Multi-Task Averaging: Theory and Practice

Sergey Feldman

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Maya R. Gupta, Chair

Maryam Fazel Sarjoui

Daniela Witten

Program Authorized to Offer Degree:
Electrical Engineering

University of Washington

Abstract

Multi-Task Averaging: Theory and Practice

Sergey Feldman

Chair of the Supervisory Committee:
Professor Maya R. Gupta
Department of Electrical Engineering

This dissertation addresses the problem of estimating the means of multiple distributions. I begin with a brief history of the mean, leading to a discussion and literature review of Stein estimation and multi-task learning. Using a multi-task regularized empirical risk formulation, an algorithm called *multi-task averaging* (MTA) is derived and analyzed. Two main results are discussed. First, I prove that the MTA solution matrix is right-stochastic, that is, the multi-task mean estimates are always convex combinations of single-task mean estimates. Second, in the two-task case, analysis shows that the MTA estimates have smaller risk than single-task estimates for a range of task similarity values. I use this analysis to derive a theoretically optimal similarity, which has an intuitive form. I then proceed to derive two practical and efficient MTA estimators for real data of any number of tasks: constant MTA and minimax MTA. Extensive simulations and four applications demonstrate that MTA often outperforms the battle-tested James-Stein estimator, as well as single-task estimation.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to express sincere gratitude to my advisor and friend Maya Gupta for her inexhaustible streams of encouragement, patience, knowledge, wisdom, and commitment to her students. I would also like to thank the professors who have helped me on my way: Maryam Fazel, Vladimir Minin, Daniela Witten, Mike MacCoss, Mari Ostendorf, and Jeff Bilmes.

I owe a great deal to the dozens of amazing graduate students[1] I've come to know in my time at the University of Washington. In particular I'd like to thank my friends Hyrum Anderson, James Chen, and Eric Garcia for their indispensable mentoring, and everyone else I've known at the IDL and SSLI Lab for their collaborations and discussions, as well as for being easy-going, fun folks to work and learn with.

The staff at the electrical engineering department have made my time here a trouble-free experience, and I am grateful to them.

Finally, I would like to thank my friends and family for their unwavering encouragement and faith in me.

[1]Many of which have since graduated.

DEDICATION

to my father Oscar, and all the people that keep me going

Chapter 1

INTRODUCTION

This thesis is the study of an algorithm called *multi-task averaging* (MTA). MTA estimates the means of multiple random distributions simultaneously, and does so better (with respect to a specified metric) than the established methods both in theory and in practice. Before diving into statistics and machine learning, however, let's briefly review the history of the mean.

1.1 A Short History of the Mean

The mean is one of the most fundamental and useful tools in statistics [46], and it is *old*. According to Aristotle [5],

> "By the mean of a thing I denote a point equally distant from either extreme..."

And our modern minds would agree, but with a caveat: only for a symmetric distribution (with finite support)! To compute Aristotle's mean, we take the "extremes" $y_{\min}$ and $y_{\max}$, and solve for $\bar{y}$, the point equally distant from both of them: $y_{\min} - \bar{y} = \bar{y} - y_{\max}$, obtaining

$$\bar{y} = \frac{y_{\min} + y_{\max}}{2},$$

which is the average value of $y_{\min}$ and $y_{\max}$. The average was not explicitly extended to more than two values until the 16th century. R. L. Plackett attributes Tycho Brahe with the introduction of averaging for the reduction of measurement error [42]; Brahe repeatedly measured the same astronomical quantities, and averaged them to obtain the final values.

Three hundred years after Brahe, the mean was well established. In 1805, Legendre introduced the least squares method in his "Nouvelle methods pour la determination des orbites des comtes" [31], noting in the appendix that the mean is the solution to the following

optimization problem

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

A few years later, Gauss wrote [30],

> "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."

Indeed, the average was central[1] to Gauss's derivation of the normal distribution [31]. Gauss obtained the distribution in a backwards fashion; given observations $\{y_1, \ldots, y_N\}$, he sought a differentiable distribution[2] for which the average of the observations was the mode. Mathematically, he was looking for a density $p(Y = y|\{y_1, \ldots, y_N\})$ that satisfied

$$\frac{\partial \ln p(Y = y|\{y_1, \ldots, y_N\})}{\partial y} = 0, \quad \text{for } y = \bar{y},$$

where $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. This condition would only force one maximum (or minimum) of the distribution to be the mean. The solution Gauss chose, of course, is the famous Gaussian distribution.

Gauss's approach is an early precursor to present day maximum likelihood (ML) methods. Interestingly, the mean is often but *not always* the ML estimate of the expectation. A simple counter-example is the uniform distribution, for which the ML estimate of the mean is:

$$\frac{\max_i y_i - \min_i y_i}{2}.$$

Note that the sample mean $\bar{y}$ is a *sufficient statistic* for the expected value μ of a Gaussian distribution.[3] In the words of R. A. Fisher, once one computes the sufficient statistic $\bar{y}$,

[1] Pun intended.

[2] One that tended to zero at both plus and minus infinity.

[3] Note that $\bar{y}$ is not a sufficient statistic for the expectation of *all* distributions, but is a sufficient statistic for many common distributions such as the Laplacian and the Poisson.

"no other statistic which can be calculated from *the same sample* provides any additional information as to the value of the parameter" [28]. I (and not Fisher) place the emphasis on "the same sample" to foreshadow the punchline of this work: using samples from *other* distributions may provide additional information.

Last century, in a surprising result referred to as *Stein's paradox* [23], Stein showed that it is better, in terms of the summed squared error, 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 [51]. That is, it is beneficial to consider samples from *seemingly unrelated* distributions to estimate the mean. Why might this be the case? Consider an example where our goal is to estimate the means of two distributions: the distribution of movie ticket prices and the distribution of the ages of summer campers. It happens to be the case that the means of the two distributions are nearly identical, even though there is no meaningful correlation between them. So, if one has only a few samples from the first distribution and wishes to estimate the mean, it makes sense that the estimate is likely to be improved if samples from the second distribution are taken into account. And vice-versa for the second mean.

Stein's result is an early example of the motivating hypothesis behind multi-task learning (MTL): that leveraging data from different tasks can yield superior performance over learning from each task independently. As an example of multi-task learning, consider the problem of drug response prediction. A pharmaceutical drug is undergoing testing in different cities around the world, and you wish to estimate the probability of a desired effect on a patient. You have multiple sets of clinical trial data (tasks), one set per city. The naive approach to estimating the desired probability density is to combine the datasets from each city into a single dataset (that is, a single estimation task). This approach is unsatisfying because it ignores potentially crucial differences in patients from each country that may be too time-consuming or expensive to quantify like climate, genetics, diet, and quality of medical care. Conversely, you could estimate the densities for each city (task) separately, ignoring the data at other locations. This too is less than ideal since the remaining tasks are clearly relevant and likely informative. MTL is the middle ground between the above extremes, in that multiple estimation tasks are performed jointly and simultaneously, with-

out collapsing all tasks into one. MTL acknowledges what single-task approaches ignore: no dataset exists in a vacuum; relevant datasets may be available, and can help alleviate common estimation problems such as data paucity, the curse of dimensionality, and low signal-to-noise ratios.

Multi-task learning emerged decades after Stein's work, and the insights of MTL have yet to be applied to the problem of estimating multiple means. In this thesis I use the philosophy of MTL to design a multi-task mean estimator I call *multi-task averaging* (MTA), and study it extensively. Chapter 2 is an in-depth theoretical analysis of MTA. In Chapter 3, I report simulated data and application results. The thesis ends with conclusions and open questions in Chapter 4.

I will now review the appropriate background and related work. First, I will review Stein estimation. I will then review modern MTL approaches in the machine learning literature, which will be followed up by a look at manifold regularization and the graph Laplacian matrix.

1.2 Stein Estimation

As mentioned in the introduction, one of the earliest examples of multi-task learning is Stein's work on the estimation of the means of T distributions (or a single T-variate distribution) [51, 23]. Here, the tth task is the estimation of the tth mean.

Suppose that multiple random variables are all independent and have different means; Stein showed that it is better to estimate the mean of one random variable using data sampled from all of them. I will start with the simplest possible story: one sample per distribution, as shown in Figure 1.1. Specifically, given a sample drawn from each of T Gaussian distributions $Y_t \sim \mathcal{N}(\mu_t, \sigma^2)$ for $t = 1, \ldots, T$, Stein showed that the maximum likelihood estimator $\hat{Y}_t^{ML} = Y_t$ is inadmissible. In particular, it is dominated[4] by what is

[4]Estimator $\hat{Y}$ is said to be *dominated* by estimator $\tilde{Y}$ if $E[||\mu - \tilde{Y}||_2^2] \leq E[||\mu - \hat{Y}||_2^2]$ for all choices of parameter μ [37].

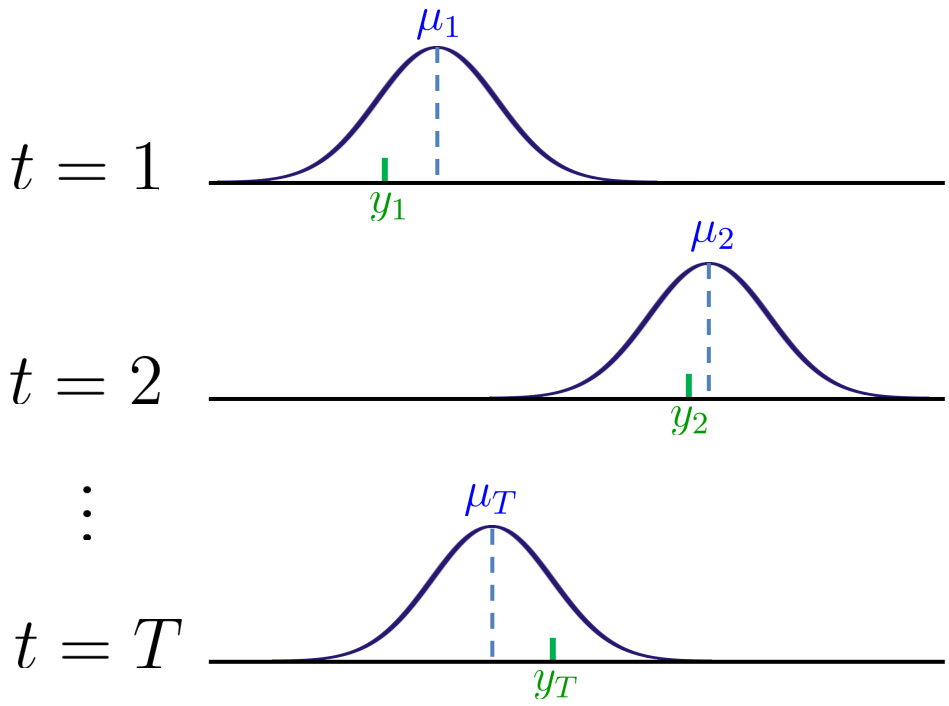

Figure 1.1: An illustration of the notation for Stein estimation with one sample per task.

known as the *James-Stein estimator*,

$$\hat{Y}_t^{JS} = \left(1 - \frac{(T-2)\sigma^2}{Y^T Y}\right) Y_t, \tag{1.1}$$

where Y is a vector with tth entry Y_t. The above estimator dominates $\hat{Y}_t^{ML}$ when $T > 2$. For $T = 2$, (1.1) reverts to the maximum likelihood estimator, which turns out to be admissible [51]. The surprising term in (1.1) is $Y^T Y$, which includes all T components. This is counter-intuitive, and that is why this estimator and its implications are often referred to as *Stein's paradox*.

How to make sense of (1.1)? It scales the unbiased ML estimate Y_t; when would this be desirable? Efron and Morris make an empirical Bayes argument to motivate the form of the James-Stein estimator [22]. The key assumption to make is that the true means are themselves drawn from a normal distribution centered at zero: $\mu_t \sim \mathcal{N}(0, \tau^2)$. In that case,

we have

$$\begin{aligned} E[\mu_t|Y_t] &= \frac{\tau^2}{\tau^2+\sigma^2}\mu_t \\ &= \left(1-\frac{\sigma^2}{\tau^2+\sigma^2}\right)\mu_t, \end{aligned}$$

where τ^2 is unknown and σ^2 is known.[5] One obtains the James-Stein estimator after noting that

$$E\left[\frac{(T-2)\sigma^2}{Y^TY}\right] = \frac{\sigma^2}{\tau^2+\sigma^2},$$

and substituting. That is, $\frac{(T-2)\sigma^2}{Y^TY}$ is an unbiased estimator of $\frac{\sigma^2}{\tau^2+\sigma^2}$.

So we see that the form of the James-Stein estimator reflects the prior belief that the μ_t lie near zero, and thus Y_t is *shrunk* towards zero (the terms "regularization" and "shrinkage" are often used interchangeably). But assuming that the μ_t are close to zero is too restrictive. More generally, the means can be assumed to be drawn as $\mu_t \sim \mathcal{N}(\xi, \tau^2)$. The James-Stein estimator then becomes

$$\hat{Y}_t^{JS} = \xi + \left(1-\frac{(T-3)\sigma^2}{(Y-\xi)^T(Y-\xi)}\right)(Y_t-\xi), \tag{1.2}$$

where ξ can be estimated as the average of means $\hat{\xi} = \frac{1}{T}\sum_{r=1}^{T}\bar{Y}_r$, and this additional estimation decreases the degrees of freedom by one.[6]

The above formulations of the James-Stein estimator assume that a single sample is drawn for each task. This assumption is unrealistically simple. Generalizing, let $Y_{ti} \sim \mathcal{N}(\mu_t, \sigma^2)$ for $t = 1, \ldots, T$ and $i = 1, \ldots, N$ (See Figure (1.2)). The James-Stein estimator becomes [22],

$$\hat{Y}_t^{JS} = \xi + \left(1-\frac{(T-3)\frac{\sigma^2}{N}}{(\bar{Y}-\xi)^T(\bar{Y}-\xi)}\right)(\bar{Y}_t-\xi), \tag{1.3}$$

---

[5] James and Stein showed that if σ^2 is unknown it can be replaced by the standard unbiased ML estimate $\hat{\sigma}^2$ [35, 15].

[6] For more details as to why $T-2$ becomes $T-3$ see Example 7.7 on page 278 in Lehmann and Casella [37].

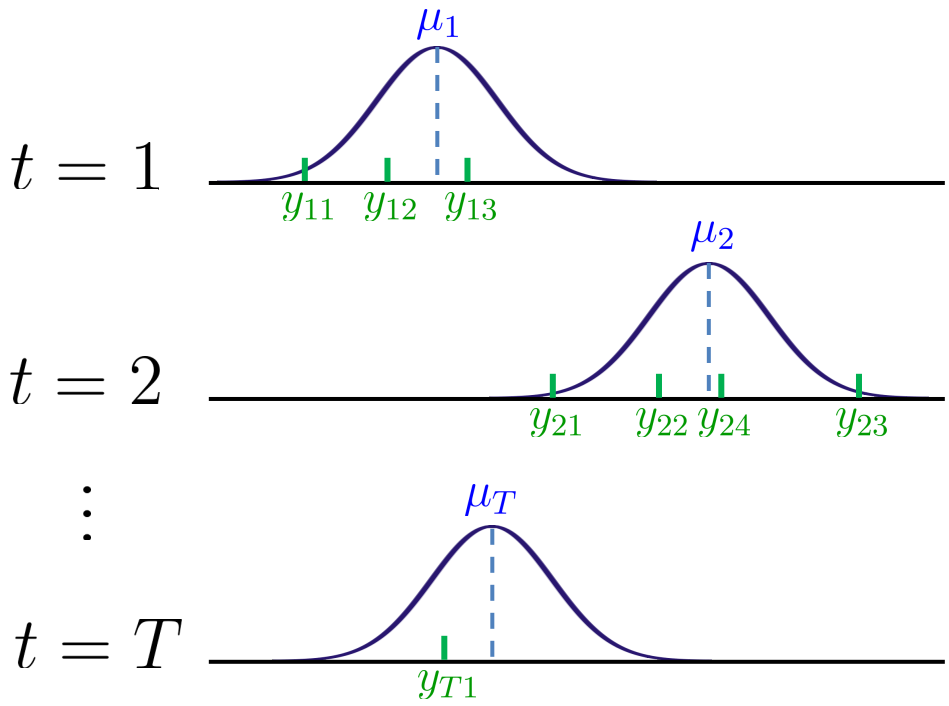

Figure 1.2: An illustration of the notation for Stein estimation with N_t sample for the tth task.

where $\bar{Y}_t = \frac{1}{N}\sum_{i=1}^{N} Y_{ti}$. To get (1.3) from (1.2) note that $\bar{Y}_t \sim \mathcal{N}(\mu_t, \frac{\sigma^2}{N})$.

Even more generally, let $Y_{ti} \sim \mathcal{N}(\mu_t, \sigma_t^2)$ for $t = 1, \ldots, T$ and $i = 1, \ldots, N_t$. Further, let Σ be the covariance matrix of $\bar{Y}$ and let $\lambda_{\max}(\Sigma)$ be the largest eigenvalue of Σ. In words, the sample are independently and identically distributed for each task, but there may be dependencies between the tasks. The James-Stein estimator becomes [11]

$$\hat{Y}_t^{JS} = \xi + \left(1 - \frac{\frac{\mathrm{tr}(\Sigma)}{\lambda_{\max}(\Sigma)} - 3}{(\bar{Y} - \xi)^T \Sigma^{-1} (\bar{Y} - \xi)}\right)(\bar{Y}_t - \xi),$$

which dominates the maximum likelihood estimate when $\frac{\mathrm{tr}(\Sigma)}{\lambda_{\max}(\Sigma)} > 3$.

It was later shown that the James-Stein estimator itself is not admissible, and is domi-

nated by the positive part James-Stein estimator [37]:

$$\hat{Y}_t^{JS} = \xi + \left(1 - \frac{\frac{\mathrm{tr}(\Sigma)}{\lambda_{\max}(\Sigma)} - 3}{(\bar{Y} - \xi)^T \Sigma^{-1} (\bar{Y} - \xi)}\right)^+ (\bar{Y}_t - \xi),$$

where $(x)^+ = \max(0, x)$. See Lehmann and Casella [37] for more general cases, such as James-Stein estimation with full covariance matrix Σ.

The term $\frac{\mathrm{tr}(\Sigma)}{\lambda_{\max}}$ is called the *effective dimension* of the estimator. In simulations where I set Σ to be the true covariance matrix and then estimated the effective dimension by estimating both the maximum eigenvalue and the trace of the sample mean covariance matrix, I found that replacing the effective dimension with the actual dimension T (when Σ is diagonal) resulted in a significant performance boost.[7] In other preliminary experiments with real data, I also found that using T rather than the effective dimension performed better due to the high variance of the estimated maximum eigenvalue in the denominator of the effective dimension.

Additionally, I chose to use the estimate $\hat{\xi} = \frac{1}{T}\sum_{r=1}^{T} \bar{Y}_r$. This choice reflects the assumption that the μ_t are drawn independently from a prior Gaussian distribution with unknown mean ξ of which $\hat{\xi}$ is an unbiased estimate [22]. One could alternatively, as done for the simplest James-Stein estimator, set $\xi = 0$ (this is akin to the ridge regularizer), but this is suboptimal choice because it is not shift-invariant and a poor general estimate of the true prior mean. Another reasonable choice is the pooled mean $\xi = \frac{1}{\sum_{t=1}^{T} N_t} \sum_{t=1}^{T} \sum_{i=1}^{N_t} Y_{ti}$, which I expect would perform very similarly to the choice I made.

These changes result in the following variant of the James-Stein estimator, which I use for all of the experiments in this thesis,

$$\hat{Y}_t^{JS} = \frac{1}{T}\sum_{r=1}^{T} \bar{Y}_r + \left(1 - \frac{T - 3}{(\bar{Y} - \xi)^T \Sigma^{-1} (\bar{Y} - \xi)}\right)^+ \left(\bar{Y}_t - \frac{1}{T}\sum_{r=1}^{T} \bar{Y}_r\right). \tag{1.4}$$

[7] For the case of a diagonal Σ, there are T separate distributions, thus the effective dimension is exactly T.

1.3 Multi-Task Learning in Machine Learning

In the machine learning community work in multi-task learning is less than two decades old. Early work explored methods such as Thrun's "learning to learn" in which the learning bias for a particular task is set based on "learning experiences" with other tasks [53]. Similarly, Baxter [8] approached multi-task problems by "first learning an appropriate internal representation for a learning environment and then using that representation to bias the learner's hypothesis space for the learning of future tasks drawn from the same environment."

Other early multi-task methods focused on neural networks. Caruana's work [14] showed that neural networks with a shared hidden layer between tasks and multiple output layers (one per task) outperformed single-task versions in which task distinctions were ignored.

1.3.1 Empirical Risk Minimization

Most recent MTL methods, as well as the algorithms studied in this thesis, are derived using the *empirical risk minimization* (ERM) principle [39]. ERM is an approach to the problem of finding a function $f(x)$ that maps inputs $x \in \mathbb{X}$ to outputs $y \in \mathbb{Y}$, where $\mathbb{X}$ and $\mathbb{Y}$ are input and output spaces, respectively. To accomplish this task, one is provided with a set of input-output pairs $\{x_i, y_i\}_{i=1}^N$, where $x_i \in \mathbb{X}$ and $y_i \in \mathbb{Y}$, and a *loss* function $L(f(x), y) \in \mathbb{R}^+$ that measures how far the estimate $f(x)$ is from y. The *risk* of function f is the expected loss:

$$R(f) = E_{X,Y}[L(f(X), Y)].$$

However, since only samples are available, the *empirical risk* is used in practice:

$$R_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^{N} L(f(x_i), y_i).$$

The ERM principle states that the function $\hat{f}(x)$ should be chosen by solving the following optimization problem,

$$\hat{f} = \arg\min_{f \in \mathcal{H}} R_{\text{emp}}(f),$$

where $\mathcal{H}$ is the hypothesis space of all considered functions. In practice, ERM often results in functions that *overfit* to the available data (that is, the function does not generalize well to heretofore unseen data). To help reduce overfitting, a regularization function $J(f) \in \mathbb{R}$ that penalizes the complexity of f (and decreases the variance of the estimate) is added to the objective. Thus, the *regularized* or *penalized* ERM objective[8] is

$$\hat{f} = \arg\min_{f \in \mathcal{H}} R_{\text{emp}}(f) + \gamma J(f),$$

where γ is a non-negative scaling parameter that controls the bias-variance trade-off [32].

1.3.2 Multi-Task Regularizers

MTL methods are usually designed for regression, classification, or feature selection, e.g. [40, 12, 3]. Estimating T means can be considered a special case of multi-task regression,[9] where one fits a constant function to each task. And, as in MTA, one of the main approaches to multi-task regression in literature is tying tasks together with an explicit multi-task parameter regularizer. See Table 1.1 for examples of recently proposed parametric regularizers $J(\{\beta_t\}_{t=1}^T)$. The parameter matrix $\beta \in \mathbb{R}^{d \times T}$ has as its tth column β_t, the parameter vector of the tth task (d is the dimensionality of the input feature space). The similarity between tasks is encoded in matrix A where A_{rs} is the similarity between task r and task s.

Abernathy et al. [1], for instance, propose to minimize the empirical loss with the following added 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. For mean estimation, $d = 0$ and the matrix β has only one row; $||\beta||_*$ reduces to the ℓ_1 norm on the outputs and this regularizer does not tie the tasks together.

---

[8] For a fuller mathematical treatment see 'structural risk minimization' in Vapnik's *The Nature of Statistical Learning Theory* [55].

