$\Sigma \in \mathbb{R}^{T\times T}$	diagonal covariance matrix of $\bar{Y}$ with $\Sigma_{tt} = \frac{\sigma_t^2}{N_t}$
$A \in \mathbb{R}^{T\times T}$	pairwise task similarity matrix
$L = D - A$	graph Laplacian of A, with diagonal D s.t. $D_{tt} = \sum_{r=1}^T A_{tr}$
W	MTA solution matrix, $W = (I + \frac{\gamma}{T}\Sigma L)^{-1}$

Table 1: Key Notation

In this paper, we judge the estimates by total squared error: given T estimates $\{\hat{Y}_t\}$ and T true means $\{\mu_t\}$:

$$\textbf{estimation error}(\{\hat{Y}_t\}) \triangleq \sum_{t=1}^{T} \left(\mu_t - \hat{Y}_t\right)^2, \tag{2}$$

Equivalently (up to an additive factor of σ^2), the metric can be expressed as the total squared expected error to a random sample Y_t from each task:

$$\textbf{regression error}(\{\hat{Y}_t\}) \triangleq \sum_{t=1}^{T} E\left[\left(Y_t - \hat{Y}_t\right)^2\right]; \tag{3}$$

we use an empirical approximation to (3) in the experiments because the true means are not known but held-out samples from the distributions are available.

Let a $T \times T$ matrix A describe the relatedness or similarity of any pair of the T tasks, with $A_{tt} = 0$ for all t without loss of generality (because the diagonal self-similarity terms are canceled in the objective below). Further we assume each task's variance σ_t^2 is known or already estimated. The proposed MTA objective is

$$\{Y_t^*\}_{t=1}^T = \underset{\{\tilde{Y}_t\}_{t=1}^T}{\arg\min} \ \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{(Y_{ti} - \tilde{Y}_t)^2}{\sigma_t^2} + \frac{\gamma}{T^2}\sum_{r=1}^{T}\sum_{s=1}^{T} A_{rs}(\tilde{Y}_r - \tilde{Y}_s)^2. \tag{4}$$

The first term of (4) minimizes the multi-task empirical loss, normalizing the contribution of each task's losses by that task's variance σ_t^2 so that high-variance tasks do not disproportionately dominate the loss term. The second term of (4) jointly regularizes the estimates

by tying them together. The regularization parameter γ balances the empirical risk and the multi-task regularizer. If $\gamma = 0$, the MTA objective decomposes into T separate minimization problems, producing the sample averages $\{\bar{Y}_t\}$. If $\gamma = 1$, the balance between empirical risk and multi-task regularizer is completely specified by the task similarity matrix A.

A more general formulation of MTA is

$$\{Y_t^*\}_{t=1}^T = \underset{\{\tilde{Y}_t\}_{t=1}^T}{\arg\min} \ \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t} L(Y_{ti}, \tilde{Y}_t) + \gamma J\left(\{\tilde{Y}_t\}_{t=1}^T\right),$$

where L is some loss function and J is some regularization function. If L is chosen to be any Bregman loss, then setting $\gamma = 0$ will produce the T sample averages (Banerjee et al., 2005). For the analysis and experiments in this paper, we restrict our focus to the tractable squared-error formulation given in (4). The MTA objective and many of the results in this paper generalize straightforwardly to samples that are vectors rather than scalars (see Section 4.2), but for most of the paper we restrict our focus to scalar samples $Y_{ti} \in \mathbb{R}$.

The task similarity matrix A can be specified as side information (e.g. from a domain expert), but often this side information is not available, or it may not be clear how to convert semantic notions of task similarity into appropriate numerical values for the task-similarity values in A. In such cases, A can be treated as a matrix parameter of the MTA objective, and in Section 4 we fix $\gamma = 1$ and derive two optimal choices of A for the $T = 2$ case: the A that minimizes expected squared error, and a minimax A. We use the $T = 2$ analysis to propose practical estimators of A for any number of tasks, removing the need to cross-validate the amount of regularization.

3. Related Work

In this section, we review related and background material: James-Stein estimation, multi-task learning, manifold regularization, and the graph Laplacian.

3.1 James-Stein Estimation

A closely related body of work to MTA is Stein estimation (James and Stein, 1961; Bock, 1975; Efron and Morris, 1977; Casella, 1985), which can be derived as an empirical Bayes strategy for estimating multiple means simultaneously (Efron and Morris, 1972). James and Stein (1961) showed that the maximum likelihood estimate of the task mean can be dominated by a shrinkage estimate given Gaussian assumptions. Specifically, given a single sample drawn from T normal distributions $Y_t \sim \mathcal{N}(\mu_t, \sigma^2)$ for $t = 1, \ldots, T$, Stein showed that the maximum likelihood estimator $\bar{Y}_t = Y_t$ is inadmissible, and is dominated by the James-Stein estimator:

$$\hat{Y}_t^{JS} = \left(1 - \frac{(T-2)\sigma^2}{\bar{Y}^\top \bar{Y}}\right)\bar{Y}_t, \tag{5}$$

where $\bar{Y}$ is a vector with tth entry $\bar{Y}_t$. The above estimator dominates $\bar{Y}_t$ when $T > 2$. For $T = 2$, (5) reverts to the maximum likelihood estimator, which turns out to be admissible (Stein, 1956). James and Stein showed that if σ^2 is unknown it can be replaced by a standard unbiased estimate $\hat{\sigma}^2$ (James and Stein, 1961; Casella, 1985).

Note that in (5) the James-Stein estimator *shrinks* the sample means towards zero (the terms "regularization" and "shrinkage" are often used interchangeably). The form of (5) and its shrinkage towards zero points to the implicit assumption that the μ_t are themselves drawn from a standard normal distribution centered at 0. More generally, the means are assumed to be drawn as $\mu_t \sim \mathcal{N}(\xi, 1)$. The James-Stein estimator then becomes

$$\hat{Y}_t^{JS} = \xi + \left(1 - \frac{(T-3)\sigma^2}{(\bar{Y} - \xi\mathbf{1})^\top(\bar{Y} - \xi\mathbf{1})}\right)(\bar{Y}_t - \xi), \tag{6}$$

where ξ can be estimated (as we do in this work) as the average of means $\xi = \bar{\bar{Y}} = \frac{1}{T}\sum_{r=1}^{T}\bar{Y}_r$, and this additional estimation decreases the degrees of freedom by one.[1] Note that (6) shrinks the estimates towards the mean-of-means ξ rather than shrinking towards zero. Also, the more similar the multiple tasks are (in the sense that individual task means are closer to the mean-of-means ξ), the more regularization occurs in (6).

There have been a number of extensions to the original James-Stein estimator. The James-Stein estimator given in (6) is itself not admissible, and is dominated by the positive part James-Stein estimator (Lehmann and Casella, 1998), which was further theoretically improved by Bock's James-Stein estimator (Bock, 1975). Throughout this work, we compare to Bock's well-regarded positive-part James-Stein estimator for multiple data points per task and independent unequal variances (Bock, 1975; Lehmann and Casella, 1998). In particular, let $Y_{ti} \sim \mathcal{N}(\mu_t, \sigma_t^2)$ for $t = 1, \ldots, T$ and $i = 1, \ldots, N_t$, let Σ be the covariance matrix of the vector of task sample means $\bar{Y}$, and let $\lambda_{\max}(\Sigma)$ be the largest eigenvalue of Σ, then the estimator is

$$\hat{Y}_t^{JS} = \xi + \left(1 - \frac{\frac{\mathbf{tr}(\Sigma)}{\lambda_{\max}(\Sigma)} - 3}{(\bar{Y} - \xi\mathbf{1})^\top\Sigma^{-1}(\bar{Y} - \xi\mathbf{1})}\right)_+(\bar{Y}_t - \xi), \tag{7}$$

where $(x)_+ = \max(0, x)$.

3.2 Multi-Task Learning for Mean Estimation

MTA is an approach to the problem of estimating T means. We are not aware of other work in the multi-task literature that addresses this problem explicitly; most MTL methods are designed for regression, classification, or feature selection, e.g. Micchelli and Pontil (2004); Bonilla et al. (2008); Argyriou et al. (2008). Estimating T means can be considered a special case of multi-task regression, where one fits a constant function to each task, since, with a feature space of zero dimensions only the constant offset term is learned. And, similarly to MTA, one of the main approaches to multi-task regression in the literature is tying tasks together with an explicit multi-task parameter regularizer.

Abernethy et al. (2009), for instance, propose to minimize the empirical loss by adding the regularizer

$$||\beta||_*,$$

where the tth column of the matrix β is the vector of parameters for the tth task and $||\cdot||_*$ is the trace norm. Applying this approach to mean estimation, the matrix β has only one row,

1. For more details as to why $T-2$ in (5) becomes $T-3$ in (6), see Example 7.7 on page 278 of Lehmann and Casella (1998).

and $||\beta||_*$ reduces to the ℓ_2 norm on the outputs, thus for mean estimation this regularizer does not actually tie the tasks together.

Argyriou et al. (2008) propose a a different regularizer,

$$\mathbf{tr}(\beta^\top D^{-1}\beta),$$

where D is a learned, shared *feature* covariance matrix. With no features (as in the MTA application of constant function regression), D is just a constant and $\mathbf{tr}(\beta^\top D^{-1}\beta)$ is a ridge regularizer on the outputs. The regularizers in the work of Jacob et al. (2008) and Zhang and Yeung (2010) reduce similarly when applied to mean estimation. These regularizers are similar to the *original* James Stein estimator in that they shrink the estimates towards zero; though more modern James Stein estimators shrink towards a pooled mean (see Sec 3.1).

The most closely related work is that of Sheldon (2008) and Kato et al. (2008), where the regularizer or constraint, respectively, is

$$\sum_{r=1}^{T}\sum_{s=1}^{T} A_{rs}\|\beta_r - \beta_s\|_2^2,$$

which is the MTA regularizer if applied to mean estimation. In this paper we do just that: apply this regularizer to mean estimation, show that this special case enables new and useful analytic results, and demonstrate its performance with simulated and real data.

3.3 Multi-Task Learning and the Similarity Between Tasks

A key issue for MTA and many other multi-task learning methods is how to estimate some notion of similarity (or task relatedness) between tasks and/or samples if it is not provided. A common approach is to estimate the similarity matrix jointly with the task parameters (Argyriou et al., 2007; Xue et al., 2007; Bonilla et al., 2008; Jacob et al., 2008; Zhang and Yeung, 2010). For example, Zhang and Yeung (2010) assume that there exists a covariance matrix for the task relatedness, and proposed a convex optimization approach to estimate the task covariance matrix and the task parameters in a joint, alternating way. Applying such joint and alternating approaches to the MTA objective given in (4) leads to a degenerate solution with zero similarity. However, the simplicity of MTA enables us to specify the optimal task similarity matrix for $T = 2$ (see Sec. 4), which we use to obtain closed-form estimators for the general $T > 1$ case.

3.4 Manifold Regularization

MTA is similar in form to *manifold regularization* (Belkin et al., 2006). For example, Belkin et al.'s Laplacian-regularized least squares objective for semi-supervised regression solves

$$\arg\min_{f\in\mathcal{H}} \quad \sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda||f||_{\mathcal{H}}^2 + \gamma\sum_{i,j=1}^{N+M} A_{ij}(f(x_i) - f(x_j))^2,$$

where f is the regression function to be estimated, $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), N is the number of labeled training samples, M is the number of unlabeled training samples, A_{ij} is the similarity (or weight in an adjacency graph) between feature samples

x_i and x_j, and $||f||_{\mathcal{H}}$ is the norm of the function f in the RKHS. In MTA, as opposed to manifold regularization, we are estimating a different function (that is, the constant function that is the mean) for each of the T tasks, rather than a single global function. One can interpret MTA as regularizing the individual task estimates over the task-similarity manifold, which is defined for the T tasks by the $T \times T$ matrix A.

3.5 Background on the Graph Laplacian Matrix

It will be helpful for later sections to review the graph Laplacian matrix. For graph G with T nodes, let $A \in \mathbb{R}^{T \times T}$ be a matrix where component $A_{rs} \geq 0$ is the weight of the edge between node r and node s, for all r, s. The *graph Laplacian matrix* is defined as $L = L(A) = D - A$, with diagonal matrix D such that $D_{tt} = \sum_s A_{ts}$.

The graph Laplacian matrix is analogous to the Laplacian operator, which quantifies how locally smooth a twice-differentiable function $g(x)$ is. Similarly, the graph Laplacian matrix L can be thought of as being a measure of the smoothness of a function defined on a graph (Chung, 2004). Given a function f defined over the T nodes of graph G, where $f_i \in \mathbb{R}$ is the function value at node i, the total *energy* of a graph is (for symmetric A)

$$\mathcal{E}(f) = \frac{1}{2} \sum_{i=1}^{T} \sum_{j=1}^{T} A_{ij}(f_i - f_j)^2 = f^\top L(A) f,$$

which is small when f is smooth over the graph (Zhu and Lafferty, 2005). If A is asymmetric then the energy can be written as

$$\mathcal{E}(f) = \frac{1}{2} \sum_{i=1}^{T} \sum_{j=1}^{T} A_{ij}(f_i - f_j)^2 = f^\top L((A + A^\top)/2) f. \tag{8}$$

When each $f_i \in \mathbb{R}^d$ is a vector, one can alternatively write the energy in terms of the distance matrix:

$$\mathcal{E}(f) = \frac{1}{2} \mathbf{tr}(\Delta^\top A),$$

where $\Delta_{ij} = (f_i - f_j)^\top (f_i - f_j)$.

As discussed above, the graph Laplacian can be thought of as an operator on a function, but it is useful in and of itself (i.e. without a function). The eigenvalues of the graph Laplacian are all real and non-negative, and there is a wealth of literature showing how the eigenvalues reveal the structure of the underlying graph (Chung, 2004); the eigenvalues of L are particularly useful for spectral clustering (v. Luxburg, 2007). The graph Laplacian is a common tool in semi-supervised learning literature (Zhu, 2006), and the Laplacian of a random walk probability matrix P (i.e. all the entries are non-negative and the rows sum to 1) is also of interest. For example, Saerens et al. (2004) showed that the pseudo-inverse of the Laplacian of a probability transition matrix is used to compute the square root of the average commute time (the average time taken by a random walker on graph G to reach node j for the first time when starting at node i, and coming back to node i).

