Stein effect for estimating many vector means

A "blessing of dimensionality" phenomenon

G. Blanchard

Université Paris-Saclay

Geometry and Statistics in Data Sciences
Non-Linear and High Dimensional Inference 03-07/10/2022

Joint work with: Hannah Marienwald (Technische Universität, Berlin),
Jean-Baptiste Fermanian (U. Paris-Saclay)

Hannah Marienwald

Jean-Baptiste Fermanian

An old problem – estimating many means

Consider a many-samples model (sample="bag")

$$\begin{cases} X_\bullet^{(k)} := (X_i^{(k)})_{1 \le i \le N_k} \overset{i.i.d.}{\sim} \mathbb{P}_k, \ k \in [\![B]\!] := \{1, \dots, B\}; \\ X_i^{(k)} \in \mathbb{R}^d; \\ (X_\bullet^{(1)}, \dots, X_\bullet^{(B)}) \text{ independent}, \\ \mu_k := \mathbb{E}_{X \sim \mathbb{P}_k}[X] \in \mathbb{R}^d, \ k \in [\![B]\!], \text{ unknown}. \end{cases}$$

- $\mathbb{P}_1, \dots, \mathbb{P}_B$: square integrable distributions on $\mathbb{R}^d$ ("tasks")
- **Goal:** estimation of mean vectors $(\mu_k)_{k \in [\![B]\!]}$
- Sometimes called "multi-task averaging" (MTA) (Feldman et al, 2014)

The naive estimator

- Criterion: single (M)SE,

$$L_k(\widehat{\mu}_k) := \|\widehat{\mu}_k - \mu_k\|^2; \qquad R_k(\widehat{\mu}_k) := \mathbb{E}[L_k(\widehat{\mu}_k)];$$

- and compound (M)SE

$$\overline{L}(\widehat{\mu}) := \frac{1}{B}\sum_{k=1}^{B}\|\widehat{\mu}_k - \mu_k\|^2; \qquad \overline{R}(\widehat{\mu}) := \mathbb{E}[L(\widehat{\mu})].$$

- The benchmark estimators are the "naive" bag-wise **empirical means**

$$\widehat{\mu}_k^{\mathrm{NE}} := \frac{1}{N_k}\sum_{i=1}^{N_k} X_i^{(k)}.$$

- It holds

$$R_k(\widehat{\mu}_k^{\mathrm{NE}}) = \frac{\operatorname{Tr}\Sigma_k}{N_k} =: s_k^2,$$

where $\Sigma_k := \operatorname{Cov}\left[X_1^{(k)}\right]$ is the covariance of $\mathbb{P}_k$.

Motivations and goals

Motivations:

- Many large databases have such a structure.

- Kernel mean embedding of distributions (possibly with random features).

Goals

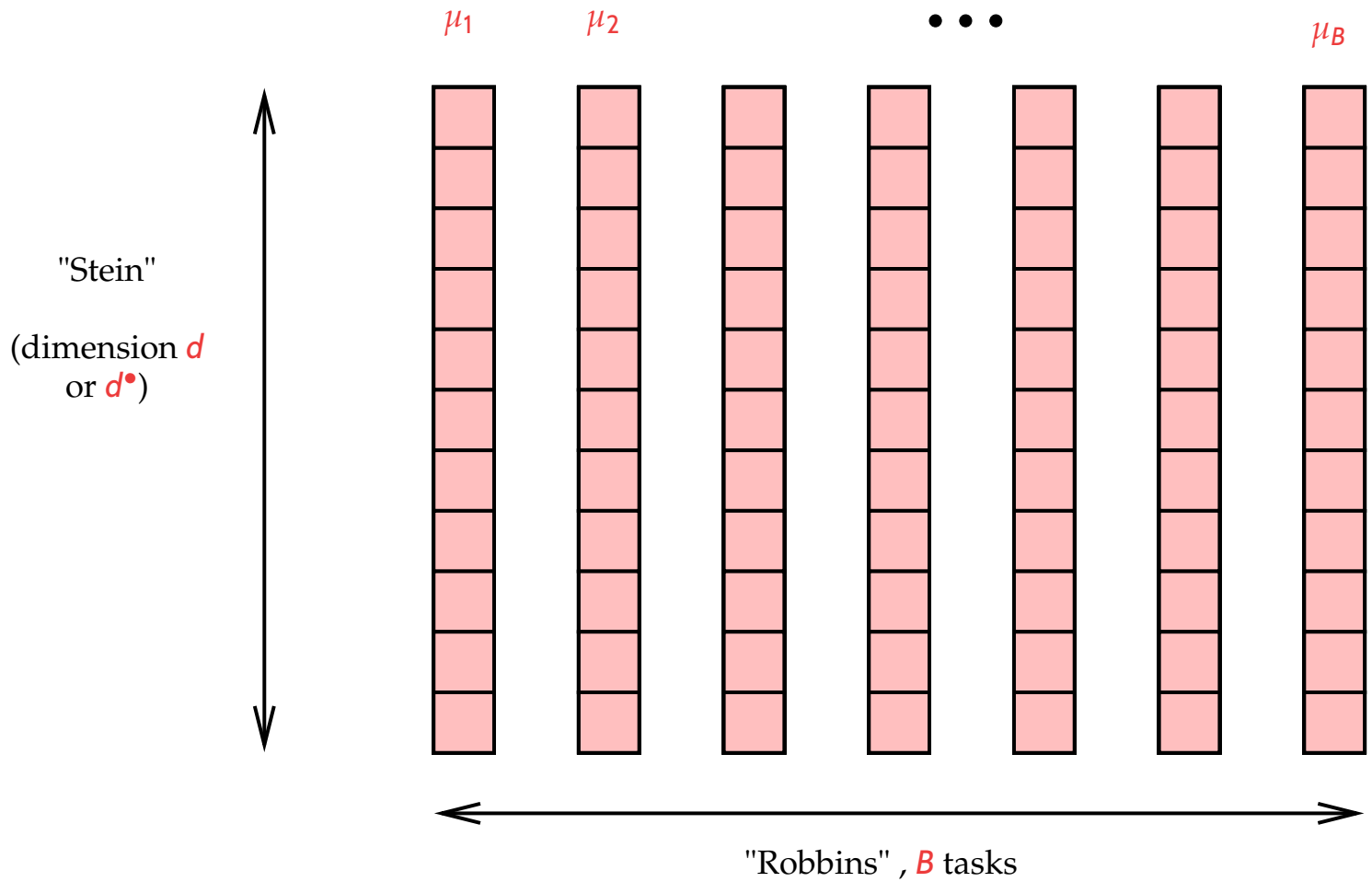
- General (old) question: can we improve over the naive estimator?
  $\rightarrow$ **the performance of estimators will be measured via ratios to the naive estimator risk s_k^2.**

- It will turn out that the improvement can be larger in large (effective) dimension,

$$d_k^\bullet := \frac{(\operatorname{Tr}\Sigma_k)^2}{\operatorname{Tr}\Sigma_k^2} = \frac{\|\Sigma_k\|_1^2}{\|\Sigma_k\|_2^2}.$$

  $\rightarrow$ **"Large dimensional asymptotics"** rather than large sample asymptotics.

