Stein unbiased risk estimators for tuning hyperparameters of distributed regression algorithms

Damiano Varagnolo, ITK

Gianluigi Pillonetto

Luca Schenato

Purposes of this seminar

1. discuss about a useful tool
2. connect with you

Roadmap

- Stein's lemma
- Stein's unbiased risk estimator (SURE)
- Model selection through SURE
  (will follow "Stein's Unbiased Risk Estimate, Statistical Machine Learning 2015, Tibshirani & Wasserman")
- RKHS-based regression
- average-consensus algorithms
- SURE in our distributed regression context

Stein's lemma

Stein's univariate lemma

If:

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{1}$$

$$f : \mathbb{R} \mapsto \mathbb{R} \text{ absolutely continuous} \tag{2}$$

$$f' \text{ exists and is s.t. } \mathbb{E}\left[|f'(X)|\right] < +\infty \tag{3}$$

Stein's univariate lemma

If:

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{1}$$

$$f : \mathbb{R} \mapsto \mathbb{R} \text{ absolutely continuous} \tag{2}$$

$$f' \text{ exists and is s.t. } \mathbb{E}\left[|f'(X)|\right] < +\infty \tag{3}$$

then:

$$\mathbb{E}\left[(X-\mu)\, f(X)\right] = \sigma^2 \mathbb{E}\left[f'(X)\right] \tag{4}$$

Implications

(with $\sigma^2 = 1$ for notational simplicity)

$$\mathbb{E}\left[(X - \mu) f(X)\right] = \operatorname{cov}\left(X, f(X)\right) = \mathbb{E}\left[f'(X)\right] \tag{5}$$

Implications

(with $\sigma^2 = 1$ for notational simplicity)

$$\mathbb{E}\left[(X-\mu)\, f(X)\right] = \operatorname{cov}\left(X, f(X)\right) = \mathbb{E}\left[f'(X)\right] \tag{5}$$

- if f is an estimator of μ, then estimating $\operatorname{cov}(X, f(X))$ through $\mathbb{E}\left[(X-\mu)\, f(X)\right]$ requires knowing $\mu \implies$ not a feasible path!

Implications

(with $\sigma^2 = 1$ for notational simplicity)

$$\mathbb{E}\left[(X-\mu)\, f(X)\right] = \operatorname{cov}\left(X, f(X)\right) = \mathbb{E}\left[f'(X)\right] \tag{5}$$

- if f is an estimator of μ, then estimating $\operatorname{cov}\left(X, f(X)\right)$ through $\mathbb{E}\left[(X-\mu)\, f(X)\right]$ requires knowing $\mu \implies$ not a feasible path!
- alternative strategy: estimate $\operatorname{cov}\left(X, f(X)\right)$ through computing $\widehat{\mathbb{E}}\left[f'(X)\right]$

Stein's multivariate lemma

If:

$$X \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \tag{6}$$

$$\begin{aligned} &f : \mathbb{R}^n \mapsto \mathbb{R} \textit{ almost differentiable}\text{, i.e.:} \\ &f\left(\cdot, x_{-i}\right) : \mathbb{R} \mapsto \mathbb{R} \text{ absolutely continuous for a.e. } x_i \in \mathbb{R}^{n-1} \text{ and } i = 1, \ldots, n \end{aligned} \tag{7}$$

$$\frac{\partial f}{\partial x_i} \text{ exists and is s.t. } \mathbb{E}\left[\left|\frac{\partial f}{\partial x_i}(X)\right|\right] < +\infty \text{ for each } i = 1, \ldots, n \tag{8}$$

Stein's multivariate lemma

If:

$$X \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \tag{6}$$

$$\begin{aligned} &f : \mathbb{R}^n \mapsto \mathbb{R} \text{ *almost differentiable*, i.e.:} \\ &f\left(\cdot, x_{-i}\right) : \mathbb{R} \mapsto \mathbb{R} \text{ absolutely continuous for a.e. } x_i \in \mathbb{R}^{n-1} \text{ and } i = 1, \ldots, n \end{aligned} \tag{7}$$

$$\frac{\partial f}{\partial x_i} \text{ exists and is s.t. } \mathbb{E}\left[\left|\frac{\partial f}{\partial x_i}(X)\right|\right] < +\infty \text{ for each } i = 1, \ldots, n \tag{8}$$

then:

$$\mathbb{E}\left[(X - \mu) f(X)\right] = \sigma^2 \mathbb{E}\left[\nabla f(X)\right] \tag{9}$$