Finally, since we will be using this fact occasionally, we note that the graph Laplacian is *homogenous*, i.e. $L(\gamma A) = \gamma L(A)$, where A is a matrix and γ is a scalar.

4. MTA Theory and Estimators

First, we give a general closed-form solution for the MTA mean estimates and characterize the MTA objective in Sections 4.1 – 4.3. Then in Section 4.4 we analyze the estimation error for the two task $T = 2$ case and give conditions for when MTA is better than the sample means. In Section 4.5, we derive the optimal similarity matrix A for the two task case.

Then in Section 4.7, we generalize our $T = 2$ analysis to propose practical estimators for any number of tasks T, and analyze their computational efficiency. In Section 4.8, we analyze the relationship of different estimators formed by linearly combining the sample means, including the MTA estimators, James Stein, and other estimators that regularize sample means towards a pooled mean. Lastly, we discuss the Bayesian interpretation of MTA in Section 4.9.

Proofs and derivations are in the appendix.

4.1 Closed-form MTA Solution

Without loss of generality, we only deal with symmetric A because in the case of asymmetric A it is equivalent to consider instead the symmetrized matrix $(A^\top + A)/2$.

For symmetric A with non-negative components, the MTA objective given in (4) is continuous, differentiable, and convex; and (4) has closed-form solution (derivation in appendix):

$$Y^* = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1} \bar{Y}, \tag{9}$$

where $\bar{Y}$ is the vector of sample averages with tth entry $\bar{Y}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} Y_{ti}$, L is the graph Laplacian of A, and Σ is the diagonal covariance matrix of the sample mean vector $\bar{Y}$ such that $\Sigma_{tt} = \frac{\sigma_t^2}{N_t}$. The inverse $\left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ in (9) always exists:

Lemma 1 *Suppose that $0 \le A_{rs} < \infty$ for all r, s, $\gamma \ge 0$, and $0 < \frac{\sigma_t^2}{N_t} < \infty$ for all t. The MTA solution matrix $W = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ exists.*

The MTA estimates Y^* converge to the vector of true means μ:

Proposition 2 *As $N_t \to \infty \ \forall\ t$, $Y^* \to \mu$.*

4.2 MTA for Vectors

MTA can also be applied to vectors. Let $\mathbb{Y}^* \in \mathbb{R}^{T\times d}$ be a matrix with Y_t^* as its tth row and let $\bar{\mathbb{Y}} \in \mathbb{R}^{T\times d}$ be a matrix with $\bar{Y}_t \in \mathbb{R}^d$ as its tth row. One can simply perform MTA on the vectorized form of $\mathbb{Y}^*$.

$$\mathbf{vec}(\mathbb{Y}^*) = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1} \mathbf{vec}(\bar{\mathbb{Y}}),$$

as long as (the now block-diagonal) $\Sigma \in R^{Td\times Td}$ is invertible. An equivalent formulation for MTA for vectors was proposed in Martínez-Rego and Pontil (2013).

4.3 Convexity of MTA Solution

One sees from (9) that the MTA estimates are linear combinations of the sample averages:

$$Y^* = W\bar{Y}, \text{ where } W = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}.$$

Moreover, and less obviously, each MTA estimate is a *convex* combination of the single-task sample averages:

Theorem 3 *If $\gamma \geq 0$, $0 \leq A_{rs} < \infty$ for all r, s and $0 < \frac{\sigma_t^2}{N_t} < \infty$ for all t, then the MTA estimates $\{Y_t^*\}$ given in (9) are convex combinations of the task sample averages $\{\bar{Y}_t\}$.*

This theorem generalizes a result of Chebotarev and Shamis (2006) that the matrix $(I + \gamma L)^{-1}$ is right-stochastic (i.e. the rows are non-negative and sum to 1) if the entries of A are strictly positive. Our proof (given in the appendix) uses a different approach, and extends the result both to the more general form of the MTA solution matrix $\left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ and to A with non-negative entries.

4.4 MSE Analysis for the Two Task Case

In this section we analyze the $T = 2$ task case, with N_1 and N_2 samples for tasks 1 and 2 respectively. Suppose random samples drawn for the first task $\{Y_{1i}\}$ are iid with finite mean μ_1 and finite variance σ_1^2, and random samples drawn for the second task $\{Y_{2i}\}$ are iid with finite mean $\mu_2 = \mu_1 + \Delta$ and finite variance σ_2^2. Let the task-relatedness matrix be $A = [0\ a; a\ 0]$, and without loss of generality, we fix $\gamma = 1$. Then the closed-form solution (9) can be simplified:

$$Y_1^* = \left(\frac{2 + \frac{\sigma_2^2}{N_2}a}{2 + \frac{\sigma_1^2}{N_1}a + \frac{\sigma_2^2}{N_2}a}\right)\bar{Y}_1 + \left(\frac{\frac{\sigma_1^2}{N_1}a}{2 + \frac{\sigma_1^2}{N_1}a + \frac{\sigma_2^2}{N_2}a}\right)\bar{Y}_2. \tag{10}$$

The mean squared error of Y_1^* is

$$\mathrm{MSE}[Y_1^*] = \frac{\sigma_1^2}{N_1}\left(\frac{4 + 4\frac{\sigma_2^2}{N_2}a + \frac{\sigma_1^2\sigma_2^2}{N_1N_2}a^2 + \frac{\sigma_2^4}{N_2^2}a^2}{\left(2 + \frac{\sigma_1^2}{N_1}a + \frac{\sigma_2^2}{N_2}a\right)^2}\right) + \frac{\Delta^2\frac{\sigma_1^4}{N_1^2}a^2}{\left(2 + \frac{\sigma_1^2}{N_1}a + \frac{\sigma_2^2}{N_2}a\right)^2}. \tag{11}$$

Next, we compare the MTA estimate Y_1^* to the sample average $\bar{Y}_1$, which is the maximum likelihood estimate of the true mean μ_1 for many distributions.[2] The MSE of the single-task sample average $\bar{Y}_1$ is $\frac{\sigma_1^2}{N_1}$, and comparing that to (11) and simplifying some tedious algebra establishes that

$$\mathrm{MSE}[Y_1^*] < \mathrm{MSE}[\bar{Y}_1] \text{ if } \Delta^2 < \frac{4}{a} + \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}. \tag{12}$$

Thus the MTA estimate of the first mean has lower MSE than the sample average estimate if the squared mean-separation Δ^2 is small compared to the summed variances of the sample means. See Figure 1 for an illustration.

2. The uniform distribution is perhaps the simplest example where the sample average is *not* the maximum likelihood estimate of the mean. For more examples, see Sec. 8.18 of Romano and Siegel (1986).

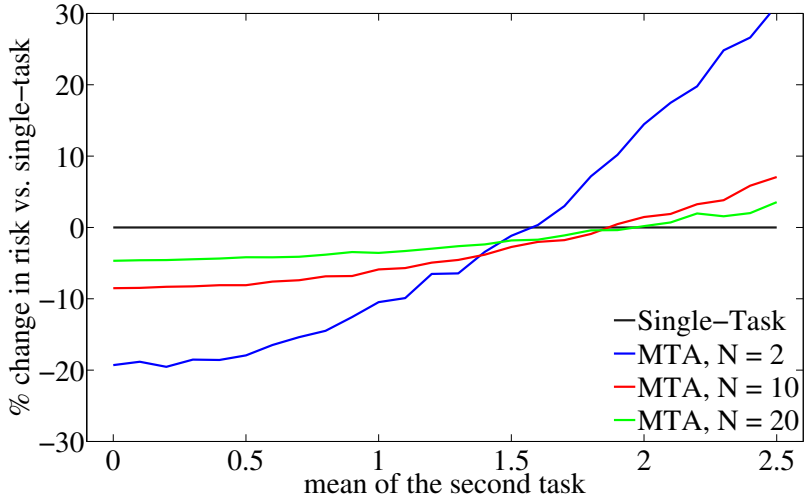

Figure 1: Plot shows the percent change in average risk for two tasks (averaged over 10,000 runs of the simulation). For each task there are N iid samples, for $N = 2, 10, 20$. The first task generates samples from a standard Gaussian. The second task generates samples from a Gaussian with $\sigma^2 = 1$ and different mean value, which is varied as marked on the x-axis. The symmetric task-relatedness value was fixed at $a = 1$ (note this is generally *not* the optimal value). One sees that given $N = 2$ samples from each Gaussian, the MTA estimate is better than the single-task sample if the difference between the true means is less than 1.5. Given $N = 20$ samples from each Gaussian, the MTA estimate is better if the distance between the means is less than 2. In the extreme case that the two Gaussians have the same mean ($\mu_1 = \mu_2 = 0$), then even with this suboptimal choice of $a = 1$, MTA provides a 20% win for $N = 2$ samples, and a 5% win for $N = 20$ samples.

Further, because of the symmetry of (12), if the condition of (12) holds, it is also true that $\mathrm{MSE}[Y_2^*] < \mathrm{MSE}[\bar{Y}_2]$, such that the MSE of each task individually is reduced.

The condition (12) shows that even when the true means are far apart such that Δ is large, there is some tiny amount of MTA regularization a that will improve the estimates.

4.5 Optimal Task Relatedness A for $T = 2$

We analyze the optimal choice of a in the task-similarity matrix $A = [0\ a; a\ 0]$. The risk is the sum of the mean squared errors:

$$R(\mu, Y^*) = \mathrm{MSE}[Y_1^*] + \mathrm{MSE}[Y_2^*], \tag{13}$$

which is a convex, continuous, and differentiable function of a, and therefore the first derivative can be used to specify the optimal value a^*, when all the other variables are fixed.

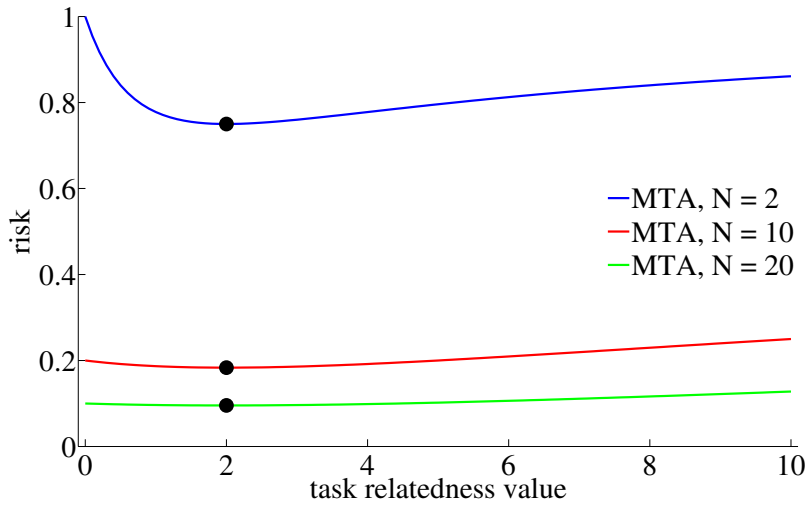

Figure 2: Plot shows the risk for two tasks, where the task samples were drawn iid from Gaussians $\mathcal{N}(0,1)$ and $\mathcal{N}(1,1)$. The task-relatedness value a was varied as shown on the x-axis. The minimum expected squared error is marked by a dot, and occurs for the choice of a given by (14), and is independent of N.

Minimizing (13) w.r.t. a one obtains the optimal:

$$a^* = \frac{2}{\Delta^2}, \tag{14}$$

which is always non-negative, as was assumed. This result is key because it specifies that the optimal task-similarity a^* ideally should measure the inverse of the squared task mean-difference. Further, the optimal task-similarity is independent of the number of samples N_t or the sample variance σ_t^2, as these are accounted for in Σ of the MTA objective. Note that a^* also minimizes the functions MSE$[Y_1^*]$ and MSE$[Y_2^*]$, separately.

The effect on the risk on the choice of a and the optimal a^* is illustrated in Figure 2.

Analysis of the second derivative shows that this minimizer always holds for $N_1, N_2 \geq 1$.

In the limit case, when the difference in the task means Δ goes to zero (while σ_t^2 stay constant), the optimal task-relatedness a^* goes to infinity, and the weights in (10) on $\bar{Y}_1$ and $\bar{Y}_2$ become $1/2$ each.

4.6 Estimating Task Similarity from Data for $T = 2$ Tasks

The optimal two-task similarity given in (14) requires knowledge of the true means μ_1 and μ_2. These are, in practice, unavailable. What similarity should be used then? A straightforward approach is to use single-task estimates instead:

$$\hat{a}^* = \frac{2}{(\bar{y}_1 - \bar{y}_2)^2},$$

and to use maximum likelihood estimates $\hat{\sigma}_t^2$ to form the matrix $\hat{\Sigma}$. This data-dependent approach is analogous to empirical Bayesian methods in which prior parameters are estimated from data (Casella, 1985).

4.7 Estimating Task Similarity from Data for Arbitrary T Tasks

Based on our analysis in the preceding sections of the optimal A for the two-task case, we propose two methods to estimate A from data for arbitrary $T > 1$. The first method is designed to minimize the approximate risk using a constant similarity matrix. The second method provides a minimax estimator. With both methods one can take advantage of the Sherman-Morrison formula (Sherman and Morrison, 1950) to avoid taking the matrix inverse or solving a set of linear equations in (9), resulting in an $O(T)$ computation time for Y^* (detailed in Section 4.7.3).

4.7.1 MTA Constant

The *risk* of estimator $\hat{Y} = W\bar{Y}$ is

$$R(\mu, W\bar{Y}) = E[(W\bar{Y} - \mu)^\top (W\bar{Y} - \mu)] \tag{15}$$

$$= \mathbf{tr}(W\Sigma W^\top) + \mu^\top (I - W)^\top (I - W)\mu, \tag{16}$$

where (16) uses the fact that $E[\bar{Y}\bar{Y}^\top] = \mu\mu^\top + \Sigma$.