Stein and Robbins



$B = 1$: the Stein effect

- Isotropic Gaussian setting: $\mathbb{P}_1 = \mathcal{N}(\mu, \sigma^2)$, known σ.
- Shrinkage estimator: $\omega \in [0, 1]$,

$$\widehat{\mu}_\omega := \omega \widehat{\mu}^{\mathrm{NE}},$$

with risk

$$R(\widehat{\mu}_\omega) = (1-\omega)^2 \|\mu\|^2 + \omega^2 s^2,$$

- **"oracle" Stein improvement factor** relative to naive estimator

$$\min_{\omega \in [0,1]} \frac{R(\widehat{\mu}_\omega)}{s^2} = \frac{\tau^{\mathrm{JS}}(\mu)}{1 + \tau^{\mathrm{JS}}(\mu)},$$

where

$$\tau^{\mathrm{JS}}(\mu) := \frac{\|\mu\|^2}{s^2}.$$

$B = 1$, JS estimator, dimensional asymptotics

- **James-Stein (1961) estimator:**

$$\widehat{\mu}^{\mathrm{JS}} := \widehat{\mu}_{\widehat{\omega}_{\mathrm{JS}}}, \qquad \text{with} \quad \widehat{\omega}_{\mathrm{JS}} := \left(1 - \frac{(d-2)\sigma^2/N_1}{\|\widehat{\mu}^{\mathrm{NE}}\|^2}\right).$$

- **Nonasymptotic risk bound:** (see e.g. Tsybakov 2003,2009)

$$\frac{R(\widehat{\mu}^{\mathrm{JS}})}{s^2} < \min\left(\frac{\tau^{\mathrm{JS}}(\mu)}{1+\tau^{\mathrm{JS}}(\mu)} + \frac{4}{d}, 1\right)$$

- **Minimax result of Pinsker (1980):** Fix $\tau \in \mathbb{R}_+$.

$$\lim_{d\to\infty} \inf_{\widehat{\mu}} \sup_{P:\tau^{\mathrm{JS}}(\mu)\le\tau} \frac{R(\widehat{\mu})}{s^2} = \frac{\tau}{1+\tau}.$$

- The JS estimator **asymptotically attains this optimal factor as $d \to \infty$** (without knowing τ)

$B > 1$: "Robbins" point of view, and naive Bayes

- One can assume an a priori distribution of the means $\mu_k \overset{i.i.d}{\sim} \mathbb{Q}$
- ...and a model for sample distributions (e.g. $\mathcal{N}(\mu_k, \sigma^2 I_d)$)
- **Independence** of bags conditional to their means is essential in this point of view (more than in Stein's!)
- The Bayes estimator is bag-wise posterior mean, of the form

$$\widehat{\mu}_{\bullet}^{\mathbb{Q}-\mathrm{Bayes}} = \left(\varphi_{\mathbb{Q}}(\widehat{\mu}_1^{\mathrm{NE}}), \ldots, \varphi_{\mathbb{Q}}(\widehat{\mu}_B^{\mathrm{NE}}) \right).$$

- **Empirical Bayes** approach initiated by the landmark works of Efron and Morris (Gaussian prior, 70s), to modern nonparametric extensions (Zhang 1997,2003; Jin and Zhang, 2009; Brown and Greenshtein, 2009): try to "estimate" $\varphi_{\mathbb{Q}}$. See also George et al. (2012).
- Most efforts in this direction appear to consider $d = 1$

Returning to many vector means: some goals

- Can we take avantage of having many distributions/samples?
  What if the true means have some (unknown) "structure", e.g.:
  - Clustered
  - Low dimensional support eg. manifold
  - Sparse (support is union of low dimensional structures)
  - **Small covering number at some scale**

- What is the effect of high (effective) dimension?
  (Effective) dimensional asymptotics?

- Improvement for each single mean and not only for compound $\overline{R}$?

- Renounce the exact Stein effect (factor always <1) but aim at improvement factor wrt. naive, with the potential of it being small.

Plan

- Consider simplified situation where some "oracle" information is known:
  - Estimation error s_k^2 of naive estimators
  - "$\tau-$"neighboring means

- Generic bound error for plug-in principle when oracle information is estimated

- Derive suitable estimates for the oracle information

A simple idea

- We want to estimate μ_1. For fixed $\tau > 0$ assume an oracle tells us about "τ-neighboring tasks"

$$V_\tau := \left\{ j \in [\![B]\!] : \left\| \mu_1 - \mu_j \right\|^2 \leq \tau s_1^2 \right\}.$$

- Consider **local shrinkage estimator** $\widehat{\mu}_\omega := \sum_{k \in V_\tau} \omega_k \widehat{\mu}_k^{\mathrm{NE}}$, $\omega \in \mathcal{S}_{V_\tau}$ ($|V_\tau|$-simplex), it holds

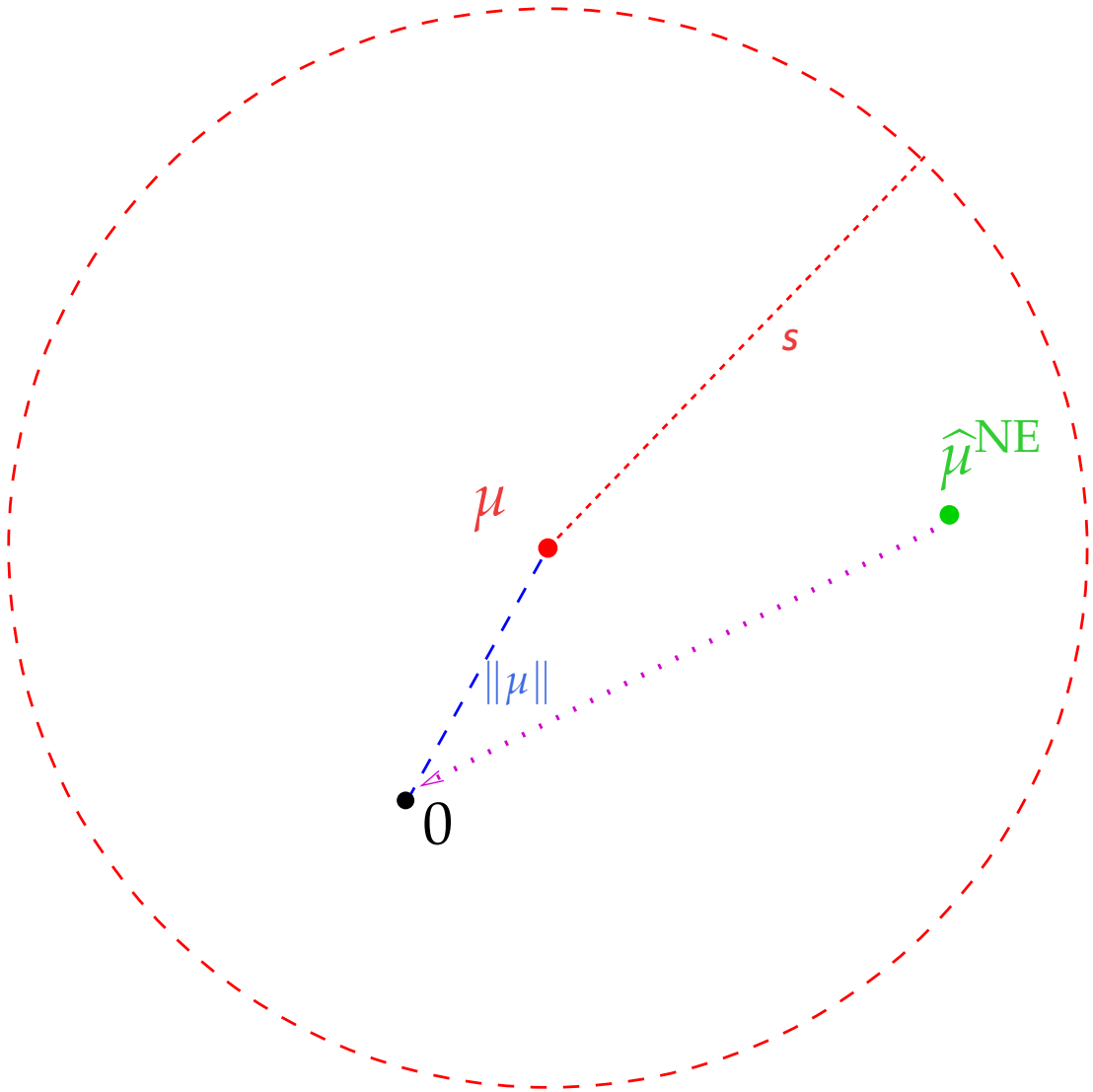
$$R_1(\widehat{\mu}_\omega) \leq \tau(1-\omega_1)^2 s_1^2 + \sum_{k \in V_\tau} \omega_k^2 s_k^2.$$

