Multi-task Averaging

Maya R. Gupta

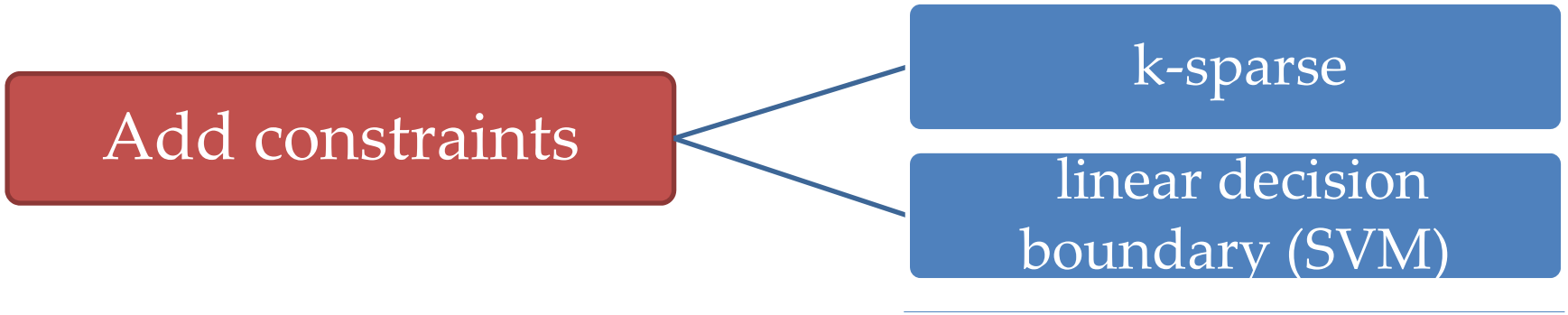
UNIVERSITY OF WASHINGTON

Sergey Feldman

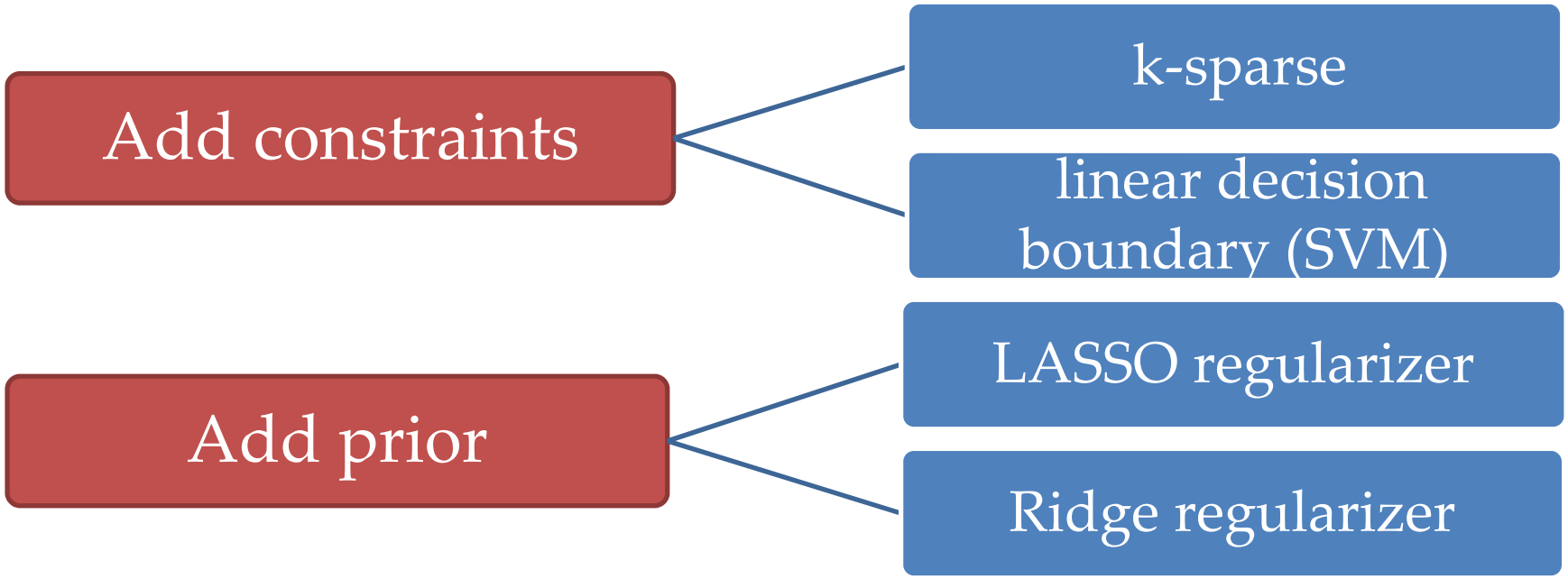
Bela Frigyik

Peter Sadowski

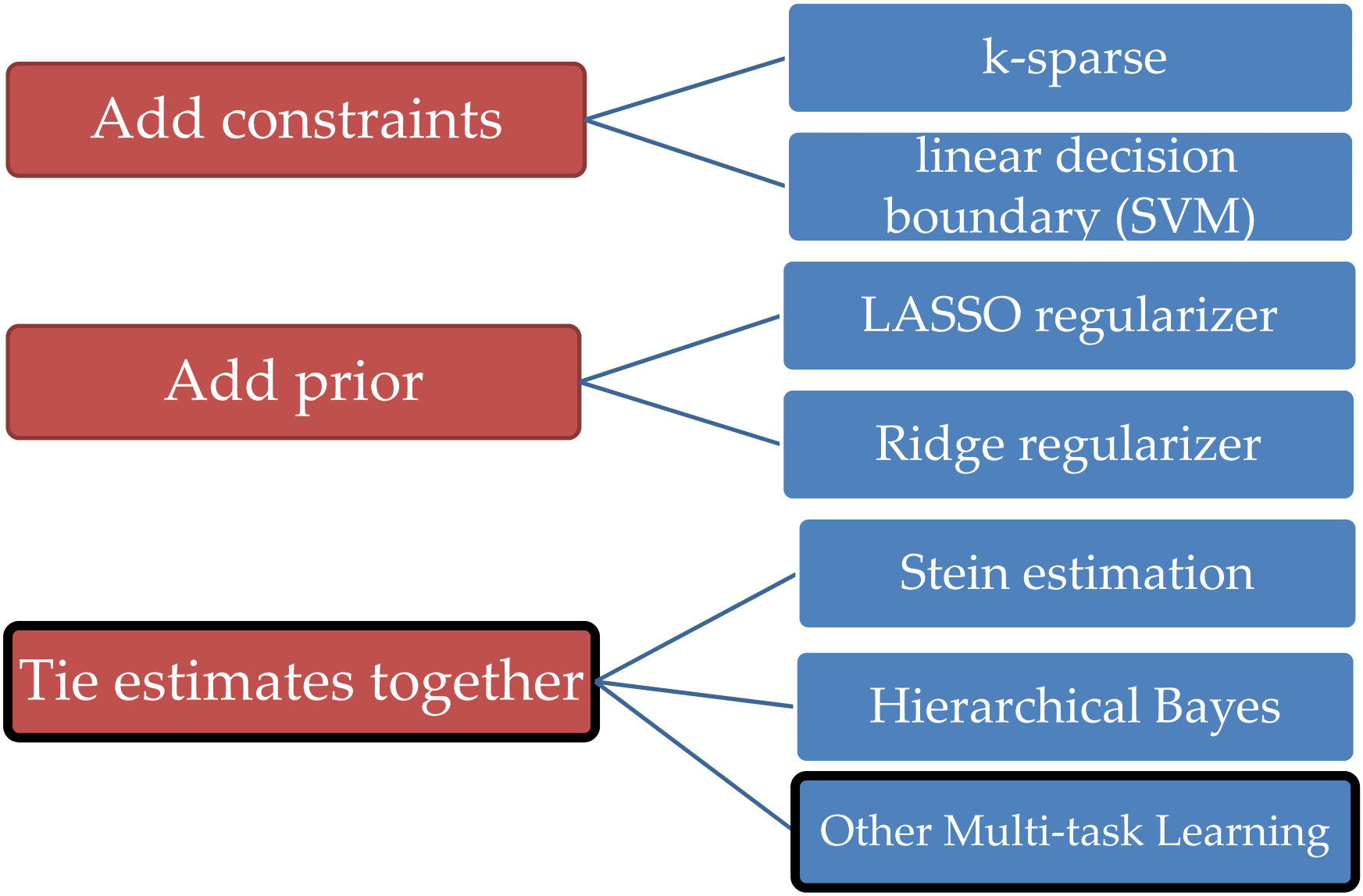
Approaches to Regularize Estimates



Approaches to Regularize Estimates



Approaches to Regularize Estimates



Multi-Task Learning

Goal: Estimate T functions $f_t(x; \beta_t)$

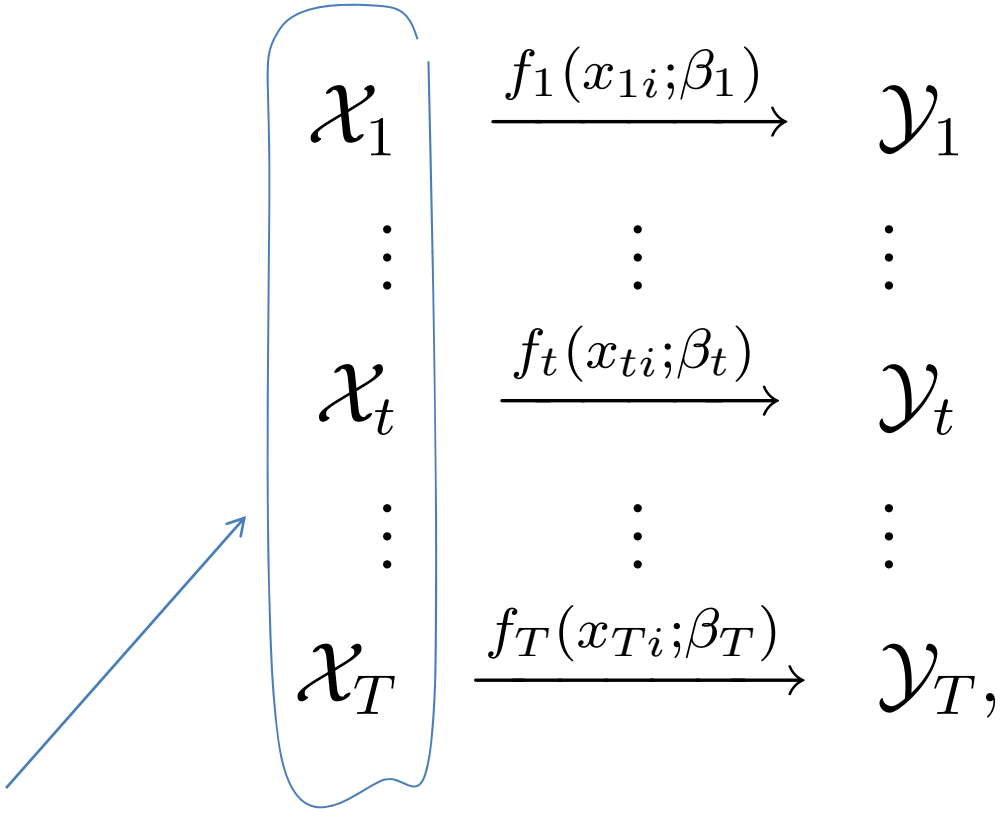
Given: $\{(x_{ti}, y_{ti})\}_{i=1}^{n_t}$ for $t = 1, \ldots, T$.

$$
\begin{array}{ccc}
\mathcal{X}_1 & \xrightarrow{f_1(x_{1i};\beta_1)} & \mathcal{Y}_1 \\
\vdots & \vdots & \vdots \\
\mathcal{X}_t & \xrightarrow{f_t(x_{ti};\beta_t)} & \mathcal{Y}_t \\