Stein's multivariate lemma (2)

$$X \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad f : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad f = [f_1, \ldots, f_n] \tag{10}$$

$$\implies \quad \mathbb{E}\left[(X - \mu)\, f_i(X)\right] = \sigma^2 \mathbb{E}\left[\nabla f_i(X)\right] \tag{11}$$

Stein's multivariate lemma (2)

$$X \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad f : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad f = [f_1, \ldots, f_n] \tag{10}$$

$$\implies \quad \mathbb{E}\left[(X - \mu)\, f_i(X)\right] = \sigma^2 \mathbb{E}\left[\nabla f_i(X)\right] \tag{11}$$

$$\implies \quad \sum_{i=1}^{n} \operatorname{cov}\left(X_i, f_i(X)\right) = \sigma^2 \mathbb{E}\left[\sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i}(X)\right] \tag{12}$$

From Stein's multivariate lemma to Stein's unbiased risk estimate (SURE)

$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\left(y\right) = \text{estimate of } \mu \text{ at } y \tag{13}$$

From Stein's multivariate lemma
to Stein's unbiased risk estimate (SURE)

$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\,(y) = \text{estimate of } \mu \text{ at } y \tag{13}$$

$$R \quad = \mathbb{E}\left[\,\|\mu - \widehat{\mu}\|_2^2\right]$$

From Stein's multivariate lemma to Stein's unbiased risk estimate (SURE)

$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\left(y\right) = \text{estimate of } \mu \text{ at } y \tag{13}$$

$$\begin{aligned} R &= \mathbb{E}\left[\|\mu - \widehat{\mu}\|_2^2\right] \\ &= \mathbb{E}\left[\|\mu - y + y - \widehat{\mu}\|_2^2\right] \end{aligned}$$

From Stein's multivariate lemma to Stein's unbiased risk estimate (SURE)

$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\left(y\right) = \text{estimate of } \mu \text{ at } y \tag{13}$$

$$\begin{aligned} R &= \mathbb{E}\left[\|\mu - \widehat{\mu}\|_2^2\right] \\ &= \mathbb{E}\left[\|\mu - y + y - \widehat{\mu}\|_2^2\right] \\ &= \mathbb{E}\left[\|\mu - y\|_2^2\right] + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\mathbb{E}\left[\left(\mu - y\right)^T\left(y - \widehat{\mu}\right)\right] \end{aligned}$$

From Stein's multivariate lemma to Stein's unbiased risk estimate (SURE)

$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\left(y\right) = \text{estimate of } \mu \text{ at } y \tag{13}$$

$$\begin{aligned}
R &= \mathbb{E}\left[\|\mu - \widehat{\mu}\|_2^2\right] \\
&= \mathbb{E}\left[\|\mu - y + y - \widehat{\mu}\|_2^2\right] \\
&= \mathbb{E}\left[\|\mu - y\|_2^2\right] + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\mathbb{E}\left[(\mu - y)^T (y - \widehat{\mu})\right] \\
&= n\sigma^2 + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\mathbb{E}\left[(\mu - y)^T (y - \widehat{\mu})\right]
\end{aligned}$$

From Stein's multivariate lemma to Stein's unbiased risk estimate (SURE)

$$y \sim \mathcal{N}\left(\mu, \sigma^2 I\right) \qquad \widehat{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \widehat{\mu}\,(y) = \text{estimate of } \mu \text{ at } y \tag{13}$$

$$\begin{aligned}
R &= \mathbb{E}\left[\|\mu - \widehat{\mu}\|_2^2\right] \\
&= \mathbb{E}\left[\|\mu - y + y - \widehat{\mu}\|_2^2\right] \\
&= \mathbb{E}\left[\|\mu - y\|_2^2\right] + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\mathbb{E}\left[(\mu - y)^T (y - \widehat{\mu})\right] \\
&= n\sigma^2 + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\mathbb{E}\left[(\mu - y)^T (y - \widehat{\mu})\right] \\
&= -n\sigma^2 + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\sum_{i=1}^{n} \operatorname{cov}\left(y_i, \widehat{\mu}_i\right)
\end{aligned} \tag{14}$$