[9] With a feature space of zero dimensions only the constant offset term is learned.

Table 1.1: Examples of MTL regularizers $J(\{\beta_t\}_{t=1}^T)$

$\sum_{r=1}^T \|\beta_r - \frac{1}{T}\sum_{s=1}^T \beta_s\|_2^2$	Distance to mean [26].
$\|\|\beta\|\|_*$	Trace norm [1].
$\mathrm{tr}(\beta^T D^{-1}\beta)$	Learned, shared feature covariance matrix [3].
$\mathrm{tr}(\beta\Sigma^{-1}\beta^T)$	Learned task covariance matrix [34, 58].
$\sum_{r=1}^T \sum_{s=1}^T A_{rs}\|\beta_r - \beta_s\|_2^2$	Pairwise distance regularizer [47] or constraint [36].

Argyriou et al. [3] propose an alternating approach with a different regularizer:

$$\mathrm{tr}(\beta^T D^{-1}\beta),$$

where D is a learned, shared *feature* covariance matrix. With no features, D is just a constant and $\mathrm{tr}(\beta^T D^{-1}\beta)$ is a ridge regularizer on the outputs. The regularizers in the work of Jacobs et al. [34] and Zhang and Yeung [58] reduce similarly when in the context of mean estimation.

The most closely related work to MTA is that of Sheldon [47] and Kato et al. [36], 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 thesis I do just that: apply this regularizer to mean estimation, show that this special case enables new analytic results, and demonstrate its performance with simulated and real data.

1.3.3 Other MTL Approaches

Also common for MTL are Bayesian frameworks in which shared statistical structures of the parameters are learned or imposed. In the work of Bakker and Heskes [6] some of the model

parameters are shared amongst tasks, while others are connected through a joint prior. In Xue et al.'s work [56], the β_t are drawn from a Dirichlet process prior, and in the work of Liu et al. [38] a joint distribution is placed over the tasks' parameters in a semi-supervised framework. A related multi-task approach that focuses on function outputs is the work of Bonilla et al. [12], where a Gaussian process prior is placed over the latent task functions to constrain their inner products. This method is limited to Gaussian process regression, and requires that the inputs be identical amongst all tasks.

Another approach is the construction of multi-task kernels. Micchelli and Pontil [40] explored a linear combination of kernels for multi-task learning, and others [25] have shown that the problem of estimating T task functions with certain types of regularization can be cast as a single-task problem using multi-task kernels.

Still other work focused on tying tasks together by jointly selecting features or learning subspaces. Argyriou et al. [3] used a $(2,1)$-norm on the matrix β to encourage a small-number of non-zero rows. Their work is built on by Obozinski et al. [41] to obtain a computationally efficient joint subspace selection.

Another major question in multi-task learning is how to estimate the similarity (or task relatedness) between tasks and/or samples, if it is not provided. The standard approach, taken by many of the above-cited papers, is to estimate the similarity matrix jointly with the task parameters [58, 4, 12, 56, 34]. As a more detailed example, Zhang and Yeung [58] assumed 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.

1.4 Background on Manifold Regularization

MTA is similar to *manifold regularization* [9]. Manifold regularization can be viewed as a one-task, multiple dimension version of MTA. For example, Belkin et al.'s Laplacian-regularized least squares objective for semi-supervised regression solves

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

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}^F 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.

A different approach to using manifold regularization concepts for MTL was recently proposed by Agarwal et al. [2] for parametric multi-task learning. They assume that the task *parameters* lie on a low-dimensional manifold, and alternate estimating the manifold with learning the tasks in an iterative way. Their approach is restricted to a RKHS and assumes a global manifold structure for all tasks, which share a common parametric model.

1.5 The Graph Laplacian Matrix

Next, I provide some background on the graph Laplacian matrix, which will later be connected to MTA.

For graph G with T nodes, Let $A \in \mathbb{R}^{T \times T}$ be the matrix of graph weights, where $A_{rs} \geq 0$ 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 $\Delta g(x) = \text{tr}(H(g(x))) = \frac{\partial^2 g(x)}{\partial x_1^2} + \frac{\partial^2 g(x)}{\partial x_2^2} + \ldots + \frac{\partial^2 g(x)}{\partial x_M^2}$, where H is the Hessian matrix, 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 [21]. Given a function 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)

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

which is small when f is smooth over the graph [60]. If A is asymmetric then the energy can be written as

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

Note that the above formulation of the energy in terms of the graph Laplacian holds for the scalar case. More generally, when $f_i \in \mathbb{R}^d$ is a vector, one can alternatively write the energy in terms of the distance matrix:

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

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

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 [21]; the eigenvalues of L are particularly useful for spectral clustering [54]. The graph Laplacian is a common tool in semi-supervised learning literature [59], 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. [45] 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).

The graph Laplacian can be written as a sum of matrix products. Let $e_i \in \mathbb{R}^T$ be a vector of all zeros except with a 1 in the ith position. Then,

$$L = \sum_{r=1}^{T}(e_r^T A\mathbf{1})e_r e_r^T - \sum_{r=1}^{T}\sum_{s=1}^{T}(e_r^T A e_s)e_r e_s^T.$$

Chapter 2

THEORY

Some of the results in this chapter are drawn from joint work with M. R. Gupta and B. A. Frigyik [27], and will be noted as such where they occur.

In this chapter, I study the *multi-task averaging* (MTA) algorithm, which is formulated using the a regularized empirical risk minimization approach. With the squared-loss, MTA has a closed-form solution, and admits some analytic results that, while important in their own right, also provide intuition for more complicated problems.

The main theorem is that the MTA estimates are a *convex* combination of the individual tasks' samples averages. I also show that MTA is more general than many common estimators. An analysis of the two-task case yields the intuitive result that the optimal similarity between the two tasks is the inverse of the squared difference between the task means. Using this analysis, two practical and computationally efficient MTA estimators for $T > 2$ are derived: constant MTA and minimax MTA. Key notation is given in Table 2.1. In some place it will be more useful to write in terms of particular samples, denoted by lower-case y, but in others it will be more useful to write in terms of random variables, denoted by upper-case Y.

2.1 MTA Objective

Consider the T-task problem of estimating the means of T random variables. Assume that for the tth random variable one is given N_t samples $\{y_{ti}\}_{i=1}^{N_t}$, where each $y_{ti} \in \mathbb{R}$ is an independent draw from the task-specific random variable Y_t. In addition, assume that the $T \times T$ matrix A describes 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

Table 2.1: Key Notation

T	Number of tasks
N_t	Number of samples in task t
γ	Regularization parameter (non-negative scalar)
Y_t	Random variable of the tth task
$\mu_t = E[Y_t]$	True mean of the tth task
y_{ti}	ith Sample from tth task
$\bar{y}_t$	Single-task sample average for tth task
$\hat{y}_t$	An estimate of the mean of tth task
y_t^*	Multi-task averaging (MTA) estimate for tth task
σ_t^2	Variance of tth distribution
$\hat{\sigma}_t^2$	Estimate of the variance
Σ	Diagonal matrix with $\Sigma_{tt} = \frac{\sigma_t^2}{N_t}$
A	Similarity matrix with $A_{ts} \geq 0 \; \forall \; t, s$
D	Diagonal matrix with $D_{tt} = \sum_s A_{ts}$
$L = L(A) = D - A$	Graph Laplacian of A
$W = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$	MTA solution matrix
$Q = (I + \gamma L)^{-1}$	Regularized Laplacian kernel (RLK) matrix

canceled out in the MTA objective). The proposed MTA objective is

$$\{y_t^*\}_{t=1}^T = \underset{\{\hat{y}_t\}_{t=1}^T}{\arg\min} \; \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{(y_{ti}-\hat{y}_t)^2}{\sigma_t^2} + \frac{\gamma}{T^2}\sum_{r=1}^{T}\sum_{s=1}^{T}A_{rs}(\hat{y}_r - \hat{y}_s)^2. \tag{2.1}$$

The first term minimizes the multi-task empirical loss, and the second term jointly regularizes the estimates (i.e. ties them together). Note that if $\gamma = 0$, (2.1) decomposes to T separate minimization problems, producing the sample averages $\bar{y}_t$.

The normalization of each error term in (2.1) by its task-specific variance σ_t^2 (which may be estimated) scales the T empirical loss terms relative to the variance of their distribution; this ensures that high-variance tasks do not disproportionately dominate the loss term.

A more general formulation of MTA is

$$\{y_t^*\}_{t=1}^T = \underset{\{\hat{y}_t\}_{t=1}^T}{\arg\min} \ \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t} L(y_{ti}, \hat{y}_t) + \frac{\gamma}{T^2} J\left(\{\hat{y}_t\}_{t=1}^T\right),$$

where L is a loss function and J is a regularization function. If L is chosen to be any Bregman loss, then setting $\gamma = 0$ will produce the T sample averages [7]. For the analysis and experiments in this thesis, I restrict my focus to the tractable squared-error formulation given in (2.1).

2.1.1 The MTA Regularizer

To gain an intuitive understanding of the MTA regularizer, consider its expansion:

$$\begin{aligned}
\sum_{r=1}^{T}\sum_{s=1}^{T} A_{rs}(\hat{y}_r - \hat{y}_s)^2 &= \sum_{r=1}^{T}\sum_{s=1}^{T} A_{rs}(\hat{y}_r^2 + \hat{y}_s^2 - 2\hat{y}_r\hat{y}_s) \\
&= \sum_{r=1}^{T}\left(\sum_{s=1}^{T} A_{rs} + A_{sr}\right)\hat{y}_r^2 - 2\sum_{r=1}^{T}\sum_{s=1}^{T} A_{rs}\hat{y}_r\hat{y}_s \\
&= \sum_{r=1}^{T}\left(\sum_{s\neq r} A_{rs} + A_{sr}\right)\hat{y}_r^2 - 2\sum_{r=1}^{T}\sum_{s\neq r} A_{rs}\hat{y}_r\hat{y}_s.
\end{aligned}$$

The ridge regularizer (for the one-dimensional case) is usually defined on the function parameters: β^2 [32]. It is designed to encourage small parameter estimates. Similarly, the first term of the MTA regularizer above is a sum of weighted (single-task) ridge regularizers *of the outputs*. This term punishes large outputs. The second sum term, being negative, rewards large inner products between outputs of different tasks, weighted by their respective similarity. The second term, not being separable, ties the tasks together, but it also encourages uncontrolled growth of the estimated outputs. The two terms work in concert to tie the tasks together, while keeping the estimates from becoming too large or too small.

2.2 Closed-form Solution for the Scalar Case

When all A_{rs} are non-negative, the differentiable MTA objective is convex[1], and admits closed-form solutions, as well as an extensive analysis. Positive similarities encode attraction between means, while negative similarities encode repulsion. One might want two means to be far apart from one another if they represent, for example, locations of two objects which cannot occupy the same space. However, for the rest of this thesis I will constrain $A_{rs} \geq 0$ for all r, s, gaining analytic tractability and losing the ability to specify degrees of repulsion. With such a constraint, I will now find the closed-form optimal solution. First, I rewrite the objective in (2.1) in matrix notation:

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

where $L = D - (A + A^T)/2$ is the graph Laplacian matrix $(A + A^T)/2$ (see Section 1.5 for more on the graph Laplacian), Σ is a diagonal matrix with $\Sigma_{tt} = \frac{\sigma_t^2}{N_t}$, and $\hat{y}$ and $\bar{y}$ are column vectors with tth entries $\hat{y}_t$ and $\bar{y}_t$, respectively.

Note that the (t, t)th entry of the matrix Σ is the variance of $\bar{y}$.

Note further that the Laplacian is of the symmetrized $(A + A^T)/2$ and not of A. For simplicity of notation, I will assume from now on that A is symmetric. If, in practice, an

[1]This is sufficient but not necessary for convexity; as long as the graph Laplacian of A is positive-semi definite, the MTA objective is convex. However, non-negativity of A_{rs} is necessary for later results.

asymmetric A is provided, it can simply be symmetrized.

To find the closed-form solution, I now take the partial derivative of the above objective and equate to zero, obtaining

$$\begin{aligned} 0 &= \frac{1}{T}\left(2\Sigma^{-1}y^* - 2\Sigma^{-1}\bar{y}\right) + 2\frac{\gamma}{T^2}Ly^* \\ &= y^* - \bar{y} + \frac{\gamma}{T}\Sigma L y^* \\ \Leftrightarrow \bar{y} &= \left(I + \frac{\gamma}{T}\Sigma L\right) y^*, \end{aligned} \tag{2.2}$$

which yields the following optimal closed-form solution for y^*:

$$y^* = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1} \bar{y}, \tag{2.3}$$

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

Lemma 1: Assume that $0 \leq A_{rs} < \infty$ for all r, s, $\gamma \geq 0$, 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.

Proof (Due to B. A. Frigyik): 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 disc [33] $\mathcal{D}(B_{tt}, R_t)$ is the closed disc in the complex plane $\mathbb{C}$ 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 in the lemma statement. Also, it is clear that $B_{tt} > R_t$ for all t. Therefore, every Gershgorin disc is contained within the positive half-plane of $\mathbb{C}$, and, by the Gershgorin Circle Theorem [33], the real part of every eigenvalue of matrix B is positive. Its determinant is therefore positive, and the matrix B is invertible: $W = B^{-1}$. $\square$

Note that the (r, s)th entry of $\frac{\gamma}{T}\Sigma L$ goes to 0 as N_t approaches infinity, and since matrix

inversion is a continuous operation, $\left(I + \frac{\gamma}{T}\Sigma L\right)^{-1} \to I$ in the norm. By the law of large numbers one can conclude that y^* asymptotically approaches the true mean μ.

2.2.1 Closed-Form Solution for the Vector Case

MTA can also be applied to vectors. Why might one want to do this? If, for example, the measurements within each task are spatial coordinates (x, y, z), then the underlying dimensionality is three as opposed to one.

Let $\bar{\mathbb{Y}}^* \in \mathbb{R}^{T\times d}$ 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 $\Sigma \in R^{Td\times Td}$ is invertible.

2.2.2 Regularized Laplacian Kernel

The MTA solution matrix $W = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ is similar to the *regularized Laplacian kernel* (RLK): $Q = (I + \gamma L)^{-1}$, introduced by Smola and Kondor [50]. In the RLK, the graph Laplacian matrix L is assumed to be symmetric, but ΣL is not necessarily symmetric. The MTA solution matrix therefore generalizes the RLK.

Note that the term *kernel* refers to the positive-definite kernel used in, for example, support vector machines [32]. The (r, s)th entry of any kernel matrix can be thought of as a similarity between the rth and sth samples. In this section, I will discuss and motivate the kind of similarity that is encoded by the RLK.

Chebotarev and Shamis [19] studied matrices of the form $Q = (I+\gamma L)^{-1}$ in the context of answering the question "given a graph, how should one evaluate the proximity between its vertices?" They prove a number of properties that lead them to conclude that Q_{ij} is a good measure of how *accessible* j is from i when taking all possible paths into account (as opposed to just the direct path that A_{ij} encodes). In their own words, "Q_{ij} may be interpreted as the fraction of the connectivity of vertices i and j in the total connectivity of i with all vertices." The following is a list of interesting properties of Q from the work

of Chebotarev and Shamis when A is symmetric and its entries are "strictly positive" [19]:

- Q exists and is right-stochastic.
- $Q_{ii} > Q_{ij}$.
- Triangle inequality: $Q_{ij} + Q_{ik} - Q_{jk} \leq Q_{ii}$.
- The distance $d_{ij}^{\alpha} = \alpha(Q_{ii}^{\alpha} + Q_{jj}^{\alpha} - Q_{ij}^{\alpha} - Q_{ji}^{\alpha})$ is a valid metric distance over vertices.
- $Q_{ij} = 0$ if and only if there exists no path between i and j.

For intuition as to why Q measures connectivity, consider the following expansion[2] [10]:

$$(I + \gamma L)^{-1} = \sum_{k=0}^{\infty} (-\gamma L)^k.$$

Thus, the RLK is a type of path counting with $-L$ instead of A as the adjacency matrix, where paths of all possible lengths are taken into account, and longer paths are weighted: equally ($\gamma = 1$), less heavily ($\gamma < 1$), or more heavily ($\gamma > 1$).

As mentioned before, the MTA solution $\left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ is a more general version of the RLK; the diagonal matrix Σ left-multiplies the Laplacian, making it asymmetric. Using a different approach than Chebotarev and Shamis, I will prove that the right-stochasticity of W still holds in the next section, assuming only non-negativity of the entries of A (instead of strict positivity as in Chebotarev and Shamis). I defer the proofs that the remaining four properties still hold to a future work.

The RLK is one of many possible graph kernels. To find the best one for a collaborative recommendation task, Fouss et al. [29] empirically compared seven graph kernels. They found that the best three kernels were the RLK, $L^{\dagger}$, and the Markov diffusion kernel. Yajima and Kuo [57] tested various graph kernels in the context of a one-class SVM for the application of recommendation tasks. They also found that the RLK was one of the top performers.

---

[2] This equality holds only if the right-hand side is convergent.

2.2.3 Right-Stochasticity of the MTA Solution

From inspection of (2.3), it is clear that each of the elements of y^* is a linear combination of the sample single-task means in $\bar{y}$. In this section, I will prove that the weights are not just linear but *convex*. Recall that $W = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$.

Assumptions for the Theorem and Lemmas 2 and 3: $\gamma \geq 0$, $0 \leq A_{rs} < \infty$ for all r, s and $0 < \frac{\sigma_t^2}{N_t} < \infty$ for all t.

Theorem:[3] The weight matrix W is right stochastic.

Proof: The theorem requires showing that W exists (which is true from Lemma 1), is entry-wise non-negative and has rows that sum to 1. The last two properties are proven in the following lemmas.

Lemma 2: W has all non-negative entries.

Proof: By inspection it is clear that $W^{-1} = \left(I + \frac{\gamma}{T}\Sigma L\right)$ is a *Z-matrix*, defined to be a matrix with non-positive off-diagonal entries [10]. 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 $W \geq 0$ (element-wise)" (Chapter 6, Theorem 2.3, G_{20} and N_{38}, [10]). 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 3: The rows of W sum to 1, i.e. $W\mathbf{1} = \mathbf{1}$.

[3]Theorem hypothesized to be true by Maya Gupta. An earlier proof of this theorem is due to Bela Frigyik. That proof used cofactor expansions of matrix inverses. Here, I give a shorter proof that instead leverages known results about non-negative matrices.

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

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

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

2.3 Equivalent Formulations of MTA

It is straightforward to rewrite (2.1) in terms of the sample means rather than the samples themselves:

$$\arg\min_{\hat{y}\in\mathbb{R}^T} \ \frac{1}{T}(\bar{y}-\hat{y})^T\Sigma^{-1}(\bar{y}-\hat{y}) + \frac{\gamma}{T^2}\hat{y}^T L\hat{y}. \tag{2.4}$$

Additionally, from (2.2) (and assuming that Σ is invertible), one sees that the minimizer of (2.1) is also the minimizer of

$$\arg\min_{\hat{y}\in\mathbb{R}^T} \ \frac{1}{T}(\bar{y}-\hat{y})^T(\bar{y}-\hat{y}) + \frac{\gamma}{T^2}\hat{y}^T\Sigma L\hat{y}. \tag{2.5}$$

In other words, one can account for the different task-specific variances in the loss term or by scaling the rows of the similarity matrix A. This enables two different interpretations of the minimizer. Formulation (2.4) scales individual losses by the inverse of the variance, and formulation (2.5) results in weaker regularization for tasks with many samples, stronger regularization for tasks with a larger variance. Also, note that formulation (2.5) effectively results in an asymmetric Laplacian (since $\Sigma L(A) = L(\Sigma A)$ for diagonal Σ).

The MTA weight matrix $W = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ (here written as W^*) is also the solution to the following constrained objective, the formulation of which assumes that the estimator

is a convex combination of the samples means, $\hat{y} = W\bar{y}$:

$$W^* = \underset{W \in \mathbb{R}^{T\times T}}{\arg\min} \ \frac{1}{2T}(\bar{y} - W\bar{y})^T\Sigma^{-1}(\bar{y} - W\bar{y}) + \frac{\gamma}{2T^2}\bar{y}^T W^T L W\bar{y} \tag{2.6}$$
$$s.t.\ W\mathbf{1} = \mathbf{1}$$
$$W \geq 0.$$

To show that the optimal solution to this constrained convex and continuous objective is indeed the MTA solution $W^* = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$, I will first find the unconstrained optimal point and then check to see that it satisfies the constraints. Taking derivatives and setting equal to zero,

$$\begin{aligned}
0 &= \frac{\partial}{\partial W^*}\frac{1}{2T}(\bar{y} - W^*\bar{y})^T\Sigma^{-1}(\bar{y} - W^*\bar{y}) + \frac{\gamma}{2T^2}\bar{y}^T W^{*T} L^S W^*\bar{y} \\
&= -2\Sigma^{-1}\bar{y}\bar{y}^T + 2\Sigma^{-1}\bar{y}\bar{y}^T + \frac{\gamma}{T}LW^*\bar{y}\bar{y}^T \\
&= -\bar{y}\bar{y}^T + W^*\bar{y}\bar{y}^T + \frac{\gamma}{T}\Sigma L W^*\bar{y}\bar{y}^T \\
&= \left(\left(I + \frac{\gamma}{T}\Sigma L\right)W^* - I\right)\bar{y}\bar{y}^T.
\end{aligned}$$

One solution (of many) to this equation is the matrix W^* such that $\left(I + \frac{\gamma}{T}\Sigma L\right)W^* = I$, which requirement the MTA matrix $\left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ fulfills. Since I have already established that $\left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ exists and satisfies the right stochastic constraints of (2.6), it is thus an optimal solution of (2.6). Note that any $W^* + Z$ such that the columns of Z are perpendicular to $\bar{y}$ is also a potential solution, but may not satisfy the right-stochasticity constraint. Also, plugging $W^* + Z$ into the original objective will yield the same numerical minimum. Thus, I set $Z = 0$ without loss of generality.

2.4 MTA Formulation Variant

The MTA objective is

$$\{y_t^*\}_{t=1}^T = \underset{\{\hat{y}_t\}_{t=1}^T}{\arg\min} \ \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\left(\frac{y_{ti} - \hat{y}_t}{\sigma_t}\right)^2 + \frac{\gamma}{T^2}\sum_{r=1}^{T}\sum_{s=1}^{T}A_{rs}(\hat{y}_r - \hat{y}_s)^2.$$

Note that in the empirical loss, the terms are normalized by the task-specific standard deviation, but not so in the regularizer. One can also write a variant of MTA where the regularizer scales the estimates by their respective standard deviations:

$$\{y_t^*\}_{t=1}^T = \underset{\{\hat{y}_t\}_{t=1}^T}{\arg\min} \ \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\left(\frac{y_{ti}-\hat{y}_t}{\sigma_t}\right)^2 + \frac{\gamma}{T^2}\sum_{r=1}^{T}\sum_{s=1}^{T}A_{rs}\left(\frac{\hat{y}_r}{\sigma_r/\sqrt{N_r}} - \frac{\hat{y}_s}{\sigma_s/\sqrt{N_s}}\right)^2,$$

where $\sigma_r/\sqrt{N_r}$ is the standard deviation of the rth sample mean. The closed-form solution to this MTA formulation is

$$y^* = \Sigma^{1/2}\left(I + \frac{\gamma}{T}L\right)^{-1}\Sigma^{-1/2}\bar{y}. \tag{2.7}$$

In effect, the sample means are standardized to have identity covariance, MTA is applied, and then the sample means are unstandardized.