- Optimizing over $\omega \in \mathcal{S}_{V_\tau}$ yields

$$\frac{R_1(\widehat{\mu}_{\omega_\tau^*})}{s_1^2} \leq \frac{\nu_1(V_\tau) + \tau}{1+\tau} =: \mathcal{B}(\tau, V_\tau), \qquad \nu_1(V) := \frac{s_1^{-2}}{\sum_{k \in V} s_k^{-2}}.$$

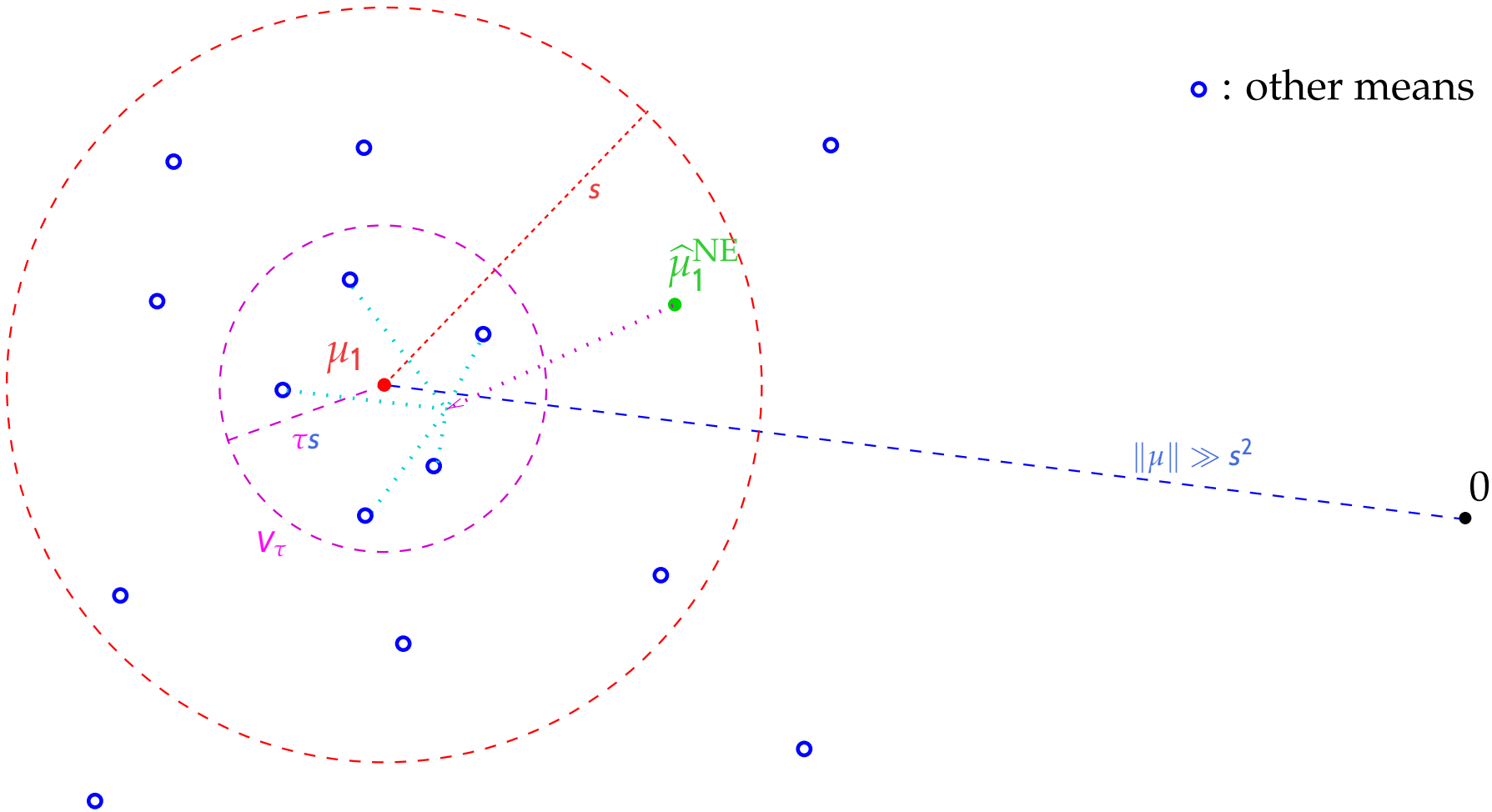
- In the homogeneous setting, $s_k^2 \equiv s_1^2$: $\nu_1(V_\tau) = |V_\tau|^{-1}$.

Stein setting vs. MTA setting



$$\tau = \frac{\|\mu\|^2}{s^2}$$

Stein setting vs. MTA setting



Minimax lower bound

- Consider the Gaussian isotropic model in dimension d: $X_i \sim \mathcal{N}(\mu_i, \sigma^2 I_d)$
- **Individual error bound:** model $\mathcal{P}_d(\tau)$:
  V means (including the first one) are close to an (unknown) vector μ:

$$\exists \nu : \forall k \in [\![B]\!] \qquad \|\mu_k - \nu\|^2 \leq \tau s^2.$$

- Then for the estimation of the first mean

$$\lim_{d\to\infty} \inf_{\widehat{\mu}} \sup_{P\in\mathcal{P}_d(\tau)} \frac{R_1(\widehat{\mu})}{s_1^2} \geq \frac{\tau + V^{-1}}{1+\tau}.$$

- Does not quite match the oracle upper bound, but up to a factor 2

Plug-in procedure

- Slight cheating: suppose we have esimates $\widetilde{V}, \widetilde{s}_k^2$ of V_τ, s_k^2 on "tilde" **independent** data (e.g. use data splitting). Results conditional on "tilde" data and in expectation with respect to "main"data"...

Proposition

Let $\tau > 0$ be fixed. Assume $\widetilde{V} \subseteq [\![B]\!], \widetilde{s}^2 = (\widetilde{s}_k^2)_{k\in[\![B]\!]} \in \mathbb{R}_+^B$ are quantities that are possibly random but independent of the samples in model (3)*. Then conditionally to the event*

$$\mathcal{A} := \begin{cases} V_{\tau'} \subseteq \widetilde{V} \subseteq V_\tau, \\ |\widetilde{s}_k^2 - s_k^2| \leq \varepsilon_k, \text{ for all } k \in \widetilde{V}, \end{cases} \tag{1}$$

if we plug in $(\widetilde{V}, \widetilde{s}^2)$ for (V_τ, s^2) into the oracle formula giving rise to weight vector $\widetilde{\omega}$, it holds

$$\frac{R_1(\widehat{\mu}_{\widetilde{\omega}}|\mathcal{A})}{s_1^2} \leq \mathcal{B}(\tau, \widetilde{V}) + \zeta s_1^{-2} \leq \mathcal{B}(\tau, V_{\tau'}) + \zeta s_1^{-2} \tag{2}$$

where $\zeta := \tau\varepsilon_1 + 2\max_{k\in\widetilde{V}} \varepsilon_k$.