\vdots & \vdots & \vdots \\
\mathcal{X}_T & \xrightarrow{f_T(x_{Ti};\beta_T)} & \mathcal{Y}_T,
\end{array}
$$

Multi-Task Learning

Goal: Estimate T functions $f_t(x; \beta_t)$

Given: $\{(x_{ti}, y_{ti})\}_{i=1}^{n_t}$ for $t = 1, \ldots, T$.

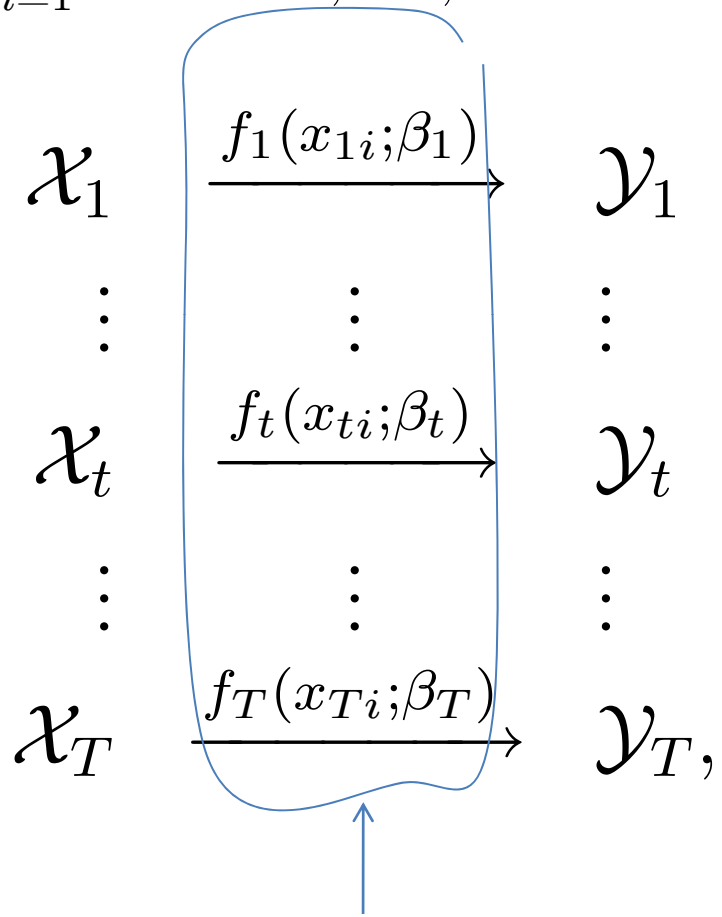


regularize the inputs
e.g. joint feature
selection

Multi-Task Learning

Goal: Estimate T functions $f_t(x; \beta_t)$

Given: $\{(x_{ti}, y_{ti})\}_{i=1}^{n_t}$ for $t = 1, \ldots, T$.

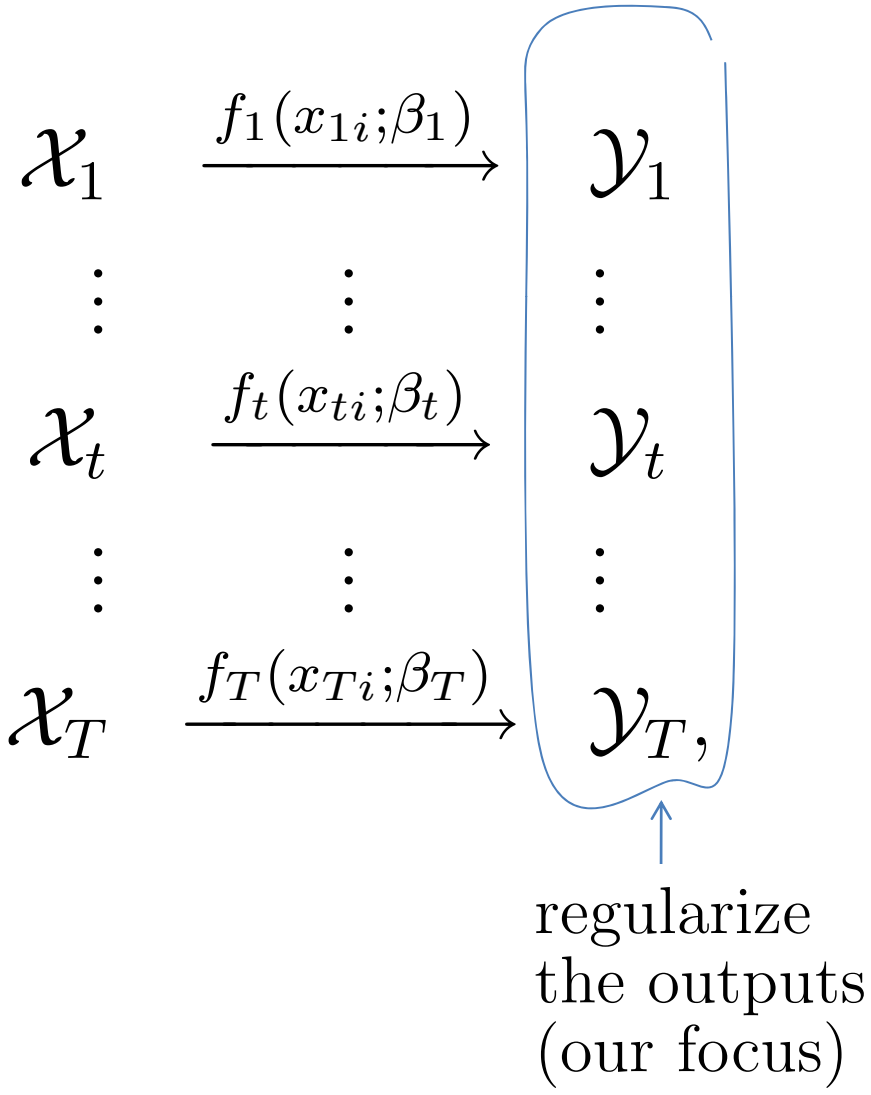


regularize the model selection
e.g. jointly shrink the parameters

Multi-Task Learning

Goal: Estimate T functions $f_t(x;\beta_t)$

Given: $\{(x_{ti}, y_{ti})\}_{i=1}^{n_t}$ for $t = 1, \ldots, T$.