This formulation in practice performs worse than the plug-in approach; see Section 4.6.

2.5 Bayesian Interpretation of MTA

In this section, I consider MTA in a Bayesian framework. The sample means solve the following objective

$$\{\bar{y}_t\}_{t=1}^T = \underset{\{\hat{y}_t\}_{t=1}^T}{\arg\min} \ \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{(y_{ti}-\hat{y}_t)^2}{\sigma_t^2},$$

and can be interpreted as separately maximizing the likelihood of T Gaussian distributions. Extending this reasoning, the MTA estimates of the mean solve

$$\{y_t^*\}_{t=1}^T = \underset{\{\hat{y}_t\}_{t=1}^T}{\arg\min} \ \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{(y_{ti}-\hat{y}_t)^2}{\sigma_t^2} + \frac{\gamma}{T^2}\sum_{r=1}^{T}\sum_{s=1}^{T}A_{rs}(\hat{y}_r-\hat{y}_s)^2,$$

which can be interpreted as jointly maximizing the likelihood of T Gaussian distributions with a joint Gaussian Markov random field (GMRF) prior [43] on the solution. In MTA, the precision matrix Σ^{-1} 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 to MTA.

Additionally, none of the results I show in this paper require a Gaussian assumption nor any other assumption about the parametric form of the underlying distribution; I only require that $E[\bar{Y}\bar{Y}^T] = \mu\mu^T + \Sigma$ in (3.3) for the derivation of the risk expression for an arbitrary number of tasks.

2.6 Generality of Matrices of MTA Form

In this section, I will show with a proposition that a variety of common regularized estimators can be written using the form of the MTA solution matrix. For example, it is standard practice (as discussed in the introduction, see Efron and Morris [22]) to regularize each of the T single-task estimates towards the average of means in a convex way:

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

where $\lambda \in (0, 1]$. I will show that this estimator, among a set of many, can be written as MTA.

I use the expression 'matrices of MTA form' to refer to matrices that can be written

$$(I + \Gamma L(A))^{-1}, \tag{2.8}$$

where A is a matrix with all non-negative entries, and Γ is a diagonal matrix with all non-negative entries.

Figure 2.1 is a Venn diagram of the sets of estimators of type $\hat{Y} = W\bar{Y}$, where W is a $T \times T$ matrix. The pink region represents estimators of the form $\hat{Y} = W\bar{Y}$, with right-

stochastic W. The green region represents estimators with matrices of MTA form. The purple region includes many well-know estimators such as the James-Stein estimator (and its variants), and estimators that regularize single-task estimates of the mean to the pooled mean or the average-of-means.

In the following proposition, I will prove that the purple region is a strict subset of the the green region. That is, all purple-region estimators can be rewritten in MTA form for specific choices of (or assumptions about) A, γ, and Σ.[4] Before formally stating and proving the proposition, let us more closely examine the weight matrix of the purple region:

$$\left(\frac{1}{\gamma}I + \mathbf{1}\alpha^T\right)\bar{Y}.$$

This matrix is right-stochastic, just like the MTA weight matrix, but each of the rows is identical. Indeed, many common estimators regularize all single-task estimates toward the *same* value. The general form of MTA studied in this thesis results in a weight matrix where each row can be different (although the practical estimators presented in later sections have simpler structures), and thus allows regularizing each single-task estimate towards different values.

Proposition: The set of estimators $W\bar{Y}$ where W is of MTA form as per (2.8) is strictly larger than the set of estimators that regularize the single-task estimates as follows:

$$\hat{Y} = \left(\frac{1}{\gamma}I + \mathbf{1}\alpha^T\right)\bar{Y}, \tag{2.9}$$

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

Proof: First I will show that estimators $\hat{Y}$ in 2.9 can be written in MTA form. Using the

[4]Note that the covariance Σ is also a "choice" as, for example, some classic estimators assume $\Sigma = I$.

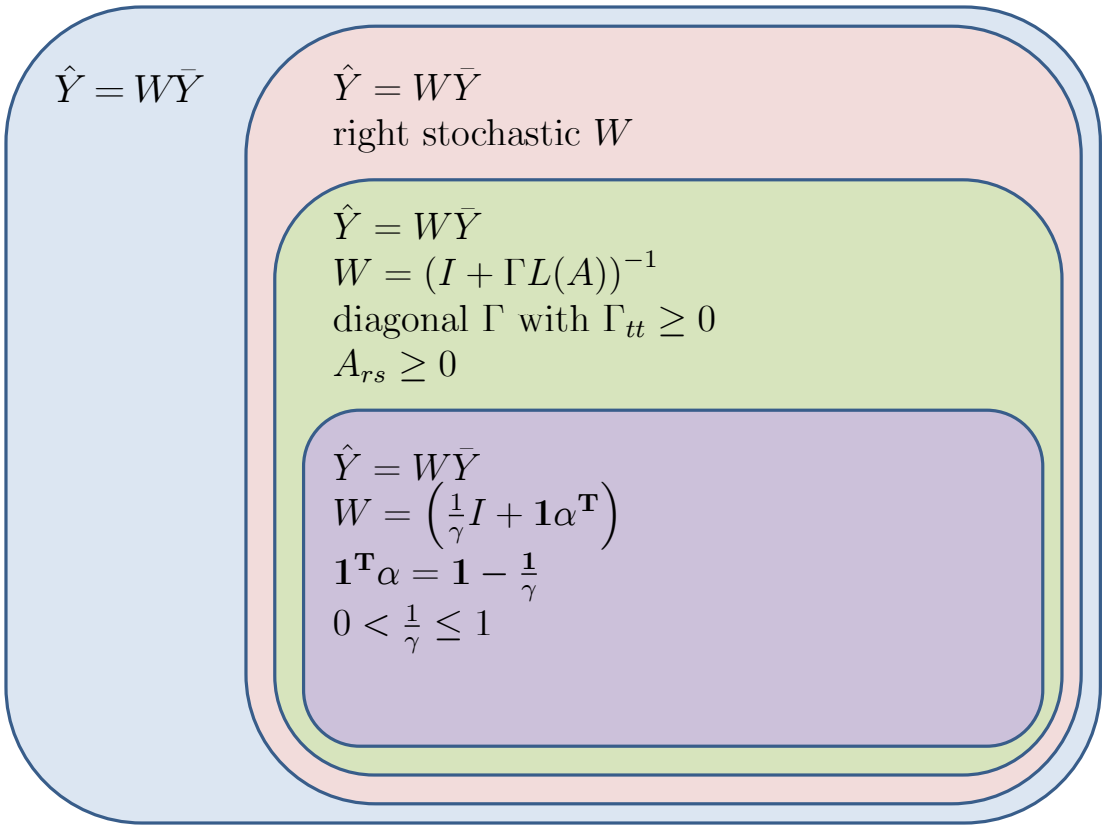

Figure 2.1: A Venn diagram of the set membership properties of various estimators of the type $\hat{Y} = W\bar{Y}$.

Sherman-Morrison [48] formula,

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

which is a matrix of MTA form for appropriate choices of γ, Σ, and A. Thus, estimators $\hat{Y}_t$

can be written in MTA form:

$$\hat{Y} = (I + \gamma L(\mathbf{1}\alpha^T))^{-1}\bar{Y}. \tag{2.10}$$

By inspection it is clear that not all matrices of the form $(I + \Gamma L(A))^{-1}$ can be written as (2.10) - one reason is that the matrix A has more degrees of freedom than afforded by the $\mathbf{1}\alpha^T$ term in (2.10). This implies that matrices of MTA form are strictly more general than matrices of the form in (2.10). □

Corollary 1: Estimators which 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

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

where $\mathbf{N}$ is a T by 1 vector with N_t as its tth entry, with corresponding choices of A and Γ obtained by visual pattern matching to (2.8).

Corollary 2: 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

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

with corresponding choices of A and Γ obtained by visual matching to (2.8).

Note that the proof of the proposition uses MTA form with *asymmetric* similarity matrix $A = \mathbf{1}\alpha^T$. The MTA form with asymmetric A arises if you replace the symmetric MTA regularization term in (2.1) with the following asymmetric regularization term:

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

The first term is just the MTA regularizer. The second term (for asymmetric similarity matrices) either adds a penalty or subtracts one, depending on the balance of the rows and columns. If the sum of the rows is larger than the sum of the columns, then smaller norms are encouraged. This would have to be taken into account when designing the matrix A, so as to reflect prior information.

2.7 Mean-Squared Error Analysis of MTA for $T = 2$

In this section I present analysis of MTA, which is drawn from joint work with M. R. Gupta and B. A. Frigyik [27]. Throughout, $\mu \in \mathbb{R}^T$ denotes the T-length vector of means for the T tasks, assumed to be finite, and capital Y's will be used to denote random variables. I take as given N_t IID (independent and identically distributed) random variables $\{Y_{ti}\}$ for the tth task with $i = 1, \ldots, N_t$. Let the vector of sample averages be $\bar{Y} \in \mathbb{R}^T$, where

$$\bar{Y}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} Y_{ti}.$$

No parametric form is assumed for the underlying sample distributions, but note that for many common parametric assumptions (such as Gaussian or Laplacian), the sample average is the maximum likelihood estimate.

I analyze the simple case of $T = 2$ tasks, with N_t samples per task. Specifically, suppose Y_1 is distributed with finite mean μ_1 and finite variance σ_1^2, and Y_2 is distributed with finite mean $\mu_2 = \mu_1 + \Delta$ and finite variance σ_2^2. W.l.o.g (without loss of generality), let the task-relatedness matrix be $A = [0\; a; a\; 0]$. For a fixed A, the regularization parameter γ in (2.1) is a useful degree of freedom to control the power of the regularizer, but in this analysis I treat A as a variable (that is, I treat a as a variable), and thus γ can be set to 1 w.l.o.g. For

this simple choice of A, the matrix inversion in the closed-form solution (2.3) is tractable, producing the following properties.

Regularized Estimate: The MTA estimate is a convex combination of the sample averages (for $a \geq 0$):

$$Y_1^* = \left(\frac{T + \frac{\sigma_2^2}{N_2} a}{T + \frac{\sigma_1^2}{N_1} a + \frac{\sigma_2^2}{N_2} a} \right) \bar{Y}_1 + \left(\frac{\frac{\sigma_1^2}{N_1} a}{T + \frac{\sigma_1^2}{N_1} a + \frac{\sigma_2^2}{N_2} a} \right) \bar{Y}_2. \tag{2.11}$$

Bias: From (2.11) it follows that the MTA estimate is biased:

$$\mathrm{E}[Y_1^*] - \mu_1 = \left(\frac{\frac{\sigma_1^2}{N_1} a}{T + \frac{\sigma_1^2}{N_1} a + \frac{\sigma_2^2}{N_2} a} \right) \Delta. \tag{2.12}$$

Variance: From (2.11) it follows that the MTA estimate has a smaller variance than the sample average (for $a > 0$):

$$\begin{aligned} \mathrm{Var}[Y_1^*] &= \frac{\sigma_1^2}{N_1} \left(\frac{T^2 + 2T\frac{\sigma_2^2}{N_2} a + \frac{\sigma_1^2 \sigma_2^2}{N_1 N_2} a^2 + \frac{\sigma_2^4}{N_2^2} a^2}{(T + \frac{\sigma_1^2}{N_1} a + \frac{\sigma_2^2}{N_2} a)^2} \right) \\ &= \frac{\sigma_1^2}{N_1} \left(\frac{T^2 + 2T\frac{\sigma_2^2}{N_2} a + \frac{\sigma_1^2 \sigma_2^2}{N_1 N_2} a^2 + \frac{\sigma_2^4}{N_2^2} a^2}{T^2 + 2T\frac{\sigma_1^2}{N_1} a + 2T\frac{\sigma_2^2}{N_2} a + \frac{\sigma_1^4}{N_1^2} a^2 + 2\frac{\sigma_1^2 \sigma_2^2}{N_1 N_2} a^2 + \frac{\sigma_2^4}{N_2^2} a^2} \right) \\ &< \frac{\sigma_1^2}{N_1} = \mathrm{Var}[\bar{Y}_1]. \end{aligned} \tag{2.13}$$

Mean-Squared Error: From (2.12) and (2.13), the mean-squared error for estimating only μ_1 is

$$\mathrm{MSE}[Y_1^*] = \frac{\sigma_1^2}{N_1} \left(\frac{T^2 + 2T\frac{\sigma_2^2}{N_2} a + \frac{\sigma_1^2 \sigma_2^2}{N_1 N_2} a^2 + \frac{\sigma_2^4}{N_2^2} a^2}{(T + \frac{\sigma_1^2}{N_1} a + \frac{\sigma_2^2}{N_2} a)^2} \right) + \frac{\Delta^2 \frac{\sigma_1^4}{N_1^2} a^2}{(T + \frac{\sigma_1^2}{N_1} a + \frac{\sigma_2^2}{N_2} a)^2}. \tag{2.14}$$

Comparing to $\text{MSE}[\bar{Y}_1] = \frac{\sigma_1^2}{N_1}$ and plugging in $T = 2$, one can conclude from (2.14) that

$$\text{MSE}[Y_1^*] < \text{MSE}[\bar{Y}_1] \text{ if } \Delta^2 - \frac{\sigma_1^2}{N_1} - \frac{\sigma_2^2}{N_2} < \frac{4}{a}, \tag{2.15}$$

with the assumption that $a > 0$. Thus, with a properly chosen a, the risk (sum of the mean squared errors over all tasks) will be smaller for MTA than for single-task averaging. Specifically, if one had the true (oracle) μ_t and σ_t^2, one could set a in the following range:

$$a \begin{cases} < \frac{4}{\Delta^2 - \frac{\sigma_1^2}{N_1} - \frac{\sigma_2^2}{N_2}} & \text{if } \Delta^2 - \frac{\sigma_1^2}{N_1} - \frac{\sigma_2^2}{N_2} > 0 \\ > 0 & \text{if } \Delta^2 - \frac{\sigma_1^2}{N_1} - \frac{\sigma_2^2}{N_2} \leq 0. \end{cases}$$

Thus the MTA estimate has lower MSE if the mean-separation Δ^2 is small compared to the variances of the sample mean. See Figure (2.2) for an illustration. Even in cases where Δ^2 is very large, a small amount of regularization can still be helpful.

2.8 Optimal Task Relatedness for $T = 2$

I 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:

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

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. Minimizing the risk $\text{MSE}[Y_1^*] + \text{MSE}[Y_2^*]$ w.r.t. a one obtains the following solution:

$$a^* = \frac{2}{(\mu_1 - \mu_2)^2}, \tag{2.16}$$

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

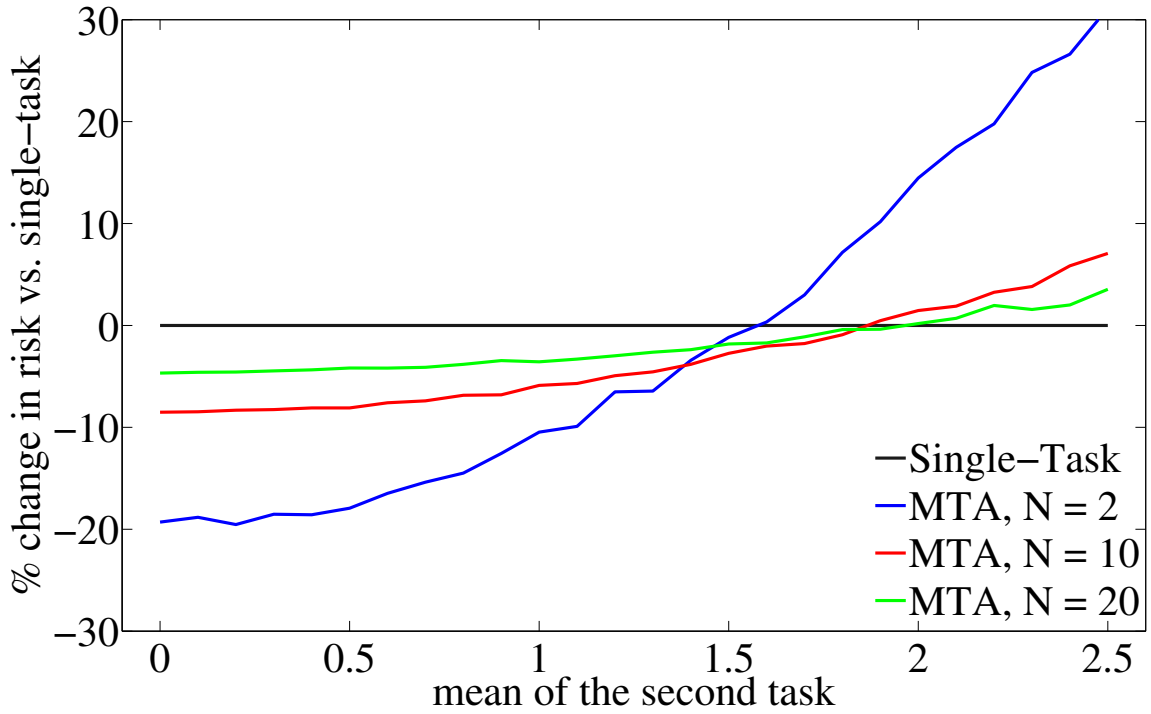

Figure 2.2: Plot shows the percent change in average risk for two tasks (averaged over 10,000 runs of the simulation). For each task there are N IID samples, for $N = 2, 10, 20$. The first task generates samples from a standard Gaussian. The second task generates samples from a Gaussian with $\sigma^2 = 1$ and varying mean value, 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 if the Gaussians are closer than 2 units apart. Given $N = 20$ samples from each Gaussian, the MTA estimate is better if the Gaussians are closer than 1.5 units apart. In the extreme case that the two Gaussians have the same mean ($\mu_1 = \mu_2 = 0$), then with this suboptimal choice of $a = 1$, MTA provides a 20% win for $N = 2$ samples, and a 5% win for $N = 20$ samples.

or the sample variance σ_t^2, as these are accounted for in Σ. Note that a^* also minimizes the functions MSE$[Y_1^*]$ and MSE$[Y_2^*]$, separately.

Analysis of the second derivative shows that the minimizer in (2.16) always holds for the cases of interest (that is, for $N_1, N_2 \geq 1$). The effect on the risk of the choice of a and the optimal a^* is illustrated in Figure 2.3.

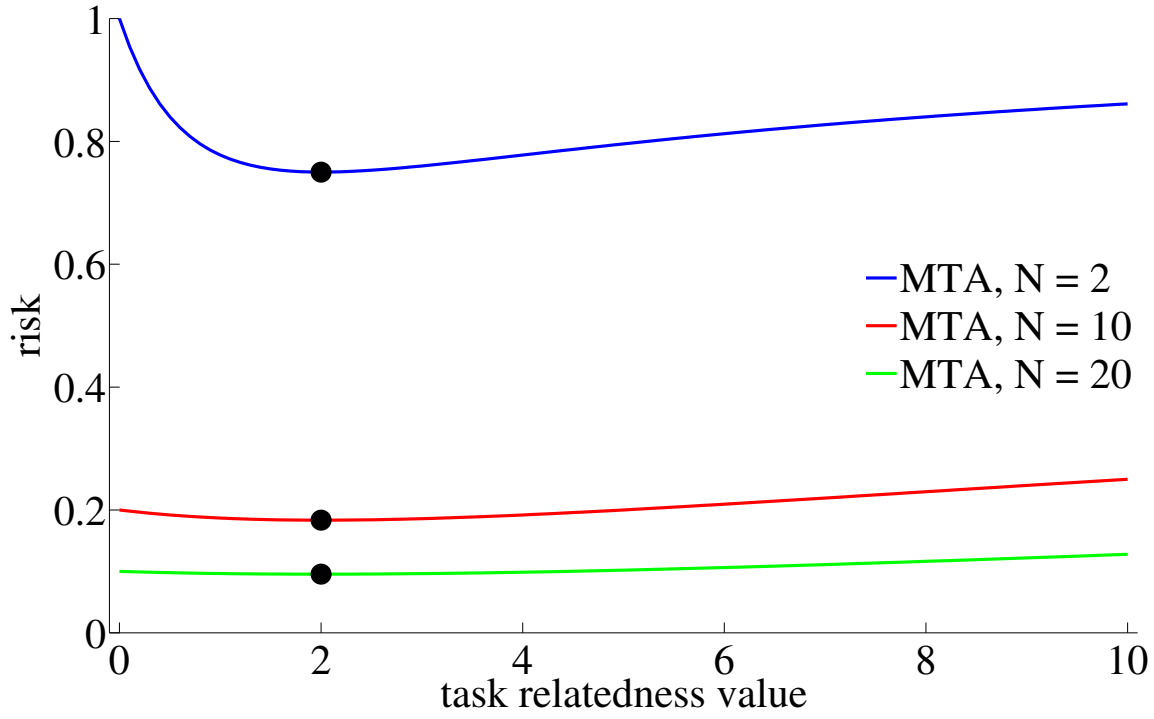

Figure 2.3: Plot shows the risk for two tasks as given in (2.14), where the task samples were drawn IID from Gaussians $\mathcal{N}(0,1)$ and $\mathcal{N}(1,1)$. The task relatedness value a was varied as shown on the x-axis. The minimum expected squared error is marked by a $*$, is independent of N and matches the optimal task-relatedness value given by (2.16).

Chapter 3

ESTIMATING THE SIMILARITY MATRIX FROM DATA

The optimal two-task similarity given in (2.16) requires knowledge of the true means μ_1 and μ_2. These are, in practice, unavailable. What similarity should be used then?

In this chapter, I will first discuss two approaches to the estimation of the optimal similarity for $T = 2$. Using this as launching off point, I will then derive closed-form solutions for two practical and efficient forms of MTA for arbitrary T: constant MTA in Section 3.3 and minimax MTA in Section 3.4. In addition I will discuss three other approaches: pairwise MTA, pooled MTA, and average-of-means MTA.

3.1 Estimating the Optimal Similarity for $T = 2$

A straightforward approach to approximating the two-task similarity in (2.16) is to use the 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 [15].

A second approach to setting a^* in practice is to minimize Stein's unbiased risk estimate (SURE) [52]. For an estimator $h(U)$, where $U = \Sigma^{-1}\bar{Y}$ is a sufficient statistic for estimating μ, the SURE estimate of the risk is a function $S(h(U))$ such that

$$E[S(h(U))] = R(\mu, h(U)).$$

In other words $S(h(U))$ is an unbiased estimate of the risk that depends only on the available data, and not on the unknown true parameters. The approach is to minimize $S(h(U))$ instead of the risk. Eldar derived the SURE for exponential families [24]. Specifically, for

the linear Gaussian model where

$$\bar{Y} = \mu + \epsilon,$$

where ϵ is zero-mean Gaussian noise with covariance Σ, the SURE for $h(U)$ is

$$S(h(U)) = ||\mu||^2 + ||h(U)||^2 - 2h^T(U)\bar{Y} + 2\text{tr}\left(\frac{\partial h(U)}{\partial u}\right),$$

where $\bar{Y}$ is the maximum likelihood estimate of μ. In the general case that I am studying in this thesis $h(U) = W\bar{Y}$ and

$$S(W\bar{Y}) = \mu^T\mu + \bar{Y}^T W^T W\bar{Y} - 2Y^T WY + 2\text{tr}(W\Sigma).$$

Note that minimizing $S(W\bar{Y})$ w.r.t. W is equivalent to minimizing

$$\begin{aligned}\arg\min_W S(W\bar{Y}) &= \arg\min_W \mu^T\mu + \bar{Y}^T W^T W\bar{Y} - 2Y^T WY + 2\text{tr}(W\Sigma) \\ &= \arg\min_W \bar{Y}^T(W-I)^T(W-I)\bar{Y} + 2\text{tr}(W\Sigma), \end{aligned} \tag{3.1}$$

which does not depend on μ, but only on the known $\bar{Y}$. Note that because W is right stochastic (3.1) can be written

$$S(W\bar{Y}) = \bar{Y}^T L(W)^T L(W)\bar{Y} + 2\text{tr}(W\Sigma),$$

where $L(W)$ is the graph Laplacian of W. I minimize this expression for the $T = 2$ case when W is the MTA solution matrix in (2.3) using maximum likelihood estimates for the variances. In a derivation analogous that of previous chapter, one obtains the optimal solution:

$$a^*_{\text{SURE}} = \frac{2}{(\bar{Y}_1 - \bar{Y}_2)^2 - \frac{\hat{\sigma}_1^2}{N_1} - \frac{\hat{\sigma}_2^2}{N_2}},$$

of which I take the positive-part to ensure the non-negativity of A:

$$a^*_{\text{SURE}} = \left(\frac{2}{(\bar{Y}_1 - \bar{Y}_2)^2 - \frac{\hat{\sigma}_1^2}{N_1} - \frac{\hat{\sigma}_2^2}{N_2}}\right)^+, \tag{3.2}$$

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

Note that compared to the pairwise plug-in estimate $\hat{a}^* = \frac{2}{(\bar{Y}_1 - \bar{Y}_2)^2}$, (3.2) is *less* conservative, that is, always larger (unless the positive-part boundary is not triggered). Also, the plug-in estimate does not require a positive-part adjustment as it never becomes negative.