Compound estimation error

Corollary

Let

$$\bar{s}^2 = \max_{k \in [\![B]\!]} s_k^2.$$

Then under the same event $\mathcal{A}$ (for independent estimates) as previously, it holds

$$\frac{\overline{R}(\widehat{\mu}_{\widetilde{\omega}} | \mathcal{A})}{\bar{s}^2} \leq \left(\frac{\tau + \frac{\mathcal{N}}{B}}{1 + \tau} \right) + \zeta / \bar{s}^2,$$

where $\mathcal{N}$ is the covering number of the means at scale $\sqrt{\tau' \bar{s}^2}/2$,

$$\mathcal{N} = \mathcal{N}\left(\{\mu_i, i \in [\![B]\!]\}, \sqrt{\tau' \bar{s}^2}/2 \right) :$$

Proof: $\sum_{k \in [\![B]\!]} |V_{\tau'}|^{-1} \leq \mathcal{N}$

Minimax lower bound for compound error

- Consider the Gaussian isotropic model in dimension d: $X_i \sim \mathcal{N}(\mu_i, \sigma^2 I_d)$
- **Compound error bound:** model $\overline{\mathcal{P}_d}(\mathcal{N}, \tau)$:
  there exist (unknown) vectors $(\nu_i)_{i \in [\![N]\!]}$, s.t. $\{\mu_1, \ldots, \mu_B\} \subseteq \bigcup_{i=1}^{\mathcal{N}} \mathcal{B}(\nu_i, \sqrt{\tau s^2})$,

$$\lim_{d \to \infty} \inf_{(\widehat{\mu}_\bullet)} \sup_{P \in \overline{\mathcal{P}_d}(\mathcal{N}, \tau)} \frac{\overline{R}(\widehat{\mu}_\bullet)}{s^2} \geq \frac{\tau + \frac{\mathcal{N}}{B}}{1 + \tau}.$$

No oracle? Use tests

- Use $(T_{ij})_{i,j\in[\![B]\!]}$ family of tests for

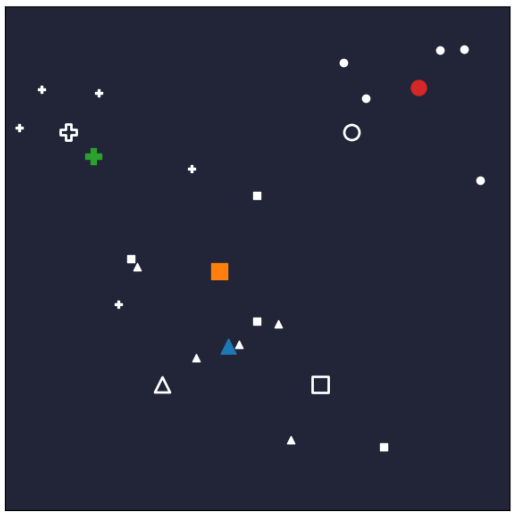
$$(H_0) : \left\|\mu_i - \mu_j\right\|^2 > \tau s_i^2, \quad \text{against} \quad (H_1) : \left\|\mu_i - \mu_j\right\|^2 \leq \tau' s_i^2$$

  ($\tau' \leq \tau$)
- The estimated neighboring tasks for task i are

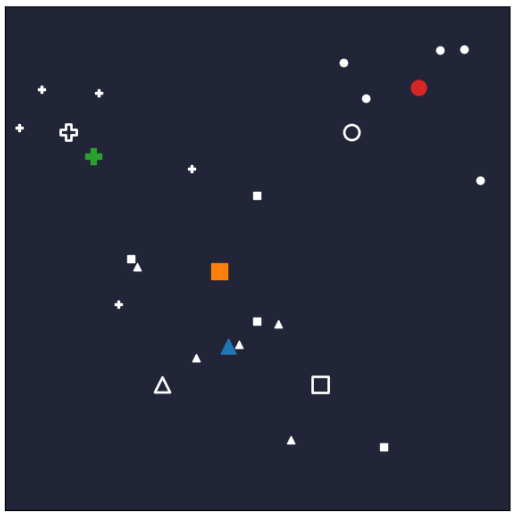
$$\widehat{V}_i := \left\{j \in [\![B]\!] : T_{ij} = 1\right\}$$

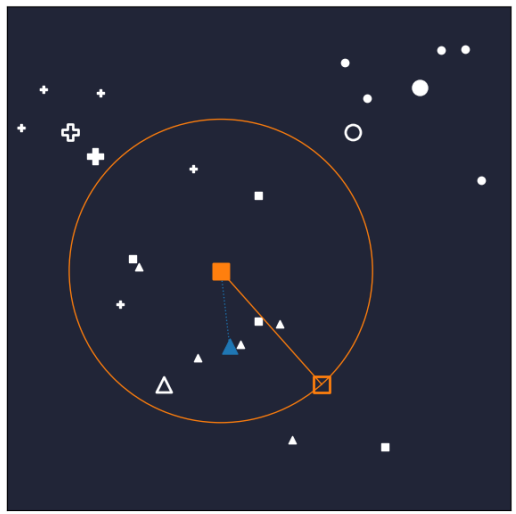
- (Note: we assume $T_{ii} \equiv 1$ so $i \in \widehat{V}_i$ always holds)
- Apply shrinkage estimator for estimated neighbors $\widehat{V}_i$.

Local shrinkage



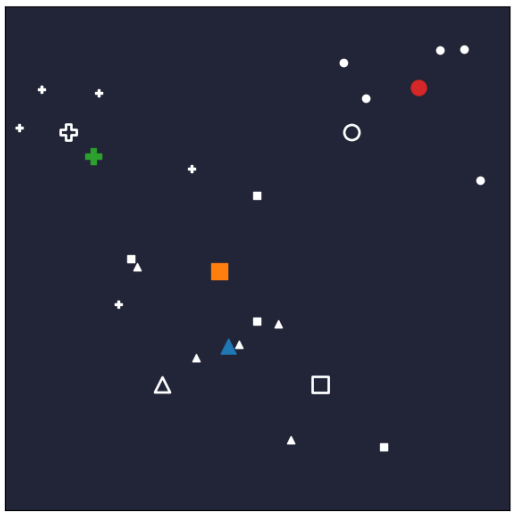
Local shrinkage

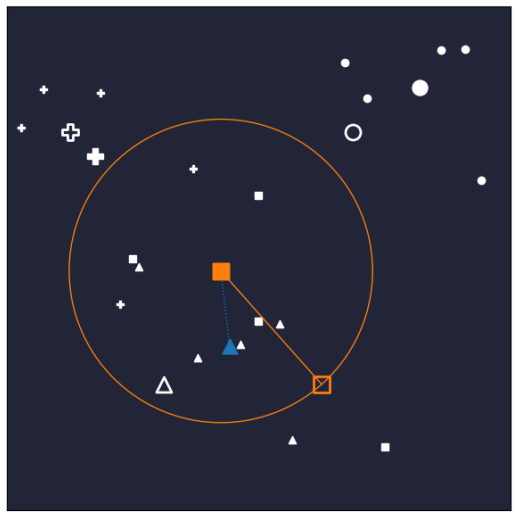




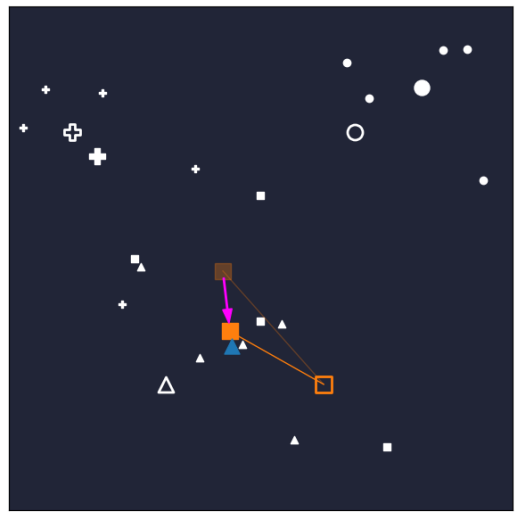
Estimate neighbors
by testing

Local shrinkage





Estimate neighbors by testing



Shrink

- Conceptually similar algorithm (using task clustering): Martinez-Rego and Pontil, 2013.