One approach to generalizing the results of Section 4.4 to arbitrary T is to try to find a symmetric, non-negative matrix A such that the (convex, differentiable) risk $R(\mu, W\bar{Y})$ is minimized for $W = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ (recall L is the graph Laplacian of A). The problem with this approach is two-fold: (i) the solution is not analytically tractable for $T > 2$ and (ii) an arbitrary A has $T(T-1)$ degrees of freedom, which is considerably more than the T means we are trying to estimate in the first place. To avoid these problems, we generalize the two-task results by constraining A to be a scaled constant matrix $A = a\mathbf{1}\mathbf{1}^\top$, and find the optimal a^* that minimizes the risk given by (16). As in Section 4.4, we fix $\gamma = 1$. For analytic tractability, we add the assumption that all the Y_t have the same variance, estimating Σ as $\frac{\mathbf{tr}(\Sigma)}{T} I$. Then minimizing (15) becomes:

$$a^* = \arg\min_a R\left(\mu, \left(I + \frac{1}{T}\frac{\mathbf{tr}(\Sigma)}{T} L(a\mathbf{1}\mathbf{1}^\top)\right)^{-1} \bar{Y}\right),$$

which has the solution

$$a^* = \frac{2}{\frac{1}{T(T-1)} \sum_{r=1}^{T} \sum_{s=1}^{T} (\mu_r - \mu_s)^2}, \tag{17}$$

which does reduce to the optimal two task MTA solution (14) when $T = 2$.

While (17) is theoretically interesting, in practice, one of course does not have $\{\mu_r\}$ as these are precisely the values one is trying to estimate, and thus cannot use (17) directly. Instead, we propose estimating a^* using the sample means $\{\bar{y}_r\}$:

$$\hat{a}^* = \frac{2}{\frac{1}{T(T-1)} \sum_{r=1}^{T} \sum_{s=1}^{T} (\bar{y}_r - \bar{y}_s)^2}. \tag{18}$$

Using the optimal estimated *constant* similarity given in (18) and an estimated covariance matrix $\hat{\Sigma}$ produces what we refer to as the *MTA Constant* estimate

$$Y^* = \left(I + \frac{1}{T}\hat{\Sigma}L(\hat{a}^*\mathbf{1}\mathbf{1}^\top)\right)^{-1}\bar{Y}. \tag{19}$$

Note that we made the assumption that the entries of Σ were the same in order to be able to derive (17), but we do not need nor necessarily suggest that assumption on the $\hat{\Sigma}$ be used in practice with $\hat{a}^*$ in (19).

4.7.2 MTA Minimax

Bock's James-Stein estimator is *minimax* (Lehmann and Casella, 1998)). In this section, we derive a minimax version of MTA for arbitrary T that prescribes less regularization than MTA Constant. Formally, an estimator Y^M of μ is called minimax if it minimizes the maximum risk:

$$\inf_{\tilde{Y}} \sup_{\mu} R(\mu, \tilde{Y}) = \sup_{\mu} R(\mu, Y^M).$$

Let $r(\pi, \hat{Y})$ be the average risk of estimator $\hat{Y}$ w.r.t. a prior $\pi(\mu)$ such that $r(\pi, \hat{Y}) = \int R(\mu, \hat{Y})\pi(\mu)d\mu$. The Bayes estimator Y^π is the estimator that minimizes the average risk, and the Bayes risk $r(\pi, Y^\pi)$ is the average risk of the Bayes estimator. A prior distribution π is called least favorable if $r(\pi, Y^\pi) > r(\pi', Y^{\pi'})$ for all priors π'.

First, we will specify MTA Minimax for the $T = 2$ case. To find a minimax estimator Y^M it is sufficient to show that *(i)* Y^M is a Bayes estimator w.r.t. the least favorable prior (LFP) and *(ii)* it has constant risk (Lehmann and Casella, 1998). To find a LFP, we first need to specify a constraint set for μ_t; we use an interval: $\mu_t \in [b_l, b_u]$, for all t, where $b_l \in \mathbb{R}$ and $b_u \in \mathbb{R}$. With this constraint set the minimax estimator is (see appendix for details):

$$Y^M = \left(I + \frac{2}{T(b_u - b_l)^2}\Sigma L(\mathbf{1}\mathbf{1}^\top)\right)^{-1}\bar{Y},$$

which reduces to (14) when $T = 2$. This minimax analysis is only valid for the case when $T = 2$, but we found that the following extension of MTA Minimax to larger T worked well in simulations and applications for any $T \geq 2$. To estimate b_u and b_l from data we assume the unknown T means are drawn from a uniform distribution and use maximum likelihood estimates of the lower and upper endpoints for the support:

$$\hat{b}_l = \min_t \bar{y}_t \quad \text{and} \quad \hat{b}_u = \max_t \bar{y}_t.$$

Thus, in practice, *MTA Minimax* is

$$Y^M = \left(I + \frac{2}{T(\hat{b}_u - \hat{b}_l)^2}\hat{\Sigma}L(\mathbf{1}\mathbf{1}^\top)\right)^{-1}\bar{Y}.$$

4.7.3 Computational Efficiency of MTA Constant and Minimax

Both MTA Constant and MTA Minimax weight matrices can be written as

$$\begin{aligned}(I + c\Sigma L(\mathbf{1}\mathbf{1}^\top))^{-1} &= (I + c\Sigma(TI - \mathbf{1}\mathbf{1}^\top))^{-1} \\ &= (I + cT\Sigma - c\Sigma\mathbf{1}\mathbf{1}^\top)^{-1} \\ &= (Z - z\mathbf{1}^\top)^{-1},\end{aligned}$$

where c is different for MTA Constant and MTA Minimax, $Z = I + cT\Sigma$, $z = c\Sigma\mathbf{1}$. The Sherman-Morrison formula (Sherman and Morrison, 1950) can be used to find the inverse:

$$(Z - z\mathbf{1}^\top)^{-1} = Z^{-1} + \frac{Z^{-1}z\mathbf{1}^\top Z^{-1}}{1 - \mathbf{1}^\top Z^{-1}z}.$$

Since Z is diagonal, Z^{-1} can be computed in $O(T)$ time, and so can $Z^{-1}z$. Thus, the entire computation $W\bar{Y}$ can be done in $O(T)$ time for MTA Constant and MTA Minimax.

4.8 Generality of MTA

In this section, we use the expression 'matrices of MTA form' to refer to matrices that can be written

$$(I + \Gamma L(A))^{-1}, \tag{20}$$

where A is a matrix with all non-negative entries, and Γ is a diagonal matrix with all non-negative entries. Matrices of the form $(I + \gamma L)^{-1}$ have been used as graph kernels (Fouss et al., 2006; Yajima and Kuo, 2006), and were termed regularized Laplacian kernels (RLKs) by Smola and Kondor (2003). The RLK assumes that A (and L) are symmetric, and thus MTA and (20) strictly generalizes the RLK because ΓL is only symmetric for some special cases such as when Γ is a scaled identity matrix. Thus, one might also refer to matrices of the form (20) as generalized regularized Laplacian kernels, but in this section we focus on their role as estimators and in understanding relationships with the proposed MTA estimator.

Figure 3 is a Venn diagram of the sets of estimators that can be expressed $\hat{Y} = W\bar{Y}$, where W is some $T \times T$ matrix. The first subset (the pink region) is all estimators where W is right-stochastic. The second subset (the green region) is estimators of MTA form as per (20). The innermost subset (the purple region) includes many well-known estimators such as the James-Stein estimator, and estimators that regularize single-task estimates of the mean to the pooled mean or the average of means. In this section we will prove that the innermost purple subset is a *strict* subset of the green MTA subset, such that any innermost estimator can be written in MTA form for specific choices of A, γ, and Σ. Note that the covariance Σ is treated as a "choice" because some classic estimators assume $\Sigma = I$.

Proposition 4 *The set of estimators $W\bar{Y}$ where W is of MTA form as per (20) is strictly larger than the set of estimators that regularize the single-task estimates as follows:*

$$\hat{Y} = \left(\frac{1}{\gamma}I + \mathbf{1}\alpha^\top\right)\bar{Y},$$

where $\sum_{r=1}^{T}\alpha_r = 1 - \frac{1}{\gamma}$, $\gamma \geq 1$, and $\alpha_r \geq 0$, $\forall r$.

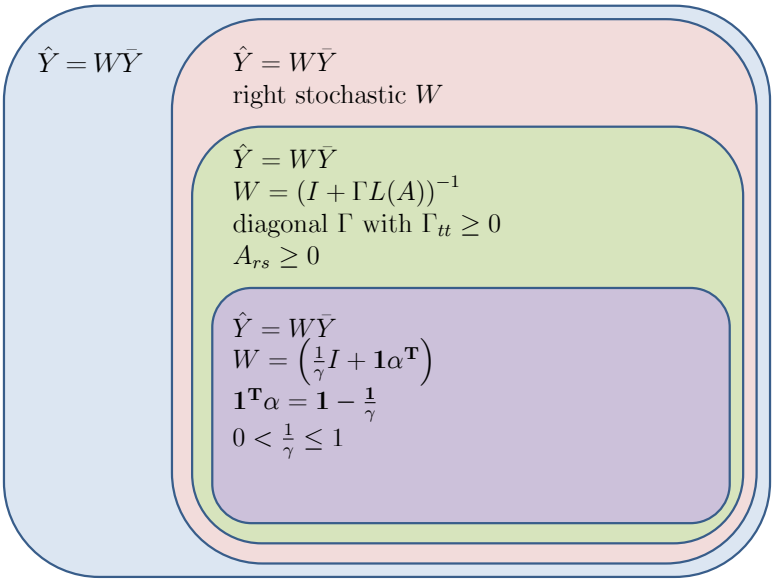

Figure 3: A Venn diagram of the set membership properties of various estimators of the type $\hat{Y} = W\bar{Y}$.

Corollary 5 *Estimators that regularize the single task estimate towards the pooled mean such that they can be written*

$$\check{Y}_t = \lambda \bar{Y}_t + \frac{1-\lambda}{\sum_{r=1}^{T} N_r} \sum_{s=1}^{T} \sum_{i=1}^{N_s} Y_{si},$$

for $\lambda \in (0, 1]$ *can also be written in MTA form as*

$$\check{Y} = \left(I + \frac{1-\lambda}{\lambda \mathbf{N}^\top \mathbf{1}} L(\mathbf{1}\mathbf{N}^\top) \right)^{-1} \bar{Y},$$

where $\mathbf{N}$ *is a* T *by* 1 *vector with* N_t *as its* t*th entry since in Proposition 4 we can choose* $\gamma = \frac{1}{\lambda}$ *and* $\alpha = \frac{1-\lambda}{\mathbf{N}^T \mathbf{1}} \mathbf{N}$*, which matches (20) with* $\Gamma = \frac{1-\lambda}{\lambda \mathbf{N}^\top \mathbf{1}} I$ *and* $A = \mathbf{1}\mathbf{N}^\top$*.*

Corollary 6 *Estimators which regularize the single task estimate towards the average of means such that they can be written*

$$\breve{Y}_t = \lambda \bar{Y}_t + \frac{1-\lambda}{T} \sum_{t=1}^{T} \bar{Y}_t,$$

for $\lambda \in (0, 1]$*, can also be written in MTA form as*

$$\breve{Y} = \left(I + \frac{1-\lambda}{\lambda T} L(\mathbf{1}\mathbf{1}^\top) \right)^{-1} \bar{Y},$$

since in Proposition 4 we can choose $\gamma = \frac{1}{\lambda}$ *and* $\alpha = \frac{1-\lambda}{T}\mathbf{1}$*, which matches (20) with* $\Gamma = \frac{1-\lambda}{\lambda T} I$ *and* $A = \mathbf{1}\mathbf{1}^\top$.

Note that the proof of the proposition in the appendix uses MTA form with *asymmetric* similarity matrix A. The MTA form with asymmetric A arises if you replace the symmetric MTA regularization term in (4) with the following asymmetric regularization term as follows:

$$\frac{1}{2}\sum_{r=1}^{T}\sum_{s=1}^{T} A_{rs}(\tilde{Y}_r - \tilde{Y}_s)^2 + \frac{1}{2}\sum_{r=1}^{T}\left(\sum_{s=1}^{T} A_{rs}\right)\tilde{Y}_r^2 - \frac{1}{2}\sum_{r=1}^{T}\left(\sum_{s=1}^{T} A_{sr}\right)\tilde{Y}_r^2.$$

Lastly, we make a note about the sum of the mean estimates for the different estimators of Figure 3. In general, the sum of the estimates $\hat{Y} = W\bar{Y}$ for right-stochastic W may differ from the sum of the sample means, because $\mathbf{1}^\top W\bar{Y} \neq \mathbf{1}^\top\bar{Y}$ for all right-stochastic W. But in the special case of Bock's positive-part James-Stein estimator the sum is preserved:

Proposition 7

$$\mathbf{1}^\top\hat{Y}^{JS} = \mathbf{1}^\top\bar{Y}, \tag{21}$$

where $\hat{Y}^{JS}$ *is given in (7).*

We illustrate this property in the Kings' reigns experiments in Table 6.6.

4.9 Bayesian Interpretation of MTA

The MTA estimates from (4) can be interpreted as jointly maximizing the likelihood of T Gaussian distributions with a joint Gaussian Markov random field (GMRF) prior (Rue and Held, 2005) on the solution. In MTA, the precision matrix (the inverse covariance of the GMRF prior) is L, the graph Laplacian of the similarity matrix, and is thus positive semi-definite (and not strictly positive definite); GMRFs with PSD inverse covariances are called intrinsic GMRFs (IGMRFs).

GMRFs and IGMRFs are commonly used in graphical models, wherein the sparsity structure of the precision matrix (which corresponds to conditional independence between variables) is exploited for computational tractability. Because MTA allows for arbitrary but non-negative similarities between any two tasks, the precision matrix does not (in general) have zeros on the off-diagonal, and it is not obvious how additional sparsity structure of L would be of help computationally.