Example: Parametric vs. Output MT Regularization

Linear model: $f_t(x) = \beta_t^T x$, noisy samples $\{(x_{ti}, y_{ti})\}$

Least-squares Fit:

$$\arg \min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \left(y_{ti} - \beta_t^T x_{ti}\right)^2$$

sum over all tasks

sum over all samples

Example: Parametric vs. Output MT Regularization

Linear model: $f_t(x) = \beta_t^T x$, noisy samples $\{(x_{ti,y_{ti}})\}$

Least-squares Fit:

$$\arg \min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \left(y_{ti} - \beta_t^T x_{ti}\right)^2$$

Parametric Regularizer:

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

(A_{rs} is side info)

Example: Parametric vs. Output MT Regularization

Linear model: $f_t(x) = \beta_t^T x$, noisy samples $\{(x_{ti,y_{ti}})\}$

Least-squares Fit:

$$\arg \min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \left(y_{ti} - \beta_t^T x_{ti}\right)^2$$

Parametric Regularizer:

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

(A_{rs} and A_{risj} are side info)

Output Regularizer:

$$+ \sum_{r=1}^{T} \sum_{s=1}^{T} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} A_{risj} \|f_r(x_{ri}) - f_s(x_{sj})\|_2^2$$

General: Parametric vs. Output Regularization

Parametric MT Regularization:

$$\arg\min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^{T}\sum_{i=1}^{N_t} L(y_{ti}, f_t(x_{ti};\beta_t)) + \gamma J(\{\beta_r\}_{r=1}^T).$$

L: loss

J: regularize parameters

Output-Regularized MT:

$$\arg\min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^{T}\sum_{i=1}^{N_t} L(y_{ti}, f_t(x_{ti};\beta_t)) + \gamma J(\{\{f_r(x_{rj};\beta_r)\}_{j=1}^{N_r}\}_{r=1}^T).$$

L: loss

J: regularize output values

General Parametric vs. Output Regularization

Parametric MT Regularization:

$$\arg\min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^{T}\sum_{i=1}^{N_t} L(y_{ti}, f_t(x_{ti};\beta_t)) + \gamma J(\{\beta_r\}_{r=1}^T).$$

L: loss

J: regularize parameters

Output-Regularized MT:

$$\arg\min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^{T}\sum_{i=1}^{N_t} L(y_{ti}, f_t(x_{ti};\beta_t)) + \gamma J(\{\{f_r(x_{rj};\beta_r)\}_{j=1}^{N_r}\}_{r=1}^T).$$

L: loss

J: regularize output values

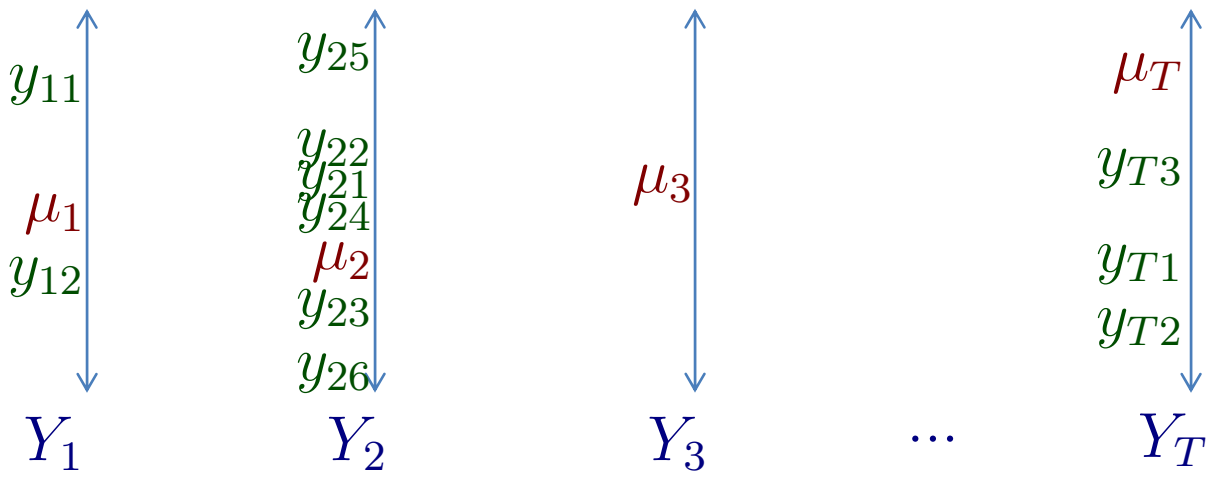
Rest of this talk: Focus on a very simple case

Multi-Task Averaging

Problem: estimate means $\{\mu_t\}$ of T random variables Y_t

Constant model: $y_{ti} = \mu_t + \text{noise}$

Given: N_t IID samples $\{y_{ti}\}_{i=1}^{N_t}$ from each Y_t,
and $T \times T$ task-similarity matrix A.



Multi-Task Averaging

Problem: estimate means $\{\mu_t\}$ of T random variables Y_t

Constant model: $y_{ti} = \mu_t + \text{noise}$

Given: N_t IID samples $\{y_{ti}\}_{i=1}^{N_t}$ from each Y_t,
and $T \times T$ task-similarity matrix A.

Single-task averaging - minimize total squared error:

$$y_t^* = \arg\min_{\hat{y}_t} \sum_{i=1}^{N_t} (y_{ti} - \hat{y}_t)^2$$

Multi-Task Averaging

Problem: estimate means $\{\mu_t\}$ of T random variables Y_t

Constant model: $y_{ti} = \mu_t + \text{noise}$

Given: N_t IID samples $\{y_{ti}\}_{i=1}^{N_t}$ from each Y_t,
and $T \times T$ task-similarity matrix A.

Single-task averaging - minimize total squared error:

$$y_t^* = \arg\min_{\hat{y}_t} \sum_{i=1}^{N_t} (y_{ti} - \hat{y}_t)^2$$

Multi-task averaging - regularize outputs:

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

MTA Closed Form Solution

Multi-task averaging - jointly estimate:

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

Closed-form solution:

$$y^* = (I - \tilde{A})^{-1}\tilde{y}$$

$$\tilde{y}_t = \frac{\sum_{i=1}^{N_t} y_{ti}}{N_t + \gamma \sum_s (A_{ts} + A_{st})}$$

$$\tilde{A}_{tr} = \frac{\gamma(A_{tr} + A_{rt})}{N_t + \gamma \sum_s (A_{ts} + A_{st})}$$

MTA Closed Form Solution

Multi-task averaging - jointly estimate:

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

Closed-form solution:

$$y^* = \boxed{(I - \tilde{A})^{-1}}\tilde{y}$$

$(I - \tilde{A})^{-1}$ exists:
for $T = 2$;
or if A has
non-negative entries
(sufficient, not-necess)

$$\tilde{y}_t = \frac{\sum_{i=1}^{N_t} y_{ti}}{N_t + \gamma \sum_s (A_{ts} + A_{st})}$$

$$\tilde{A}_{tr} = \frac{\gamma(A_{tr} + A_{rt})}{N_t + \gamma \sum_s (A_{ts} + A_{st})}$$

MTA Closed Form Solution

Multi-task averaging - jointly estimate:

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

Closed-form solution:

$$y^* = (I - \tilde{A})^{-1}\tilde{y}$$

Linear combo of sample averages

Scaled sample averages

$$\tilde{y}_t = \frac{\sum_{i=1}^{N_t} y_{ti}}{N_t + \gamma \sum_s (A_{ts} + A_{st})}$$

$$\tilde{A}_{tr} = \frac{\gamma(A_{tr} + A_{rt})}{N_t + \gamma \sum_s (A_{ts} + A_{st})}$$

MTA Closed Form Solution

Multi-task averaging - jointly estimate:

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

Closed-form solution:

$$y^* = (I - \tilde{A})^{-1}\tilde{y}$$

Theorem: **convex** combo of sample averages

Scaled sample averages

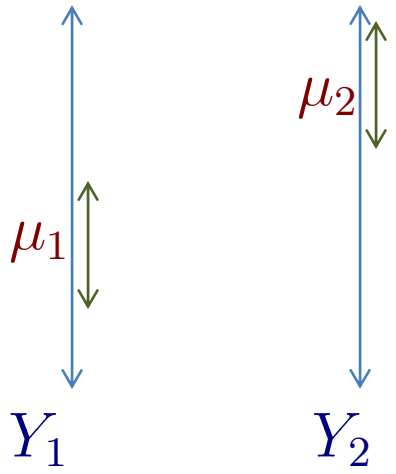
$$\tilde{y}_t = \frac{\sum_{i=1}^{N_t} y_{ti}}{N_t + \gamma \sum_s (A_{ts} + A_{st})}$$

$$\tilde{A}_{tr} = \frac{\gamma(A_{tr} + A_{rt})}{N_t + \gamma \sum_s (A_{ts} + A_{st})}$$

Key: If all $y_{ti} \in C$, then $y^* \in C$.