The story so far

- Possible to improve naive estimation for each **individual** mean if $\tau \ll 1$
- Needed: **test** to estimate V_τ, **estimation** of s_k^2
- We want high test power for $\tau' = c\tau$ with c close to 1
- In low dimension detection distance is of the the order same as estimation error $\rightarrow$ hopeless strategy
- But in high dimension, squared detection distance is **smaller by a factor of** $1/\sqrt{d}$ than naive MSE! (Nonasymptotic point of view: Baraud, 2002)

Tests

- We consider tests based on the following U-statistics (here for $N_i = N_j = N$)

$$U_{ij} := \frac{1}{N(N-1)} \sum_{k \neq l \in [\![N]\!]} \left(\langle X_k^{(i)}, X_l^{(i)} \rangle + \langle X_k^{(j)}, X_l^{(j)} \rangle \right) - \frac{2}{N^2} \sum_{k,l \in [\![N]\!]} \langle X_k^{(i)}, X_l^{(j)} \rangle.$$

- Unbiased estimate of $\left\| \mu_i - \mu_j \right\|^2$
- Same statistic used in so-called **MMD tests** in the kernel setting
- (Also possible to use the biased estimate $\left\| \widehat{\mu}_i^{\mathrm{NE}} - \widehat{\mu}_j^{\mathrm{NE}} \right\|^2$)

Tests – Gaussian case

Proposition

Assume the Gaussian setting, and equal covariances & sample sizes across tasks.
Let $\alpha \in (0,1)$, $\tau > 0$ *and* $C \geq 16$ *be fixed; put* $u_\alpha := \log(8B/\alpha)$. *Let* $\widetilde{T}_k$ *be given by*

$$\widetilde{T}_k := \mathbf{1}\Big\{\widetilde{U}_k \leq (1 - 2C^{-1})\tau s_1^2\Big\}. \tag{3}$$

Assume $\tau \geq \tau_{\min} := 4C^2 u_\alpha / \sqrt{d_1^\bullet}$; *Defining*

$$\widetilde{V} := \Big\{k \in [\![B]\!] : \widetilde{T}_k = 1\Big\},$$

then with probability at least $1 - \alpha$ *it holds* $V_{\kappa\tau} \subseteq \widetilde{V} \subseteq V_\tau$, *where* $\kappa := (1 - 3C^{-1})^2$.
(Note that $\kappa \geq \frac{1}{2}$ *under the assumption on* C.*)*

Gaussian case – Estimating $\|\Sigma_k\|_p (p = 1, 2)$

Proposition

Let $\widetilde{s}_k^2 := \frac{1}{N_k(N_k-1)} \sum_{i=1}^{N_k} \|\widetilde{X}_i^{(k)} - \widetilde{\mu}_k\|^2$, *where* $\widetilde{\mu}_k := N_k^{-1} \sum_{i=1}^{N_k} \widetilde{X}_i^{(k)}$, *and let* $\alpha \in (0, 1)$. *Assume the Gaussian setting. Then with probability at least* $1 - \alpha$*:*

$$\forall k \in [\![B]\!] : \qquad \left|\widetilde{s}_k^2 - s_k^2\right| \leq \frac{s_k^2}{\sqrt{d_k^\bullet}} . 4 \sqrt{\frac{\log(4B\alpha^{-1})}{N_k}} . \tag{4}$$

Also there exists a U*-statistic* $\widetilde{W}_k$ *such that with probability* $1 - \alpha$

$$\left| \sqrt{\widetilde{W}_k} - \|\Sigma_k\|_2 \right| \lesssim \|\Sigma\|_2 \frac{(\log(55B\alpha^{-1}))^2}{\sqrt{N}} .$$

Important take-out: it suffices $N_k \gtrsim \log(B\alpha^{-1})$ to ensure from (0.4) that the remainder term in the risk is of lower order ($1/\sqrt{d_1^\bullet}$) than the oracle risk when $\tau \geq \tau_{\min}$.

Extending the Gaussian setting

- Qualitatively comparable results (for tests and quantile estimation) in the bounded case relevant for KME with bounded kernel. Requires $N \gtrsim d^\bullet$.
  - Annoyingly bounded does not imply "strongly sub-Gaussian" in high dimension $\rightarrow$ specific analysis required

- Similar results seem obtainable in the heavy-tailed case (requiring only moments of order 4) using the Median of Means principle (ongoing)

Results on Gaussian isotropic data

Ratio of MSE of different methods compared to `NE` on Gaussian data (lower is better).

Model	Dimension	JS+	MTA	MTA stb	STB-0	STB
UNIF	100	0.980	0.930	0.992	0.998	0.992
	250	0.952	0.843	0.566	0.578	0.446
	1000	0.832	0.573	0.361	0.203	0.187
CLUSTER	25	0.543	0.492	0.301	0.141	0.141
	50	0.522	0.490	0.127	0.044	0.044
	100	0.509	0.487	0.038	0.022	0.022
SPHERE	100	0.962	0.964	0.571	0.661	0.453
	250	0.911	0.910	0.421	0.300	0.256
	1000	0.715	0.715	0.259	0.106	0.102
SPARSE	50	0.777	0.838	0.633	0.596	0.558
	250	0.463	0.514	0.515	0.469	0.429
	1000	0.198	0.211	0.248	0.250	0.215

Application to Kernel mean embedding (KME) estimation

- If $\phi : \mathcal{Z} \to \mathcal{H}$ is a feature mapping into Hilbert space, define the KME of probability distribution $\mathbb{P}$ on $\mathcal{Z}$ as

$$\mu_{\mathbb{P}} = \phi(\mathbb{P}) := \mathbb{E}_{Z \sim \mathbb{P}}[\phi(Z)]$$

- (The kernel trick is unimportant here, the above can be seen as a generalized method of moments; see also: random feature approximation)
- Used in many applications for various purposes (e.g. Distribution Regression)
- With many samples $Z_{\bullet}^{(k)}$ from distributions $\mathbb{P}_1, \ldots, \mathbb{P}_B$, estimating the KMEs $\phi(\mathbb{P})$ is an instance of the MTA problem with $X = \phi(Z)$
- Common assumption: ϕ is bounded $\to$ we are interested in the bounded setup; effective dimension is more important than dimension.