Additionally, none of the results we show in this paper require a Gaussian assumption nor any other assumption about the parametric form of the underlying distribution.

5. Simulations

As we have shown in the previous section, MTA is a theoretically rich formulation. In the next two sections we test the usefulness of MTA Constant and MTA Minimax given data, first with simulations, then with real data. In these sections we use lower-case notation to indicate that we are dealing with actual data as opposed to random variables.

In this section, we test estimators using simulations so that comparisons to ground truth can be made. The simulated data was generated from either a Gaussian or uniform

hierarchical process with many sources of randomness (detailed below), in an attempt to imitate the uncertainty of real applications, and thereby determine if these are good general-purpose estimators. The reported results demonstrate that MTA works well averaged over many different draws of means, variances, and numbers of samples.

Simulations are run for $T = \{2, 5, 25, 500\}$ tasks, and parameters were set so that the variances of the distribution of the true means are the same in both uniform and Gaussian simulations. Simulation results are reported in Figures 4 and 5 for the Gaussian experiments, and Figures 6 and 7 for the uniform experiments. The Gaussian simulations were run as follows:

1. Fix σ_μ^2, the variance of the distribution from which $\{\mu_t\}$ are drawn.
2. For $t = 1, \ldots, T$:
   (a) Draw the mean of the tth distribution μ_t from a Gaussian with mean 0 and variance σ_μ^2.
   (b) Draw the variance of the tth distribution $\sigma_t^2 \sim \text{Gamma}(0.9, 1.0) + 0.1$, where the 0.1 is added to ensure that variance is never zero.
   (c) Draw the number of samples to be drawn from the tth distribution N_t from an integer uniform distribution in the range of 2 to 100.
   (d) Draw N_t samples $Y_{ti} \sim \mathcal{N}(\mu_t, \sigma_t^2)$.

The uniform simulations were run as follows:

1. Fix σ_μ^2, the variance of the distribution from which $\{\mu_t\}$ are drawn.
2. For $t = 1, \ldots, T$:
   (a) Draw the mean of the tth distribution μ_t from a uniform distribution with mean 0 and variance σ_μ^2.
   (b) Draw the variance of the tth distribution $\sigma_t^2 \sim U(0.1, 2.0)$.
   (c) Draw the number of samples to be drawn from the tth distribution N_t from an integer uniform distribution in the range of 2 to 100.
   (d) Draw N_t samples $Y_{ti} \sim U[\mu_t - \sqrt{3\sigma_t^2}, \mu_t + \sqrt{3\sigma_t^2}]$.

We compared MTA Constant and MTA Minimax to single-task sample averages and to Bock's James-Stein estimator (Bock, 1975) given in (7), with a slight adaptation for better performance. The term $\frac{\mathbf{tr}(\Sigma)}{\lambda_{\max}}$ in (7) is called the *effective dimension* of the estimator. In simulations where we set Σ to be the true covariance matrix and then estimated the effective dimension by estimating the maximum eigenvalue and trace of the sample mean covariance matrix, we found that replacing the effective dimension with the number of tasks T (when Σ is diagonal) resulted in a significant performance boost for Bock's estimator, due to the high variance of the estimated maximum eigenvalue in the denominator of the effective dimension. Preliminary experiments with real data also showed a performance advantage to using T rather than the effective dimension. Consequently, to present James-Stein estimation in its best light, for all of the experiments in this paper, the James-Stein comparison refers to (7) using T instead of the effective dimension.

James-Stein, MTA Constant and MTA Minimax all self-estimate the amount of regularization to use (for MTA Constant and MTA Minimax the paramtere $\gamma = 1$). So we also compared to a 50-50 randomized cross-validated (CV) version of each. For the cross-validated versions, we randomly subsampled $N_t/2$ samples and chose the value of γ for MTA Constant/Minimax or λ for James-Stein that resulted in the lowest average left-out risk compared to the sample mean estimated with *all* N_t samples. In the optimal versions of MTA Constant/Minimax γ was set to 1, as this was the case during derivation. Note that the James-Stein formulation with a cross-validated regularization parameter λ is simply a convex regularization towards the average of the sample means:

$$\lambda \bar{y}_t + (1 - \lambda)\bar{\bar{y}}.$$

We used the following parameters for CV: $\gamma \in \{2^{-5}, 2^{-4}, \ldots, 2^5\}$ for the MTA estimators and for cross-validated James-Stein a comparable set of λ spanning $(0, 1)$ by the transformation $\lambda = \frac{\gamma}{\gamma+1}$. Even when cross-validating the regularization parameter for MTA, an advantage of using the proposed MTA Constant or MTA Minimax is that these estimators provide a data-adaptive scale for γ, where $\gamma = 1$ sets the regularization parameter to be $\frac{a^*}{T}$ or $\frac{1}{T(b_u - b_l)^2}$, respectively.

Some observations from Figures 4-7:

- Further to the right on the x-axis the means are more likely to be further apart, and multi-task approaches help less on average compared to the single-task sample means.
- For $T = 2$, the James-Stein estimator reduces to the single-task estimator. The MTA estimators provide a gain while the means are close with high probability (that is, when $\sigma_\mu^2 < 1$) but deteriorate quickly thereafter.
- For $T = 5$, MTA Constant dominates in the Gaussian case, but in the uniform case does worse than single-task when the means are far apart. For all $T > 2$, MTA Minimax almost always outperforms James-Stein and always outperforms single-task, which is to be expected as it was designed conservatively.
- The $T = 25$ and $T = 500$ cases illustrate that all estimators benefit from an increase in the number of tasks. The difference between $T = 25$ performance and $T = 500$ performance is minor, indicating that benefit from jointly estimating a larger number of tasks together levels off early on.
- For MTA Constant, cross-validation is always worse than the estimated optimal regularization, while the opposite is true for MTA Minimax. This is to be expected, as minimax estimators are not designed to minimizes the average risk, but average risk is the metric optimized during cross-validation and is the metric reported.
- Cross-validating MTA Constant or MTA Minimax should result in similar performance, and this can be seen in the figures where the green and blue dotted lines are superimposed. The performance differs slightly because the discrete set of γ choices multiply different a's for the MTA Constant and MTA Minimax.

In summary, when the tasks are close to each other compared to their variances, MTA Constant is the best estimator to use by a wide margin. When the tasks are farther apart, MTA Minimax provides a win over both James-Stein and sample averages.

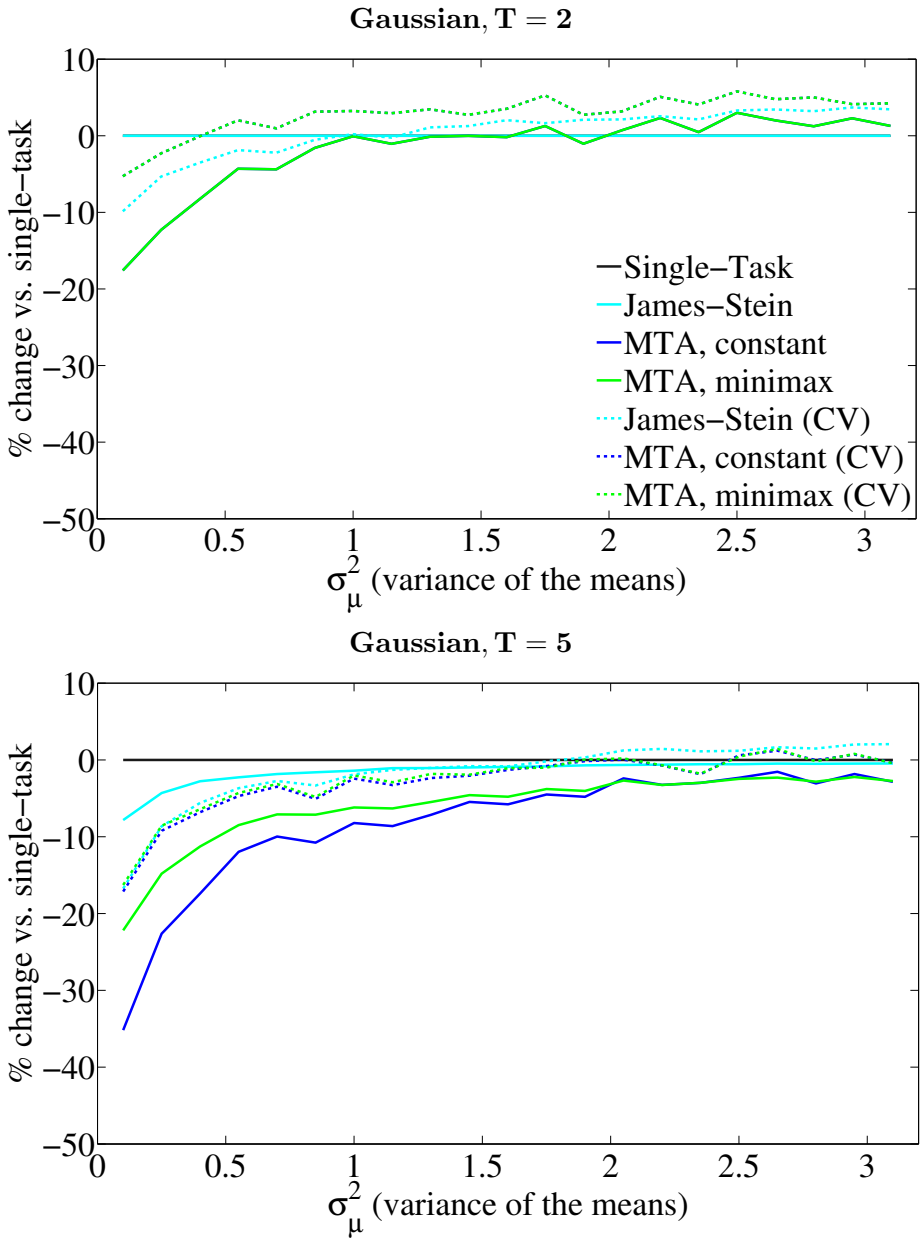

Figure 4: Gaussian experiment results for $T = \{2, 5\}$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task. Note: for $T = 2$ the James-Stein estimator reduces to single-task, and so the cyan and black lines overlap. Similarly, for $T = 2$, MTA Constant and MTA Minimax are identical, and so the blue and green lines overlap.

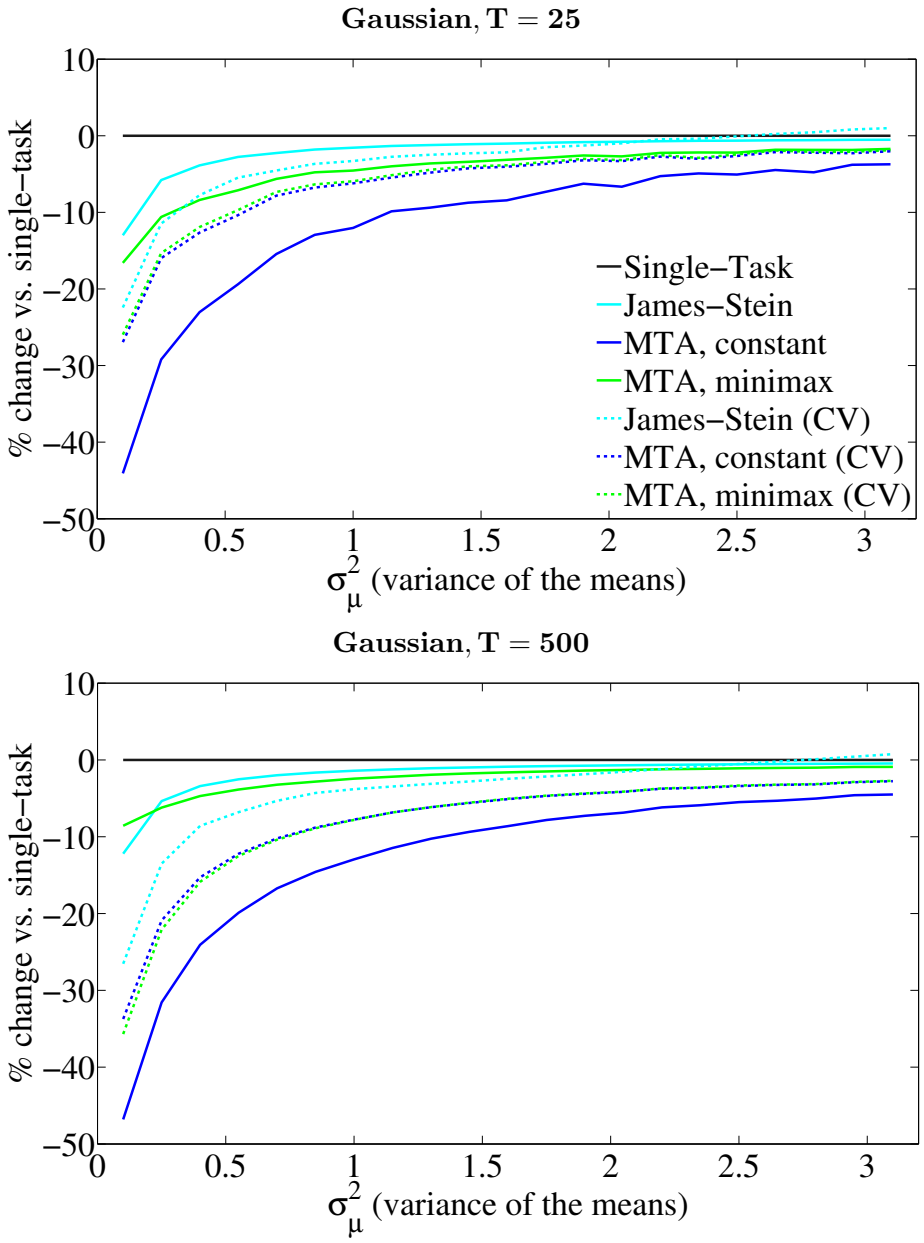

Figure 5: Gaussian experiment results for $T = \{25, 500\}$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task.