From Stein's multivariate lemma to Stein's unbiased risk estimate (SURE)

$$R = \mathbb{E}\left[\|\mu - \widehat{\mu}\|_2^2\right] = -n\sigma^2 + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\sum_{i=1}^{n} \operatorname{cov}\left(y_i, \widehat{\mu}_i\right)$$

From Stein's multivariate lemma to Stein's unbiased risk estimate (SURE)

$$
\begin{aligned}
R = \mathbb{E}\left[\|\mu - \widehat{\mu}\|_2^2\right] &= -n\sigma^2 + \mathbb{E}\left[\|y - \widehat{\mu}\|_2^2\right] + 2\sum_{i=1}^{n} \operatorname{cov}\left(y_i, \widehat{\mu}_i\right) \\
\implies \qquad \widehat{R} &= -n\sigma^2 + \|y - \widehat{\mu}\|_2^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \widehat{\mu}_i}{\partial y_y}(y)
\end{aligned} \tag{15}
$$

$$
\text{with } \mathbb{E}\left[\widehat{R}\right] = R \tag{16}
$$

Model selection through SURE in general

Model selection through SURE in general

$$\widehat{\mu} \mapsto \widehat{\mu}_\lambda \tag{17}$$

Model selection through SURE in general

$$\widehat{\mu} \mapsto \widehat{\mu}_\lambda \tag{17}$$

$$\widehat{R}_\lambda = -n\sigma^2 + \|y - \widehat{\mu}_\lambda\|_2^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \widehat{\mu}_{\lambda,i}}{\partial y_i}(y) \tag{18}$$

Model selection through SURE in general

$$\widehat{\mu} \mapsto \widehat{\mu}_\lambda \tag{17}$$

$$\widehat{R}_\lambda = -n\sigma^2 + \|y - \widehat{\mu}_\lambda\|_2^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \widehat{\mu}_{\lambda,i}}{\partial y_i}(y) \tag{18}$$

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \|y - \widehat{\mu}_\lambda\|_2^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \widehat{\mu}_{\lambda,i}}{\partial y_i}(y) \tag{19}$$

requires:

- to verify that $\widehat{\mu}_\lambda$ is almost differentiable

Model selection through SURE in general

$$\widehat{\mu} \mapsto \widehat{\mu}_\lambda \tag{17}$$

$$\widehat{R}_\lambda = -n\sigma^2 + \|y - \widehat{\mu}_\lambda\|_2^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \widehat{\mu}_{\lambda,i}}{\partial y_i}(y) \tag{18}$$

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \|y - \widehat{\mu}_\lambda\|_2^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \widehat{\mu}_{\lambda,i}}{\partial y_i}(y) \tag{19}$$

requires:

- to verify that $\widehat{\mu}_\lambda$ is almost differentiable
- to compute the *divergence* of $\widehat{\mu}_\lambda$, i.e., $\dfrac{\partial \widehat{\mu}_{\lambda,i}}{\partial y_i}(y)$

Model selection through SURE: the linear case

If:

$$y = \mu + e \qquad \mu \text{ deterministic} \tag{20}$$

$$y^* = \mu + e^* \text{ future measurements on the same input locations} \tag{21}$$

$$e, e^* \text{ uncorrelated, zero mean, with covariance } \Sigma \tag{22}$$

$$\widehat{y} = Sy \text{ linear estimator of } y^* \tag{23}$$

Model selection through SURE: the linear case

If:

$$y = \mu + e \qquad \mu \text{ deterministic} \tag{20}$$

$$y^* = \mu + e^* \text{ future measurements on the same input locations} \tag{21}$$

$$e, e^* \text{ uncorrelated, zero mean, with covariance } \Sigma \tag{22}$$

$$\widehat{y} = Sy \text{ linear estimator of } y^* \tag{23}$$

Then:

$$\|y - \widehat{y}\|^2 + 2\mathrm{tr}\left(S\Sigma\right) \text{ is an unbiased estimator of the risk } \mathbb{E}\left[\|y^* - \widehat{y}\|_2^2\right] \tag{24}$$