3.2 The Risk Expression for Arbitrary T

The *risk* of estimator Y^* of unknown parameter vector μ for the squared loss is the sum of the mean squared errors:

$$R(\mu, Y^*) = \sum_{t=1}^{T} \mathrm{MSE}[Y_t^*] = E[(Y^* - \mu)^T (Y^* - \mu)].$$

Simplifying for the MTA estimator $Y^* = W\bar{Y}$ and noting that $E[\bar{Y}\bar{Y}^T] = \mu\mu^T + \Sigma$, we have

$$\begin{aligned}
R(\mu, W\bar{Y}) &= E[(W\bar{Y} - \mu)^T (W\bar{Y} - \mu)] \\
&= E[\bar{Y}^T W^T W \bar{Y}] - 2\mu^T W E[\bar{Y}] - \mu^T \mu \\
&= E[\mathrm{tr}(\bar{Y}\bar{Y}^T W^T W)] - 2\mu^T W \mu - \mu^T \mu \\
&= \mathrm{tr}(E[\bar{Y}\bar{Y}^T] W^T W) - 2\mu^T W \mu - \mu^T \mu \\
&= \mathrm{tr}((\mu\mu^T + \Sigma) W^T W) - 2\mu^T W \mu - \mu^T \mu \\
&= \mathrm{tr}(W\Sigma W^T) + \mathrm{tr}(W\mu\mu^T W^T) - 2\mu^T W \mu + \mu^T \mu \qquad (3.3) \\
&= \mathrm{tr}(W\Sigma W^T) + \mu^T W^T W \mu - 2\mu^T W \mu + \mu^T \mu \\
&= \mathrm{tr}(W\Sigma W^T) + \mu^T W^T W \mu - \mu^T W \mu - \mu^T W^T \mu + \mu^T \mu \\
&= \mathrm{tr}(W\Sigma W^T) + \mu^T (I - W)^T (I - W) \mu, \qquad (3.4)
\end{aligned}$$

where the penultimate equality is true because $\mu^T W \mu = \mu^T W^T \mu$. Note that for the case where W is right stochastic,

$$R(\mu, W\bar{Y}) = \mathrm{tr}(W\Sigma W^T) + \mu^T L(W)^T L(W) \mu,$$

where $L(W)$ is the graph Laplacian of W.

3.3 Constant MTA

One approach to generalizing the results of Section 2.8 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 number of means we are trying to estimate in the first place. To avoid these problems, I generalize the two-task results by constraining A to be a scaled constant matrix $A = a\mathbf{1}\mathbf{1}^T$, resulting in the following weight matrix (γ is set to 1 w.l.o.g.):

$$W^{\text{cnst}} = \left(I + \frac{1}{T}\Sigma L(a\mathbf{1}\mathbf{1}^T)\right)^{-1}. \tag{3.5}$$

I will find the optimal a^* that minimizes the risk of the estimator $W^{\text{cnst}}\bar{Y}$ in (3.4). In addition, for analytic tractability of the derivation, I assume that all the tasks have the same variance, estimating Σ as $\frac{\text{tr}(\Sigma)}{T}I$. Then it remains to solve:

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

First, I simplify $\left(I + \frac{1}{T}\frac{\mathrm{tr}(\Sigma)}{T}L(a\mathbf{1}\mathbf{1}^T)\right)^{-1}$ using the Sherman-Morrison formula [48],

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

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

$$\begin{aligned}
R(\mu, Y^*) &= \mathrm{tr}\left(\frac{1}{a\frac{\mathrm{tr}(\Sigma)}{T}+1}\left(I + a\frac{\mathrm{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^T\right)\Sigma I \frac{1}{a\frac{\mathrm{tr}(\Sigma)}{T}+1}\left(I + a\frac{\mathrm{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^T\right)^T\right) \\
&\quad + \mu^T\left(\frac{1}{a\frac{\mathrm{tr}(\Sigma)}{T}+1}\left(I + a\frac{\mathrm{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^T\right) - I\right)^T\left(\frac{1}{a\frac{\mathrm{tr}(\Sigma)}{T}+1}\left(I + a\frac{\mathrm{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^T\right) - I\right)\mu \\
&= \frac{1}{(a\frac{\mathrm{tr}(\Sigma)}{T}+1)^2}\mathrm{tr}\left(\left(I + a\frac{\mathrm{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^T\right)\Sigma\left(I + a\frac{\mathrm{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^T\right)\right) \\
&\quad + \mu^T\left(\frac{-a\frac{\mathrm{tr}(\Sigma)}{T}}{a\frac{\mathrm{tr}(\Sigma)}{T}+1}I + \frac{a\frac{\mathrm{tr}(\Sigma)}{T}}{a\frac{\mathrm{tr}(\Sigma)}{T}+1}\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)^T\left(\frac{-a\frac{\mathrm{tr}(\Sigma)}{T}}{a\frac{\mathrm{tr}(\Sigma)}{T}+1}I + \frac{a\frac{\mathrm{tr}(\Sigma)}{T}}{a\frac{\mathrm{tr}(\Sigma)}{T}+1}\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)\mu \\
&= \frac{1}{(a\frac{\mathrm{tr}(\Sigma)}{T}+1)^2}\mathrm{tr}\left(\Sigma + 2a\frac{\mathrm{tr}(\Sigma)}{T^2}\mathbf{1}\mathbf{1}^T\Sigma + a^2\frac{\mathrm{tr}(\Sigma)^2}{T^4}\mathbf{1}\mathbf{1}^T\Sigma\mathbf{1}\mathbf{1}^T\right) \\
&\quad + \frac{(a\frac{\mathrm{tr}(\Sigma)}{T})^2}{(a\frac{\mathrm{tr}(\Sigma)}{T}+1)^2}\mu^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)\mu \\
&= \frac{1}{(a\frac{\mathrm{tr}(\Sigma)}{T}+1)^2}\left(\mathrm{tr}(\Sigma) + 2a\frac{\mathrm{tr}(\Sigma)^2}{T^2} + a^2\frac{\mathrm{tr}(\Sigma)^3}{T^3}\right) \\
&\quad + \frac{(a\frac{\mathrm{tr}(\Sigma)}{T})^2}{(a\frac{\mathrm{tr}(\Sigma)}{T}+1)^2}\mu^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)\mu \\
&= \frac{\frac{\mathrm{tr}(\Sigma)}{T}}{(a\frac{\mathrm{tr}(\Sigma)}{T}+1)^2}\left(T + 2a\frac{\mathrm{tr}(\Sigma)}{T} + \left(a\frac{\mathrm{tr}(\Sigma)}{T}\right)^2\right) \\
&\quad + \frac{(a\frac{\mathrm{tr}(\Sigma)}{T})^2}{(a\frac{\mathrm{tr}(\Sigma)}{T}+1)^2}\mu^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\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

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

and omitting some tedious algebra,

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

For $T = 2$, this result exactly matches the result obtained in Section 2.8: $a^* = \frac{2}{(\mu_1-\mu_2)^2}$. Intuitively, the denominator of a^* is the average squared distance between means. Plugging this a^* into W^{cnst} in (3.5), the constant MTA weight matrix is

$$W^{\mathrm{cnst}} = \left(I + \frac{1}{T}\Sigma L(a^*\mathbf{1}\mathbf{1}^T)\right)^{-1}.$$

In practice, one of course does not have $\{\mu_r\}$ as these are precisely the values one is trying to estimate. So, to estimate a^* I use the sample means $\{\bar{Y}_r\}$:

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

Using this optimal estimated constant similarity and an estimated covariance matrix $\hat{\Sigma}$ produces what we refer to as the *constant MTA* estimate

$$Y^* = \left(I + \frac{1}{T}\hat{\Sigma}L(\hat{a}^*\mathbf{1}\mathbf{1}^T)\right)^{-1}\bar{Y}. \tag{3.6}$$

Note that I made the assumption that the entries of Σ were the same in order to be able to derive the constant similarity a^*, but I do not need nor suggest that assumption on the $\hat{\Sigma}$ used with $\hat{a}^*$ in (3.6).

3.4 Minimax MTA

In preliminary experiments I found that Bock's James-Stein type estimator [11] outperformed constant MTA in some statistical settings. The reason is that Bock's estimator is minimax, i.e. it minimizes the worst-case loss, instead of the expected loss, in order to ensure that the worst possible error is relatively small. Minimizing the risk (expected loss) provides no such guarantees - performance is good only *on average*. So, to ensure that MTA remains competitive with minimax estimators, in this section I will derive a minimax-inspired MTA. First, some definitions are in order.

- The *risk* of an estimator $\hat{Y}$ of unknown parameter vector μ for the squared loss is the sum of the mean squared errors:

  $$R(\mu, \hat{Y}) = E[(\mu - \hat{Y})^T(\mu - \hat{Y})].$$

  In the MTA case, with $\hat{Y} = Y^* = W\bar{Y}$, the risk is

  $$R(\mu, Y^*) = \mathrm{tr}(W\Sigma W^T) + \mu^T(W - I)^T(W - I)\mu. \tag{3.7}$$

- An estimator Y^M of μ which minimizes the maximum risk

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

  is called a *minimax* estimator.

- The *average risk* for estimator $\hat{Y}$ is

  $$r(\pi, \hat{Y}) = \int R(\mu, \hat{Y})\pi(\mu)d\mu, \tag{3.8}$$

  where π is a prior on μ.

- The estimator that minimizes the average risk is called the *Bayes estimator* and is

written

$$Y_\pi = \arg\min_{\hat{Y}} r(\pi, \hat{Y}).$$

- The *Bayes risk* is the risk of the Bayes estimator and is written

$$r_\pi = r(\pi, Y_\pi) = \int R(\mu, Y_\pi)\pi(\mu)d\mu. \tag{3.9}$$

- A prior distribution π is *least favorable* if $r_\pi \geq r_{\pi'}$ for all priors π'.

To find a minimax MTA, I will need the following theorem and corollary (Theorem 1.4, Chapter 5 [37]).

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

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

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 (3.9)), 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. I will use the box constraint set, $\mu_t \in [b_l, b_u]^T$, where $b_l \in \mathbb{R}$ and $b_u \in \mathbb{R}$. It is straightforward to show that the

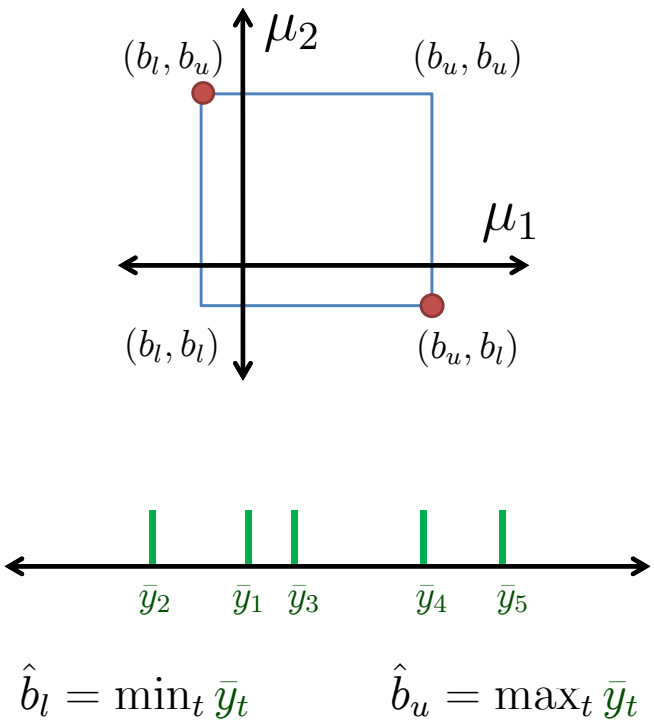

Figure 3.1: The top portion of the figure is an illustration of the square constraint set and the LFP - the red dots are the locations of non-zero probability. The bottom portion of the figure illustrates how b_u and b_l are estimated when $T > 2$, that is, the max and min, respectively, of all T sample means.

corresponding LFP is

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

See the top portion of Figure 3.1 for an illustration. 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 it is minimax. For the $T = 2$ case, I guess W^M to be

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

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 I 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:

$$\begin{aligned} &\frac{1}{2} \left([b_l \ b_u](W - I)^T (W - I) \begin{bmatrix} b_l \\ b_u \end{bmatrix} + \text{tr}(W \Sigma W^T) \right) \\ &+ \frac{1}{2} \left([b_u \ b_l](W - I)^T (W - I) \begin{bmatrix} b_u \\ b_l \end{bmatrix} + \text{tr}(W \Sigma W^T) \right). \end{aligned} \tag{3.10}$$

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

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

is $W^* = \left(I + \frac{2}{T(\mu_1 - \mu_2)^2} \Sigma L(\mathbf{1}\mathbf{1}^T) \right)^{-1}$. Thus, the minimizer of either term in (3.10) (and thus the minimizer of (3.10)) is

$$W^M = \left(I + \frac{2}{T(b_u - b_l)^2} \Sigma L(\mathbf{1}\mathbf{1}^T) \right)^{-1} \tag{3.11}$$

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.

This minimax analysis is only valid for the case when $T = 2$, but I found by experimentation that the following extension of minimax MTA to larger T worked well in simulations and applications for any $T \geq 2$. To estimate b_u and b_l from data I assume the unknown T means are drawn from a uniform distribution and use maximum likelihood estimates for

the lower and upper endpoints of the support:

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

See the bottom portion of Figure 3.1 for an illustration. Thus, in practice, *minimax MTA*[1] is

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

3.5 Computational Efficiency of Constant and Minimax MTA

Both the constant MTA and minimax MTA weight matrices can be written as

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

where c is different for constant MTA and minimax MTA, $Z = I + cT\Sigma$, and $z = c\Sigma\mathbf{1}$. The Sherman-Morrison formula [48] can be used to find the inverse:

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

Since Z is diagonal, Z^{-1} can be computed in $O(T)$ time, and so can $Z^{-1}z$.

The c for constant MTA is $\hat{a}^* = \frac{2}{\frac{1}{T(T-1)} \sum_{r=1}^{T} \sum_{s=1}^{T} (\bar{Y}_r - \bar{Y}_s)^2}$. The denominator contains a double sum that would take, naively, $O(T^2)$ time to compute. However, it can be decom-

[1]For $T > 2$ this estimator can not rightly be called minimax. Its full name is *minimax-inspired MTA*, but I call it minimax MTA for short.

posed fruitfully as follows:

$$\begin{aligned}\sum_{r=1}^{T}\sum_{s=1}^{T}(\bar{Y}_r-\bar{Y}_s)^2 &= \sum_{r=1}^{T}\sum_{s=1}^{T}(\bar{Y}_r^2+\bar{Y}_s^2-2\bar{Y}_r\bar{Y}_s)\\ &= \sum_{r=1}^{T}\bar{Y}_r^2\sum_{s=1}^{T}+\sum_{s=1}^{T}\bar{Y}_s^2\sum_{r=1}^{T}-2\sum_{r=1}^{T}\sum_{s=1}^{T}\bar{Y}_r\bar{Y}_s\\ &= 2T\sum_{r=1}^{T}\bar{Y}_r^2-2\sum_{r=1}^{T}\bar{Y}_r\sum_{s=1}^{T}\bar{Y}_s\\ &= 2T\sum_{r=1}^{T}\bar{Y}_r^2-2\left(\sum_{r=1}^{T}\bar{Y}_r\right)^2,\end{aligned}$$

both terms of which can be computed in $O(T)$ time.

Similarly, the c for constant MTA requires two one-dimensional searches:

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

both of which can also be computed in $O(T)$ time.

Thus, the entire computation $W\bar{Y}$ can be done in $O(T)$ time for constant MTA and minimax MTA.

3.6 Pairwise MTA

Another intuitive extension of MTA for more than two tasks is to simply use the optimal two-task similarity to populate the similarity matrix for any $T \geq 2$ as per (2.16):

$$A_{rs}^{\text{pr}}=\frac{2}{(\mu_r-\mu_s)^2},$$

with the estimated similarity using the maximum likelihood estimates of the means:

$$\hat{A}_{rs}^{\text{pr}}=\frac{2}{(\bar{Y}_r-\bar{Y}_s)^2}.$$

But as will be shown in the simulations section, pairwise MTA is strictly worse than constant MTA, its closest competitor.

3.7 Pooled MTA

According to Corollary 1 from Section 2.6, the estimator that regularizes single-task estimates towards the pooled mean

$$\check{Y}_t = \lambda \bar{Y}_t + \frac{1-\lambda}{\sum_{r=1}^{T} N_r} \mathbf{N}^T \bar{Y}$$

can be written as $\check{Y} = W^{\text{pl}} \bar{Y}$ where

$$W^{\text{pl}} = \left(\lambda I + \frac{1-\lambda}{\mathbf{N}^T \mathbf{1}} \mathbf{1} \mathbf{N}^T \right).$$

To compare this estimator to constant MTA and minimax MTA, I will now find the optimal value of λ by minimizing the risk (assuming that $\bar{Y}_r$ and $\bar{Y}_s$ are independent for all r, s):

$$\begin{aligned}
R(\mu, \check{Y}) &= \operatorname{tr}\left(\left(\lambda I + \frac{1-\lambda}{\mathbf{N}^T \mathbf{1}} \mathbf{1} \mathbf{N}^T \right) \Sigma \left(\lambda I + \frac{1-\lambda}{\mathbf{N}^T \mathbf{1}} \mathbf{1} \mathbf{N}^T \right)^T \right) \\
&\quad + \mu^T \left(\lambda I + \frac{1-\lambda}{\mathbf{N}^T \mathbf{1}} \mathbf{1} \mathbf{N}^T - I \right)^T \left(\lambda I + \frac{1-\lambda}{\mathbf{N}^T \mathbf{1}} \mathbf{1} \mathbf{N}^T - I \right) \mu \\
&= \operatorname{tr}\left(\lambda^2 \Sigma + \frac{\lambda(1-\lambda)}{\mathbf{N}^T \mathbf{1}} \mathbf{N} \mathbf{1}^T \Sigma + \frac{\lambda(1-\lambda)}{\mathbf{N}^T \mathbf{1}} \mathbf{1} \mathbf{N}^T \Sigma + \frac{(1-\lambda)^2}{(\mathbf{N}^T \mathbf{1})^2} \mathbf{N} \mathbf{1}^T \mathbf{1} \mathbf{N}^T \Sigma \right) \\
&\quad + (\lambda - 1)^2 \mu^T \left(\frac{1}{\mathbf{N}^T \mathbf{1}} \mathbf{1} \mathbf{N}^T - I \right)^T \left(\frac{1}{\mathbf{N}^T \mathbf{1}} \mathbf{1} \mathbf{N}^T - I \right) \mu.
\end{aligned}$$

The partial derivative set at zero is

$$\begin{aligned}
\frac{\partial R(\mu, \check{Y})}{\partial \lambda_{\text{pl}}} = 0 &= 2\lambda_{\text{pl}} \operatorname{tr}(\Sigma) + \frac{1 - 2\lambda_{\text{pl}}}{\mathbf{N}^T \mathbf{1}} \operatorname{tr}(\mathbf{N} \mathbf{1}^T \Sigma + \mathbf{1} \mathbf{N}^T \Sigma) + 2T \frac{\lambda_{\text{pl}} - 1}{(\mathbf{N}^T \mathbf{1})^2} \operatorname{tr}(\mathbf{N} \mathbf{N}^T \Sigma) \\
&\quad + 2(\lambda_{\text{pl}} - 1) \mu^T \left(\frac{1}{\mathbf{N}^T \mathbf{1}} \mathbf{1} \mathbf{N}^T - I \right)^T \left(\frac{1}{\mathbf{N}^T \mathbf{1}} \mathbf{1} \mathbf{N}^T - I \right) \mu. \\
&= 2\lambda_{\text{pl}} \operatorname{tr}(\Sigma) + 2 \frac{1 - 2\lambda_{\text{pl}}}{\mathbf{N}^T \mathbf{1}} \mathbf{1}^T \Sigma \mathbf{N} + 2T \frac{\lambda_{\text{pl}} - 1}{(\mathbf{N}^T \mathbf{1})^2} \mathbf{N}^T \Sigma \mathbf{N} \\
&\quad + 2(1 - \lambda_{\text{pl}}) \mu^T L \left(\frac{1}{\mathbf{N}^T \mathbf{1}} \mathbf{1} \mathbf{N}^T \right)^T L \left(\frac{1}{\mathbf{N}^T \mathbf{1}} \mathbf{1} \mathbf{N}^T \right) \mu,
\end{aligned}$$

from which I obtain

$$\begin{aligned}
\lambda_{\text{pl}} &= \frac{\mu^T L \left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)^T L \left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)\mu + \frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}^T\Sigma\mathbf{N} - \frac{T}{(\mathbf{N}^T\mathbf{1})^2}\mathbf{N}^T\Sigma\mathbf{N}}{\mu^T L \left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)^T L \left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)\mu + \frac{2}{\mathbf{N}^T\mathbf{1}}\mathbf{1}^T\Sigma\mathbf{N} - \frac{T}{(\mathbf{N}^T\mathbf{1})^2}\mathbf{N}^T\Sigma\mathbf{N} + \text{tr}(\Sigma)} \\
&= \frac{\mu^T L \left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)^T L \left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)\mu + \frac{1}{\mathbf{N}^T\mathbf{1}}\left(\sum_{t=1}^T \sigma_t^2 - T\frac{\sum_{t=1}^T N_t\sigma_t^2}{\sum_{t=1}^T N_t}\right)}{\mu^T L \left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)^T L \left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)\mu + \frac{1}{\mathbf{N}^T\mathbf{1}}\left(2\sum_{t=1}^T \sigma_t^2 - T\frac{\sum_{t=1}^T N_t\sigma_t^2}{\sum_{t=1}^T N_t}\right) + \text{tr}(\Sigma)} \\
&= \frac{\mu^T L \left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)^T L \left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)\mu + \frac{1}{\mathbf{N}^T\mathbf{1}}\sum_{t=1}^T\left(\sigma_t^2 - \sigma_{\text{pl}}^2\right)}{\mu^T L \left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)^T L \left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)\mu + \frac{1}{\mathbf{N}^T\mathbf{1}}\sum_{t=1}^T\left(2\sigma_t^2 - \sigma_{\text{pl}}^2\right) + \text{tr}(\Sigma)},
\end{aligned}$$

where $\sigma_{\text{pl}}^2 = \frac{\sum_{t=1}^T N_t\sigma_t^2}{\sum_{t=1}^T N_t}$ is the pooled variance. The unbiased pooled variance is $\sigma_{\text{pl}}^2 = \frac{\sum_{t=1}^T (N_t-1)\sigma_t^2}{\sum_{t=1}^T (N_t-1)}$. I also take the positive part of the solution to ensure non-negativity: $(\lambda_{\text{pl}})^+$.