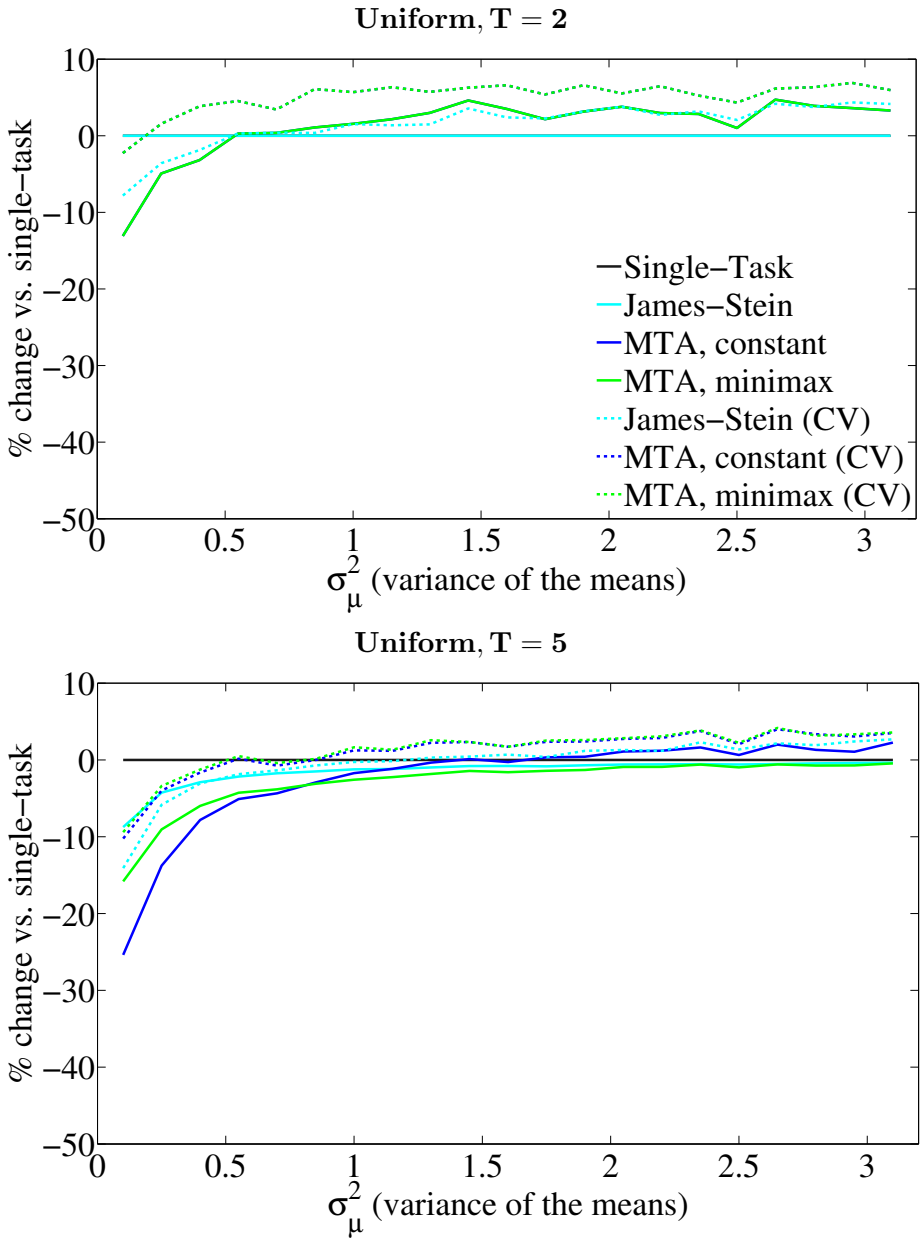

Figure 6: Uniform experiment results for $T = \{2, 5\}$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task. Note: for $T = 2$ the James-Stein estimator reduces to single-task, and so the cyan and black lines overlap. Similarly, for $T = 2$, MTA Constant and MTA Minimax are identical, and so the blue and green lines overlap.

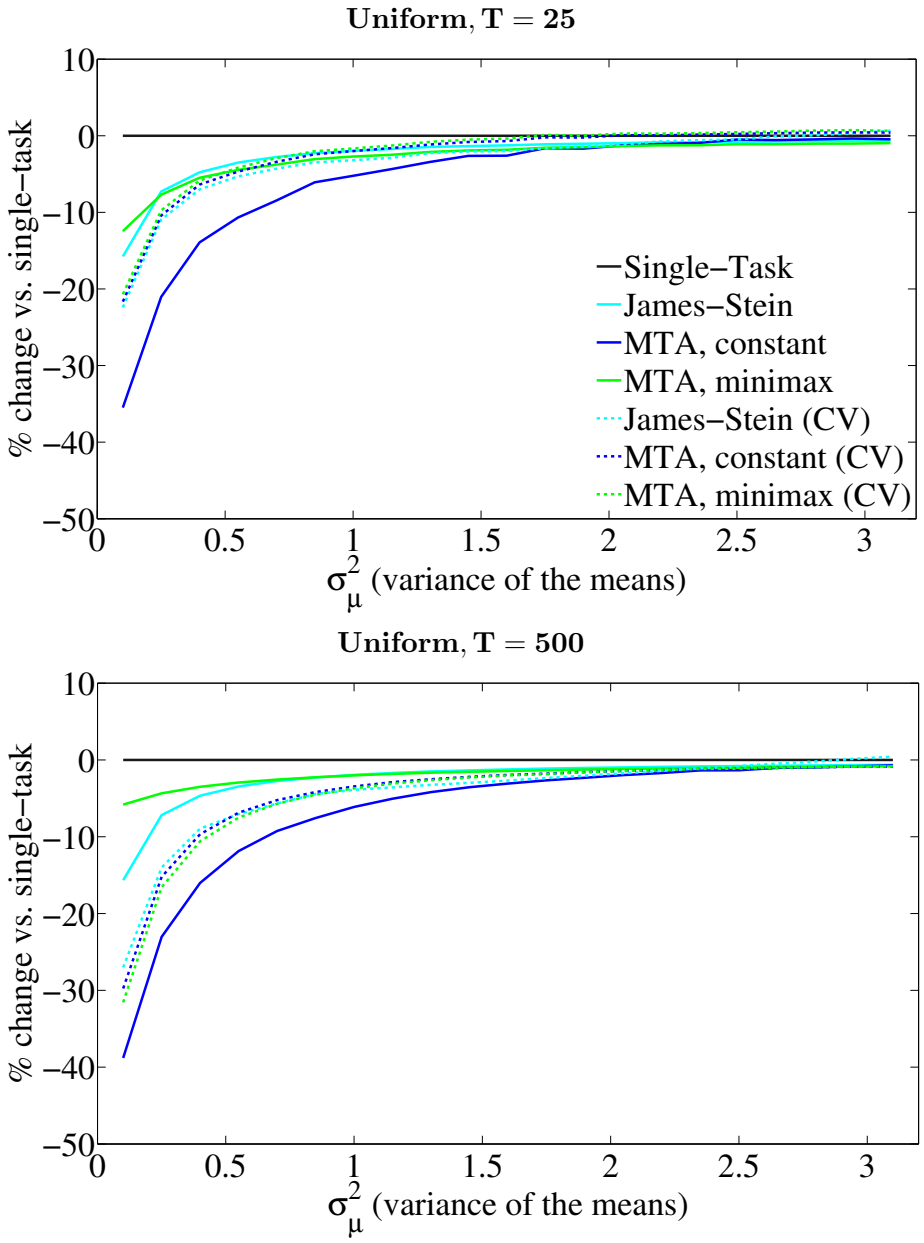

Figure 7: Uniform experiment results for for $T = \{25, 500\}$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task.

5.1 Oracle Performance

To illustrate the best performance we know is possible to achieve with MTA, Figure 8 shows the effect of using the true "oracle" means and variances for the calculation of optimal pairwise similarities for $T > 2$:

$$A_{rs}^{\text{orcl}} = \frac{2}{(\mu_r - \mu_s)^2}. \tag{22}$$

This matrix is the best *pairwise* oracle, but does not generally minimize the risk over all possible A for $T > 2$. However, comparing to it illustrates how well the MTA formulation can do, without the added error due to estimating A from the data.[3]:

Figure 8 reproduces the results from the $T = 5$ Gaussian simulation (excluding cross-validation results), and compares to the performance of oracle pairwise MTA using (22). Oracle MTA is over 30% better than MTA Constant, indicating that practical estimates of the similarity are highly suboptimal compared to the best possible MTA performance, and motivating better estimates of A as a direction for future research.

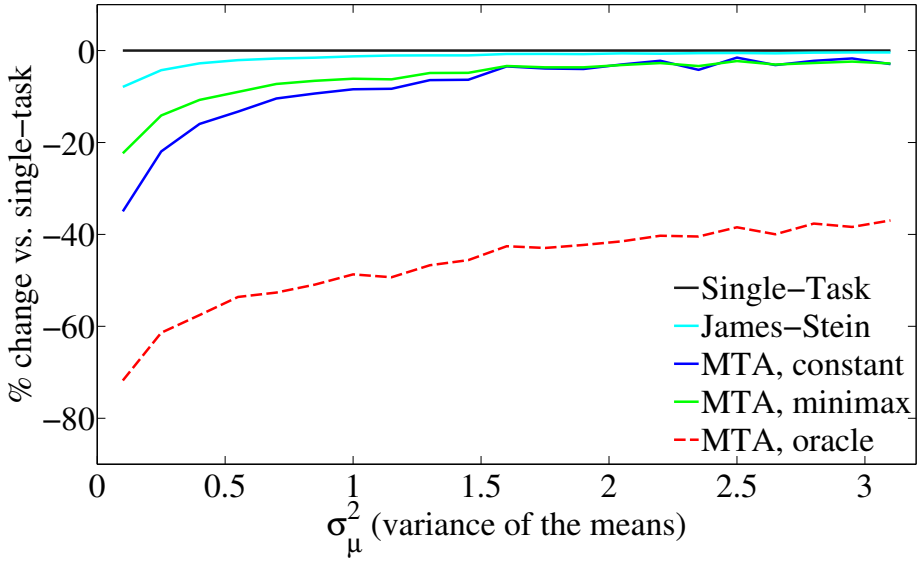

Figure 8: Average (over 10000 random draws) percent change in risk vs. single-task with $T = 5$ for the Gaussian simulation. Oracle MTA uses the true means and variance to specify the weight matrix W.

6. Real Data Experiments

We present four real data experiments,[4] comparing eight estimators on both goals (2) and (3). The first experiment estimates future customer reviews based on past customer

3. Preliminary experiments (not reported) showed that for $T > 2$ estimating A pairwise as $\hat{A}_{rs} = \frac{2}{(\bar{y}_r - \bar{y}_s)^2}$ was almost always worse than constant MTA.
4. Research-grade Matlab code and the data used in these experiments can be found here.

reviews. The second experiment estimates final grades based on homework grades. The third experiment forecasts a customer's future order size based on the size of their past orders. The fourth experiment takes a more in-depth look at the estimates produced by these methods for the historical problem of estimating the length of a king's reign.

6.1 Metrics

For all the experiments except estimating final grades, we only have sample data, and so we compare the estimators using a metric that is an empirical approximation to the regression error defined in (3). First, we replace the expectation in (3) with a sum over the samples. Second, we measure the squared error between a sample y_{ti} and an estimator formed without that sample, $\hat{y}_{t\setminus y_{ti}}$. That is, the empirical risk we measure is:

$$\sum_{t=1}^{T}\left(\frac{1}{N_t}\sum_{i=1}^{N_t}\left[(y_{ti}-\hat{y}_{t\setminus y_{ti}})^2\right]\right). \tag{23}$$

To make the results more comparable across datasets, we present the results as the percent the error given in (23) is reduced compared to the single-task sample mean estimate.

6.2 Experimental Details

For the cross-validation estimators, we cross-validate the regularization parameter from the set $\{2^{-15}, 2^{-14}, \ldots, 2^{14}, 2^{15}\}$. This is a larger range of cross-validation values than used in the simulations, but we found that necessary to achieve good results with cross-validation in the real data experiments. Cross-validation parameters were chosen using double-leave-one-out cross-validation (for each sample left out for test, the remaining N-1 samples undergo leave-one-out cross-validation to optimize (23)). For real-data experiments with more than 50 tasks, to make the double leave-one-out cross-validation fast enough to be feasible, we randomly sub-sampled uniformly and independently for each held-out sample 50 tasks for the estimation of the regularization parameter (but all tasks were used in all cases for the actual estimates).

In addition to James-Stein, MTA, and their variants, we also compare to the completely-regularized baseline, the pooled mean estimator:

$$\hat{y}_t^{\text{pooled}} = \bar{\bar{y}} = \frac{1}{TN}\sum_{s=1}^{T}\sum_{i=1}^{N} y_{si}, \tag{24}$$

which estimates the same value for each task.

For each experiment, a single pooled variance estimate when needed was used for all tasks: $\sigma_t^2 = \sigma^2$, for all t. We found that using a pooled variance estimate improved performance for all the estimators compared.

6.3 Estimating Customer Reviews for Amazon Products

We model amazon.com customer reviews for a product as iid random draws from an unknown distribution. We scraped customer review scores (ranging from 1 to 5) for four different product types from the amazon.com website, as detailed in Table 6.3. We treat

each product as a task, and jointly estimate the mean reviews for all products of the same type. The eight estimators are compared to see how well they predict held-out customer reviews, as per (23); a lower (more negative) score corresponds to greater percent reduction in risk compared to the sample mean estimates.

	# of Products	Mean # of Reviews	Range of # of Reviews
Machine Learning Books	156	7.7	2–80
Blue Suede Shoes	37	16.2	2–143
Espresso Machines	277	47.1	2–1788
Robot Vacuums	59	137.1	3–883

Table 2: Products used in customer reviews experiments, ordered by mean number of reviews (that is, mean sample size).

Table 6.3 shows the percent risk reduction for each estimator compared to single-task estimates. Some observations:

- MTA Constant (no cross-validation) has the best risk reduction averaged across the products at 11.9% average risk reduction over the single-task estimates, slightly better than the cross-validated forms of MTA.