KME and Relation to Gram matrix of tasks

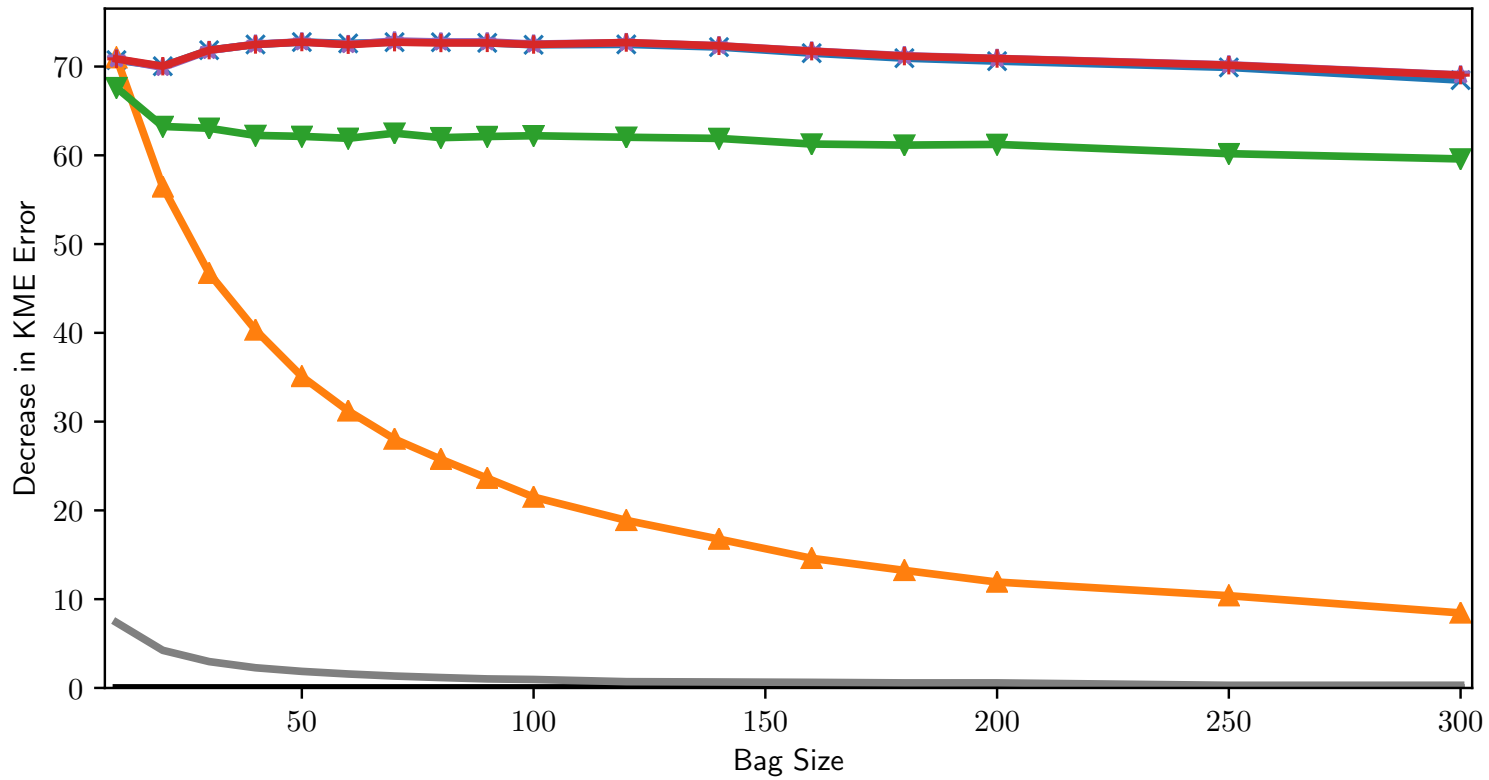
- The (kernel) Gram matrix of the KMEs plays an important role in many applications (e.g. distribution regression)

$$G = \left(\left\langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \right\rangle_{\mathcal{H}} \right)_{i,j \in [\![B]\!]}.$$

- If it is estimated by $\widehat{G} = \left(\left\langle \widehat{\mu}_{\mathbb{P}_i}, \widehat{\mu}_{\mathbb{P}_j} \right\rangle_{\mathcal{H}} \right)_{i,j \in [\![B]\!]}$, then

$$\left\| \frac{1}{B} \left(G - \widehat{G} \right) \right\|_{Fr}^2 \leq 4 \|\phi\|_\infty^2 \overline{R}_{\mathcal{H}}(\widehat{\mu}_\bullet).$$

Results on toy data for KME

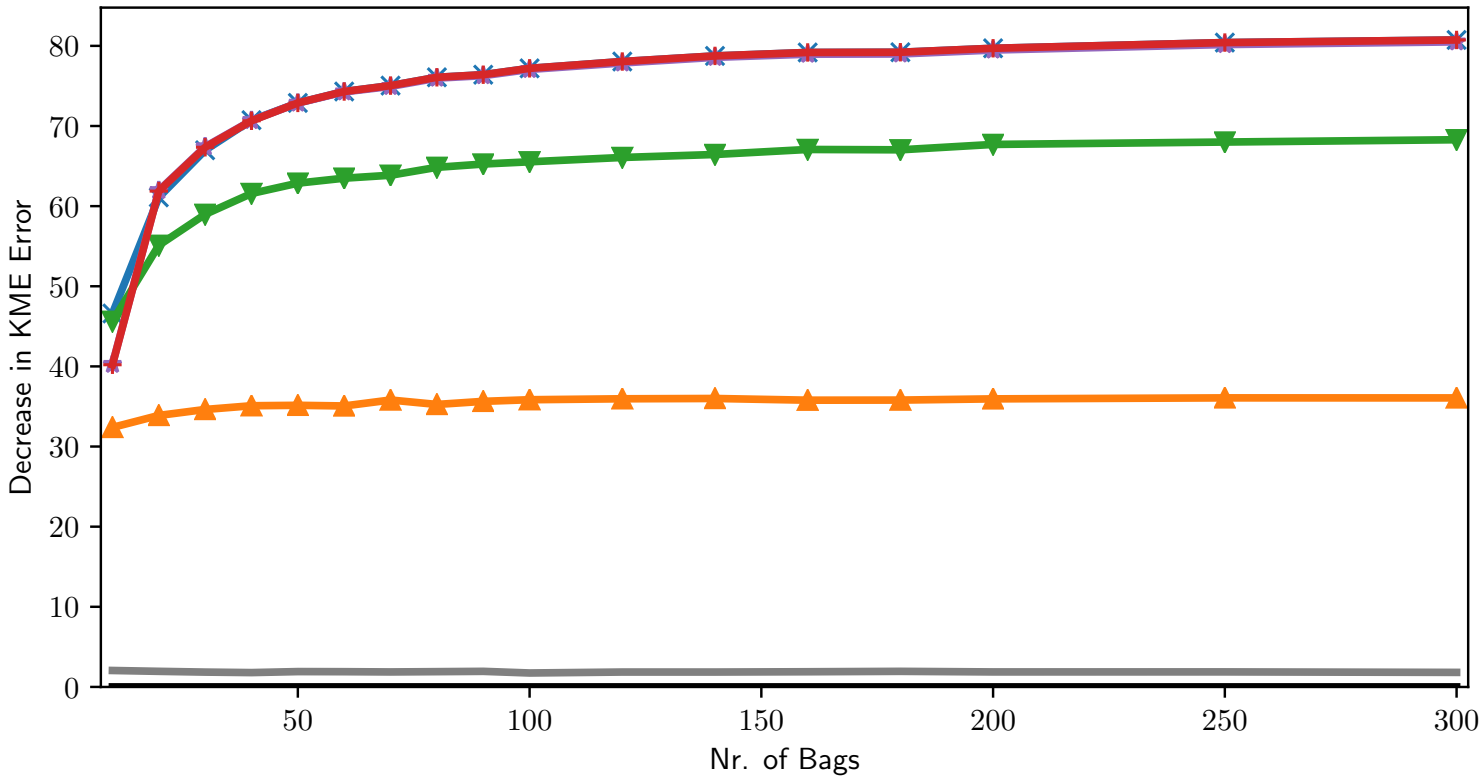


Different bagsizes

Naive
R-KMSE
STB-0
MTA const
STB theory
MTA stb
STB weight

Decrease in KME estimation error compared to `NE` in percent. Higher is better.