3.8 Average-of-Means MTA

From Corollary 2 of Section 2.6, the estimator that regularizes single-task estimates towards the average-of-means (AM)

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

can also be written $\breve{Y} = W^{\text{am}}\bar{Y}$, where

$$W^{\text{am}} = \left(\lambda I + \frac{1-\lambda}{T}\mathbf{1}\mathbf{1}^T\right).$$

To compare this estimator to other MTA forms, I will repeat the procedure of the previous subsection and find the optimal λ w.r.t. the risk.

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

The partial derivative set at zero is

$$\begin{aligned}
\frac{\partial R(\mu, \breve{Y})}{\partial \lambda_{\text{am}}} = 0 &= 2\lambda_{\text{am}}\operatorname{tr}(\Sigma) - 2\frac{\lambda_{\text{am}}}{T}\operatorname{tr}(\Sigma) + 2(\lambda_{\text{am}} - 1)\mu^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)\mu \\
\Leftrightarrow \lambda_{\text{am}} &= \frac{\mu^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)\mu}{\mu^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)\mu + (1 - \frac{1}{T})\operatorname{tr}(\Sigma)} \\
&= \frac{\mu^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)\mu}{\mu^T L\left(\frac{1}{T}\mathbf{1}\mathbf{1}^T\right)\mu + (1 - \frac{1}{T})\operatorname{tr}(\Sigma)} \\
&= \frac{\mu^T L\left(\mathbf{1}\mathbf{1}^T\right)\mu}{\mu^T L\left(\mathbf{1}\mathbf{1}^T\right)\mu + (T-1)\operatorname{tr}(\Sigma)},
\end{aligned}$$

where the second-to-last equality is true because

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

3.9 A Summary of Proposed Estimators

I summarize the various forms of MTA discussed in Table 3.1. Note that $\bar{Y}_{\text{am}} \in \mathbb{R}^T$ is a constant vector of the average-of-means, where each entry is $\frac{1}{T}\mathbf{1}^T\bar{Y}$, and $\bar{Y}_{\text{pl}} \in \mathbb{R}^T$ is a constant vector where each is the pooled mean $\frac{1}{\sum_{r=1}^{T} N_r}\sum_{t=1}^{T}\sum_{i=1}^{N_t} y_{ti}$. Note further that the optimal parameters are now written with hats, indicating that they are based on estimates: specifically, the unknown means μ are replaced with maximum likelihood (ML) estimates $\bar{Y}$, and the unknown matrix of variances Σ is also replaced by its ML estimate $\hat{\Sigma}$.

Table 3.1: Summary of the Considered Estimators

Single-Task	$\bar{Y}$
Robust James-Stein	$\hat{\lambda}_{\text{js}}\bar{Y} + (1-\hat{\lambda}_{\text{js}})\bar{Y}_{\text{am}}$ $\hat{\lambda}_{\text{js}} = \left(1 - \frac{T-3}{(\bar{Y}-\bar{Y}_{\text{am}})^T\hat{\Sigma}^{-1}(\bar{Y}-\bar{Y}_{\text{am}})}\right)^+$
Constant MTA	$\left(I + \frac{\hat{a}^*\gamma}{T}\hat{\Sigma}L(\mathbf{1}\mathbf{1}^T)\right)^{-1}\bar{Y}$ $\hat{a}^* = \frac{2T(T-1)}{\bar{Y}^T L(\mathbf{1}\mathbf{1}^T)\bar{Y}}$
Minimax MTA	$\left(I + \frac{2\gamma}{T(\hat{b}_u-\hat{b}_l)^2}\hat{\Sigma}L(\mathbf{1}\mathbf{1}^T)\right)^{-1}\bar{Y}$ $\hat{b}_u = \max_t \bar{Y}_t,\ \hat{b}_l = \min_t \bar{Y}_t$
Pairwise MTA	$\left(I + \frac{\gamma}{T}\hat{\Sigma}L(\hat{A}^{\text{pr}})\right)^{-1}\bar{Y}$ $\hat{A}^{\text{pr}}_{rs} = \frac{2}{(\bar{Y}_r-\bar{Y}_s)^2}$
Pooled MTA	$\hat{\lambda}_{\text{pl}}\bar{Y} + (1-\hat{\lambda}_{\text{pl}})\bar{Y}_{\text{pl}}$ $\hat{\lambda}_{\text{pl}} = \left(\frac{\bar{Y}^T L\left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)^T L\left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)\bar{Y} + \frac{1}{\mathbf{N}^T\mathbf{1}}\sum_{t=1}^T\left(\hat{\sigma}_t^2-\hat{\sigma}_{\text{pl}}^2\right)}{\bar{Y}^T L\left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)^T L\left(\frac{1}{\mathbf{N}^T\mathbf{1}}\mathbf{1}\mathbf{N}^T\right)\bar{Y} + \frac{1}{\mathbf{N}^T\mathbf{1}}\sum_{t=1}^T\left(2\hat{\sigma}_t^2-\hat{\sigma}_{\text{pl}}^2\right)+\text{tr}(\hat{\Sigma})}\right)^+$
Average-of-Means MTA	$\hat{\lambda}_{\text{am}}\bar{Y} + (1-\hat{\lambda}_{\text{am}})\bar{Y}_{\text{am}}$ $\hat{\lambda}_{\text{am}} = \frac{\bar{Y}^T L(\mathbf{1}\mathbf{1}^T)\bar{Y}}{\bar{Y}^T L(\mathbf{1}\mathbf{1}^T)\bar{Y} + (T-1)\text{tr}(\hat{\Sigma})}$

Chapter 4

SIMULATIONS

In this chapter I present a variety of simulated data experiments to test the usefulness of constant MTA and minimax MTA formulations given in the theory chapter.[1]

I 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 constant MTA and minimax MTA work well when averaged over many different draws of means, variances, and numbers of samples.

4.1 Varying Distance Between Means

Simulations in this section test the performance of MTA as the average distance between the true means μ_t is varied. Simulations were 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 4.1 and 4.2 for the Gaussian experiments, and Figures 4.3 and 4.4 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.

[1] In preliminary experiments, pairwise MTA, pooled MTA, and average-of-means MTA performed nearly always worse than constant MTA and minimax MTA.

(b) Draw the variance of the tth distribution $\sigma_t^2 \sim \text{Gamma}(0.9, 1.0) + 0.1$.[2]

(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 $y_{ti} \sim U[\mu_t - \sqrt{3\sigma_t^2}, \mu_t + \sqrt{3\sigma_t^2}]$.

Simulation details are also shown in Table 4.1.

I compared constant MTA and minimax MTA to single-task sample averages and to the James-Stein estimator given in (1.4). I also compared to randomized 5-fold cross-validated (CV) versions of James-Stein and MTA. For the CV versions, I randomly subsampled $N_t/2$

[2]The 0.1 is added to ensure that variance is never zero.

Gaussian Simulations	**Uniform Simulations**
$\mu_t \sim \mathcal{N}(0, \sigma_\mu^2)$	$\mu_t \sim U(-\sqrt{3\sigma_\mu^2}, \sqrt{3\sigma_\mu^2})$
$\sigma_t^2 \sim \text{Gamma}(0.9, 1.0) + 0.1$	$\sigma_t^2 \sim U(0.1, 2.0)$
$N_t \sim U\{2, \ldots, 100\}$	$N_t \sim U\{2, \ldots, 100\}$
$y_{ti} \sim \mathcal{N}(\mu_t, \sigma_t^2)$	$y_{ti} \sim U(\mu_t - \sqrt{3\sigma_t^2}, \mu_t + \sqrt{3\sigma_t^2})$

Table 4.1: Simulation details.

samples and chose the value of γ for constant/minimax MTA 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 constant/minimax MTA, 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:

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

where $\bar{\bar{y}} = \frac{1}{T}\sum_{r=1}^{T} \bar{y}_r$.

I used the following parameters for CV: $\gamma \in \{2^{-5}, 2^{-4}, \ldots, 2^5\}$ for the MTA estimators and a comparable set of λ spanning $(0, 1)$ by the transformation $\lambda = \frac{\gamma}{\gamma+1}$. When cross-validating MTA, I keyed the range of γ by the value of $\hat{a}^*$ from consant MT, creating a data-adaptive scale for γ, where $\gamma = 1$ sets the regularization parameter to be $\hat{a}^*$ or $\frac{1}{(\hat{b}_u - \hat{b}_l)^2}$, respectively.

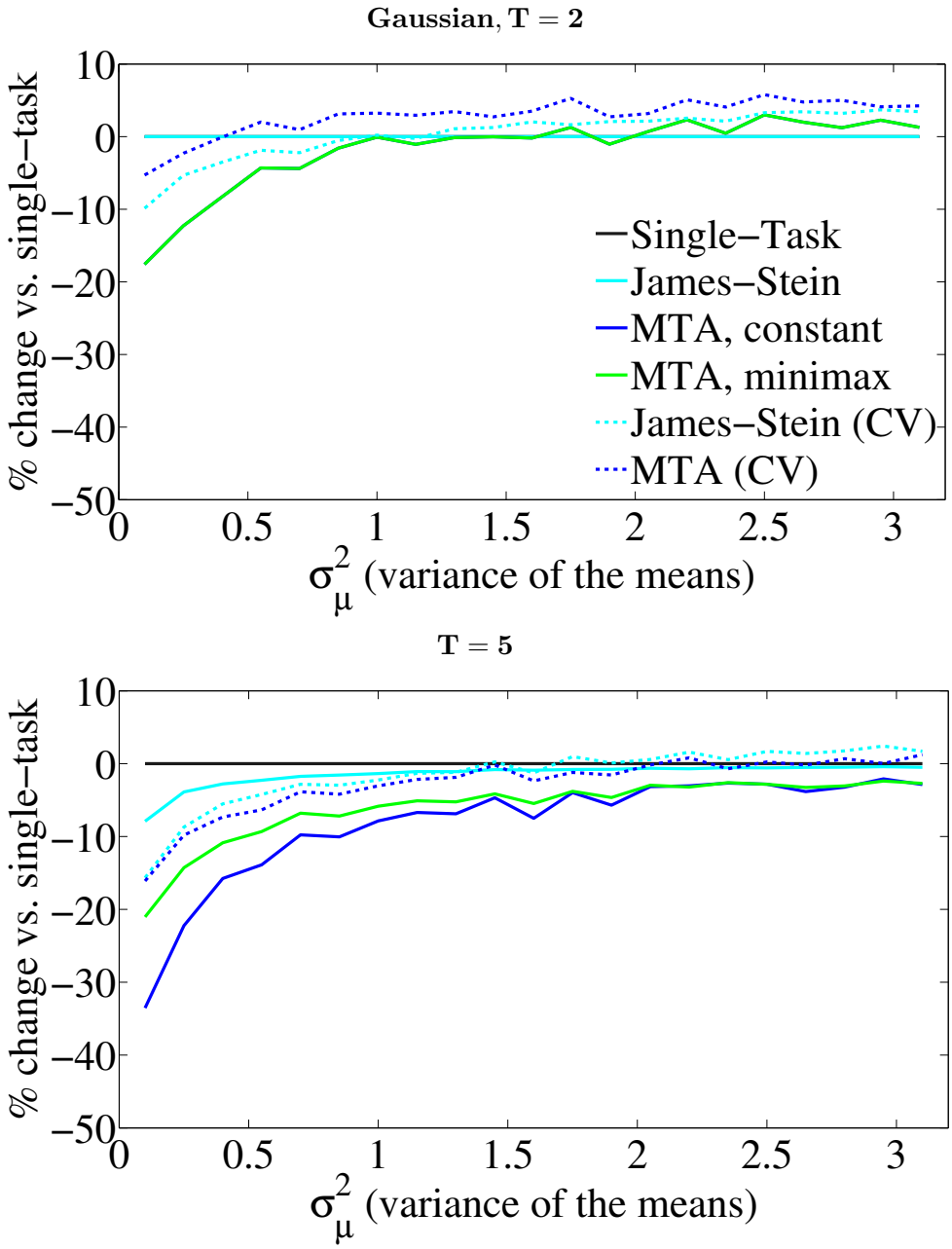

Figure 4.1: Gaussian experiment results for $T = \{2, 5\}$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task. Note: for $T = 2$ the James-Stein estimator reduces to single-task, and so the cyan and black lines overlap. Similarly, for $T = 2$, constant MTA and minimax MTA are identical, and so the blue and green lines overlap.

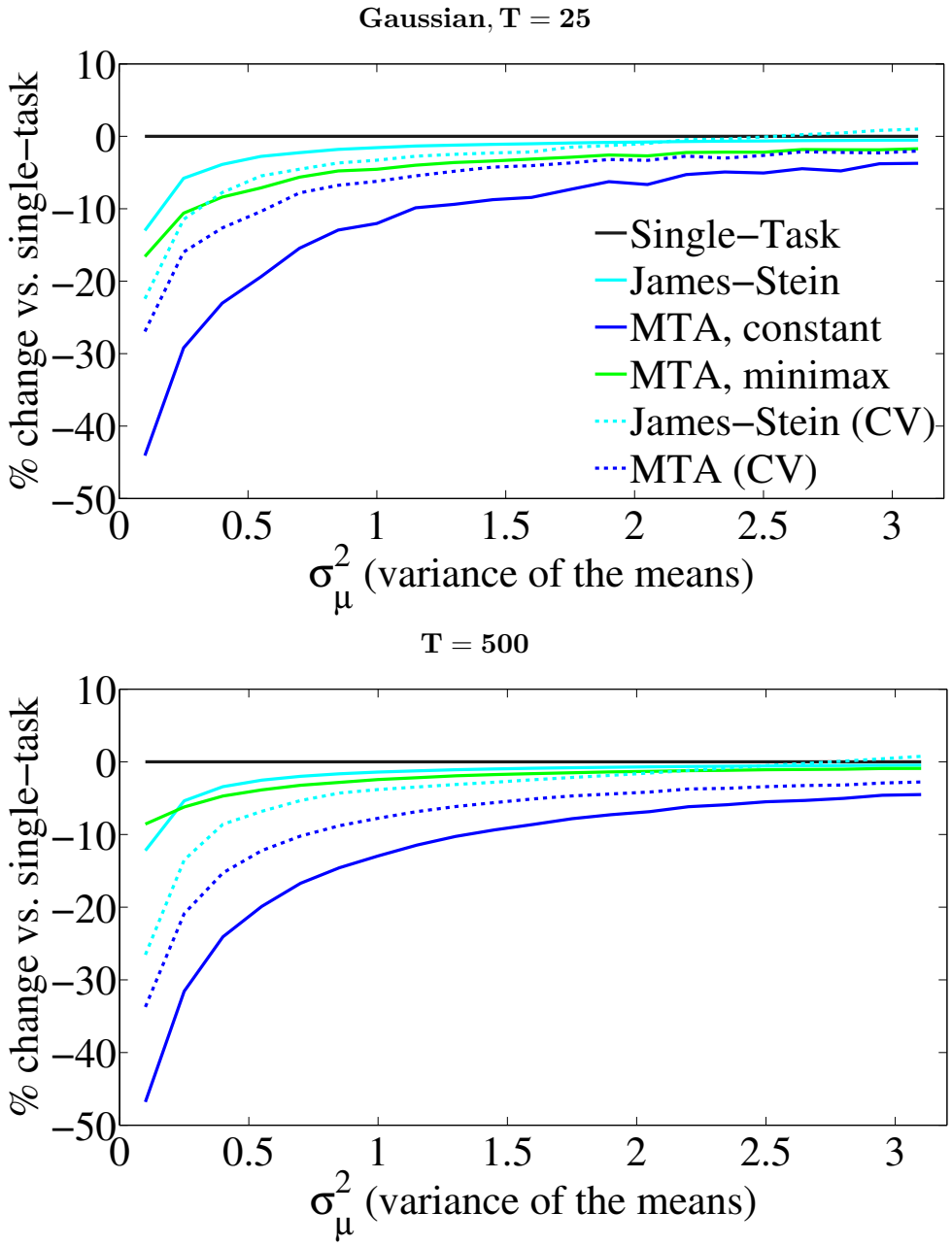

Figure 4.2: Gaussian experiment results for $T = \{25, 500\}$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task.

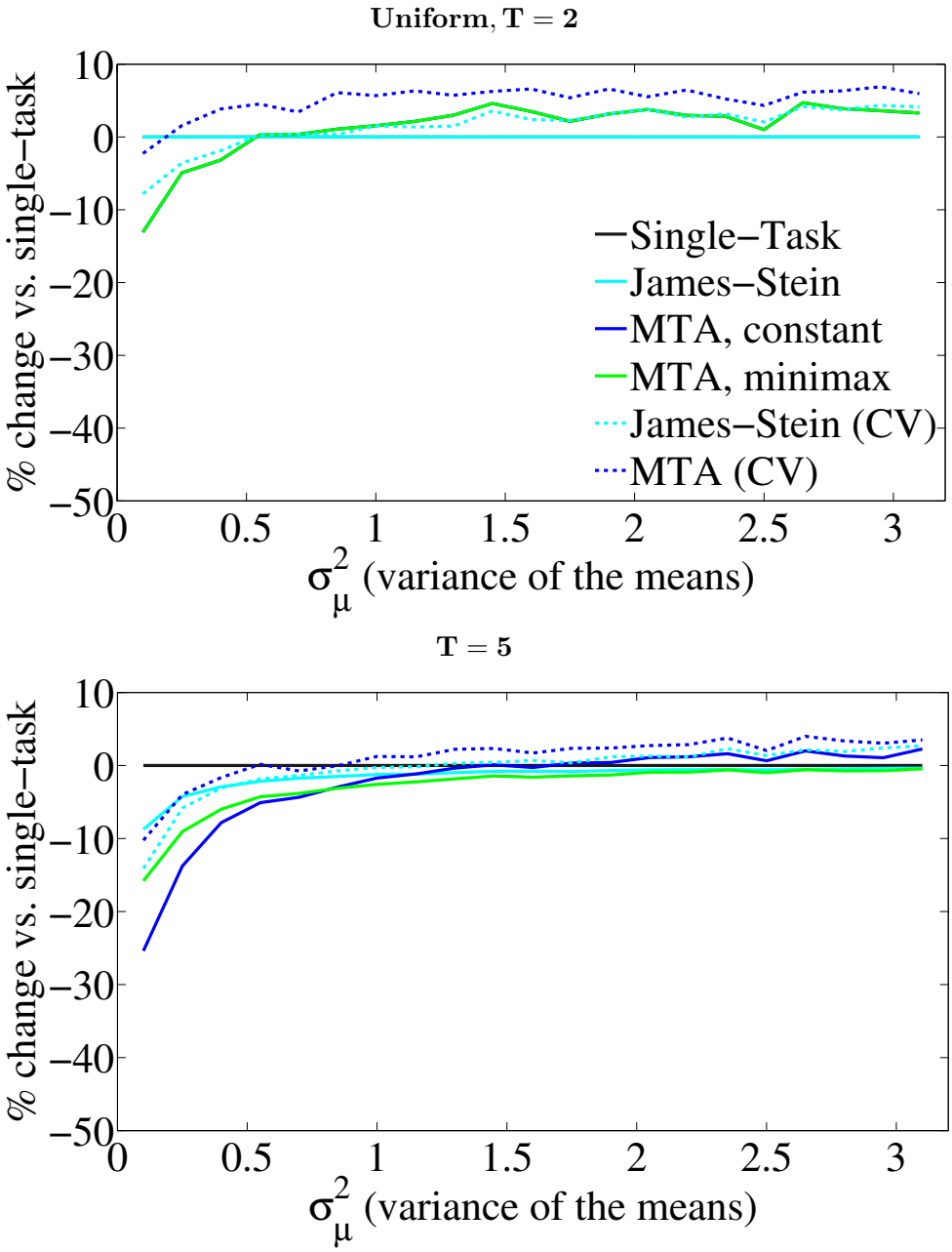

Figure 4.3: Uniform experiment results for $T = \{2, 5\}$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task. Note: for $T = 2$ the James-Stein estimator reduces to single-task, and so the cyan and black lines overlap. Similarly, for $T = 2$, constant MTA and minimax MTA are identical, and so the blue and green lines overlap.

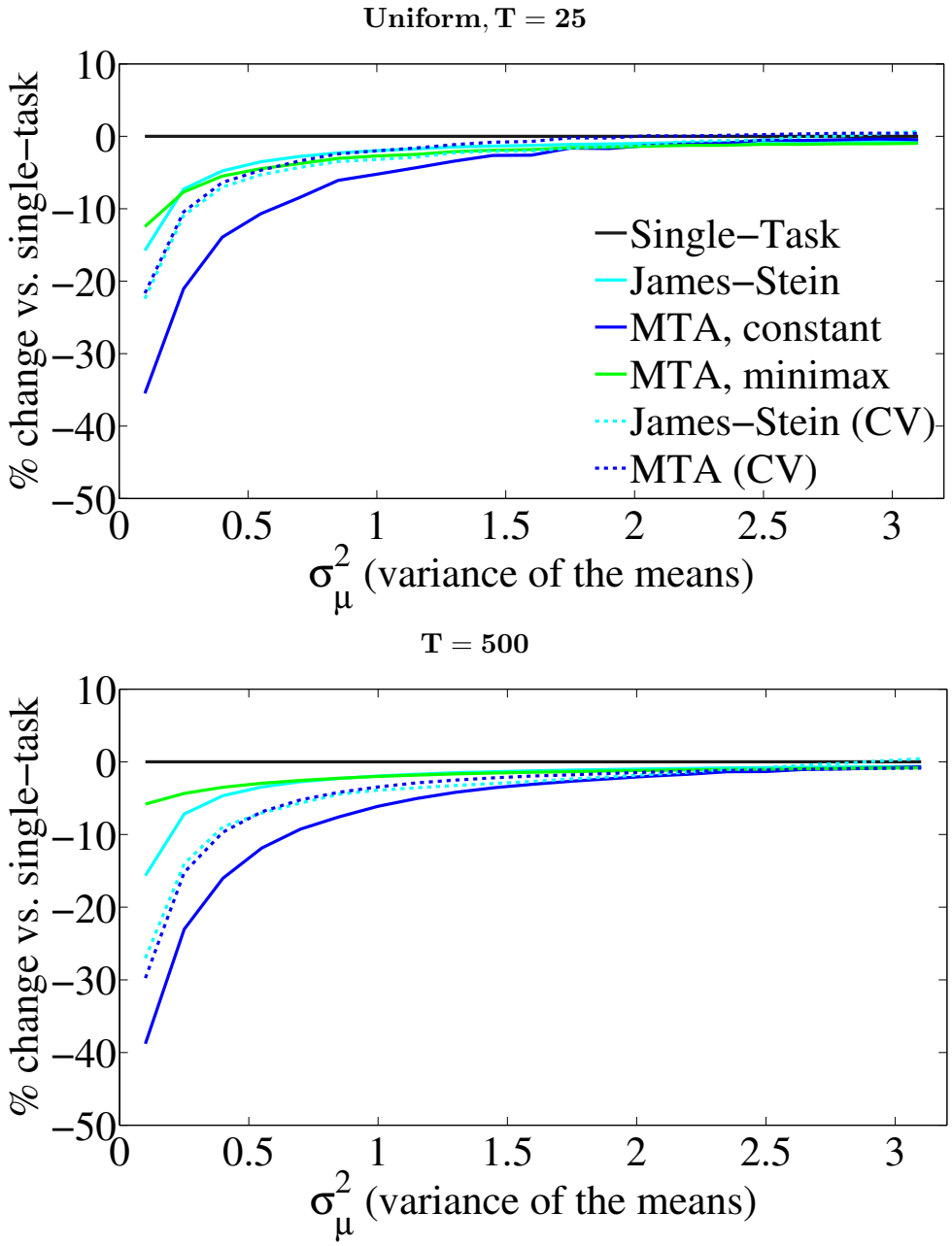

Figure 4.4: Uniform experiment results for for $T = \{25, 500\}$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task.