Analysis: Symmetric two-task case

$T = 2$ with N samples per task
Y_1, Y_2 both have variance σ^2,

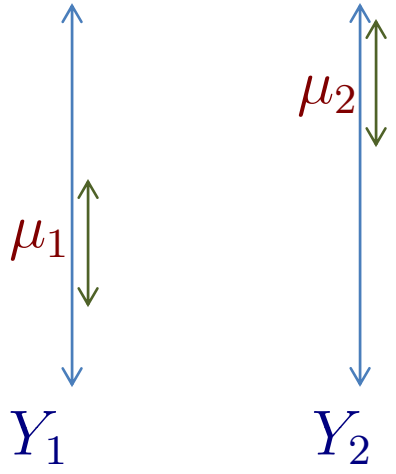


Analysis: Symmetric two-task case

$T = 2$ with N samples per task
Y_1, Y_2 both have variance σ^2,

MTA estimate:

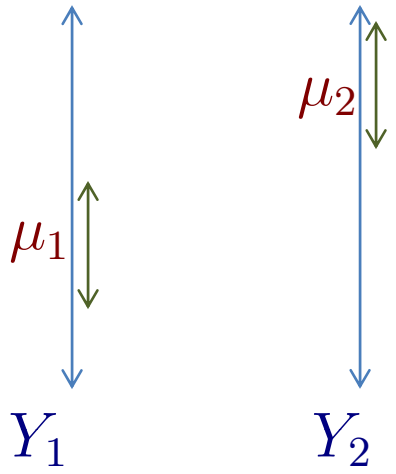
$$Y_1^* = \left(\frac{N + A_{12}}{N + 2A_{12}}\right)\bar{Y}_1 + \left(\frac{A_{12}}{N + 2A_{12}}\right)\bar{Y}_2$$



Biased, but smaller variance than sample averages.

Analysis: Symmetric two-task case

$T = 2$ with N samples per task
Y_1, Y_2 both have variance σ^2,



MTA estimate:

$$Y_1^* = \left(\frac{N + A_{12}}{N + 2A_{12}}\right)\bar{Y}_1 + \left(\frac{A_{12}}{N + 2A_{12}}\right)\bar{Y}_2$$

Biased, but smaller variance than sample averages.

$$\text{MSE}[Y_1^*] < \text{MSE}[\bar{Y}_1] \text{ if } (\mu_1 - \mu_2)^2 < 2\sigma^2\left(\frac{1}{N} + \frac{1}{A_{12}}\right)$$

MTA better than sample averages if:

means close compared to variance

fewer samples N

you don't regularize too hard (small A_{12})

Analysis: Symmetric two-task case

What should the task-similarity matrix values A mean?

Example:

Task 1:
Estimate
average movie
ticket price

Task 2:
Estimate
mean age
of kids at
summer camp

Analysis: Symmetric two-task case

What should the task-similarity matrix values A mean?

Example:

Answer: the optimal task similarity in terms of MSE:

$$A^*_{12} = \frac{\sigma^2}{(\mu_1 - \mu_2)^2}$$

Analysis: Asymmetric Two-task Case

N_1 samples, σ_1^2

N_2 samples, σ_2^2

Analysis: Asymmetric Two-task Case

N_1 samples, σ_1^2

N_2 samples, σ_2^2

Optimal Task-Similarity to Minimize MSE For Task 1:

$$A_{12}^* = \frac{\sigma_1^2}{\Delta^2 - \frac{\sigma_1^2+\sigma_2^2}{N_2}}$$

Analysis: Asymmetric Two-task Case

N_1 samples, σ_1^2

N_2 samples, σ_2^2

Optimal Task-Similarity to Minimize MSE For Task 1:

$$A_{12}^* = \frac{\sigma_1^2}{\Delta^2 - \frac{\sigma_1^2+\sigma_2^2}{N_2}}$$

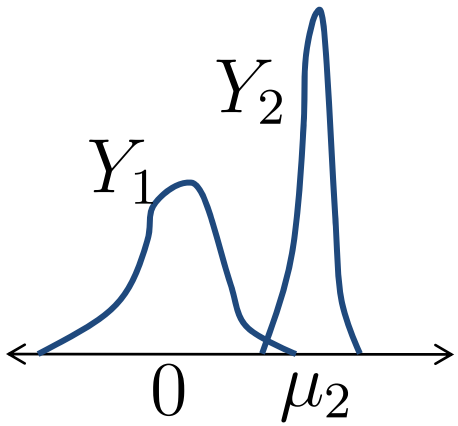
Optimal Task-Similarity to Minimize MSE For Both Tasks:

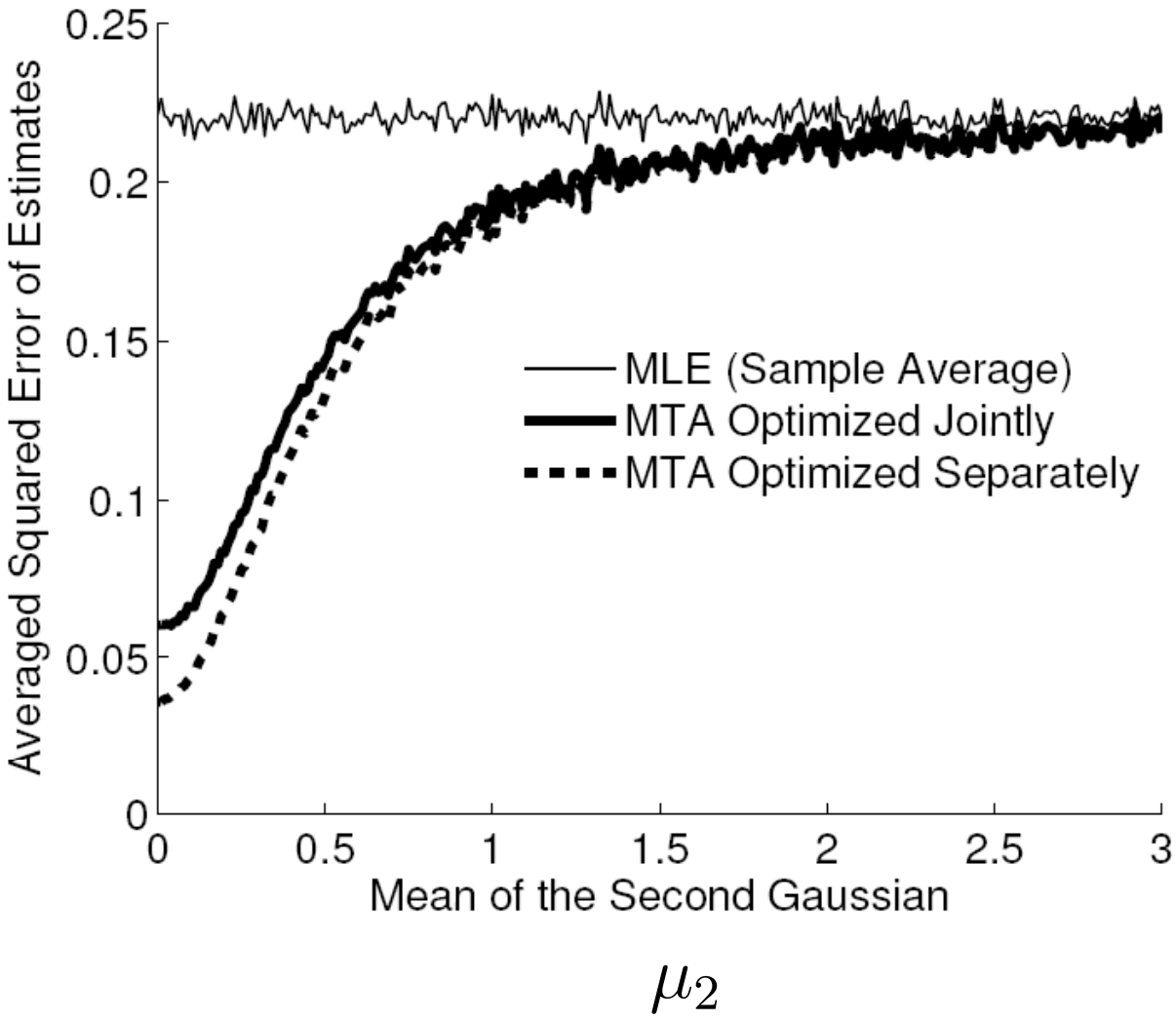
$$A_{12}^* = \frac{N_2^2\sigma_1^2 + N_1^2\sigma_2^2}{\Delta^2(N_1^2+N_2^2) + \sigma_1^2(N_1 - N_2) + \sigma_2^2(N_2 - N_1)}$$