Model selection through SURE: examples of literature

literature on other cases:

- Li (1985), *From Stein's unbiased risk estimates to the method of generalized cross-validation*, Annals of Statistics
- Li (1986), *Asymptotic optimality of c_l and generalized cross-validation in ridge regression with application to spline smoothing*, Annals of Statistics
- Johnstone (1986), *On imadmissibility of some unbiased estimates of loss*, technical report
- Kneip (1994), *Ordered linear smoothers*, Annals of Statistics
- Donoho & Johnstone (1995), *Adapting to unknown smoothness via wavelet shrinkage*, Journal of the American Statistical Association
- Efron et al. (2004), *Least angle regression*, Annals of Statistics
- Zou et al. (2007), *On the degrees of freedom of the lasso*, Annals of Statistics
- Tibshirani & Taylor (2011), *The solution path of the generalized lasso*, Annals of Statistics
- Tibshirani & Taylor (2012), *Degrees of freedom in lasso problems*, Annals of Statistics

Model selection through SURE for a specific case

The original practical problem: function estimation

$$f : \mathcal{X} \to \mathbb{R} \tag{25}$$

$$y_m = f\left(x_m\right) + \nu_m \qquad m = 1, \ldots, M \tag{26}$$

$$x_m \sim \mu(\mathcal{X}) \;\; \text{i.i.d.} \qquad \nu_m \sim \mathcal{N}\left(0, \sigma_\nu^2\right) \qquad m = 1, \ldots, M \tag{27}$$

$$\{x_m\}_{m=1}^{M} \qquad \{\nu_m\}_{m=1}^{M} \qquad \text{mutually independent} \tag{28}$$

The original practical problem: function estimation

$$f : \mathcal{X} \to \mathbb{R} \tag{25}$$

$$y_m = f\left(x_m\right) + \nu_m \qquad m = 1, \ldots, M \tag{26}$$

$$x_m \sim \mu(\mathcal{X}) \;\; \text{i.i.d.} \qquad \nu_m \sim \mathcal{N}\left(0, \sigma_\nu^2\right) \qquad m = 1, \ldots, M \tag{27}$$

$$\{x_m\}_{m=1}^{M} \qquad \{\nu_m\}_{m=1}^{M} \qquad \text{mutually independent} \tag{28}$$

Problem: estimate f starting from $\{x_m, y_m\}$

Nonparametric approach *(cast the problem as a Gaussian regression)*

$$f \sim \mathcal{N}(0, K) \qquad K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad \text{so that} \quad \mathbb{E}\left[f(x)f(x')\right] = K\left(x, x'\right) \tag{29}$$

Nonparametric approach *(cast the problem as a Gaussian regression)*

$$f \sim \mathcal{N}(0, K) \qquad K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad \text{so that} \quad \mathbb{E}\left[f(x) f(x')\right] = K\left(x, x'\right) \tag{29}$$

Examples:

- Brownian motion: $K\left(x, x'\right) = \min\left(x, x'\right) \qquad \mathcal{X} = [0, 1]$
- Radial basis: $K\left(x, x'\right) = \exp\left(-\left\|x - x'\right\|^2\right) \qquad \mathcal{X} \subseteq \mathbb{R}^m$

Maximum a posteriori estimator

$$\widehat{f}_{\mathrm{MAP}}(x) = \mathbb{E}\left[f(x) \,\middle|\, \{x_m, y_m\} \right] \qquad \textit{(also MV)}$$

Maximum a posteriori estimator

$$
\begin{aligned}
\widehat{f}_{\mathrm{MAP}}(x) &= \mathbb{E}\Big[f(x) \;\Big|\; \{x_m, y_m\} \Big] && \textit{(also MV)} \\
&= \sum_{m=1}^{M} K(x, x_m) c_m && \textit{(a.k.a. regularization network)}
\end{aligned}
\tag{30}
$$