Some observations from Figures 4.1-4.4:

- 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.
- 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 $T = 5$, constant MTA dominates in the Gaussian case, but in the uniform case does worse than single-task when the means are far apart. Note that for all $T > 2$ minimax MTA almost always outperforms James-Stein and always outperforms single-task, which is to be expected as it was designed conservatively.
- For $T = 25$ and $T = 500$, one sees the trend 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 further tasks levels off early on.
- For constant MTA, cross-validation is always worse than the estimated optimal regularization, while the opposite is true for minimax MTA. This is to be expected, as minimax estimators are not designed to minimizes the average risk, which is both what I report as well as the metric optimized during cross-validation.
- Performance on the uniform experiments is generally worse because samples are more likely (compared to the Gaussian distribution) to appear at the edges.

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

4.2 Varying the Number of Samples

Simulations in this section test the performance of MTA as the average number of samples per task is varied. In the previous section, σ_μ^2 was varied, and N_t were always drawn from

an integer uniform distribution between 2 and 100. In this section, σ_μ^2 is fixed (at 1 or 2), and the upper boundary of the integer uniform is varied from 3 to 253, in increments of 25.

The simulations were run for $T = \{5, 500\}$ tasks for both uniform and Gaussian simulations. The other simulation details are identical to those of the preceding section; cross-validation results were left out to enhance the clarity of the plots. Results are reported in Figures 4.5 and 4.6 for the Gaussian experiments, and Figures 4.7 and 4.8 for the uniform experiments.

The plots reveal that having more samples does not significantly decrease the potential benefit from MTA. Constant MTA, minimax MTA, and the James-Stein estimator have a relatively slow decrease in performance after the upper range of N_t reaches 25.

Other observations:

- The MTA estimators on average outperform the James-Stein estimator for the entire range of N_t in the Gaussian simulations.
- In the uniform simulations, James-Stein sometimes outperforms minimax MTA when the N_t is a maximum of 100, but this ordering switches thereafter.
- As in the preceding section, when σ_μ^2 is high and the number of tasks is low, constant MTA does worse than single-task for the uniform experiments.
- Having many more tasks helps constant MTA considerably, but has little effect on minimax MTA or James-Stein.

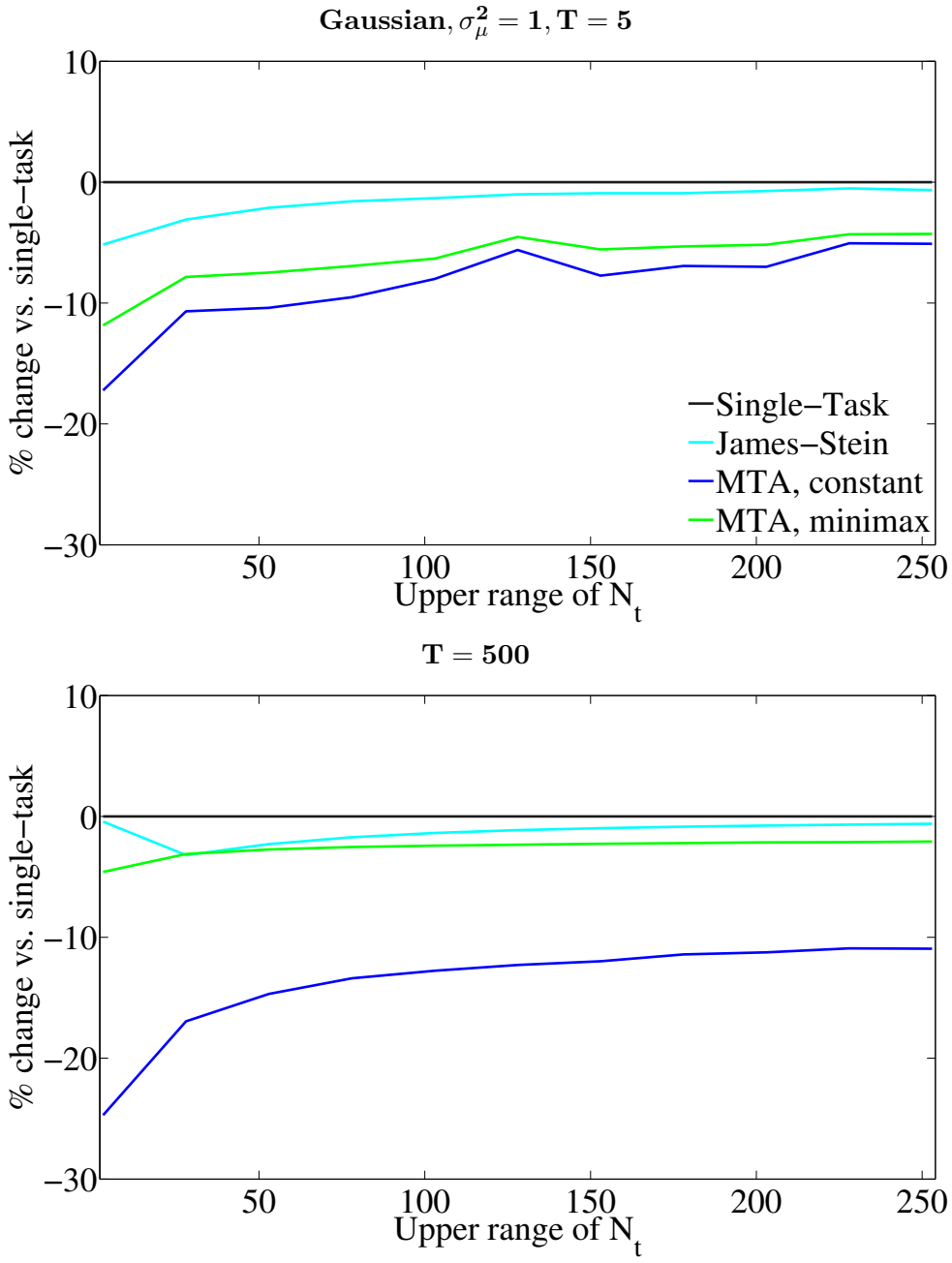

Figure 4.5: Gaussian experiment results for $T = \{5, 500\}$ with $\sigma_\mu^2 = 1$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task.

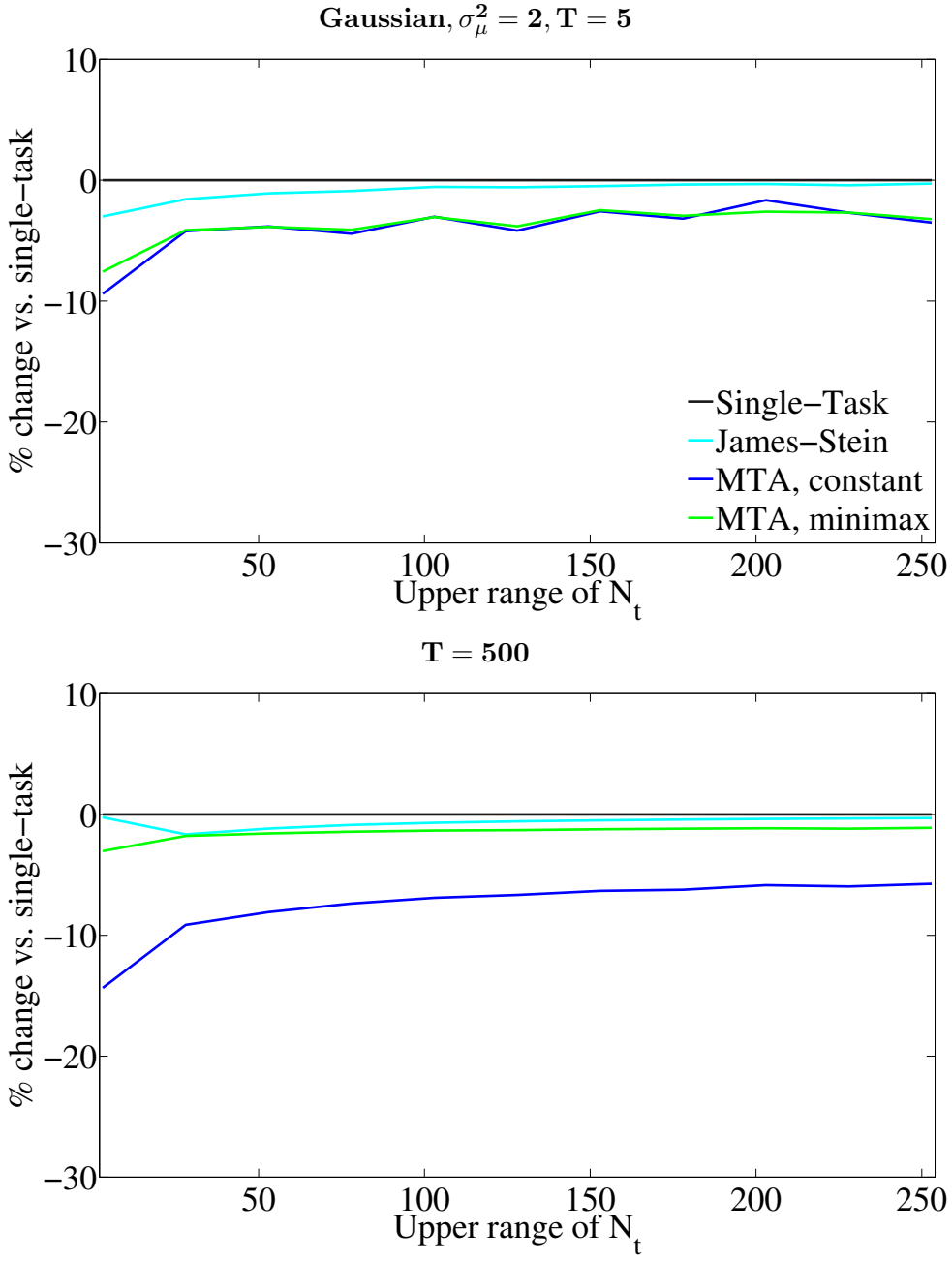

Figure 4.6: Gaussian experiment results for $T = \{5, 500\}$ with $\sigma_\mu^2 = 2$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task.

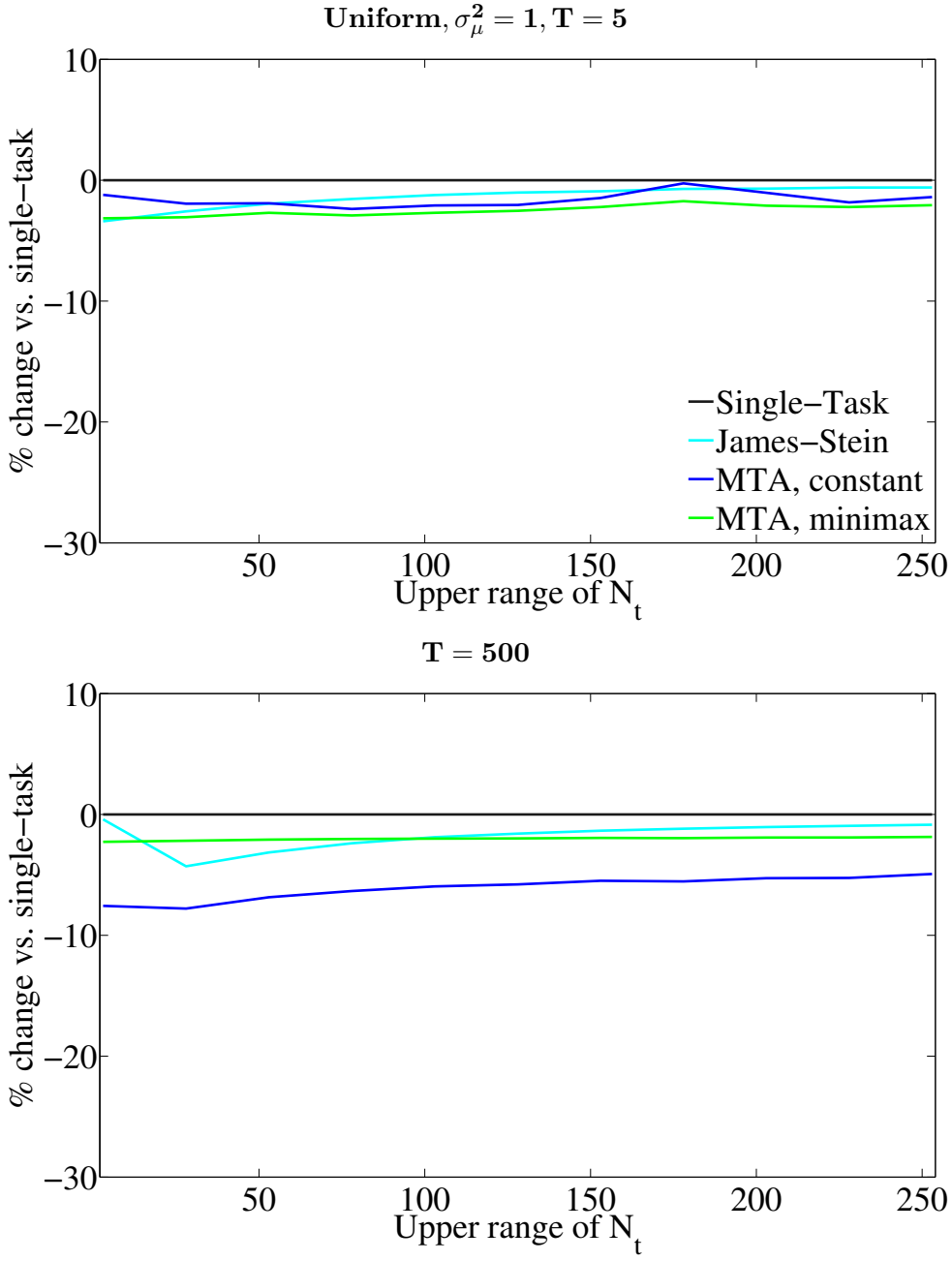

Figure 4.7: Uniform experiment results for $T = \{5, 500\}$ with $\sigma_\mu^2 = 1$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task.

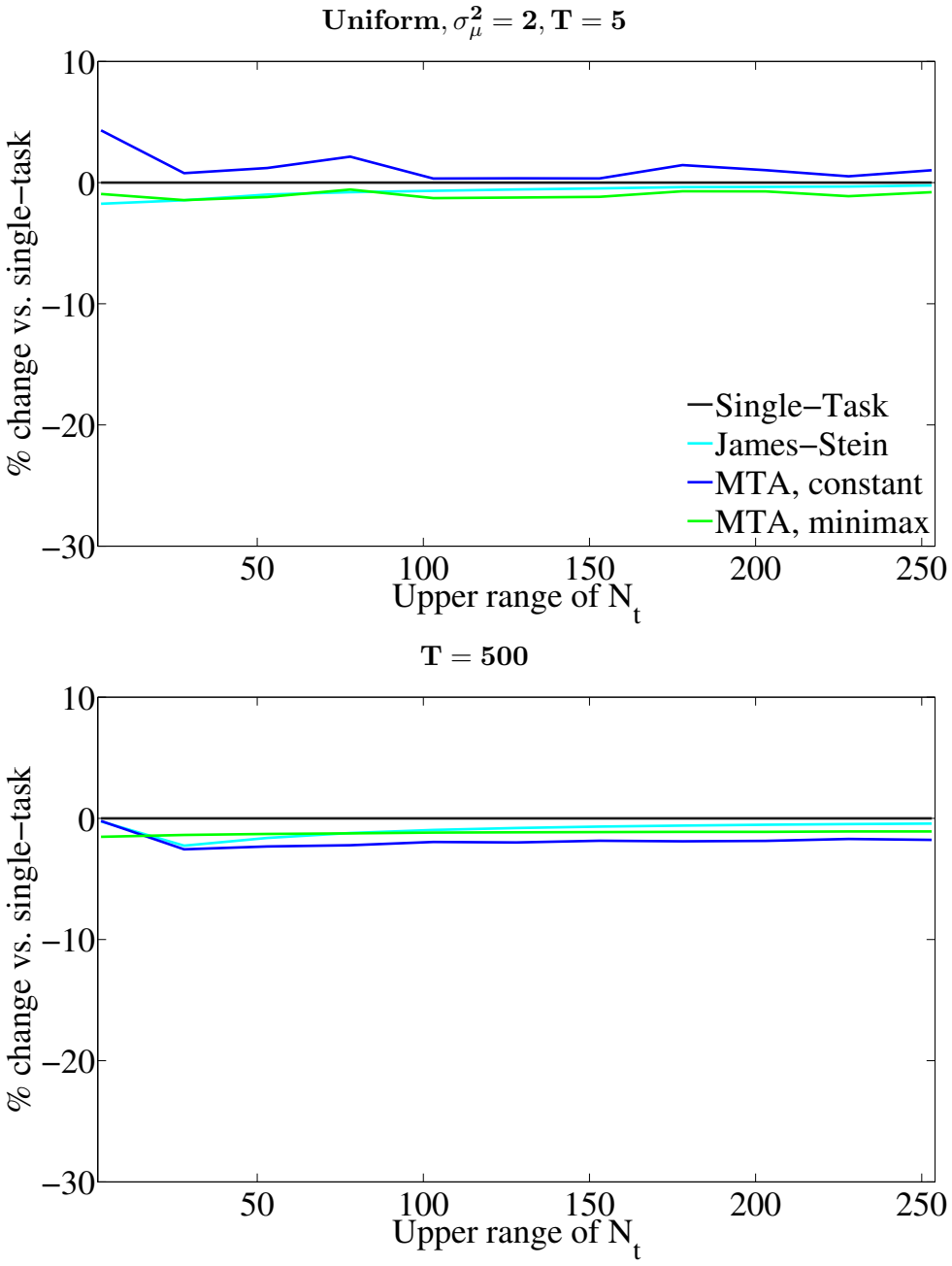

Figure 4.8: Uniform experiment results for $T = \{5, 500\}$ with $\sigma_\mu^2 = 2$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task.

4.3 *SURE Performance*

In Section 2.8 I discussed two different approaches to making practical the optimal pairwise similarity (for the $T = 2$ case). The first approach, plugging sample averages $\bar{y}_t$ in for the means μ_t in a^* to obtain $\hat{a}^*$, was adopted for all subsequent experiments. The second approach was to set the similarity by minimizing Stein's unbiased risk estimate (SURE) [52]. The SURE approach performs worse than the plug-in approach, and I do not recommend its use; see the Gaussian simulation results for $T = 2$ in Figure 4.9.

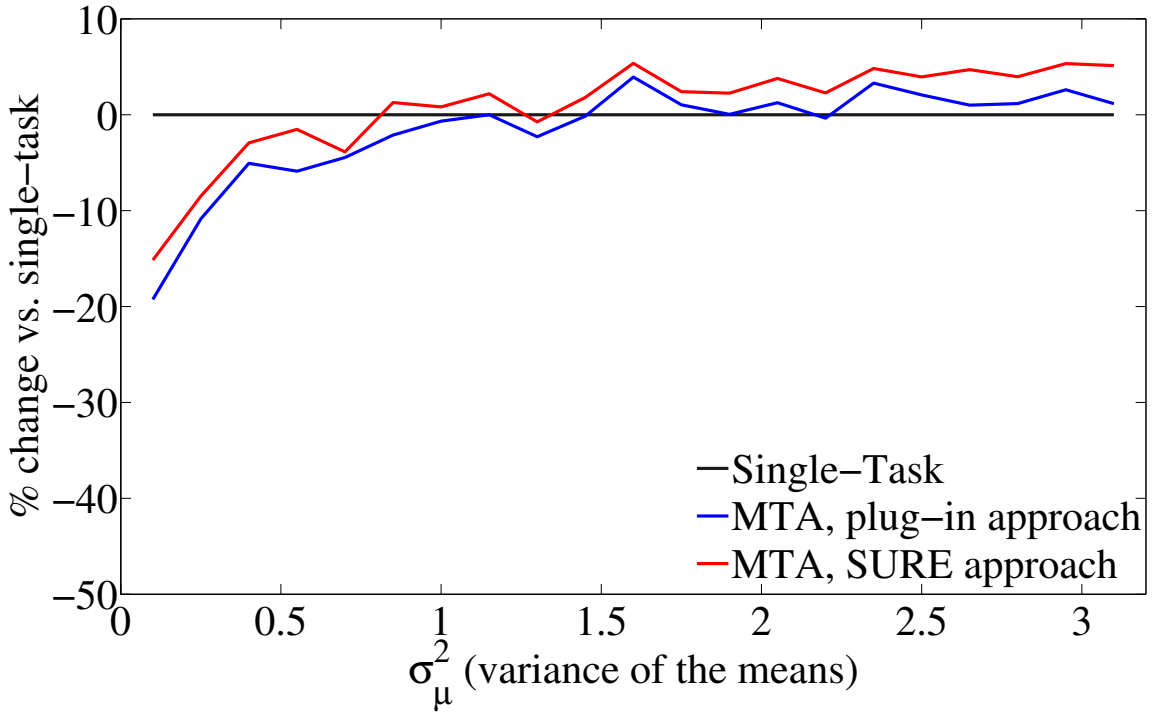

Figure 4.9: Average (over 10000 random draws) percent change in risk vs. single-task for $T = 2$. The SURE approach to setting the similarities is on average outperformed by MTA with ML estimates of the means.

4.4 *Pairwise MTA for* $T > 2$

In Section 3.6 I described pairwise MTA, for which the two task similarity is used to populate the similarity matrix A for any $T \geq 2$. In practice, this results in the following estimated

similarity:

$$\hat{A}^{\mathrm{pr}}_{rs} = \frac{2}{(\bar{Y}_r - \bar{Y}_s)^2}.$$

Simulation results for this similarity are reported in Figure 4.10 for $T = 5$. The plot reproduces the results from the $T = 5$ Gaussian simulation in the preceding sections, but includes only constant MTA and pairwise MTA performance. Clearly, pairwise MTA is inferior to constant MTA throughout. Indeed, this plot was the original motivation for the formulation and derivation of constant MTA.

Why does constant MTA outperform pairwise MTA? Some possible reasons:

- Constant MTA is scaled by a theoretically determined constant, while pairwise MTA heuristically generalizes the $T = 2$ scaling of $\frac{1}{T}$. I have tried various other sensible scalings for pairwise MTA, but none improved performance. On the other hand, using non-optimal scalings for constant MTA did not seriously diminish its performance, suggesting that constant MTA is a low-variance estimator.