Results on toy data for KME

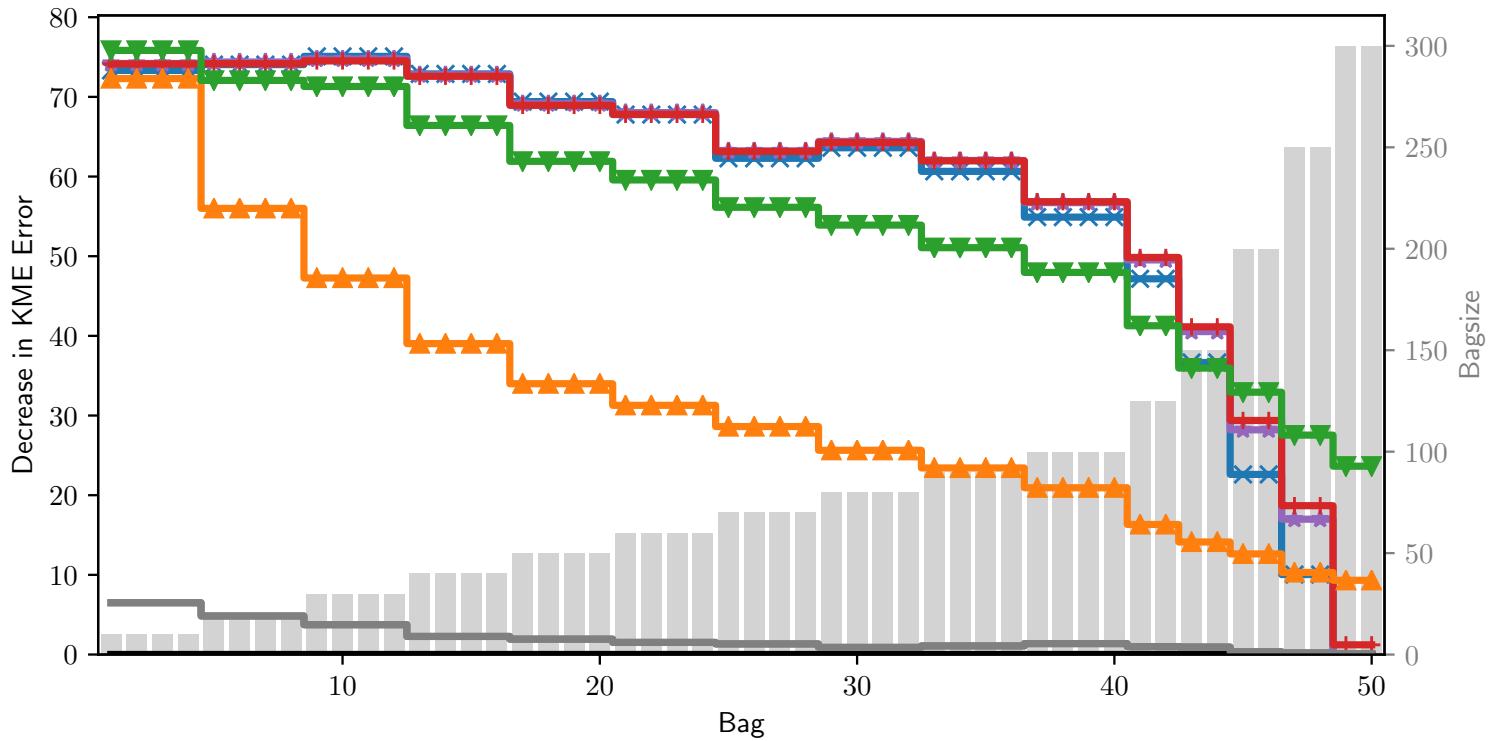


Different Number of Bags

Naive
STB-0
STB theory
STB weight
R-KMSE
MTA const
MTA stb

Decrease in KME estimation error compared to `NE` in percent. Higher is better.

Results on toy data for KME

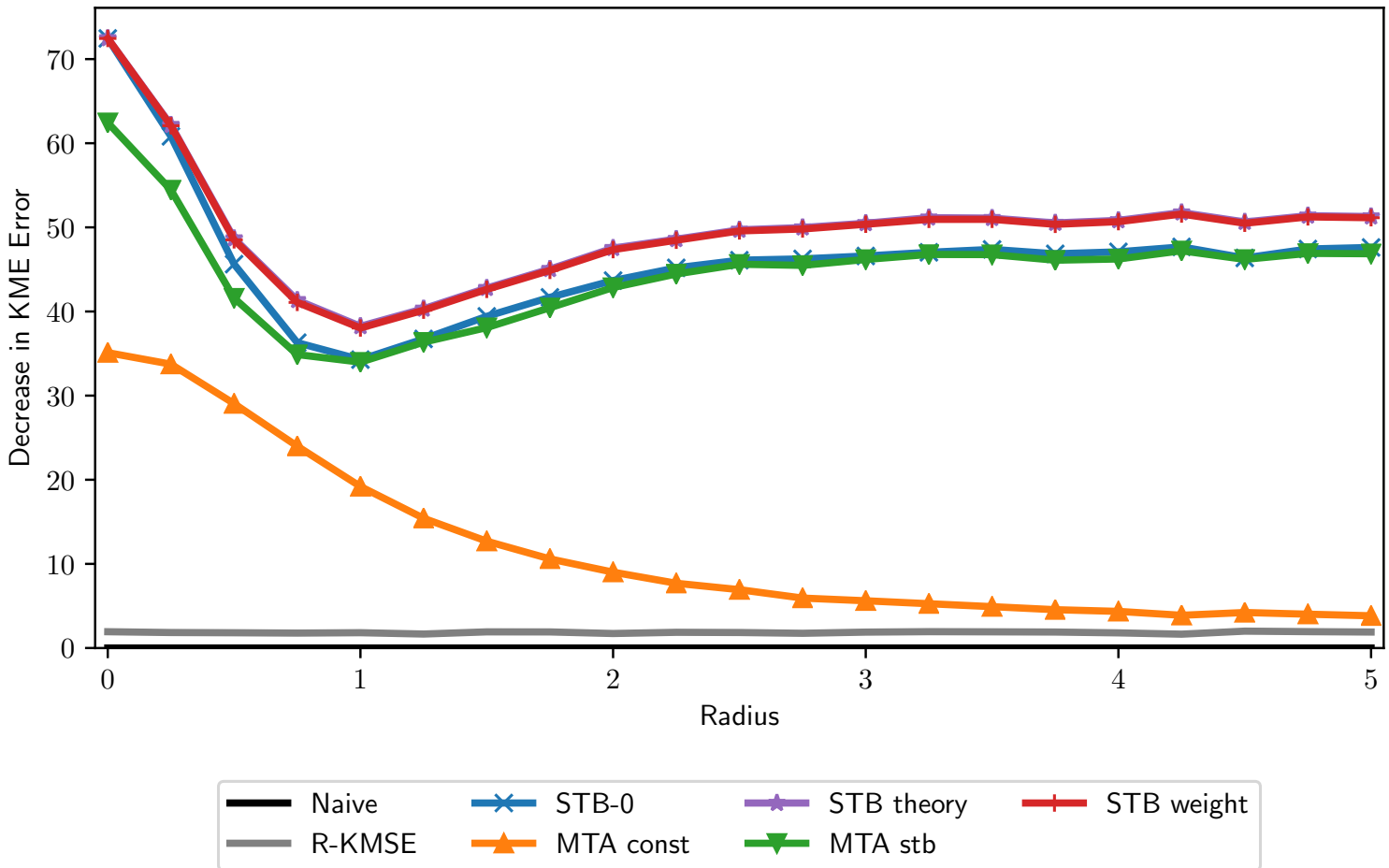


Imbalanced Bags

Naive
R-KMSE
STB-0
MTA const
STB theory
MTA stb
STB weight

Decrease in KME estimation error compared to `NE` in percent. Higher is better.