- The average MTA Constant risk reduction is 34% better than JS (11.9% vs 8.9%), and MTA Constant is better than JS on all the datasets.

- On all datasets, all the joint estimators (not including the pooled mean baseline) do better than the single-task estimates except JS CV on the robot vacuums dataset, showing that joint estimation usually helps.

- MTA Minimax consistently provides small gains over single-task, on average reducing risk by 4.0%, with the lowest standard deviation of improvement of 2.1.

- The JS estimator is more sensitive to the quality of the pooled mean estimate than the MTA Constant estimator.

- JS does better on average than its cross-validated counterpart JS CV, and MTA Constant does better on average than its cross-validated counterpart MTA Constant CV.

- The rows in Table 6.3 are ordered by the average number of reviews (that is, the average number of samples per task). As one would expect from theory, the gains are larger if there are fewer reviews per task.

- Mixing un-related products (the last row of Table 6.3) still produces substantial gains over single-task estimates.

	Pooled Mean	JS	JS CV	MTA Constant	MTA Constant CV	MTA Minimax	MTA Minimax CV
ML Books	**-24.6**	-23.1	-22.9	**-24.6**	-23.3	-6.5	-23.1
Blue Suede Shoes	-12.4	-11.5	-10.6	**-12.5**	-11.6	-4.8	-11.6
Espresso Machines	2.7	-3.7	-6.3	**-8.4**	-7.8	-3.6	-8.3
Robot Vacuums	8.7	-0.7	7.3	**-2.5**	-2.2	-0.8	-1.8
All Products	-1.9	-5.4	-9.3	**-11.3**	-11.0	-4.3	-10.7
Average	-5.5	-8.9	-8.4	**-11.9**	-11.2	-4.0	-11.1
STD	13.2	8.9	10.8	8.1	7.7	**2.1**	7.7

Table 3: Percent change in risk vs. single-task for customer reviews experiment (lower is better). 'JS' denotes James-Stein, 'CV' denotes cross-validation, and 'STD' denotes standard deviation.

6.4 Estimating Final Grades from Homework Grades

We model homework grades as random samples drawn iid from an unknown distribution where the mean for each student is that student's final class grade. We compare the eight estimators to see how well they predict each student's final grade given only their homework grades. Final class grades are based on the homeworks, but also on projects, labs, quizzes, exams and sometimes class participation, with the mix varying by class. We collected 22 anonymized datasets from six different instructors at three different universities for undergraduate electrical engineering classes. Further experimental details:

- Each of the 22 datasets is for a different class, and constitutes a single experiment, where each student corresponds to a task.
- We treat the ith homework grade of the tth student as sample y_{ti}.
- For each class and each cross-validation method, cross-validation parameters were chosen independently using leave-one-out cross-validation on the homework grades.
- For each class, the error measurement for estimator $\hat{y}$ is the sum of squared errors across all T students:
$$\sum_{t=1}^{T} (\mu_t - \hat{y}_t)^2,$$
where μ_t is the given tth student's final grade.

Table 6.4 compares the estimators in terms of the percent change in error compared to the single task estimate $\bar{y}_t$. A lower (more negative) score corresponds to greater percent reduction in risk compared to the single task estimates.

Class Size	Pooled Mean	JS	JS CV	MTA Constant	MTA Constant CV	MTA Minimax	MTA Minimax CV
16	26.3	0.7	**-0.0**	0.6	**-0.0**	**-0.0**	**-0.0**
20	71.2	−3.2	**-5.2**	−4.7	−3.4	−1.7	−4.6
25	776.9	−12.2	**-12.3**	−12.2	−12.2	−2.7	−12.1
29	−7.6	−11.6	−31.2	−11.4	**-35.2**	−1.8	−29.6
34	373.6	−4.9	−12.4	−5.0	−12.7	−1.1	**-13.3**
36	**-28.3**	−17.4	−0.0	−16.0	−0.0	−2.8	−0.0
39	42.0	**-5.8**	−0.0	−5.6	−0.0	−0.9	−0.0
44	3.0	−47.6	−64.5	−42.7	−68.0	−7.0	**-69.0**
45	127.6	−3.0	−0.0	**-19.2**	−0.0	−4.6	−0.0
47	**-12.8**	−8.0	−0.0	−7.1	−0.0	−0.7	−0.0
48	**-21.0**	−20.5	−0.0	−18.5	−0.0	−2.5	−0.0
50	63.5	63.5	−0.0	9.3	−0.0	**-4.4**	−0.0
50	3.7	−33.6	−41.5	−29.7	−42.4	−3.2	**-47.4**
57	23.3	**-3.8**	−0.0	−3.6	−0.0	−0.4	−0.0
58	−0.2	**-16.3**	−0.0	−15.6	−0.0	−2.8	−0.0
65	45.0	**-29.4**	−0.0	−26.2	−0.0	−4.2	−0.0
68	−16.9	**-45.5**	−16.5	−39.0	−17.0	−6.1	−19.8
69	−14.7	**-41.0**	−14.7	−39.8	−14.7	−4.5	−14.8
72	34.6	−32.9	−27.3	−29.0	−27.8	−4.0	**-34.8**
73	224.2	−28.1	−41.1	−26.4	−39.6	−2.4	**-41.2**
110	5.7	−14.8	**-25.3**	−13.4	−20.6	−1.2	−22.0
149	**-16.6**	−11.7	−0.0	−10.1	−0.0	−0.8	−0.0
Average	77.4	−14.9	−13.3	**-16.6**	−13.3	−2.7	−14.0
STD	182.0	22.7	18.1	13.7	18.7	**1.9**	19.4

Table 4: Percent change in risk vs. single-task for the grade estimation experiment (lower is better). 'JS' denotes James-Stein, 'CV' denotes cross-validation, and 'STD' denotes standard deviation.

Some observations:

- MTA Constant (no cross-validation) has the best average risk reduction, at 16.6% better on average than the standard single-task estimate. The standard deviation of the win over single task for MTA Constant is 13.7% - also lower than any of the other estimators except MTA Minimax. This shows MTA Constant is consistently providing good error reduction.
- MTA Minimax consistently provides small gains, as designed, with low variance.
- Once again, the higher variance of the James-Stein estimator compared to the others is because of the positive-part aspect of the JS estimator – when the positive-part boundary is triggered, JS reduces to the one-task (average-of-means) estimator, which can have poor performance.
- JS does better on average than its cross-validated counterpart JS CV, and MTA Constant does better on average than its cross-validated counterpart MTA Constant CV.

6.5 Estimating Customer Spending

We collaborated with the wooden jigsaw puzzle company Artifact Puzzles to estimate how much each repeat customer would spend on their next order. We treated each customer as a task; in the time period spanned by the data there are $T = 1355$ unique customers who have each purchased at least twice. We modelled each order by a customer as an iid draw from that customer's unknown spending distribution. The number of orders per customer (that is, samples per task) ranged from 2-23, with a mean of 3.03 orders per customer. The amount spent on a given order had a rather long tail distribution, ranging from \$9-\$2403, with a mean of \$82.16.

Results are shown in Table 6.5, showing the percentage reduction in (23) compared to the single-task sample means.

Some observations from Table 6.5:

- MTA Constant performed slightly better than the James-Stein estimator, reducing the empirical risk by 22.4% rather than 21.1%.
- JS does better than its cross-validated counterpart JS CV, and MTA Constant does better than its cross-validated counterpart MTA Constant CV.

6.6 Estimating the Length of Kings' Reigns

To illustrate the differences between the actual estimates, we re-visit an estimation problem studied by Isaac Newton, "How long does the average king reign?" (Newton, 1728; Stigler, 1999). Newton considered 9 different kingdoms, from the Kings of Judah to more recent French kings. Our dataset covers 30 well-known dynasties, listed in Table 6.6, from ancient to modern times, and spread across the globe. All data was taken from wikipedia.org in August and September 2013 (see the linked data files for the raw data and more historical details).

	Pooled Mean	JS	JS CV	MTA Constant	MTA Constant CV	MTA Minimax	MTA Minimax CV
Customer Spending	-10.6	-21.1	-17.6	**-22.4**	-19.7	-0.6	-19.5

Table 5: Percent change in risk vs. single-task for the customer spending experiments (lower is better). 'JS' denotes James-Stein, 'CV' denotes cross-validation.

	Pooled Mean	JS	JS CV	MTA Constant	MTA Constant CV	MTA Minimax	MTA Minimax CV
Kings' Reigns	-8.2	-8.7	-4.7	**-8.9**	-2.9	-3.1	-3.2

Table 6: Percent change in risk vs. single-task for the kings' reigns experiments (lower is better). 'JS' denotes James-Stein, 'CV' denotes cross-validation.

Results are shown in Table 6.6, showing the percentage reduction in (23) compared to the single-task sample means. Some observations about these results:

- The pooled mean is 8.2% better than estimating each dynasty's average separately. We found it surprising that pooling across cultures and history forms overall better estimates: the fate of man is apparently the fate of man, regardless of whether it is 1000 BC in Babylon or 19th century Denmark.
- The JS and MTA Constant estimators achieve a slightly bigger reduction in squared error compared to the pooled mean.
- The MTA Contant estimator is very slightly better than the JS estimator, -8.9% vs -8.7%.
- The JS and MTA estimators do better than their cross-validated counterparts.

Dynasty	# Kings	Avg.	Pooled Mean	JS	JS CV	MTA Const.	MTA Const. CV	MTA MM	MTA MM CV
Larsa	15	17.7	19.5	19.2	18.5	18.3	18.1	17.8	18.1
Amorite	11	26.9	19.5	22.3	24.6	24.6	25.5	26.5	25.6
Assyrian	27	17.3	19.5	19.1	18.2	17.8	17.6	17.4	17.6
Israel	21	13.4	19.5	17.7	15.6	14.8	14.2	13.6	14.1
Judah	22	21.5	19.5	20.5	21.0	21.2	21.3	21.5	21.4
Achaemenid	9	24.3	19.5	21.4	22.9	22.4	23.1	23.9	23.2
Khmer	33	20.0	19.5	20.0	20.0	20.0	20.0	20.0	20.0
Song	18	17.7	19.5	19.2	18.5	18.3	18.0	17.8	18.0
Mongol	4	10.8	19.5	16.8	13.8	16.1	14.5	12.0	14.3
Ming	17	16.3	19.5	18.7	17.5	17.2	16.8	16.4	16.8
Qing	12	24.6	19.5	21.6	23.0	23.1	23.7	24.4	23.8
Mamluk	10	10.1	19.5	16.6	13.4	13.6	12.3	10.7	12.1
Ottoman	36	17.0	19.5	19.0	18.0	17.4	17.2	17.1	17.2
Normandy	3	23.0	19.5	21.0	22.0	21.1	21.6	22.5	21.7
Plantagenet	8	30.8	19.5	23.7	27.2	26.4	28.0	30.0	28.2
Lancaster	3	20.3	19.5	20.1	20.2	20.1	20.2	20.3	20.2
York	3	8.0	19.5	15.9	12.0	15.8	13.8	10.1	13.4
Tudor	5	23.4	19.5	21.1	22.3	21.6	22.2	23.0	22.3
Stuart	6	16.8	19.5	18.9	17.9	18.4	17.8	17.1	17.8
Hanover	6	31.0	19.5	23.7	27.3	25.7	27.5	30.0	27.8
Windsor	3	14.0	19.5	17.9	16.0	17.9	16.9	15.0	16.7
Capet	15	22.7	19.5	20.9	21.8	21.9	22.3	22.6	22.3
Valois	7	24.3	19.5	21.4	22.9	22.4	23.1	23.9	23.2
Habsburg	5	34.4	19.5	24.9	29.6	26.8	29.3	32.8	29.6
Bourbon	10	21.8	19.5	20.6	21.2	21.2	21.4	21.7	21.4
Oldenburg	16	25.8	19.5	22.0	23.9	24.3	25.0	25.6	25.0
Mughal	20	15.7	19.5	18.5	17.1	16.6	16.2	15.8	16.2
Edo	15	18.6	19.5	19.5	19.1	19.0	18.8	18.7	18.8
Kamehameha	5	15.4	19.5	18.4	16.9	17.8	17.1	15.9	16.9
Zulu	4	15.8	19.5	18.5	17.2	18.2	17.5	16.3	17.4
Average Over Dynasties		19.98	19.49	19.98	19.98	20.00	20.03	20.01	20.04

Table 7: Sample average and eight other estimators of the expected length of the reign of a king for each dynasty, ordered chronologically. 'JS' denotes James-Stein, 'CV' denotes cross-validation, 'Const.' denotes Constant, and 'MM' denotes Minimax.

We also give the actual estimators of the average length of the reign for each kingdom in Table 6.6. Some observations from Table 6.6:

- Table 6.6 shows that while all the estimators regularize the single task mean (given in column 1) to the pooled mean (given in column 2), the actual estimates can differ quite a bit. For example, MTA Constant and MTA Minimax differ by 5 years in their estimates of the average length of reign of a king from the House of York.

- One sees that the JS estimates are regularized harder towards the pooled mean of 19.5 than the MTA Constant estimates. The MTA Minimax estimates are (as expected) least changed from the task means.

- The last row of Table 6.6 shows the estimates averaged over the different dynasties. Note that the JS and JS CV estimators have the same average across the tasks (dynasties) as the single-task average, as expected from Proposition 7.

- Based on Tables 6.5 and 6.6, we estimate the expected length of a king's reign to be the dynasty-averaged MTA Constant estimate of 20.00 years. Newton's wrote his estimate as "eighteen or twenty years" (Newton, 1728), and the analysis of Stigler (1999) of Newton's data shows that the maximum likelihood estimate from his data was a more pessimistic 19.03 years.

7. Conclusions And Open Questions

We conclude with a summary and then some open questions.

7.1 Summary

We proposed a simple additive regularizer to jointly estimate multiple means using a pairwise task similarity matrix A. Our analysis of the $T = 2$ task case establishes that both MTA estimates are better than the individual sample means when the separation between the true means is small relative to the variance of the samples from each distribution. For the two-task case, we provide a formula for the optimal pairwise task similarity matrix A, that is, one can analytically estimate the optimal amount of regularization without the need to cross-validate or tune a regularization hyper-parameter. We generalized that formula to multiple tasks to form the practical and computationally-efficient MTA Constant mean estimator, as well as a more conservative minimax variant. Simulations and four sets of real data experiments show the MTA Constant estimator can substantially reduce errors over the sample means, and generally performs slightly better than James-Stein estimation (which also does not require cross-validation).