- Constant MTA requires the estimation of only a single scaling parameter, while pairwise MTA requires estimating $T(T-1)/2$ parameters (the pairwise distances), which is larger than the total number of means to be estimated (when $T > 2$). The variance of pairwise MTA, therefore, is large, even though it is likely less biased than constant MTA. This is a common theme in machine learning - constraining the model yields better performance, even though the constrained model may be less 'correct' than the unconstrained model. In this case constraining the similarity matrix to be constant acts as a kind of regularization.

4.5 Oracle Performance

To illustrate the best achievable performance with MTA, Figure 4.11 shows the effect of using the true "oracle" means and variances for the calculation of optimal pairwise similarities. This experiment separates how well the MTA formulation can do from the issue of estimating

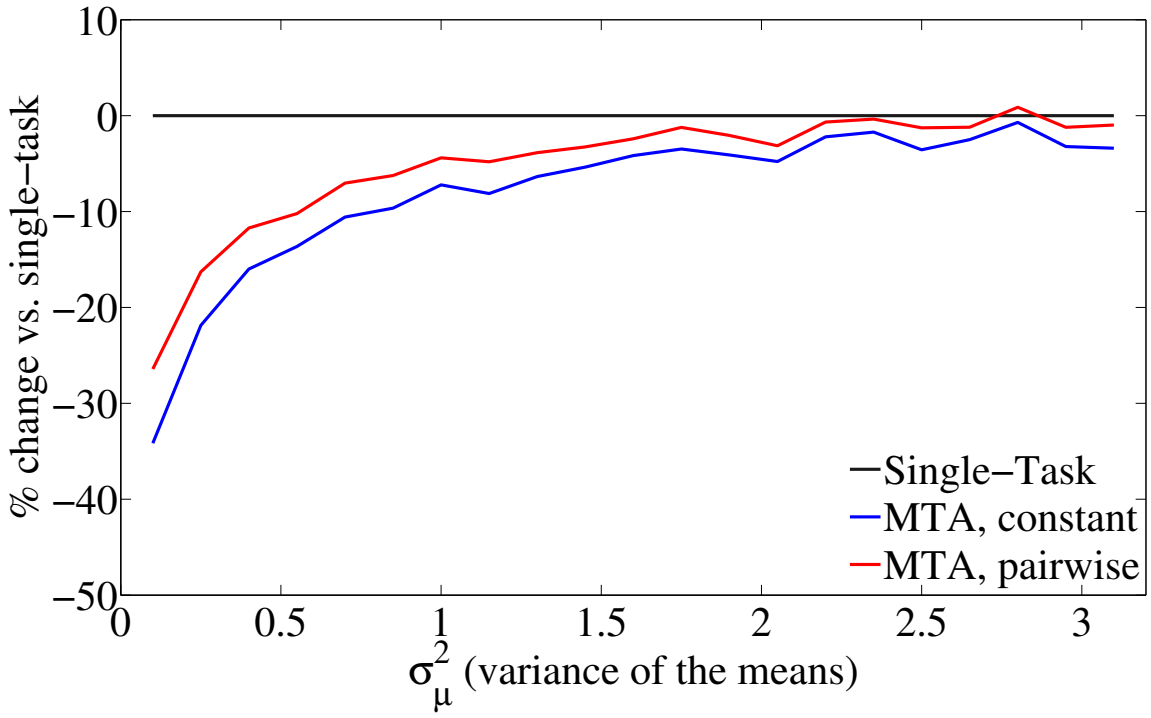

Figure 4.10: Average (over 10000 random draws) percent change in risk vs. single-task for $T = 5$. Constant MTA outperforms pairwise MTA.

the optimal similarity matrix from the data. I use the pairwise oracle matrix A:

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

which consistently bested all other oracle MTA estimators I tried. Interestingly, this is the reverse trend of the non-oracle estimators where pairwise MTA is worse than constant MTA, indicating that pairwise MTA would be the preferred estimator if only one could more accurately estimate the pairwise distances.

The figure reproduces results from the $T = 5$ Gaussian simulation (excluding cross-validation results), and includes oracle pairwise MTA. Oracle MTA is over 30% better than constant MTA, indicating that practical estimates of the similarity, while improving on single-task estimation, are far from the theoretically optimal.

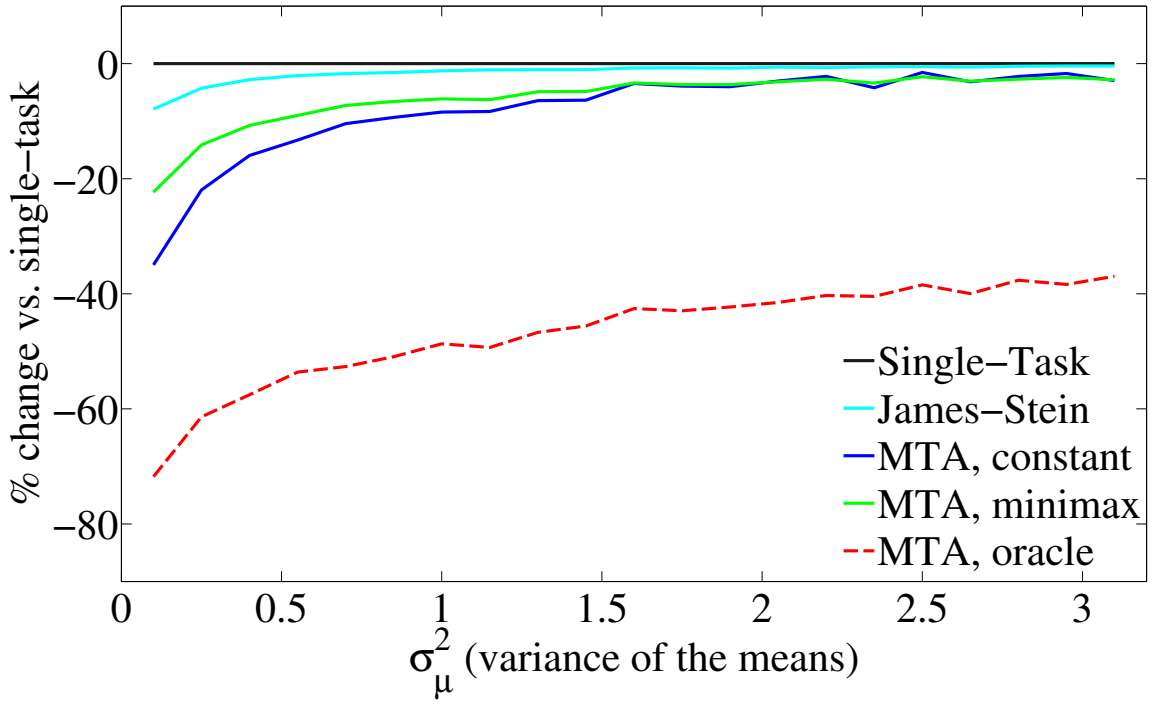

Figure 4.11: Average (over 10000 random draws) percent change in risk vs. single-task for $T = 5$. Oracle MTA uses the true means and variance to specify the weight matrix W.

4.6 MTA Variant Performance

In this section I will compare MTA to the variant discussed in Section 2.4. Recall that the MTA solution matrix is

$$\left(I + \frac{\gamma}{T}\Sigma L\right)^{-1},$$

with, for $T = 2$, optimal similarity

$$a^* = \frac{2}{(\mu_1 - \mu_2)^2}.$$

The MTA variant solution matrix is

$$\Sigma^{1/2}\left(I + \frac{\gamma}{T}L\right)^{-1}\Sigma^{-1/2},$$

with, for $T = 2$, optimal similarity

$$a^* = \frac{2}{\left(\frac{\mu_1}{\sigma_1} - \frac{\mu_2}{\sigma_2}\right)^2}.$$

The $T = 2$ Gaussian simulation experiments from Section 4 are repeated here, but comparing only constant/minimax MTA (which are the same for $T = 2$) with the variant formulation, both using sample mean plug-in estimates of the true means and variances. Results are in Figure 4.12, and show clearly that the original MTA formulation being studied in this thesis outperforms the variant. The results in the bottom figure serve to confirm that using the true oracle standard deviations to pre-normalize the sample means does not help the variant formulation beat the original MTA.

For some intuition as to why, consider the following example:

$$\mu_1 = 2, \quad \sigma_1 = 1, \quad \mu_2 = 4, \quad \sigma_2 = 2.$$

In this case

$$a^* = \frac{2}{\left(\frac{\mu_1}{\sigma_1} - \frac{\mu_2}{\sigma_2}\right)^2} = \infty,$$

which is counterintuitive behavior. It is likely that if the goal was to estimate $\frac{\mu_t}{\sigma_t}$ (instead of μ_t as is the case in this work), then the variant would outperform original MTA.

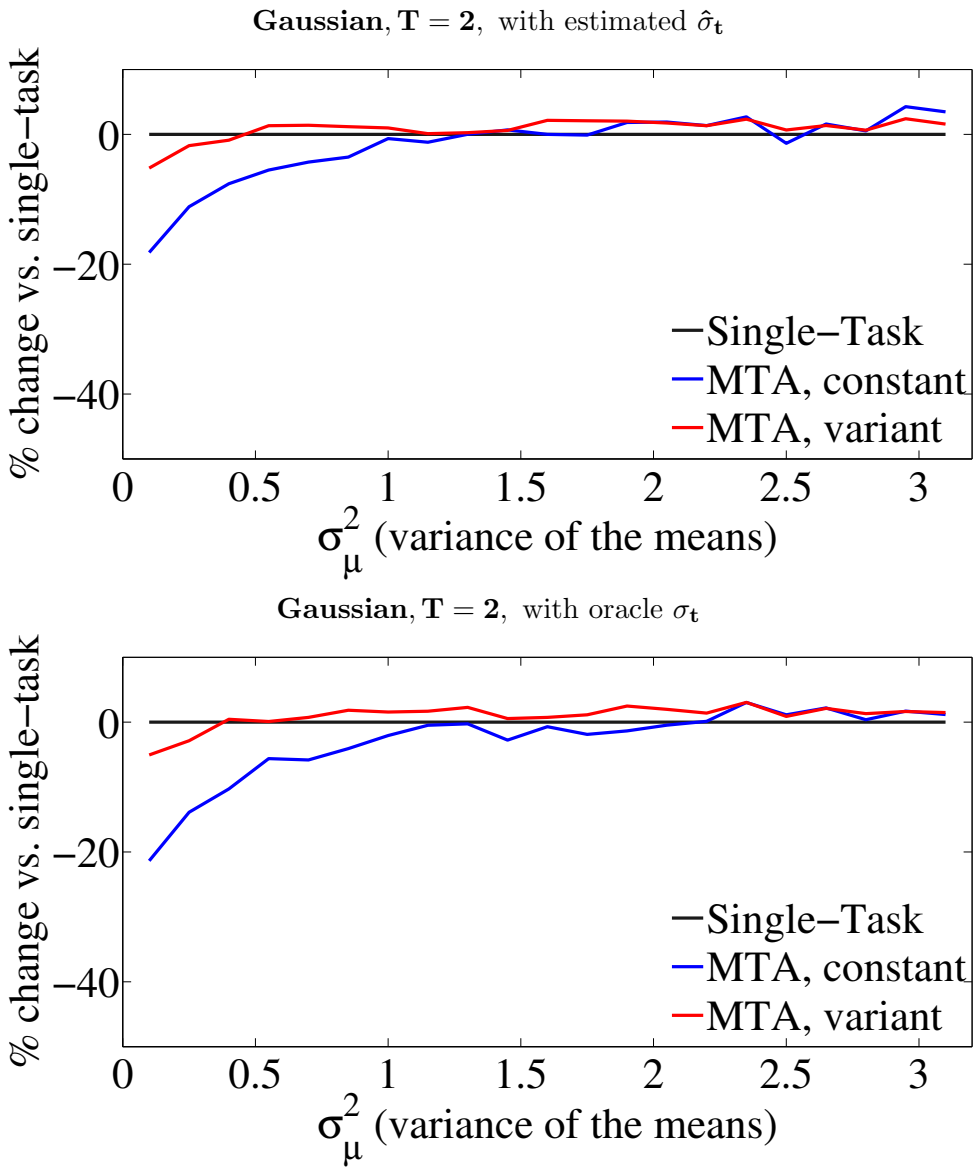

Figure 4.12: Figure shows proposed MTA formulation (Constant MTA) to be better than the variant considered in Section 2.4 using Gaussian simulation results for $T = 2$. The y-axis is average (over 10000 random draws) percent change in risk vs. single-task, such that -50% means the estimator has half the risk of single-task. The top figure shows results with estimated $\hat{\sigma}_t$, and the bottom figure shows results with oracle σ_t.

Chapter 5

APPLICATIONS

In this chapter, I present five applications with real data. The first two applications parallel the simulations: estimating expected values of final grades, and sales of related products. The third application uses MTA for multi-task kernel density estimation, highlighting the applicability of MTA to any algorithm that uses sample averages, and the fourth application uses MTA to improve the performance of the similarity discriminant analysis algorithm. I include a fifth application, estimating the results of the 2008 U.S. presidential election, to demonstrate the effect of model mismatch on multi-task algorithms.

5.1 MTA for Grade Prediction

The goal of this application is to predict the final grade $\{\mu_t\}_{t=1}^T$ of the tth student for all T students in a class, given only each student's N homework grades $\{y_{ti}\}_{i=1}^N$ (in this application $N_t = N$ for all t as every student had been assigned the same number of homeworks). The final class grades include homeworks, projects, labs, quizzes, midterms, and the final exam, but only the homework grades are used to predict the final grade. The implicit model assumption is that each grade (homework, test, etc) is drawn IID from a distribution centered at the final grade μ_t. Clearly this model is inaccurate, as, for instance, tests tend to be harder than homeworks, but the homeworks can reasonably be assumed to be of roughly equal difficulty.

The 16 anonymized datasets were provided by instructors at the University of Washington Department of Electrical Engineering, but the weightings used by each instructor to compute the final grades is not known to us. Other experimental details:

- Each of the 16 datasets (classes) constitutes a single experiment, and the students in that class are treated as the tasks.

- All the grades have been scaled to be between 0 and 100.
- The number of students across the 16 classes is between $T = 16$ and $T = 149$.
- Cross-validation parameters were chosen by training on $N/2$ of the homework grades and validating on the sample mean of all N given grades. I used randomized 5-fold cross-validation as described in the simulation section.
- For each class, a single pooled variance estimate was used for all tasks (that is, students): $\sigma_t^2 = \sigma_{\text{pl}}^2$, for all t.
- The estimator marked "one-task" is just a constant pooled mean for all tasks:

$$\hat{y}_{\text{pl}} = \frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} y_{ti}.$$

- For each class of students, the error measurement for estimator $\hat{y}$ is the risk (sum of squared errors) across all T students:

$$\sum_{t=1}^{T} (\mu_t - \hat{y}_t)^2.$$

This error metric was computed for each class (dataset) separately, and the percent changes in average risk vs. single-task are reported in Table 5.1.

Some observations:

- Constant MTA (without CV) has the biggest percent improvement averaged across all classes.
- The James-Stein estimator has the biggest percent improvement compared to single-task on 7 of the 16 classes.
- The cross-validated versions of estimators do worse than their estimated optimal counterparts.

Table 5.1: Percent change in risk vs. single-task for the grade estimation application (lower is better). 'JS' denotes James-Stein, 'MTA cnst' and 'MTA mm' denote constant MTA and minimax MTA, respectively, 'CV' denotes cross-validation, and 'STD' denotes standard deviation. Lower is better.

Class Size	One-Task	JS	JS CV	MTA cnst	MTA mm	MTA CV
16	26.3	0.7	0.2	0.6	**0**	0.1
29	−6.8	−11.0	**−13.4**	−10.8	−1.7	−5.9
36	**−28.3**	−17.4	−12.4	−16.0	−2.8	−9.1
39	42.0	**−5.8**	−2.3	−5.6	−0.9	−0.9
44	3.0	**−47.6**	−47.3	−42.7	−7.0	−42.7
47	**−12.8**	−8.0	−5.2	−7.1	−0.7	−4.1
48	**−21.0**	−20.5	−13.7	−18.5	−2.5	−5.8
50	63.5	63.5	16.3	9.3	**−4.4**	9.3
50	3.7	**−33.6**	−19.1	−29.7	−3.2	−10.1
57	23.3	−3.8	**−4.1**	−3.6	−0.4	−2.1
58	−0.2	**−16.3**	−5.9	−15.6	−2.8	−4.4
68	−16.9	**−45.5**	−38.5	−39.0	−6.1	−27.7
69	−14.7	−41.0	**−42.4**	−39.8	−4.5	−39.8
72	34.6	**−32.9**	−29.4	−29.0	−4.0	−18.3
110	5.7	**−14.8**	−11.5	−13.4	−1.2	−7.7
149	**−16.6**	−11.7	−11.8	−10.1	−0.8	−5.9
Mean	5.3	−15.4	−15.0	**−16.9**	−2.7	−10.9
σ	25.9	25.9	16.9	15.2	14.3	2.1

- For classes 5 and 8, only minimax MTA with estimated similarity does better than the single-task estimate $\bar{y}$. This is rare, but not impossible; a few of the students happened to have a low average homework grade and an even lower final grade, which resulted in an out-sized contribution to the risk. The expectation is over random samples used to form the estimate, and here the particular realizations of the all the grades were poor. Also, the model that homework grades are drawn IID from a distribution with the final grade as the mean is only a model.[1]

- Minimax MTA was never worse than single-task, robustly providing relatively small gains, as designed.

[1] "All models are wrong, but some are useful." –George E. P. Box

- The James-Stein estimator is also a minimax estimator, but its performance is as highly variable as the one-task estimator. This is because of the positive-part aspect of the JS estimator – when the positive-part boundary is triggered, JS reduces to the one-task estimator.

- Surprisingly, the one-task estimator, which pools all students' scores to estimate a single grade, does better than single-task for half of the classes, and is the best performer for 4 out of 16. When the one-task estimator outperforms single-task, I hypothesize that individual homework grades are poor estimates of final grades. Further, when the one-task estimator is the best estimator, I hypothesize that the assumed model is wrong. That is, the homework grades are *not* IID draws from the "true" distribution of grades, and, in fact, in those cases the homework grades of any individual student provide little information about the final grade. This may occur if the instructor chose to down-weigh the homeworks in the computation of the final grade, or if the tests and labs required a different skill set from the homework.

Grade estimation is a potentially high-risk application; an instructor would be hesitant to give students 'improved' estimates of their grades if those estimates were known to have high variance. However, minimax MTA consistently outperforms single-task estimates without ever doing worse (in our small sample size), and is thus an ideal candidate for this application.

5.2 MTA for Product Sales Estimation

In this section I consider two multi-task problems using sales data supplied by Artifact Puzzles, a company that sells puzzles online. For both problems, I model the given samples as being drawn IID from each task.

The first problem estimates the impact of a particular *puzzle* on repeat business: "Estimate how much any customer will spend on a given order, if on their last order they purchased the tth puzzle, for each of $T = 77$ puzzles." The samples were the dollar amounts different customers had spent on orders after buying each of the T puzzles during a given

time period, and ranged from \$0 for customers that had not re-ordered in the specified time period, to \$480. The number of samples for each puzzle ranged from $N_t = 8$ to $N_t = 348$.

The second problem estimates the monetary value of a particular *customer*: "Estimate how much the tth customer will spend on a given order, for each of $T = 477$ customers." The samples were the order amounts for each of the T customers. Order amounts varied from 15 to 480. The number of samples for each customer ranged from $N_t = 2$ to $N_t = 17$.

I have only samples, no ground truth, so to compare the estimators I treat the single-task means computed from *all* N_t samples as the truth μ_t, and use estimates computed from a uniformly randomly chosen $\frac{N_t}{2}$ of the samples as inputs to MTA, James-Stein, etc. How to motivate such an approach? I prove in Appendix A that this setup results in an unbiased estimate of the true single-task MSE

$$E[(\bar{Y}_t - \mu_t)^2],$$

and a discrete approximation of this error function is precisely the error metric reported throughout this work. There are other unbiased approximations of the MSE, but I want the best possible plug-in estimate of the ground truth, and that requires using all N_t samples to compute μ_t.

Another motivation for this way of treating the data when ground truth is not available is as follows. For the datasets in this section, the results show that MTA can act like having additional samples. In other words, MTA results in estimates that are on-average closer to the N_t-sample mean than the simple $\frac{N_t}{2}$-sample single-task mean.

A different approach is motivated by analogy to constant regression; in the MTA case the dimensionality of the feature vectors is zero – only the constant offset is estimated. When doing regression, the unbiased way to evaluate the performance of a regressor is to have disparate training and test sets. Using all of the data to generate the test statistic μ_t and half of the data to generate the training statistic $\bar{y}_t$ seems to violate this principle. Thus, one can also split the available into training and test with no overlap. See Appendices B and C for results and further discussion.

Results in Table 5.2 are averaged over 1000 random draws of the 50% used for estima-

Table 5.2: Percent change in average risk for puzzle and customer data (first two columns, lower is better), and mean reciprocal rank for terrorist data (last column, higher is better).

Estimator	Puzzles $T = 77$	Customers $T = 477$	Suicide Bombings $T = 7$
Single-Task	0	0	0.15
One-Task	181.7	109.2	0.13
James-Stein	-6.9	-14.0	0.15
James-Stein CV	-21.2	-31.0	0.19
Constant MTA	-17.5	**-32.3**	0.19
Minimax MTA	-8.4	-3.0	0.19
MTA CV	**-21.7**	-30.9	0.19
Expert MTA	-	-	0.19

tion. I used randomized 5-fold cross-validation with the same parameter choices as in the simulations section.

I bolded those entries that were the best or not statistically significantly better than the best according to two one-sided Wilcoxon rank statistical significance tests.

Some observations:

- One-task is a very poor estimator for all of the experiments in this section.

- Constant MTA provided comparable performance to the cross-validated estimators. It was the best or not statistically significantly worse than the best of all the other non-CV estimators.

- Using cross-validation with the two minimax estimators (James-Stein and minimax MTA) statistically significantly outperformed their estimated optimal minimax counterparts. This is consistent with the simulation results.

5.3 MTA for Multi-Task Kernel Density Estimation

In this section I present *multi-task kernel density estimation* (MT-KDE), a variant of MTA. This work was done in collaboration with P. Sadowski and M. R. Gupta.

MTA can be used whenever averages are taken. Recall that for standard single-task kernel density estimation (KDE) [49], a set of random samples $x_i \in \mathbb{R}^d, i \in \{1, \ldots, N\}$ are assumed to be IID from an unknown distribution p_X, and the problem is to estimate the density for a query sample, $z \in \mathbb{R}^d$. Given a kernel function $K(x_i, x_j)$, the un-normalized single-task KDE estimate is

$$\hat{p}(z) = \frac{1}{N} \sum_{i=1}^{N} K(x_i, z),$$

which is just a sample average.

When multiple kernel densities $\{\hat{p}_t(z)\}_{t=1}^{T}$ are estimated for the same domain, I replace the multiple sample averages with MTA estimates, which I refer to as multi-task kernel density estimation (MT-KDE).

I compared KDE and MT-KDE on a problem of estimating the probability of terrorist events in Jerusalem using the Naval Research Laboratory's Adversarial Modeling and Exploitation Database (NRL AMX-DB). The NRL AMX-DB combined multiple open primary sources[2] to create a rich representation of the geospatial features of urban Jerusalem and the surrounding region, and accurately geocoded locations of terrorist attacks. Density estimation models are used to analyze the behavior of such violent agents, and to allocate security and medical resources. In related work, [13] also used a Gaussian kernel density estimate to assess risk from past terrorism events.