Maximum a posteriori estimator

$$
\begin{aligned}
\widehat{f}_{\mathrm{MAP}}(x) &= \mathbb{E}\Big[f(x) \;\Big|\; \{x_m, y_m\} \Big] && \textit{(also MV)} \\
&= \sum_{m=1}^{M} K(x, x_m) c_m && \textit{(a.k.a. regularization network)}
\end{aligned}
\tag{30}
$$

$$
\begin{bmatrix} c_1 \\ \vdots \\ c_M \end{bmatrix} = H_{\mathrm{MAP}} \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}
\tag{31}
$$

$$
H_{\mathrm{MAP}} := \left(\begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_M) \\ \vdots & & \vdots \\ K(x_M, x_1) & \cdots & K(x_M, x_M) \end{bmatrix} + \sigma_\nu^2 I \right)^{-1}
\tag{32}
$$

Gaussian regression – practical issues associated to the MAP

$$\widehat{f}_{\text{MAP}}(x) = \begin{bmatrix} K(x, x_1) & \dots & K(x, x_M) \end{bmatrix} H_{\text{MAP}} \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} \tag{33}$$

$$H_{\text{MAP}} := \left(\begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_M) \\ \vdots & & \vdots \\ K(x_M, x_1) & \cdots & K(x_M, x_M) \end{bmatrix} + \sigma_\nu^2 I \right)^{-1} \tag{34}$$

computational cost $O\left(M^3\right)$

How may we tackle the $O\left(M^3\right)$ computational cost issue?

$$H_{\text{MAP}} := \left(\begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_M) \\ \vdots & & \vdots \\ K(x_M, x_1) & \cdots & K(x_M, x_M) \end{bmatrix} + \sigma_\nu^2 I \right)^{-1} \tag{35}$$

Typical approaches: low-rank / sparsification approximations

Smola & Schölkopf (2000)
Sparse greedy matrix approximations for machine learning

Quiñonero-Candela & Rasmussen (2005)
A unifying view of sparse approximate Gaussian process regression

Bach & Jordan (2005)
Predictive low-rank decompositions for kernel methods

Snelson & Ghahramani (2006)
Sparse Gaussian processes using pseudo inputs

Culis et al. (2006)
Learning low-rank kernel matrices

Zhang & Kwok (2010)
Clustered Nyström method for large scale manifold learning and dimension reduction

Ambikasaran et al. (2016)
Fast direct methods for Gaussian processes

Our approach: Karhunen-Loève expansions

$$f(x) = \sum_{e=1}^{+\infty} \alpha_e \phi_e(x)$$

Our approach: Karhunen-Loève expansions

$$f(x) = \sum_{e=1}^{+\infty} \alpha_e \phi_e(x) = \underbrace{\sum_{e=1}^{E} a_e \phi_e(x)}_{=:\text{interesting}} + \underbrace{\sum_{e=1}^{+\infty} b_e \phi_{E+e}(x)}_{=:\text{remainder}} \tag{36}$$

Our approach: Karhunen-Loève expansions

$$f(x) = \sum_{e=1}^{+\infty} \alpha_e \phi_e(x) = \underbrace{\sum_{e=1}^{E} a_e \phi_e(x)}_{=:\text{interesting}} + \underbrace{\sum_{e=1}^{+\infty} b_e \phi_{E+e}(x)}_{=:\text{remainder}} \tag{36}$$

$$\lambda_e \phi_e(x) = \int_{\mathcal{X}} K(x, x') \phi_e(x') d\mu(x') \qquad \lambda_1 \geq \lambda_2 \ldots > 0 \tag{37}$$

$$K(x, x') = \sum_{e=1}^{+\infty} \lambda_e \phi_e(x) \phi_e(x') \qquad \int_{\mathcal{X}} \phi_i(x) \phi_j(x) d\mu(x) = \delta_{ij} \tag{38}$$