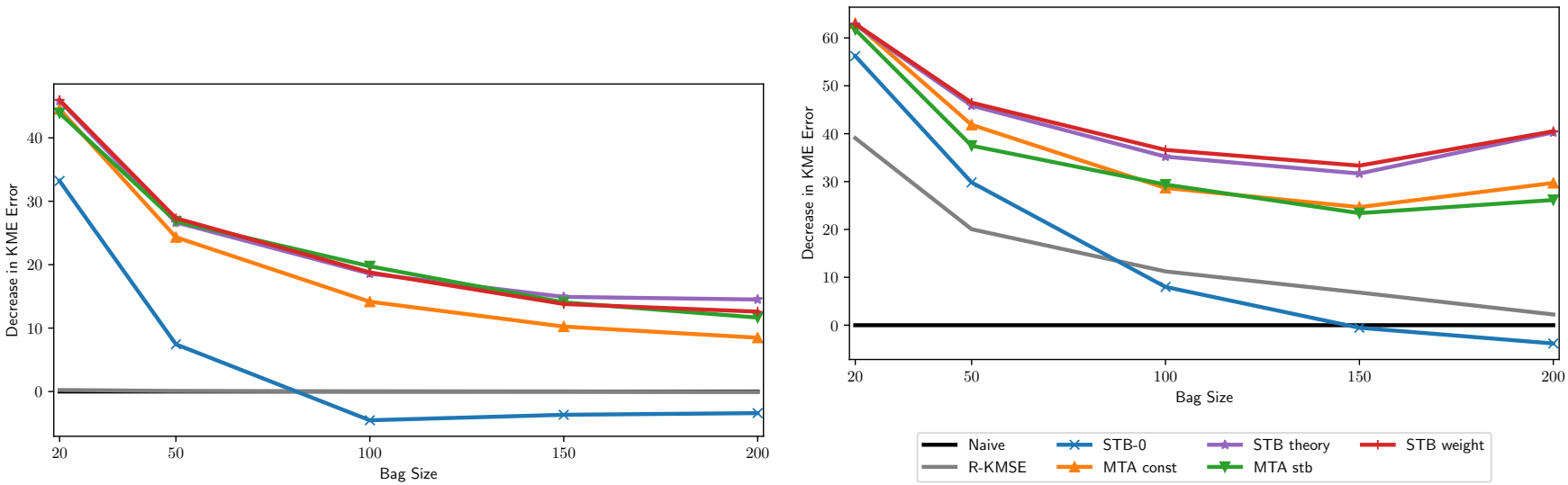
Results on toy data for KME



Clustered Bags

Decrease in KME estimation error compared to `NE` in percent. Higher is better.

Results on real data ("wine") for KME



Decrease in KME estimation error compared to `NE` in percent on the wine data set for different bag sizes and kernels. Higher is better.

Further perspective: Q-aggregation

- Aim at directly optimizing weights:

$$\widehat{\omega} := \underset{\omega \in \mathcal{S}_B}{\operatorname{Arg\,Min}} \left(\widehat{L}_1(\widehat{\mu}_\omega) + \qquad \right),$$

where

$$\widehat{L}_1(\omega) := \left\| \sum_{i=2}^{B} (\widehat{\mu}_i^{\mathrm{NE}} - \widehat{\mu}_1^{\mathrm{NE}}) \right\|^2 + (2\omega_1 - 1)\widehat{s}_1^2,$$

is an unbiased estimate of $R_1(\omega)$,

Further perspective: Q-aggregation

- Aim at directly optimizing weights:

$$\widehat{\omega} := \underset{\omega \in \mathcal{S}_B}{\operatorname{Arg\,Min}} \left(\widehat{L}_1(\widehat{\mu}_\omega) + c_0 \widehat{q}(\omega) \right),$$

where

$$\widehat{L}_1(\omega) := \left\| \sum_{i=2}^{B} (\widehat{\mu}_i^{\mathrm{NE}} - \widehat{\mu}_1^{\mathrm{NE}}) \right\|^2 + (2\omega_1 - 1)\widehat{s}_1^2,$$

is an unbiased estimate of $R_1(\omega)$, and

$$\widehat{q}(\omega) := \frac{1}{\sqrt{N_1}} \sum_{i=2}^{B} \omega_i \widehat{q}_i,$$

where

$$\widehat{q}_i^2 := \frac{1}{N_1 - 1} \sum_{k=1}^{N_1} \left\langle \widehat{\mu}_1^{\mathrm{NE}} - \widehat{\mu}_j^{\mathrm{NE}}, X_k^{(i)} - \widehat{\mu}_1^{\mathrm{NE}} \right\rangle^2.$$

is an unbiased estimate of $(\widehat{\mu}_i^{\mathrm{NE}} - \mu_1)^T \Sigma_1 (\widehat{\mu}_i^{\mathrm{NE}} - \mu_1)$ conditional to all samples except $k = 1$.

Further perspectives

- Oracle-type inequality:

$$R(\widehat{\mu}_\omega) \leq (1+\gamma_1) \min_{\omega \in \mathcal{S}_B} \left(\left\| \sum_{i=1}^{B} \omega_i \mu_i - \mu_1 \right\|^2 + \sum_{i=1}^{B} \omega_i^2 s_i^2 + \gamma_2 \frac{s_1}{\sqrt{d_1^{\text{eff}}}} \sum_{i=2}^{B} \omega_i \|\mu_1 - \mu_i\| \right) + \gamma_3 \max_{i \in [\![B]\!]} \frac{s_i^2}{\sqrt{d_i^\bullet}}$$

  (where γ_1 can be made small and γ_2, γ_3 involve a $\sqrt{\log B}$ factor)

- In the case of homogenous tasks (equal covariances and samples sizes):

$$\frac{R(\widehat{\mu}_\omega)}{s_1^2} \leq (1+\gamma_1) \min_{\tau} \left(B(\tau, V_\tau) + \gamma_2 \frac{\tau}{\sqrt{d_1^{\text{eff}}}} \right) + \frac{\gamma_3}{\sqrt{d_1^\bullet}}$$

Some take-home messages

- One can take advantage of neighbor means to improve over naive and Stein estimates for single means
- Nonasymptotic guarantees for improvement of estimation of each single mean
- Key is the high-dimensional "blessing" that testing separation is faster than estimation
- Can adapt to unknown structure of true means
  (e.g. measured via covering numbers at appropriate scale)
- Improvement capped at $\tau \asymp \sqrt{\log B / d^\bullet}$ for this effect
- Possible also to choose τ adaptively (based on tradeoff bias/nb of neighbors)
- More general approach being developed: optimize directly a suitable upper bound of the risk for convex combination weights
- Open question: more fine-grained minimax rates (so far: only $d \to \infty$, $\tau \asymp \text{cst}$.)

Thank you for your attention

H. Marienwald, J-B. Fermanian, G. Blanchard. High-Dimensional Multi-Task Averaging and Application to Kernel Mean Embedding. AISTATS 2021

G. Blanchard, J-B. Fermanian. Nonasymptotic one-and two-sample tests in high dimension with unknown covariance structure. (ArXiv/ To appear in: Festschrift in the honor of V. Spokoiny)