One can also cross-validate the amount of regularization in the MTA formula or in the James-Stein formula. Our results show that both cross-validations work well, though in both simulations and real data experiments, MTA Constant performed slightly better or comparable to the cross-validations.

7.2 Open Questions

Averaging is common, and MTA has potentially broad applicability as a subcomponent to the many algorithms that use means as a subroutine, such as k-means clustering, kernel density estimation, or non-local means denoising.

Most multi-task learning formulations contain an explicit or implicit dependence on the pairwise similarity between tasks. For MTA, this is the A matrix. Even when side information about task similarities is available, it may not be in the optimal numerical form. This paper shows good performance with the assumption that A has constant entries, where that constant is the average of pairwise similarities estimated based on the sample means (MTA Constant). However, the oracle performance plots in Section 5 show that the right choice of A can perform much better. Estimating all $T \times T$ parameters of A optimally may be difficult, but we hypothesize that other structured assumptions (e.g. low rank A) might perform better than our constant approximation. Martínez-Rego and Pontil (2013) have shown some promising results by clustering tasks in a pre-processing stage.

We focused in this paper on estimating scalar means. The extension to vectors is straightforward (see Section 4.2 and Martínez-Rego and Pontil (2013)). However, how well the vector extension works in practice, how to best estimate the block diagonal covariance matrix, and whether different regularization norms would be better remain open questions. A further extension is when the samples themselves are distributions, and the task means to be estimated are expected distributions (Frigyik et al., 2008).

We showed in Section 4 that computing the MTA Constant and MTA Minimax estimators can be done in $O(T)$ time for T tasks. Simulations showed that the achievable gains generally go up slowly with the number of tasks T, with $T = 500$ producing an average risk reduction of 40% in the extreme case that the true means for the 500 tasks were the same. In the real data experiment on customer spending, there were $T = 1355$ tasks that produced a risk reduction of 22.4%. Larger-scale experiments and analysis of the effect of large T on the error would be intriguing.

We focused on squared error loss and the graph Laplacian regularizer because they are standard, generally work well, and are easy to analyze. But re-considering the MTA objective with other loss functions and regularizers might lead to interesting new perspectives and estimates. Lastly, we hope that some of the analyses and results in this paper inspire further theoretical analysis of other multi-task learning methods.

Acknowledgments

This work was funded by a United States PECASE Award and by the United States Office of Naval Research. We thank Peter Sadowski for helpful discussions.

Appendix A: MTA Closed-form Solution

When all A_{rs} are non-negative, the differentiable MTA objective is convex, and admits closed-form solution. First, we rewrite the objective in (4) using the graph Laplacian matrix

$L = D - (A + A^\top)/2$:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sigma_t^2}\sum_{i=1}^{N_t}(Y_{ti}-\tilde{Y}_t)^2 + \frac{\gamma}{T^2}\sum_{r=1}^{T}\sum_{s=1}^{T}A_{rs}(\tilde{Y}_r-\tilde{Y}_s)^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\sigma_t^2}\sum_{i=1}^{N_t}Y_{ti}^2 + \frac{N_t}{\sigma_t^2}\tilde{Y}_t^2 - 2\frac{N_t}{\sigma_t^2}\tilde{Y}_t\bar{Y}_t\right) + \frac{\gamma}{T^2}\tilde{Y}^\top L\tilde{Y} \\
&= \frac{1}{T}\left(\sum_{t=1}^{T}\frac{1}{\sigma_t^2}\sum_{i=1}^{N_t}Y_{ti}^2 + \tilde{Y}^\top\Sigma^{-1}\tilde{Y} - 2\tilde{Y}^\top\Sigma^{-1}\bar{Y}\right) + \frac{\gamma}{T^2}\tilde{Y}^\top L\tilde{Y},
\end{aligned}$$

where, Σ is a diagonal matrix with $\Sigma_{tt} = \frac{\sigma_t^2}{N_t}$, and $\tilde{Y}$ and $\bar{Y}$ are column vectors with tth entries $\tilde{Y}_t$ and $\bar{Y}_t$, respectively.

For simplicity of notation, we assume from now on that A is symmetric. If, in practice, an asymmetric A is provided, it can be symmetrized without loss of generality.

Take the partial derivative of the above objective w.r.t. $\tilde{Y}$ and equate to zero,

$$\begin{aligned}
0 &= \frac{1}{T}\left(2\Sigma^{-1}Y^* - 2\Sigma^{-1}\bar{Y}\right) + 2\frac{\gamma}{T^2}LY^* \\
&= Y^* - \bar{Y} + \frac{\gamma}{T}\Sigma LY^* \\
\bar{Y} &= \left(I + \frac{\gamma}{T}\Sigma L\right)Y^*,
\end{aligned} \tag{25}$$

which yields the following optimal closed-form solution:

$$Y^* = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}\bar{Y}, \tag{26}$$

as long as the inverse exists, which we will prove next.

Appendix B: Proof of Lemma 1

Assumptions: $\gamma \geq 0$, $0 \leq A_{rs} < \infty$ for all r, s and $0 < \frac{\sigma_t^2}{N_t} < \infty$ for all t.

Lemma 1 *The MTA solution matrix* $W = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ exists.

Proof Let $B = W^{-1} = I + \frac{\gamma}{T}\Sigma L$. The (t,s)th entry of B is

$$B_{ts} = \begin{cases} 1 + \frac{\gamma\sigma_t^2}{TN_t}\sum_{s\neq t}A_{ts} & \text{if } t = s \\ -\frac{\gamma\sigma_t^2}{TN_t}A_{ts} & \text{if } t \neq s, \end{cases}$$

The Gershgorin disk (Horn and Johnson, 1990) $\mathcal{D}(B_{tt}, R_t)$ is the closed disk in $\mathbb{C}$ with center B_{tt} and radius

$$R_t = \sum_{s\neq t}|B_{ts}| = \frac{\gamma\sigma_t^2}{TN_t}\sum_{s\neq t}A_{ts} = B_{tt} - 1.$$

One knows that $B_{tt} \geq 1$ for non-negative A and when $\frac{\gamma\sigma_t^2}{TN_t} \geq 0$, as assumed prior to the lemma statement. Also, it is clear that $B_{tt} > R_t$ for all t. Therefore, every Gershgorin disk is contained within the positive half-plane of $\mathbb{C}$, and, by the Gershgorin Circle Theorem (Horn and Johnson, 1990), the real part of every eigenvalue of matrix B is positive. Its determinant is therefore positive, and the matrix B is invertible: $W = B^{-1}$. ■

Appendix C: Proof of Proposition 2

Proposition 2 *As $N_t \to \infty \forall\ t$, $Y^* \to \mu$.*

Proof First note that the (t,t)th diagonal entry of Σ is $\frac{\sigma_t^2}{N_t}$, which approaches 0 as $N_t \to 0$, implying that all entries of $\frac{\gamma}{T}\Sigma L \to 0$ as $N_t \to 0$ as well. Since matrix inversion is a continuous operation, $\left(I + \frac{\gamma}{T}\Sigma L\right)^{-1} \to I$ in the norm.[5] By the law of large numbers one can conclude that Y^* asymptotically approaches the true mean μ.

Note futher that the above proof is only valid for diagonal Σ, but can be easily extended for non-diagonal Σ by noting that $\Sigma_{rs} = \frac{\sigma_r\sigma_s}{\sqrt{N_r N_s}}$ also converges to 0 as $N_r, N_s \to 0$. ■

Appendix D: Proof of Theorem 3

Assumptions: $\gamma \geq 0$, $0 \leq A_{rs} < \infty$ for all r, s and $0 < \frac{\sigma_t^2}{N_t} < \infty$ for all t.

We next state and prove two lemmas that will be used to prove Theorem 3.

Lemma 8 *W has all non-negative entries.*

Proof Because the off-diagonal elements of the graph Laplacian are non-positive, $W^{-1} = \left(I + \frac{\gamma}{T}\Sigma L\right)$ is a *Z-matrix*, defined to be a matrix with non-positive off-diagonal entries (Berman and Plemmons, 1979). If W^{-1} is a Z-matrix, then the following two statements are true and equivalent: "the real part of each eigenvalue of W^{-1} is positive" and "W exists and $W \geq 0$ (elementwise)" (Berman and Plemmons, 1979, Chapter 6, Theorem 2.3, G_{20} and N_{38}). It has already been proven in Lemma 1 that the real part of every eigenvalue of W^{-1} is positive. Therefore, W exists and is element-wise non-negative. ■

Lemma 9 *The rows of W sum to 1, i.e. $W\mathbf{1} = \mathbf{1}$.*

5. Any matrix norm will do since the dimensionality is fixed, and on finite dimensional vector spaces all norms are equivalent and therefore generate the same topology.

Proof As proved in Lemma 1, W exists. Therefore, one can write:

$$\begin{aligned} W\mathbf{1} =& \mathbf{1} \\ \mathbf{1} =& W^{-1}\mathbf{1} \\ =& \left(I + \frac{\gamma}{T}\Sigma L\right)\mathbf{1} \\ =& I\mathbf{1} + \frac{\gamma}{T}\Sigma L\mathbf{1} \\ =& \mathbf{1} + \frac{\gamma}{T}\Sigma\mathbf{0} \\ =& \mathbf{1}, \end{aligned}$$

where the the third equality is true because the graph Laplacian has rows that sum to zero. The rows of W therefore sum to 1. ■

Theorem 3 *The MTA solution matrix* $W = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ *is right-stochastic.*

Proof We know that W exists (from Lemma 1), is entry-wise non-negative (from Lemma 8), and has rows that sum to 1 (from Lemma 9). ■

Appendix E: MTA Constant Derivation

For the case when $T > 2$, analytically specifying a general similarity matrix A that minimizes the risk is intractable. To address this limitation for arbitrary T, we constrain the similarity matrix to be the constant matrix $A = a\mathbf{1}\mathbf{1}^\top$, resulting in the following weight matrix:

$$W^{\text{cnst}} = \left(I + \frac{1}{T}\Sigma L(a\mathbf{1}\mathbf{1}^\top)\right)^{-1}. \tag{27}$$

For tractability, we optimize a using $\mathbf{tr}(\Sigma)I$ rather than the full Σ matrix, such that

$$a^* = \arg\min_a R\left(\mu, \left(I + \frac{1}{T}\frac{\mathbf{tr}(\Sigma)}{T}L(a\mathbf{1}\mathbf{1}^\top)\right)^{-1}\bar{Y}\right), \tag{28}$$

and then plug this a^* into (27) to obtain MTA Constant.

We simplify $\left(I + \frac{1}{T}\frac{\mathbf{tr}(\Sigma)}{T}L(a\mathbf{1}\mathbf{1}^\top)\right)^{-1}$ using the Sherman-Morrison formula,

$$\begin{aligned}
\left(I + \frac{1}{T}\frac{\mathbf{tr}(\Sigma)}{T}L(a\mathbf{1}\mathbf{1}^\top)\right)^{-1} &= \left(I + \frac{a}{T}\frac{\mathbf{tr}(\Sigma)}{T}(TI - \mathbf{1}\mathbf{1}^\top)\right)^{-1} \\
&= \left(I + a\frac{\mathbf{tr}(\Sigma)}{T} - \frac{a}{T}\frac{\mathbf{tr}(\Sigma)}{T}\mathbf{1}\mathbf{1}^\top\right)^{-1} \\
&= \frac{1}{1 + a\frac{\mathbf{tr}(\Sigma)}{T}}I + \frac{\frac{1}{1+a\frac{\mathbf{tr}(\Sigma)}{T}}\frac{a}{T}\frac{\mathbf{tr}(\Sigma)}{T}\mathbf{1}\mathbf{1}^\top\frac{1}{1+a\frac{\mathbf{tr}(\Sigma)}{T}}}{1 - \frac{a}{T}\mathbf{1}^\top\frac{1}{1+a\frac{\mathbf{tr}(\Sigma)}{T}}\frac{\mathbf{tr}(\Sigma)}{T}\mathbf{1}} \\
&= \frac{1}{a\frac{\mathbf{tr}(\Sigma)}{T} + 1}I + \frac{\frac{a\frac{\mathbf{tr}(\Sigma)}{T}}{a\frac{\mathbf{tr}(\Sigma)}{T}+1}\frac{1}{T}\mathbf{1}\mathbf{1}^\top\frac{1}{1+a\frac{\mathbf{tr}(\Sigma)}{T}}}{1 - \frac{a\frac{\mathbf{tr}(\Sigma)}{T}}{1+a\frac{\mathbf{tr}(\Sigma)}{T}}} \\
&= \frac{1}{a\frac{\mathbf{tr}(\Sigma)}{T} + 1}I + \frac{a\frac{\mathbf{tr}(\Sigma)}{T}}{a\frac{\mathbf{tr}(\Sigma)}{T} + 1}\frac{1}{T}\mathbf{1}\mathbf{1}^\top \\
&= \frac{1}{a\frac{\mathbf{tr}(\Sigma)}{T} + 1}\left(I + a\frac{\mathbf{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^\top\right).
\end{aligned}$$