Asymmetric Two-Task Case Simulation

$$Y_1 \sim \mathcal{N}(0, 1)$$
$$N_1 = 5$$

$$Y_2 \sim \mathcal{N}(\mu_2, .2)$$
$$N_2 = 10$$

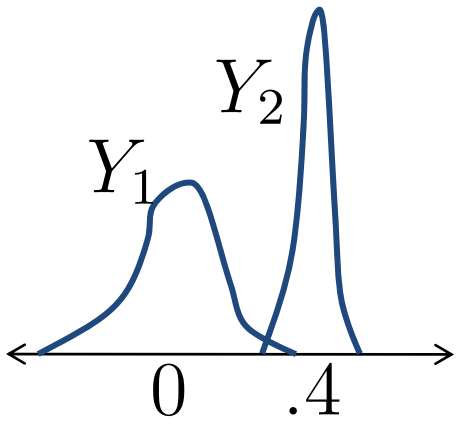


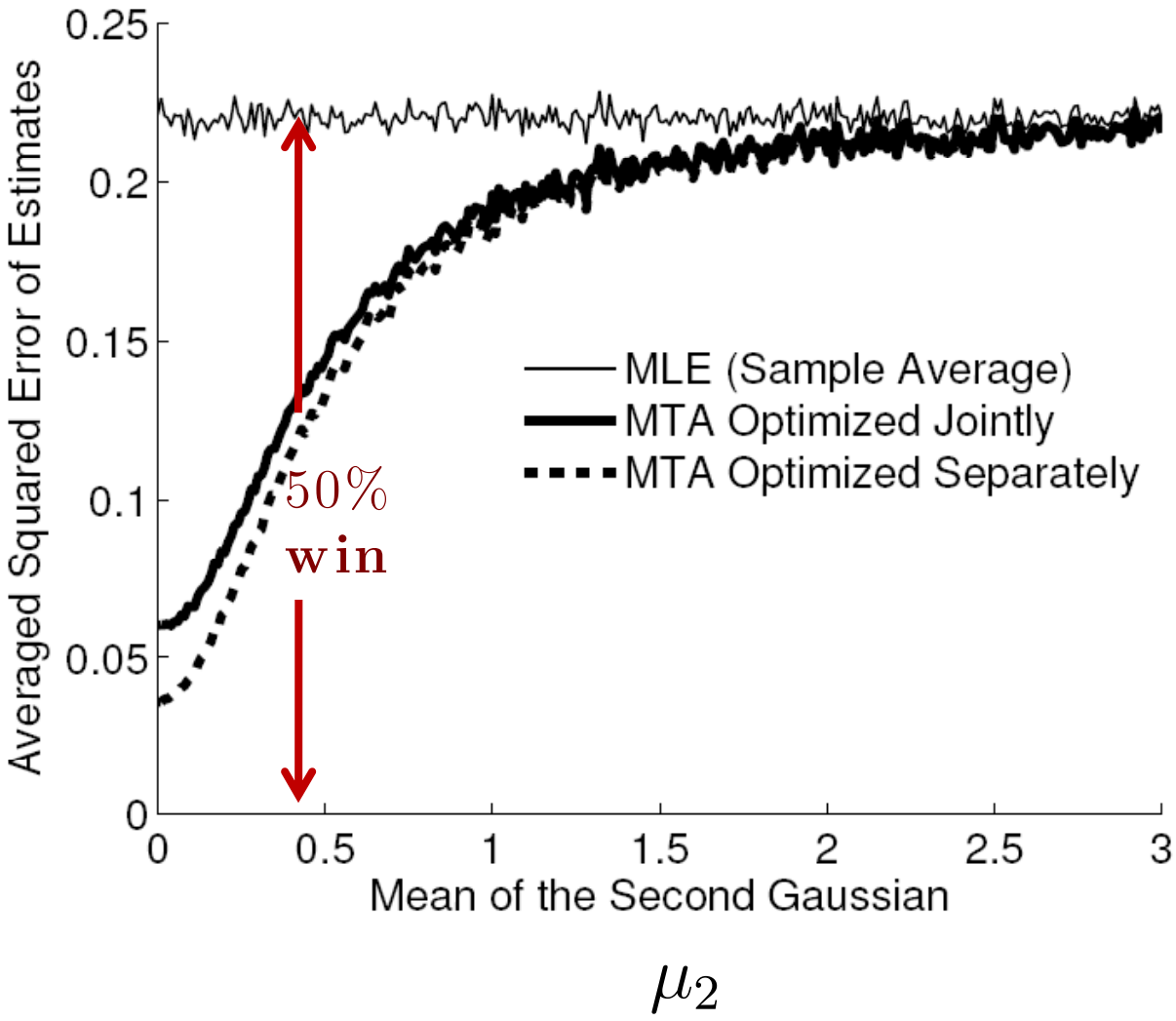


Asymmetric Two-Task Case Simulation

$$Y_1 \sim \mathcal{N}(0, 1)$$
$$N_1 = 5$$

$$Y_2 \sim \mathcal{N}(\mu_2, .2)$$
$$N_2 = 10$$

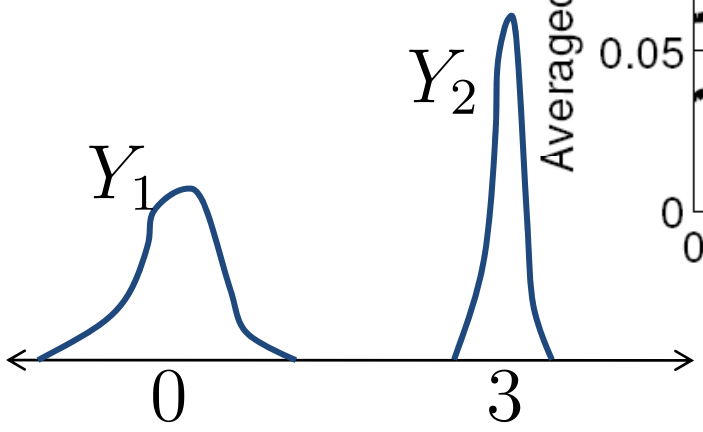


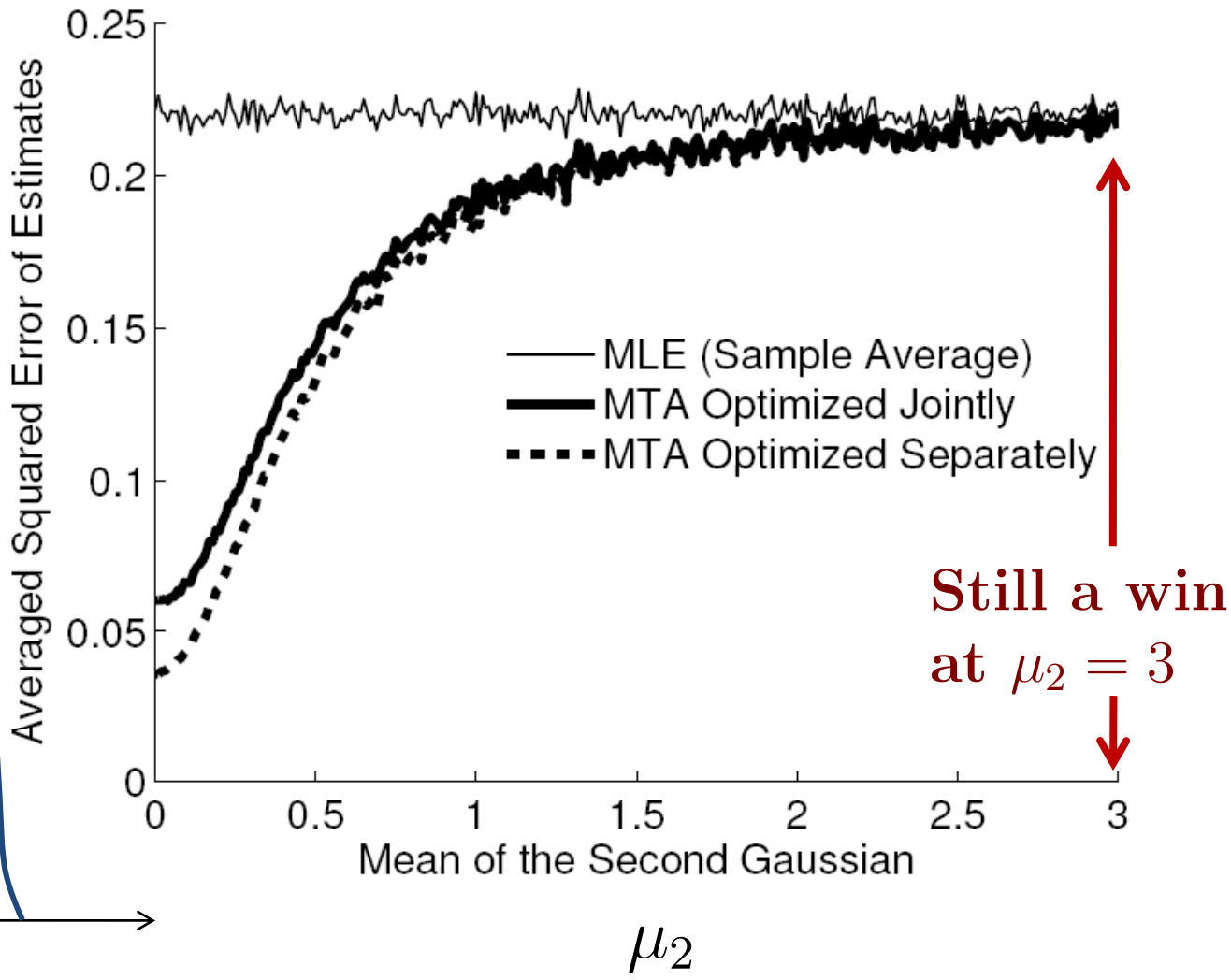


Asymmetric Two-Task Case Simulation

$Y_1 \sim \mathcal{N}(0, 1)$
$N_1 = 5$

$Y_2 \sim \mathcal{N}(\mu_2, .2)$
$N_2 = 10$





Analysis: Asymmetric Two-task Case

N_1 samples, σ_1^2

N_2 samples, σ_2^2

Optimal Task-Similarity to Minimize MSE For Task 1:

$$A_{12}^* = \frac{\sigma_1^2}{(\mu_1 - \mu_2)^2 - \frac{\sigma_1^2 + \sigma_2^2}{N_2}}$$

Optimal task similarity can be negative!

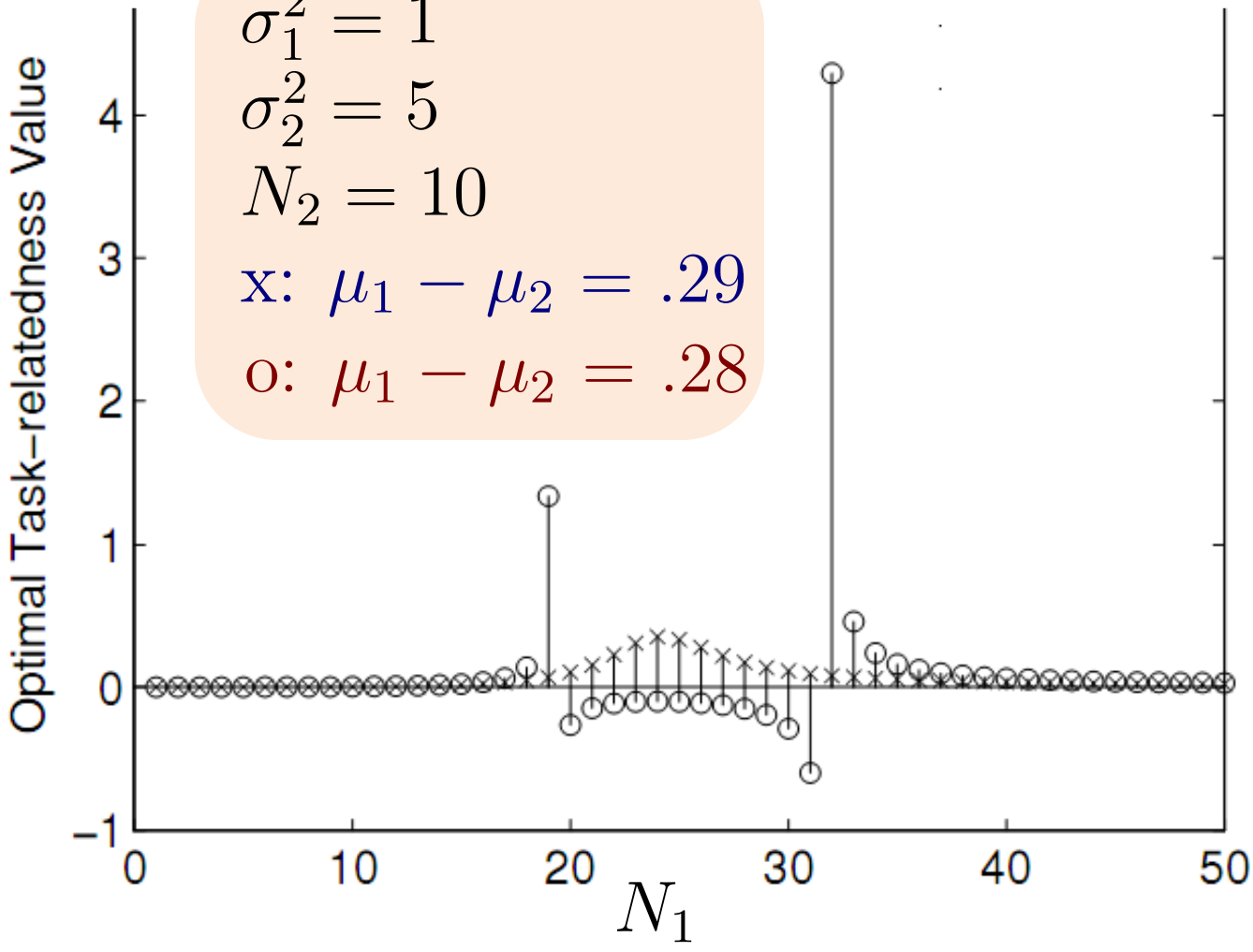
Optimal Task-Similarity to Minimize MSE For Both Tasks:

$$A_{12}^* = \frac{N_2^2\sigma_1^2 + N_1^2\sigma_2^2}{(\mu_1 - \mu_2)^2(N_1^2 + N_2^2) + \sigma_1^2(N_1 - N_2) + \sigma_2^2(N_2 - N_1)}$$

Optimal Similarity Examples

$$A_{12}^{*} = \frac{N_2^2\sigma_1^2 + N_1^2\sigma_2^2}{(\mu_1-\mu_2)^2(N_1^2+N_2^2) + \sigma_1^2(N_1-N_2) + \sigma_2^2(N_2-N_1)}$$

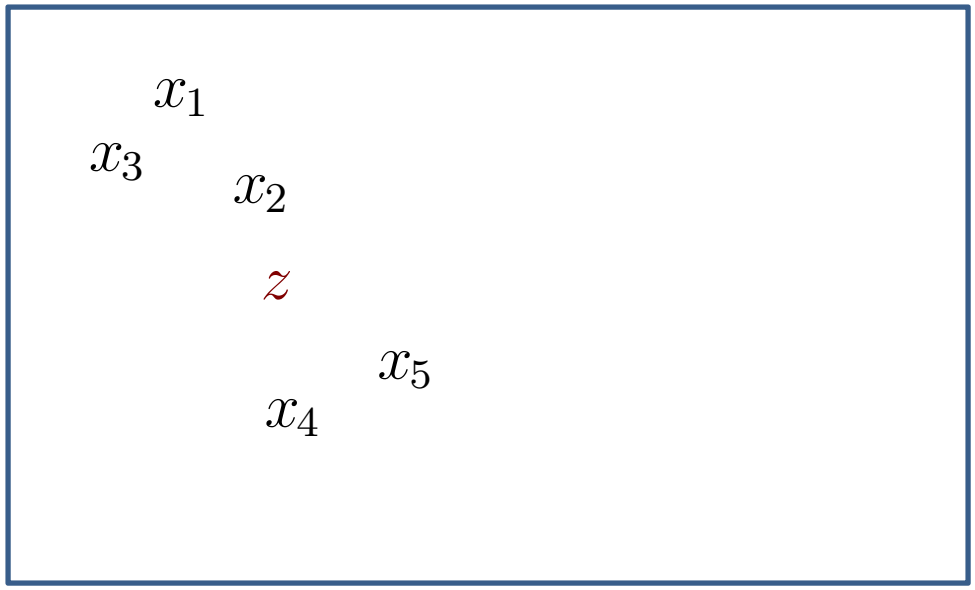
$\sigma_1^2 = 1$
$\sigma_2^2 = 5$
$N_2 = 10$
x: $\mu_1 - \mu_2 = .29$
o: $\mu_1 - \mu_2 = .28$



MTA Applied to Kernel Density Estimation

KDE: Given that events $\{x_i\}$ happened, estimate the probability of event z as