Our approach: Karhunen-Loève expansions

$$f(x) = \sum_{e=1}^{+\infty} \alpha_e \phi_e(x) = \underbrace{\sum_{e=1}^{E} a_e \phi_e(x)}_{=:\text{interesting}} + \underbrace{\sum_{e=1}^{+\infty} b_e \phi_{E+e}(x)}_{=:\text{remainder}} \tag{36}$$

$$\lambda_e \phi_e(x) = \int_{\mathcal{X}} K(x, x') \phi_e(x') d\mu(x') \qquad \lambda_1 \geq \lambda_2 \ldots > 0 \tag{37}$$

$$K(x, x') = \sum_{e=1}^{+\infty} \lambda_e \phi_e(x) \phi_e(x') \qquad \int_{\mathcal{X}} \phi_i(x) \phi_j(x) d\mu(x) = \delta_{ij} \tag{38}$$

$$a_e \sim \mathcal{N}(0, \lambda_e), \ e = 1, \ldots, E \qquad b_e \sim \mathcal{N}(0, \lambda_{E+e}), \ e = 1, 2, \ldots \tag{39}$$

Zhu et al. (1998)
Gaussian regression and optimal finite dimensional linear models

Our approach: Karhunen-Loève expansions

$$f(x) = \underbrace{\sum_{e=1}^{E} a_e \phi_e(x)}_{=:\text{interesting}} + \underbrace{\sum_{e=1}^{+\infty} b_e \phi_{E+e}(x)}_{=:\text{remainder}} \qquad \lambda_e \phi_e(x) = \int_{\mathcal{X}} K(x, x') \phi_e(x') d\mu(x') \tag{40}$$

$\Longrightarrow$ first E ϕ_e's $=$ best a-priori E-dimensional approximation in a MSE sense

Our approach: Karhunen-Loève expansions

$$\boldsymbol{y} \coloneqq [y_1, \ldots, y_M]^T \quad \boldsymbol{\nu} \coloneqq [\nu_1, \ldots, \nu_M]^T \quad \boldsymbol{a} \coloneqq [a_1, \ldots, a_E]^T \quad \boldsymbol{b} \coloneqq [b_1, b_2, \ldots]^T \quad (41)$$

Our approach: Karhunen-Loève expansions

$$\boldsymbol{y} := [y_1, \ldots, y_M]^T \quad \boldsymbol{\nu} := [\nu_1, \ldots, \nu_M]^T \quad \boldsymbol{a} := [a_1, \ldots, a_E]^T \quad \boldsymbol{b} := [b_1, b_2, \ldots]^T \tag{41}$$

$$G := \begin{bmatrix} \phi_1(x_1) & \ldots & \phi_E(x_1) \\ \vdots & & \vdots \\ \phi_1(x_M) & \ldots & \phi_E(x_M) \end{bmatrix} \qquad Z := \begin{bmatrix} \phi_{E+1}(x_1) & \phi_{E+2}(x_1) & \ldots \\ \vdots & & \vdots \\ \phi_{E+1}(x_M) & \phi_{E+2}(x_M) & \ldots \end{bmatrix} \tag{42}$$

Our approach: Karhunen-Loève expansions

$$\boldsymbol{y} := [y_1, \dots, y_M]^T \quad \boldsymbol{\nu} := [\nu_1, \dots, \nu_M]^T \quad \boldsymbol{a} := [a_1, \dots, a_E]^T \quad \boldsymbol{b} := [b_1, b_2, \dots]^T \tag{41}$$

$$G := \begin{bmatrix} \phi_1(x_1) & \dots & \phi_E(x_1) \\ \vdots & & \vdots \\ \phi_1(x_M) & \dots & \phi_E(x_M) \end{bmatrix} \qquad Z := \begin{bmatrix} \phi_{E+1}(x_1) & \phi_{E+2}(x_1) & \dots \\ \vdots & & \vdots \\ \phi_{E+1}(x_M) & \phi_{E+2}(x_M) & \dots \end{bmatrix} \tag{42}$$

$$\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu} \tag{43}$$

Our approach: Karhunen-Loève expansions

$$\boldsymbol{y} := [y_1, \ldots, y_M]^T \quad \boldsymbol{\nu} := [\nu_1, \ldots, \nu_M]^T \quad \boldsymbol{a} := [a_1, \ldots, a_E]^T \quad \boldsymbol{b} := [b_1, b_2, \ldots]^T \tag{41}$$