The goal in this application is to estimate a risk density for 40,000 geographical locations (samples) in a 20km $\times$ 20km area of interest in Jerusalem. Each geographical location is represented by a $d = 76$-dimensional feature vector. Each of the 76 features is the distance in kilometers to the nearest instance of some geographic location of interest, such as the nearest market or bus stop. Locations of past events are known for 17 suicide bombings. All the events are attributed to one of seven terrorist groups. The density estimates for these seven groups are expected to be related, and are treated as $T = 7$ tasks.

---

[2]Primary sources included the NRL Israel Suicide Terrorism Database (ISD) cross referenced with open sources (including the Israel Ministry of Foreign Affairs, BBC, CPOST, Daily Telegraph, Associated Press, Ha'aretz Daily, Jerusalem Post, Israel National News), as well as the University of New Haven Institute for the Study of Violent Groups, the University of Maryland Global Terrorism Database, and the National Counter Terrorism Center Worldwide Incident Tracking System.

Table 5.3: Hafez's Similarity Matrix A

	AAMB	Hamas	PIJ	PFLP	Fatah	Force17	Unknown
AAMB	0	.2	.2	.6	.8	.8	.6
Hamas	.2	0	.8	.2	.2	.2	.4
PIJ	.2	.8	0	.2	.2	.2	.4
PFLP	.6	.2	.2	0	.6	.6	.5
Fatah	.8	.2	.2	.6	0	1	.6
Force17	.8	.2	.2	.6	1	0	.6
Unknown	.6	.4	.4	.5	.6	.6	0

The kernel K was taken to be a Gaussian kernel with identity covariance; the bandwidth was set to 1. In addition to constant A and minimax A, I also obtained a side-information A from terrorism expert Mohammed M. Hafez of the Naval Postgraduate School; he assessed the similarity between the seven groups during the Second Intifada (the time period of the data), providing similarities between 0 and 1. The similarities are shown in Table 5.3.

The KDE estimates were computed separately for each grid point and each task. The MT-KDE estimates were obtained for one grid point at a time, but for all of the tasks simultaneously. In other words, the regularization was performed only across tasks, and not across grid points.

Leave-one-out cross validation was used to assess KDE and MT-KDE for this problem, as follows. After computing the KDE and MT-KDE density estimates using all but one of the training examples $\{x_{ti}\}$ for each task, I sort the resulting 40,000 estimated probabilities for each of the seven tasks, and extract the rank of the left-out known event. The mean reciprocal rank (MRR) metric is reported in Table 5.2. Ideally, the MRR of the left-out events would be as close to 1 as possible, and indicating that the location of the left-out event is at high-risk. The results show that the MRR for MT-KDE are lower or not worse than those for KDE for both problems; there are, however, too few samples to verify statistical significance of these results. Also, note that the solution of pooling all of the training data into one big task gives inferior performance, and I suspect that this is because each terrorist group has its own target preferences.

5.4 MTA for Similarity Discriminant Analysis

Similarity discriminant analysis (SDA) is a generative classifier that models the class-conditional probability distribution of the similarities between samples using discrete exponentials [18]. Like quadratic discriminant analysis, SDA may produce biased class models, and it has been shown that applying SDA locally (called *local SDA*) to a set of k neighboring (most similar) training samples for each test sample generally decreases model bias and improves performance [17]. However, estimating the required discrete exponential parameters from only k neighbors can have problematically high variance, and in some cases the maximum likelihood parameter estimate is infinite. Here I show that MTA can be profitably applied to this parameter estimation task. This work was done in collaboration with L. Cazzanti, M. R. Gupta, and M. Gabbay. [16].

In similarity-based classification, the training data is a $N \times N$ matrix S of pairwise similarities between N training samples, and their corresponding class labels $y_i \in \{1, 2, \ldots, G\}$ for $i = 1, \ldots, N$. The similarities are assumed to have a discrete (perhaps after quantization) domain Ω. The test data is a $N \times 1$ vector s of similarities between a test sample and the N training samples. SDA produces a probability that the test sample belongs to each of the G classes. In these experiments I use local SDA, but for notational simplicity I restrict the explanation to the SDA; the only difference is in practice the SDA model is trained anew for the k nearest-neighbors of each test sample.

The pairwise SDA model [44] requires estimating G^2 discrete exponential distributions of the probability of seeing a certain similarity value between two samples if one sample is from class g and the other sample is from class h. That is, the gth-hth distribution has the form:

$$P(s_{gh}) = \gamma_{gh} e^{\lambda_{gh} s_{gh}}, \tag{5.1}$$

where s_{gh} is a similarity between a sample from class g and a sample from class h.

I focus on the problem of estimating the parameter λ_{gh} in (5.1); given λ_{gh} the normalizer γ_{gh} will be implied. The maximum likelihood estimate $\hat{\lambda}_{gh}$ satisfies the empirical mean

constraint [44]

$$\frac{\sum_{a:y_a=g}\sum_{b:y_b=h} S(a,b)}{k_g k_h} = \sum_{s_{gh}\in\Omega} s_{gh}\gamma_{gh}e^{\lambda_{gh}s_{gh}}, \tag{5.2}$$

where the left-hand side is the empirical mean with k_g and k_h being the number of training samples in class g and class h, respectively. To find the maximum likelihood estimate $\hat{\lambda}_{gh}$, one can compute the sample average on the left-hand side of (5.2), and then numerically solve for the best $\hat{\lambda}_{gh}, \hat{\gamma}_{gh}$ pair. A particular problem arises when $S(a,b)$ is at an extreme value of the similarity domain Ω for all a, b. For example, if the similarities can only be 0 or 1, and all training sample pairs from class g and class h happen to have zero similarity, then the $\hat{\lambda}_{gh}$ will be ∞.

We applied MTA to this problem and computed multi-task regularized sample averages to replace the left-hand side of (5.2), and then numerically solved (5.2) to produce the *multi-task SDA* (MT-SDA) estimates of the discrete exponential parameters. Here, the multiple tasks are the G^2 class-pairings.

In this application, the multi-task regularization operates across pairs of classes: the average similarity of samples from class g to samples of class h is regularized toward the average similarity of the samples from class l to class m. In the experiments, I define the task-similarity matrix A as a Gaussian kernel on the pairwise class sample averages, as follows. Let $v_{gh} = \frac{\sum_{a:y_a=g}\sum_{b:y_b=h} S(a,b)}{k_g k_h}$ denote the gth-hth sample average given by the left-hand side of (5.2). The task similarity between the gth-hth task and the l-mth task is taken to be

$$A_{g-h;l-m} = e^{-(v_{gh}-v_{lm})^2/\sigma}\,.$$

5.4.1 MT-SDA Experiments

I report results provided by L. Cazzanti [16] comparing the classification performance of the local pairwise SDA classifier to the above MT-SDA variant. Other similarity-based classifiers have been proposed [20], but I restrict the comparison to local SDA to focus on the value of adding the multi-task regularization.

I report classification results for six different benchmark similarity data sets [20]. Table 5.4 shows the mean test error rates computed from 20 random train/test splits of the

Table 5.4: Percent test error averaged over 20 random test/train splits for the similarity data sets.

	Amazon 2 classes	Sonar 2 classes	Patrol 8 classes	Protein 4 classes	Voting 2 classes	FaceRec 139 classes
Local SDA	11.32	15.25	11.56	10.00	6.15	4.23
Multi-task Local SDA	8.95	14.50	11.56	9.77	5.52	3.44

data, with 20% of the data held out for testing and 80% used for training. For each split, the training set was further split into 10 disjoint cross-validation partitions to select the multi-task parameter γ and the Gaussian kernel parameter σ among the possible values $\{10^{-3}, 10^{-2}, 0.1, 1, 10\}$, and the neighborhood size k among the possible values $\{2, 4, 8, 16, 32, 64, \max(128, N)\}$). Across five data sets multi-task local SDA outperforms the standard local SDA and one data set is a tie. The improved performance is statistically significant according to a Wilcoxon sign rank test ($p = 0.05$).

5.5 Model Mismatch: Estimating Election Results

In this section, I apply multi-task averaging methods to predicting the results of the 2008 U.S. presidential election. The data is made up of $T = 51$ tasks, corresponding to the 50 states plus the District of Columbia. Each of the tasks consists of N_t pre-election poll percentages in favor of the candidates Barack Obama and John McCain. We set the ground truth μ_t to be the final voting percentage during the presidential election for the tth state.

Results are reported in Table 5.5 and show that the best results were achieved with single-task averaging. In this case we know the model is wrong – the final voting numbers are *not* the true mean of a distribution from which the polls were sampled because of many complicating factors, such as: differing probabilities of people who actually vote, differential state-specific campaigning, differing voting populations, and evolving public opinion.

This application demonstrates that all multi-task methods for mean estimation can fail. One should take care before applying MTA or the James-Stein estimator, and ensure that the assumed model is appropriate to the estimation task at hand.

Table 5.5: Percent change in average risk vs. single-task for election data (lower is better).

Estimator	Obama	McCain
Single-Task	**0.0**	**0.0**
One-Task	313.5	925.6
James-Stein	1.05	4.25
James-Stein CV	3.04	19.26
Constant MTA	0.65	8.31
Minimax MTA	0.02	0.44
MTA CV	0.03	0.88

Chapter 6

CONCLUSIONS

Though perhaps unintuitive, I showed that both in theory and in practice estimating multiple *unrelated* means in a joint MTL fashion can improve the overall risk, and under some conditions even more so than with the classic, battle-tested James-Stein estimator. Averaging is common, and MTA has potentially broad applicability as a subcomponent to many algorithms, such as k-means clustering, kernel density estimation, or non-local means denoising.

Many open questions remain for future research:

- The two most successful variants of MTA, constant MTA and minimax MTA, use a similarity matrix of all ones scaled by some constant. Is this the best one can do in practice? Can other similarity matrices with less bias but more variance provide better performance on average, under some conditions?

- In Section 5.5 I gave one practical example of when MTA and the James-Stein estimator did not outperform the standard average. However, in the absence of ground truth, it is unclear how to tell in advance whether these estimators will help or harm the single-task estimate. How to test model appropriateness in order to decide whether to use multi-task algorithms is an open question.

- The MTA formulation discussed in this thesis uses the squared-loss due to its analytic tractability:

$$\underset{\{\hat{y}_t\}_{t=1}^T}{\arg\min} \; \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{(y_{ti}-\hat{y}_t)^2}{\sigma_t^2} + \frac{\gamma}{T^2}\sum_{r=1}^{T}\sum_{s=1}^{T}A_{rs}(\hat{y}_r-\hat{y}_s)^2.$$

  However, other formulations are possible. One could use any number of Bregman divergences as a loss. In addition, the median is the solution to the following opti-

mization problem:

$$\arg\min_{\hat{y}} \sum_{i=1}^{N_t} |y_i - \hat{y}|,$$

which can be extended in a way analogous to MTA to produce the multi-task median objective:

$$\arg\min_{\{\hat{y}_t\}_{t=1}^T} \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{|y_{ti} - \hat{y}_t|}{\sigma_t} + \frac{\gamma}{T^2}\sum_{r=1}^{T}\sum_{s=1}^{T} A_{rs}(\hat{y}_r - \hat{y}_s)^2,$$

or

$$\arg\min_{\{\hat{y}_t\}_{t=1}^T} \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{|y_{ti} - \hat{y}_t|}{\sigma_t} + \frac{\gamma}{T^2}\sum_{r=1}^{T}\sum_{s=1}^{T} A_{rs}|\hat{y}_r - \hat{y}_s|,$$

depending on how the algorithm designer intends to tie the tasks together. Using a squared-loss moves all of the multi-task averages towards each other in a smooth way, while the absolute loss encourages a subset of the tasks to be identical or nearly so.

These formulations do not have obvious closed-form solutions, and the type of analysis performed in this thesis would not be an option. Thus, their study would have to be computational in nature. The formulations are convex (and remain so for any combination of convex losses), and can be solved efficiently with readily available convex solvers (albeit not in $O(T)$ time as with the closed-form solutions presented in this work). The question of setting γ and A without theoretical justification is particularly tricky – one would be limited to cross-validation or expert specified similarities.

- One can also apply the above reasoning to a multi-task extension of finding the center of the domain of a distribution. The center of a domain can be written as an optimization objective using the L_∞ loss as follows:

$$\frac{\max_i y_i - \min_i y_i}{2} = \arg\min_{\hat{y}} \max_i |y_i - \hat{y}|.$$

Generalizing this to be multi-task, one obtains:

$$\arg\min_{\{\hat{y}_t\}_{t=1}^T} \frac{1}{T}\sum_{t=1}^{T}\max_i \frac{|y_{ti}-\hat{y}_t|}{\sigma_t} + \frac{\gamma}{T^2}\sum_{r=1}^{T}\sum_{s=1}^{T} A_{rs}(\hat{y}_r-\hat{y}_s)^2.$$

- Throughout this work, I used the L_2 risk (that is, the sum of the expected squared errors) to derive the optimal similarities, but this may not always be the most appropriate function to optimize. For example, in finding a multi-task mean, we may care more the sum of the absolute errors between the estimated mean and the true mean rather than the squared estimation errors, and this would indicate that minimizing the standard L_2 risk is inappropriate. One alternative, for example, is to try to choose a constant $A = a\mathbf{1}\mathbf{1}^T$ such that the L_1 risk is minimized:

$$\arg\min_a E[\mathbf{1}^T|W\bar{Y}-\mu|].$$

- I presented asymptotic convergence results for when N_t approaches infinity. Results for T approaching infinity are still an open question. In particular, what happens when both N_t and T approach infinity at various differing rates?

BIBLIOGRAPHY

[1] 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.

[2] A. Agarwal, H. Daumé III, and S. Gerber. Learning multiple tasks using manifold regularization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 46–54. 2010.

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

[4] 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.

[5] Aristotle. *Nichomachean Ethics*. Harvard University Press, Cambridge, MA, 1994.

[6] B. Bakker and T. Heskes. Task clustering and gating for Bayesian multitask learning. *Journal Machine Learning Research*, 4:83–99, December 2003.

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

[8] J. Baxter. Learning internal representations. In *International Conference on Computational Learning Theory*, pages 311–320, 1995.

[9] 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.

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

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

[12] 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.

[13] D. Brown, J. Dalton, and H. Hoyle. Spatial forecast methods for terrorist events in urban environments. *Lecture Notes in Computer Science*, 3073:426–435, 2004.

[14] R. Caruana and D. Pomerleau. Multitask learning. In *Machine Learning*, 1997.

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

[16] L. Cazzanti, S. Feldman, M. R. Gupta, and M. Gabbay. Multi-task regularization of generative similarity models. *Lecture Notes in Computer Science*, 7005:90–103, 2011.

[17] L. Cazzanti and M. R. Gupta. Local similarity discriminant analysis. In *Proc. Int. Conf. Machine Learning*, 2007.

[18] L. Cazzanti, M. R. Gupta, and A. J. Koppal. Generative models for similarity-based classification. *Pattern Recognition*, 41(7):2289–2297, July 2008.

[19] P. Y. Chebotarev and E. V. Shamis. The matrix-forest theorem and measuring relations in small social groups. *Automation and Remote Control*, 58:1505–1514, 1997.

[20] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal Machine Learning Research*, 10:747–776, March 2009.

[21] F. R. K. Chung. *Spectral Graph Theory*. 2004.

[22] B. Efron and C. 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.

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

[24] Y. C. Eldar. Generalized SURE for exponential families: Applications to regularization. *IEEE Trans. Signal Processing*, 57(2):471–481.

[25] T. Evgeniou, C. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal Machine Learning Research*, (6), 2005.

[26] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *KDD '04*, pages 109–117, 2004.

[27] S. Feldman, B. A. Frigyik, and M. R. Gupta. Multi-task averaging. *In Submission to Journal Machine Learning Research.*

[28] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. A*, 22:133–142, 1922.

[29] 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.

[30] 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.

[31] A. Hald. *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935.* Springer, 2006.

[32] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer-Verlag, New York, 2001.

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

[34] 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.

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

[36] 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.

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

[38] Q. Liu, X. Liao, H. Li, J. R. Stack, and L. Carin. Semisupervised multitask learning. *IEEE Trans. Pattern Analysis and Machine Intelligence*, (6), June 2009.

[39] G. Lugosi and K. Zeger. Nonparametric estimation via empirical risk minimization. *IEEE Trans. Information Theory*, 41(3):677–687, 1995.

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

[41] G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, pages 231–252, 2010.

[42] 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.

[43] 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.

[44] P. Sadowski, L. Cazzanti, and M. R. Gupta. Bayesian and pairwise local similarity discriminant analysis. In *Proc. IEEE Conf. Cognitive Information Processing*, 2010.

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

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

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

[48] 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.

[49] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, 1986.

[50] A. J. Smola and I. R. Kondor. Kernels and regularization on graphs. In *Proc. Annual Conference on Computational Learning Theory*, 2003.

[51] 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.

[52] C. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

[53] S. Thrun. Is learning the n-th thing any easier than learning the first? In *Advances in Neural Information Processing Systems (NIPS)*, pages 640–646, 1995.

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

[55] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, New York, NY, USA, 1995.

[56] 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.

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

[58] 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.

[59] X. Zhu. Semi-supervised learning literature survey, 2006.

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

Appendix A

APPROXIMATING THE MEAN SQUARED ERROR

The sample mean estimate of a true, unknown mean μ given N random samples is $\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$, and it has a mean squared error

$$E[(\mu - \bar{Y}_N)^2] = \frac{\sigma^2}{N}.$$

The goal is to find an approximation to this MSE when only N samples are available (μ is not available). The approach in this thesis is to use $M = N/2$ of the samples to estimate $\bar{Y}_N$, and all N samples to estimate μ. I will show that this approximation is *unbiased*.

First, I will rewrite $\bar{Y}_N$:

$$\begin{aligned}
\bar{Y}_N &= \frac{1}{N}\sum_{i=1}^{N} Y_i \\
&= \frac{1}{N}\sum_{i=1}^{M} Y_i + \frac{1}{N}\sum_{i=M+1}^{N} Y_i \\
&= \frac{M}{N}\frac{1}{M}\sum_{i=1}^{M} Y_i + \frac{1}{N}\sum_{i=M+1}^{N} Y_i \\
&= \frac{M}{N}\bar{Y}_M + \frac{1}{N}\sum_{i=M+1}^{N} Y_i \\
&= \frac{M}{N}\bar{Y}_M + \frac{1}{N}\sum_{j=1}^{M} Z_j \\
&= \frac{M}{N}\bar{Y}_M + \frac{M}{N}\bar{Z}_M \\
&= \frac{1}{2}\left(\bar{Y}_M + \bar{Z}_M\right)
\end{aligned}$$

where the Z_j are just Y_i renamed to indicate that $\bar{Y}_M$ and $\bar{Z}_M$ are composed of *different and independent* samples. I will need the fact that $E[\bar{Z}_M\bar{Y}_M] = E[\bar{Z}_M]E[\bar{Y}_M] = \mu^2$ in the

following derivation of the approximate mean squared error:

$$\begin{aligned}
E[(\bar{Y}_N - \bar{Y}_M)^2] &= E[\bar{Y}_N^2] + E[\bar{Y}_M^2] - 2E[\bar{Y}_N \bar{Y}_M] \\
&= \frac{\sigma^2}{N} + \mu^2 + \frac{\sigma^2}{M} + \mu^2 - 2E\left[\frac{1}{2}\left(\bar{Y}_M + \bar{Z}_M\right)\bar{Y}_M\right] \\
&= \frac{\sigma^2}{N} + \frac{\sigma^2}{M} + 2\mu^2 - E[\bar{Y}_M^2] - E[\bar{Y}_M \bar{Z}_M] \\
&= \frac{\sigma^2}{N} + \frac{\sigma^2}{M} + 2\mu^2 - E[\bar{Y}_M^2] - E[\bar{Y}_M]E[\bar{Z}_M] \\
&= \frac{\sigma^2}{N} + \frac{\sigma^2}{M} + 2\mu^2 - \frac{\sigma^2}{M} - \mu^2 - \mu^2 \\
&= \frac{\sigma^2}{N}.
\end{aligned}$$

I have shown that

$$E[(\mu - \bar{Y}_N)^2] = E[(\bar{Y}_N - \bar{Y}_M)^2],$$

and thus the approximate MSE used in this thesis is an unbiased estimate of the true MSE.

Table B.1: Percent change in average risk for puzzle and customer data (lower is better) using 50/50 splits for training and test data.

Estimator	Puzzles $T = 77$	Customers $T = 477$
Single-Task	0	0
One-Task	-2.2	-24.1
James-Stein	-21.1	**-30.4**
James-Stein CV	-19.3	-22.4
Constant MTA	**-22.0**	-26.3
Constant MTA CV	-17.0	-22.3
Minimax MTA	-4.2	-1.4
Minimax MTA CV	-15.6	-15.8

Appendix B

PUZZLE DATA RESULTS WITH 50/50 SPLITS

In Section 5.2, because I had no ground truth for the puzzle datasets, I used all of the N_t samples to compute μ_t, and a random half of samples to compute $\bar{y}_t$. As mentioned in the body of the main text, analogy to constant regression motivates a different approach. With the goal of having no overlap between training and test data, one can use half of the samples for μ_t and the other half for $\bar{y}_t$.

The experimental results with this splitting approach are in Table B.1. The results are similar to that of Table 5.2, with one notable exception: the performance of the one-task estimator. In Section 5.2 the one-task estimator was useless - using a single estimate for all T tasks was not at all helpful. But here, the one-task estimator is better than single-task. This, in turn, increases the accuracy of James-Stein, which reverts to the one-task estimator when the positive-part boundary is triggered.

Constant MTA is still a top performer for both the 'Puzzles' and 'Customers' datasets, and minimax MTA gives a small but reliable win over single-task.

Appendix C

MTA AS CONSANT REGRESSSION RESULTS FOR PUZZLE DATA

But the analogy to constant regression leads to another metric modality entirely: evaluating the performance of MTA using a regression-like error metric. Currently (using the single-task sample mean as an example), MTA error is calculated for the tth task as:

$$\text{error}_t^{(\text{mean})} = (\mu_t - \bar{y}_t)^2.$$

A regression-like metric instead computes the average distance to each of the $N_t/2$ samples that were used to calculate μ_t in Appendix B:

$$\text{error}_t^{(\text{reg})} = \frac{1}{N_t/2} \sum_{i=1}^{N_t/2} (y_{ti} - \bar{y}_t)^2.$$

Results on the puzzle data when using this error metric are in Table C.1. Despite the fact that none of the multi-task algorithms were designed to optimize the regression error metric, a performance win is achieved by all of them. MTA algorithms are still among the top performers, but very little performance is gained for the puzzles-as-tasks experiments.

I leave for future work the derivation of an MTA variant that optimizes the regression-like loss.

Table C.1: Percent change in average regression error for puzzle and customer data (lower is better) using 50/50 splits for training and test data.

Estimator	Puzzles $T = 77$	Customers $T = 477$
Single-Task	0	0
One-Task	-0.3	-22.8
James-Stein	-1.9	-28.8
James-Stein CV	-1.6	-29.3
Constant MTA	**-2.0**	-24.9
Constant MTA CV	-1.7	**-29.5**
Minimax MTA	-0.4	-1.3
Minimax MTA CV	**-1.9**	-21.1

VITA

Sergey Feldman was born in L'viv, Ukraine in 1984, and moved to Skokie, IL nine years later. He received a B.Sc. in electrical engineering from the University of Illinois at Chicago in 2006, and joined Professor Maya Gupta's lab at the University of Washington in 2007. After obtaining his M.Sc. in 2009, Sergey was awarded the Bonderman Fellowship and traveled abroad for ten months.