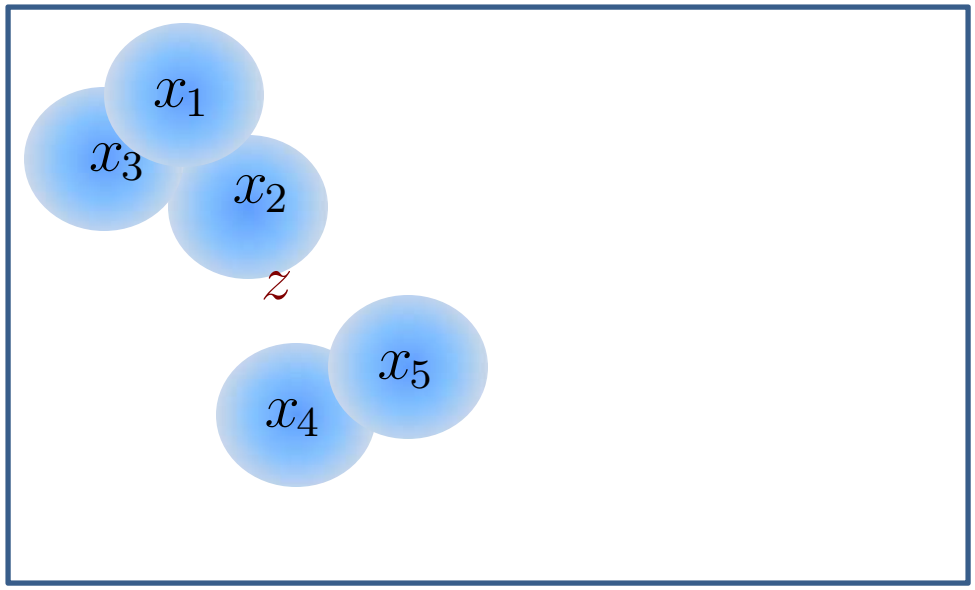
$$\hat{p}(z) = \gamma \sum_{i=1}^{N} K(x_i, z)$$



MTA Applied to Kernel Density Estimation

KDE: Given that events $\{x_i\}$ happened, estimate the probability of event z as

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



MTA Applied to Kernel Density Estimation

KDE: Given that events $\{x_i\}$ happened, estimate the probability of event z as

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

Equivalently,

$$\arg\min_{\hat{y}(z)} \sum_{i=1}^{N} \left(K(x_i, z) - \hat{y}(z)\right)^2$$

MTA Applied to Kernel Density Estimation

KDE: Given that events $\{x_i\}$ happened, estimate the probability of event z as

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

Equivalently,

$$\arg\min_{\hat{y}(z)} \sum_{i=1}^{N} \left(K(x_i, z) - \hat{y}(z)\right)^2$$

Use MTA to form a Multi-task KDE:

$$\arg\min_{\{\hat{y}_t(z_t)\}_{t=1}^{T}} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \left(K_t(x_{ti}, z_t) - \hat{y}_t(z_t)\right)^2 + \gamma \sum_{r=1}^{T} \sum_{s=1}^{T} A_{rs} \left(\hat{y}_r(z_r) - \hat{y}_s(z_s)\right)^2$$

MT-KDE for Terrorism Risk Assessment

(in collaboration with NRL)

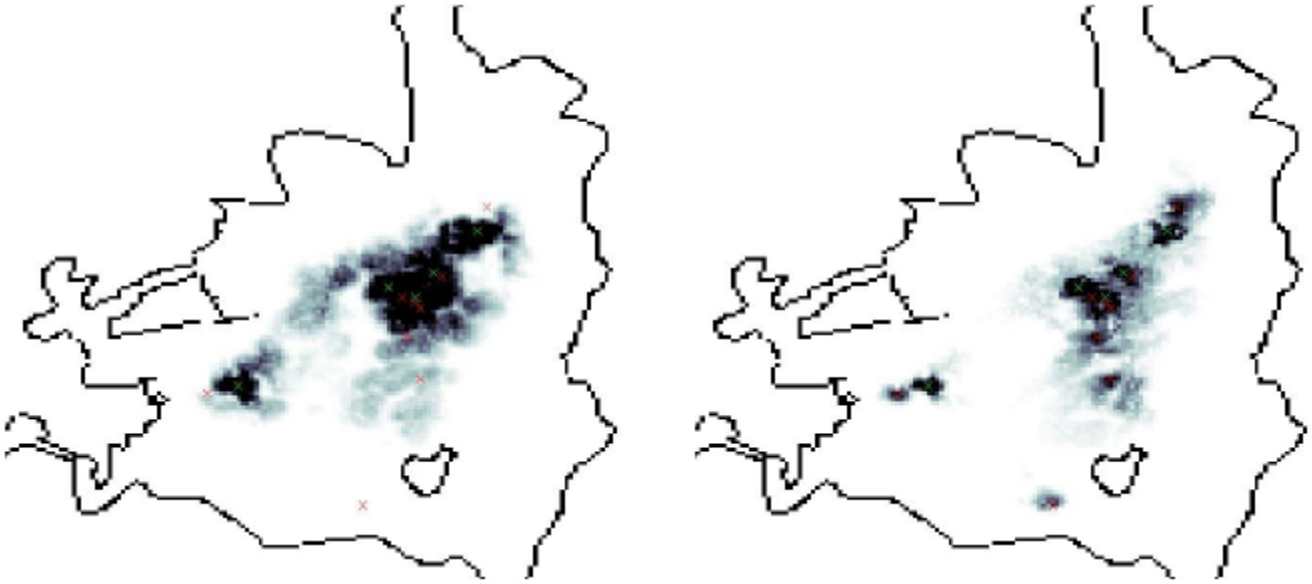
Problem: Estimate the probability of terrorist events by each of $T = 7$ terrorist groups for 40,000 locations in Jerusalem, each location mapped to a 74 dimensional feature vector.

Task similarity matrix from Mohammed Hafez:

	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

MT-KDE for Terrorism Risk Assessment

(in collaboration with NRL)



KDE for AAMB

MT-KDE for AAMB

Average Rank of Left-Out Event

	Suicide Bombings	Bombings	Shootings
KDE	308	6,512	11,160
MT-KDE	88	900	4,752

MTA Averaging

Simple, provably good.

Applicable for any smoothing/averaging.

Works great in practice.

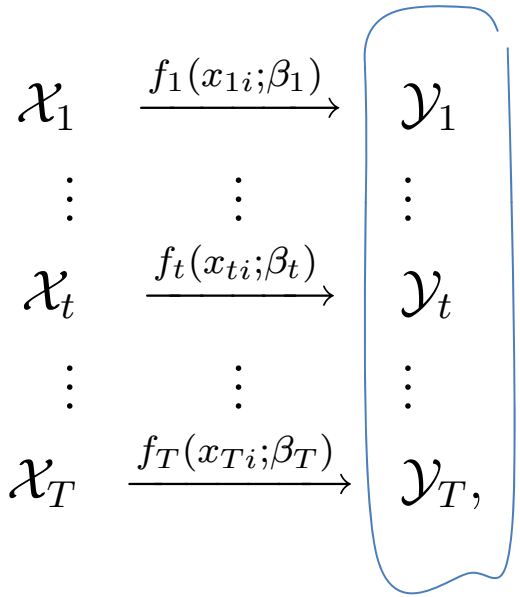
MTA Averaging

Simple, provably good.

Applicable for any smoothing/averaging.

Works great in practice.

Can we do more complicated MT-OSR?



MTA Averaging

Simple, provably good.

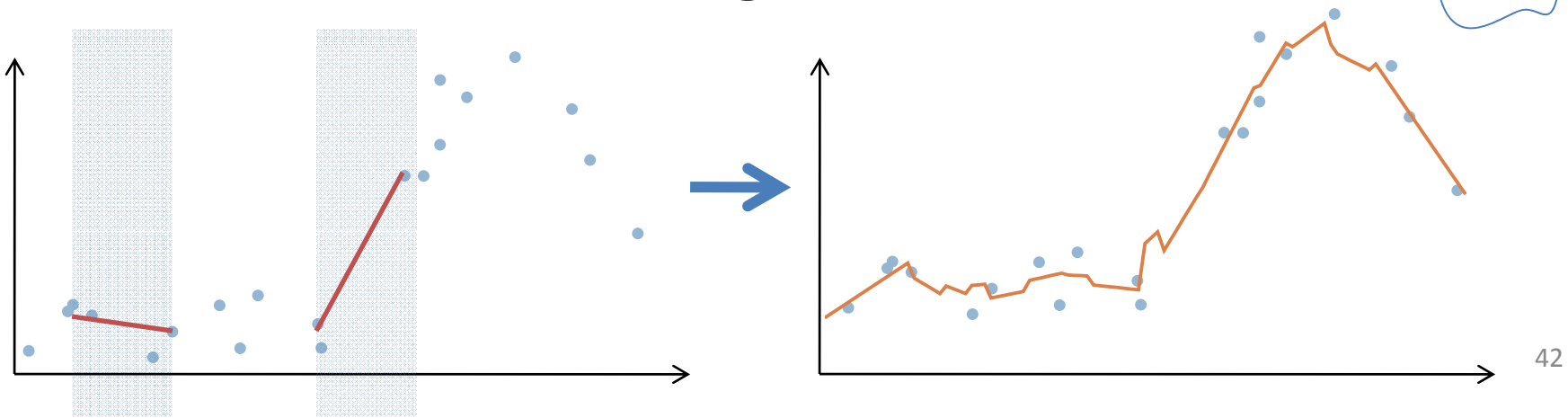
Applicable for any smoothing/averaging.

Works great in practice.