The risk of $Y^* = \frac{1}{a\frac{\mathbf{tr}(\Sigma)}{T}+1}\left(I + a\frac{\mathbf{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^\top\right)\bar{Y}$ is

$$\begin{aligned}
R(\mu, Y^*) &= \mathbf{tr}\left(\frac{1}{a\frac{\mathbf{tr}(\Sigma)}{T} + 1}\left(I + a\frac{\mathbf{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^\top\right)\Sigma I\frac{1}{a\frac{\mathbf{tr}(\Sigma)}{T} + 1}\left(I + a\frac{\mathbf{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^\top\right)^\top\right) \\
&\quad + \mu^\top\left(\frac{1}{a\frac{\mathbf{tr}(\Sigma)}{T} + 1}\left(I + a\frac{\mathbf{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^\top\right) - I\right)^\top\left(\frac{1}{a\frac{\mathbf{tr}(\Sigma)}{T} + 1}\left(I + a\frac{\mathbf{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^\top\right) - I\right)\mu \\
&= \frac{1}{(a\frac{\mathbf{tr}(\Sigma)}{T} + 1)^2}\mathbf{tr}\left(\left(I + a\frac{\mathbf{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^\top\right)\Sigma\left(I + a\frac{\mathbf{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^\top\right)\right) \\
&\quad + \mu^\top\left(\frac{-a\frac{\mathbf{tr}(\Sigma)}{T}}{a\frac{\mathbf{tr}(\Sigma)}{T} + 1}I + \frac{a\frac{\mathbf{tr}(\Sigma)}{T}}{a\frac{\mathbf{tr}(\Sigma)}{T} + 1}\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)^\top\left(\frac{-a\frac{\mathbf{tr}(\Sigma)}{T}}{a\frac{\mathbf{tr}(\Sigma)}{T} + 1}I + \frac{a\frac{\mathbf{tr}(\Sigma)}{T}}{a\frac{\mathbf{tr}(\Sigma)}{T} + 1}\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)\mu \\
&= \frac{1}{(a\frac{\mathbf{tr}(\Sigma)}{T} + 1)^2}\mathbf{tr}\left(\Sigma + 2a\frac{\mathbf{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^\top\Sigma + a^2\frac{\mathbf{tr}(\Sigma)^2}{T^4}\mathbf{1}\mathbf{1}^\top\Sigma\mathbf{1}\mathbf{1}^\top\right) \\
&\quad + \frac{(a\frac{\mathbf{tr}(\Sigma)}{T})^2}{(a\frac{\mathbf{tr}(\Sigma)}{T} + 1)^2}\mu^\top L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)^\top L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)\mu \\
&= \frac{\frac{\mathbf{tr}(\Sigma)}{T}}{(a\frac{\mathbf{tr}(\Sigma)}{T} + 1)^2}\left(T + 2a\frac{\mathbf{tr}(\Sigma)}{T} + \left(a\frac{\mathbf{tr}(\Sigma)}{T}\right)^2\right) \\
&\quad + \frac{(a\frac{\mathbf{tr}(\Sigma)}{T})^2}{(a\frac{\mathbf{tr}(\Sigma)}{T} + 1)^2}\mu^\top L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)^\top L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)\mu.
\end{aligned}$$

To find the minimum, we take the partial derivative w.r.t. a and set it equal to zero. Noting that

$$L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)^\top L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right) = L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right),$$

and omitting some tedious algebra,

$$\begin{aligned}
\frac{\partial}{\partial a^*}R(\mu, Y^*) = 0 &= \frac{2\frac{\mathbf{tr}(\Sigma)}{T}(-T+1+a^*\mu^\top L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)\mu)}{(a^*\frac{\mathbf{tr}(\Sigma)}{T}+1)^3} \\
\Leftrightarrow a^* &= \frac{T-1}{\mu^\top L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)^\top L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)^\top \mu} \\
&= \frac{T-1}{\mu^\top L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)\mu} \\
&= \frac{2}{\frac{1}{T(T-1)}\sum_{r=1}^T\sum_{s=1}^T(\mu_r-\mu_s)^2}.
\end{aligned}$$

Appendix F: MTA Minimax Derivation

Recall Lehmann and Casella (1998, Chapter 5, Theorem 1.4):

Theorem *Suppose that π is a distribution on the space of μ such that*

$$r(\pi, Y_\pi) = \sup_\mu R(\mu, Y_\pi),$$

where $r(\pi, Y_\pi) = \int R(\mu, Y_\pi)\pi(\mu)d\mu$ is the Bayes risk. Then:

1. *Y_π is minimax.*
2. *If Y_π is the* unique *Bayes solution w.r.t. π (i.e. if it is the only minimizer of the Bayes risk), then it is the unique minimax estimator.*
3. *The prior π is least favorable.*

Corollary *If a Bayes estimator Y_π has constant risk, then it is minimax.*

The first step in finding a minimax solution for the $T = 2$ case is specifying a constraint set for μ over which a least favorable prior (LFP) can be found. We will use the box constraint set, $\mu_t \in [b_l, b_u]^\top$, where $b_l \in \mathbb{R}$ and $b_u \in \mathbb{R}$. It is straightforward to show that the corresponding LFP is

$$p(\mu) = \begin{cases} \frac{1}{2}, & \text{if } \mu = [b_l, b_u]^\top \\ \frac{1}{2}, & \text{if } \mu = [b_u, b_l]^\top \\ 0, & \text{otherwise.} \end{cases}$$

The next step is to *guess* a minimax weight matrix W^M and show that the estimator $Y^M = W^M\bar{Y}$ (i) has constant risk and (ii) is a Bayes solution. According to the corollary, if both (i) and (ii) hold for the guessed W^M, then $W^M\bar{Y}$ is minimax. For the $T = 2$ case, we guess W^M to be

$$W^M = \left(I + \frac{2}{T(b_l - b_u)^2}\Sigma L(\mathbf{1}\mathbf{1}^\top)\right)^{-1},$$

which is just W^{cnst} with $a = \frac{2}{(b_l - b_u)^2}$. This choice of W is not a function of μ and thus we have shown that (i) the Bayes risk w.r.t the LFP is constant for all μ. What remains to be shown is (ii) W^M is indeed the Bayes solution, i.e. it is minimizer of the Bayes risk:

$$\begin{aligned}&\frac{1}{2}\left([b_l\ b_u](W-I)^\top(W-I)\begin{bmatrix} b_l \\ b_u\end{bmatrix} + \mathbf{tr}(W\Sigma W^\top)\right)\\ +&\frac{1}{2}\left([b_u\ b_l](W-I)^\top(W-I)\begin{bmatrix} b_u \\ b_l\end{bmatrix} + \mathbf{tr}(W\Sigma W^\top)\right).\end{aligned} \tag{29}$$

Note that this expression is the sum of two convex risks. We already know that for $T = 2$ the minimizer of the risk

$$[\mu_1\ \mu_2](W-I)^\top(W-I)\begin{bmatrix} \mu_1 \\ \mu_2\end{bmatrix} + \mathbf{tr}(W\Sigma W^\top)$$

is $W^* = \left(I + \frac{2}{T(\mu_1-\mu_2)^2}\Sigma L(\mathbf{1}\mathbf{1}^\top)\right)^{-1}$. Thus, the minimizer of either term in (29) is

$$W^M = \left(I + \frac{2}{T(b_u - b_l)^2}\Sigma L(\mathbf{1}\mathbf{1}^\top)\right)^{-1} \tag{30}$$

as was to be shown. One can conclude that W^M is minimax over all estimators of the form $\left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ for $T = 2$ for the box constraint set.

Appendix G: Proof of Proposition 4

Proposition 4 *The set of estimators $W\bar{Y}$ where W is of MTA form as per (20) is strictly larger than the set of estimators that regularize the single-task estimates as follows:*

$$\hat{Y} = \left(\frac{1}{\gamma}I + \mathbf{1}\alpha^\top\right)\bar{Y},$$

where $\sum_{r=1}^{T}\alpha_r = 1 - \frac{1}{\gamma}$, $\gamma \geq 1$, and $\alpha_r \geq 0$, $\forall r$.

Proof Using the Sherman-Morrison formula,

$$\begin{aligned}\left(\frac{1}{\gamma}I + \mathbf{1}\alpha^\top\right)^{-1} &= \gamma I - \frac{\gamma^2 \mathbf{1}\alpha^\top}{1+\gamma\alpha^\top\mathbf{1}} \\ &= \gamma I - \gamma\mathbf{1}\alpha^\top \\ &= I + (\gamma-1)I - \gamma\mathbf{1}\alpha^\top \\ &= I + \gamma\left(1-\frac{1}{\gamma}\right)I - \gamma\mathbf{1}\alpha^\top \\ &= I + \gamma L(\mathbf{1}\alpha^\top),\end{aligned}$$

which is a matrix of MTA form for $\Gamma = \gamma I$ and $A = \mathbf{1}\alpha^T$. Thus, estimators $\hat{Y}_t$ can be written in MTA form:

$$\hat{Y} = (I + \gamma L(\mathbf{1}\alpha^\top))^{-1}\bar{Y}. \tag{31}$$

The converse clearly does not hold: not all matrices $(I+\Gamma L(A))^{-1}$ can be written as (31). ■

Appendix H: Proof of Proposition 7

Proposition 7

$$\mathbf{1}^\top \hat{Y}^{JS} = \mathbf{1}^\top \bar{Y},$$

where $\hat{Y}^{JS}$ is given in (7).

Proof The tth component of $\hat{Y}^{JS}$ can be written:

$$\hat{Y}_t^{JS} = \frac{1}{T}\sum_{r=1}^{T}\bar{Y}_r + c(\bar{Y}_t - \frac{1}{T}\sum_{r=1}^{T}\bar{Y}_r),$$

for some scalar $c \in [0,1]$ that does not depend on t. Thus,

$$\hat{Y}^{JS} = \frac{1-c}{T}\left(\sum_{r=1}^{T}\bar{Y}_r\right)\mathbf{1} + c\bar{Y},$$

and the sum of the estimates is:

$$\begin{aligned}\mathbf{1}^\top\hat{Y}^{JS} &= \mathbf{1}^\top\left(\frac{1-c}{T}\left(\sum_{r=1}^{T}\bar{Y}_r\right)\mathbf{1} + c\bar{Y}\right) \\ &= \frac{1-c}{T}\left(\sum_{r=1}^{T}\bar{Y}_r\right)\mathbf{1}^\top\mathbf{1} + c\mathbf{1}^\top\bar{Y} \\ &= (1-c)\sum_{r=1}^{T}\bar{Y}_r + c\sum_{r=1}^{T}\bar{Y}_r \\ &= \mathbf{1}^\top\bar{Y}.\end{aligned}$$

■

References

J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal Machine Learning Research*, 10, 2009.

A. Argyriou, C. A. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

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

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal Machine Learning Research*, 6:1705–1749, December 2005.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal Machine Learning Research*, 7:2399–2434, 2006.

A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, 1979.

M. E. Bock. Minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics*, 3(1), 1975.

E. V. Bonilla, K. M. A. Chai, and C. K. I. Williams. Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2008.

G. Casella. An introduction to empirical Bayes data analysis. *The American Statistician*, pages 83–87, 1985.

P. Chebotarev and E. Shamis. The matrix-forest theorem and measuring relations in small social groups. *Computing Research Repository*, abs/math/0602070, 2006.

F. R. K. Chung. *Spectral Graph Theory*. 2004.

B. Efron and C. N. Morris. Limiting the risk of Bayes and empirical Bayes estimators–part II: The empirical Bayes case. *Journal of the American Statistical Association*, 67(337): 130–139, 1972.

B. Efron and C. N. Morris. Stein's paradox in statistics. *Scientific American*, 236(5): 119–127, 1977.

S. Feldman, M. R. Gupta, and B. A. Frigyik. Multi-task averaging. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

F. Fouss, L. Yen, A. Pirotte, and M. Saerens. An experimental investigation of graph kernels on a collaborative recommendation task. In *ICDM*, pages 863–868, 2006.

B. A. Frigyik, S. Srivastava, and M. R. Gupta. Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Trans. Information Theory*, 54(11):5130–5139, 2008.

C. F. Gauss. *Theory of the motion of the heavenly bodies moving about the sun in conic sections*. Little, Brown, 1857. Translated by C. H. Davis.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990. Corrected reprint of the 1985 original.

L. Jacob, F. Bach, and J.-P. Vert. Clustered multi-task learning: A convex formulation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 745–752, 2008.

W. James and C. Stein. Estimation with quadratic loss. *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379, 1961.

T. Kato, H. Kashima, M. Sugiyama, and K. Asai. Multi-task learning via conic programming. In *Advances in Neural Information Processing Systems (NIPS)*, pages 737–744. 2008.

Legendre. *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*. Appendix, Paris, 1805.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, New York, 1998.

D. Martínez-Rego and M. Pontil. Multi-task averaging via task clustering. In *Proc. SIMBAD*, 2013.

C. A. Micchelli and M. Pontil. Kernels for multi–task learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

I. Newton. *Chronology of Ancient Kingdoms Amended*. Kessinger Publishing, England, 1728.

R. L. Plackett. Studies in the history of probability and statistics: VII. the principle of the arithmetic mean. *Biometrika*, 45(1/2):130–135, 1958.

J. P. Romano and A. F. Siegel. *Counterexamples in Probability and Statistics*. Chapman and Hall, Belmont, CA USA, 1986.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 2005.

M. Saerens, F. Fouss, L. Yen, and P. Dupont. The principal components analysis of a graph, and its relationships to spectral clustering. In *In Proc. Eur. Conf. Machine Learning*, pages 371–383. Springer-Verlag, 2004.

D. Salsburg. *The Lady Tasting Tea*. Holt Paperbacks, New York, NY, 2001.

D. Sheldon. Graphical multi-task learning, 2008. Advances in Neural Information Processing Systems (NIPS) Workshops.

J. Sherman and W. J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Stat.*, 21:124–127, 1950.

A. J. Smola and I. R. Kondor. Kernels and regularization on graphs. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2003.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206, 1956.

S. M. Stigler. *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA, 1999.

U. v. Luxburg. A tutorial on spectral clustering. *Computing Research Repository*, abs/0711.0189, 2007.

Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with Dirichlet process priors. *Journal Machine Learning Research*, 8:35–63, 2007.

Y. Yajima and T.-F. Kuo. Efficient formulations for 1-SVM and their application to recommendation tasks. *JCP*, 1(3):27–34, 2006.

Y. Zhang and D.-Y. Yeung. A convex formulation for learning task relationships. In *Proc. of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

X. Zhu. Semi-supervised learning literature survey, 2006.

X. Zhu and J. Lafferty. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *In Proc. Int. Conf. Machine Learning*, pages 1052–1059. ACM Press, 2005.