Can we do more complicated MT-OSR?

$$\begin{array}{ccc} \mathcal{X}_1 & \xrightarrow{f_1(x_{1i};\beta_1)} & \mathcal{Y}_1 \\ \vdots & \vdots & \vdots \\ \mathcal{X}_t & \xrightarrow{f_t(x_{ti};\beta_t)} & \mathcal{Y}_t \\ \vdots & \vdots & \vdots \\ \mathcal{X}_T & \xrightarrow{f_T(x_{Ti};\beta_T)} & \mathcal{Y}_T, \end{array}$$

Multi-task Local Linear Regression:



Multi-Task Local Linear Regression

Linear model: $f_t(x) = \beta_t^T x + \beta_{t0}$

Tasks: the tth neighborhood is the tth task.

Objective:

$$\arg\min_{\beta} \sum_{t=1}^{T} \sum_{i=1}^{N_t} (y_{ti} - \beta_t^T x_{ti} - \beta_{t0})^2$$

Tikhonov regularizer

$$+ \lambda \sum_{t=1}^{T} (\beta_t - \beta_G)^T (\beta_t - \beta_G)$$

Output Space Regularizer

$$+ \frac{\gamma}{2} \sum_{r,s=1}^{T} A_{rs} (f_r(z_r) - f_s(z_s))^2$$

MT-LLR Closed Form Solution

$$\hat{B}^* = (XX^T + \gamma ZLZ^T + \lambda I)^{-1}(XY^T - XB_0^T - \gamma ZL\beta_0^T + \lambda B_G)\mathbf{1}$$

$X_t \in \mathbb{R}^{D \times N_t}$ the matrix with columns x_{ti}, for $i = \{1, \ldots, N_t\}$

$y_t \in \mathbb{R}^{1 \times N_t}$ the row vector with entries y_{ti}

$X \in \mathbb{R}^{DT \times N}$ a block diagonal matrix with blocks X_t, for $t = \{1, \ldots, T\}$, with $N = \sum_{t=1}^{T} N_t$

$Z \in \mathbb{R}^{DT \times T}$ a block diagonal matrix with single column blocks z_t

$Y \in \mathbb{R}^{T \times N}$ a block diagonal matrix with single row blocks y_t, for $t = \{1, \ldots, T\}$

$B \in \mathbb{R}^{DT \times T}$ a block diagonal matrix with single column blocks β_t

$L = A^d - A$ the graph Laplacian matrix (Chung, 2004) of A, where A^d is a diagonal matrix with entries $A^d_{rr} = \sum_{t=1}^{T} A_{rt}$

$B_0 \in \mathbb{R}^{T \times N}$ a block diagonal matrix with single row blocks containing β_{t0} repeated N_t times

$\beta_0 \in \mathbb{R}^{T \times T}$ a diagonal matrix, with (t,t)th entry β_{t0}

$B_G \in \mathbb{R}^{dT \times T}$ a block diagonal matrix with single column blocks β_G

$\mathbf{1}$ a column vector of ones.

MT-LLR Application: Color Management

Problem: Estimate what RGB color should be input to a printer to best reproduce a desired CIELab color for CLP300 and HP D760 printers.

Tasks: RGB value estimation for the tth gridpoint in the color lattice. $T_{CLP} = 611$ and $T_{HP} = 756$.

Results:

	Mean Error	Mean Roughness
LLR w/ Tikh. Reg.	3.07	18.25
LLR w/ Tikh. Reg + OSR	3.06	16.25

same error

smoother results

Conclusions

Regularize outputs across task:

$$\begin{array}{ccc} \mathcal{X}_1 & \xrightarrow{f_1(x_{1i};\beta_1)} & \mathcal{Y}_1 \\ \vdots & \vdots & \vdots \\ \mathcal{X}_t & \xrightarrow{f_t(x_{ti};\beta_t)} & \mathcal{Y}_t \\ \vdots & \vdots & \vdots \\ \mathcal{X}_T & \xrightarrow{f_T(x_{Ti};\beta_T)} & \mathcal{Y}_T, \end{array}$$

Conclusions

Regularize outputs across task:

$$\begin{array}{ccc} \mathcal{X}_1 & \xrightarrow{f_1(x_{1i};\beta_1)} & \mathcal{Y}_1 \\ \vdots & \vdots & \vdots \\ \mathcal{X}_t & \xrightarrow{f_t(x_{ti};\beta_t)} & \mathcal{Y}_t \\ \vdots & \vdots & \vdots \\ \mathcal{X}_T & \xrightarrow{f_T(x_{Ti};\beta_T)} & \mathcal{Y}_T, \end{array}$$

Provably good (MTA), **useful in practice** (MTA, MTLLR).

Conclusions

Regularize outputs across task:

$$\begin{array}{ccc} \mathcal{X}_1 & \xrightarrow{f_1(x_{1i};\beta_1)} & \mathcal{Y}_1 \\ \vdots & \vdots & \vdots \\ \mathcal{X}_t & \xrightarrow{f_t(x_{ti};\beta_t)} & \mathcal{Y}_t \\ \vdots & \vdots & \vdots \\ \mathcal{X}_T & \xrightarrow{f_T(x_{Ti};\beta_T)} & \mathcal{Y}_T, \end{array}$$

Provably good (MTA), **useful in practice** (MTA, MTLLR).

Output Space Regularization is flexible: input space/functions don't need to be the same. Only requires outputs to be comparable.

Conclusions

Regularize outputs across task:

$$\begin{array}{ccc} \mathcal{X}_1 & \xrightarrow{f_1(x_{1i};\beta_1)} & \mathcal{Y}_1 \\ \vdots & \vdots & \vdots \\ \mathcal{X}_t & \xrightarrow{f_t(x_{ti};\beta_t)} & \mathcal{Y}_t \\ \vdots & \vdots & \vdots \\ \mathcal{X}_T & \xrightarrow{f_T(x_{Ti};\beta_T)} & \mathcal{Y}_T, \end{array}$$

Provably good (MTA), **useful in practice** (MTA, MTLLR).

Output Space Regularization is flexible: input space/functions don't need to be the same. Only requires outputs to be comparable.

Paper available at arxiv.org/abs/1107.4390

Like manifold regularization but for multiple tasks

Belkin et al.'s Laplacian-regularized least squares objective:

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

(Function regularizer in RKHS $\mathcal{H}$)

Regularizes function outputs of a SINGLE task

In the multi-task case, we don't have a notion of manifold

Bayesian Analysis

This regularization doesn't correspond to a nice prior.

Bayesian Analysis

Assuming differences are independent:

$$p(\hat{y}) \propto \prod_{r=1}^{T} \prod_{s=1}^{T} e^{-\gamma A_{rs}(\hat{y}_r - \hat{y}_s)^2} \xrightarrow{\text{for } T=3} \begin{aligned} \hat{Y}_1 - \hat{Y}_2 &\sim \mathcal{N}(0, 1/2\gamma A_{12}) \\ \hat{Y}_2 - \hat{Y}_3 &\sim \mathcal{N}(0, 1/2\gamma A_{23}) \\ \hat{Y}_1 - \hat{Y}_3 &\sim \mathcal{N}(0, 1/2\gamma A_{13}) \end{aligned}$$

$$\hat{Y}_1 - \hat{Y}_3 = (\hat{Y}_1 - \hat{Y}_2) + (\hat{Y}_2 - \hat{Y}_3) \rightarrow \frac{1}{A_{13}} = \frac{1}{A_{12}} + \frac{1}{A_{23}}$$

$$\hat{Y}_1 - \hat{Y}_2 = (\hat{Y}_1 - \hat{Y}_3) + (\hat{Y}_3 - \hat{Y}_2) \rightarrow \frac{1}{A_{12}} = \frac{1}{A_{13}} + \frac{1}{A_{32}}$$

$$\hat{Y}_2 - \hat{Y}_3 = (\hat{Y}_2 - \hat{Y}_1) + (\hat{Y}_1 - \hat{Y}_3) \rightarrow \frac{1}{A_{23}} = \frac{1}{A_{21}} + \frac{1}{A_{13}}$$

Impossible to satisfy all RHS with any finite A!

Multi-Task Additive Regularizers

Goal: Estimate T functions $f_t(x; \beta_t)$

Given: $\{(x_{ti}, y_{ti})\}_{i=1}^{n_t}$ for $t = 1, \ldots, T$.

Standard multi-task: regularize the parameters:

$$\arg \min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^{T} \sum_{i=1}^{N_t} L(y_{ti}, f_t(x_{ti}; \beta_t)) + \gamma J(\{\beta_r\}_{r=1}^T).$$

loss (empirical risk)

Example: Linear model $f_t(x) = \beta_t^T x$ (Evgeniou et al. 2004)

$$\arg \min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \left(y_{ti} - \beta_t^T x_i\right)^2 + \gamma \sum_{r=1}^{T} \|\beta_r - \frac{1}{T} \sum_{s=1}^{T} \beta_s\|_2^2$$

least-squares fit

regularize parameters