$$G := \begin{bmatrix} \phi_1(x_1) & \ldots & \phi_E(x_1) \\ \vdots & & \vdots \\ \phi_1(x_M) & \ldots & \phi_E(x_M) \end{bmatrix} \qquad Z := \begin{bmatrix} \phi_{E+1}(x_1) & \phi_{E+2}(x_1) & \ldots \\ \vdots & & \vdots \\ \phi_{E+1}(x_M) & \phi_{E+2}(x_M) & \ldots \end{bmatrix} \tag{42}$$

$$\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu} \tag{43}$$

$$\widehat{f}_E(x) := \begin{bmatrix} \phi_1(x) & \cdots & \phi_E(x) \end{bmatrix} \widehat{\boldsymbol{a}} \qquad \widehat{\boldsymbol{a}} = H\boldsymbol{y} \qquad H := \left(\frac{G^T G}{M} + \frac{\sigma_\nu^2}{M} \Lambda_E^{-1} \right)^{-1} \frac{G^T}{M} \tag{44}$$

Summary

$$\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu} \tag{45}$$

$$\widehat{f}_E(x) := \begin{bmatrix} \phi_1(x) & \cdots & \phi_E(x) \end{bmatrix} H \boldsymbol{y} \qquad H := \left(\frac{G^T G}{M} + \frac{\sigma_\nu^2}{M} \Lambda_E^{-1} \right)^{-1} \frac{G^T}{M} \tag{46}$$

computational cost: $O(E^3)$

Summary

$$\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu} \tag{45}$$

$$\widehat{f}_E(x) := \begin{bmatrix}\phi_1(x) & \cdots & \phi_E(x)\end{bmatrix} H\boldsymbol{y} \qquad H := \left(\frac{G^T G}{M} + \frac{\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1} \frac{G^T}{M} \tag{46}$$

computational cost: $O(E^3)$

interesting for us because

$$\frac{G^T G}{M} = \frac{1}{M}\sum_{m=1}^{M} G_m^T G_m \qquad \frac{G^T \boldsymbol{y}}{M} = \frac{1}{M}\sum_{m=1}^{M} G_m^T y_m \tag{47}$$

Average consensus

(i.e., how to compute an average in a distributed fashion?)

synchronous communications

synchronous consensus: $\boldsymbol{x}(k+1) = P\boldsymbol{x}(k)$ (with P doubly stochastic) (Markov chains ('60s), Seneta 2006, . . .)

synchronous communications

synchronous consensus: $\boldsymbol{x}(k+1) = P\boldsymbol{x}(k)$ (with P doubly stochastic) (Markov chains ('60s), Seneta 2006, ...)

asynchronous communications with perfect channel feedback

ratio consensus (Bénézit et al. 2010)

synchronous communications

synchronous consensus: $\boldsymbol{x}(k+1) = P\boldsymbol{x}(k)$ (with P doubly stochastic) (Markov chains ('60s), Seneta 2006, ...)

asynchronous communications with perfect channel feedback

ratio consensus (Bénézit et al. 2010)

asynchronous communications without perfect channel feedback

robust ratio consensus (Dominguez-Garcia et al. 2011)

Ratio consensus

asynchronous communications with perfect channel feedback (Bénézit et al. 2010)

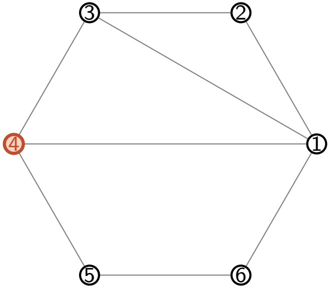
$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

Ratio consensus

asynchronous communications with perfect channel feedback (Bénézit et al. 2010)

$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

$$P(k) = \begin{bmatrix} 1 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

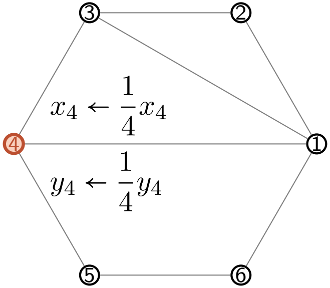


Ratio consensus

asynchronous communications with perfect channel feedback (Bénézit et al. 2010)

$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

$$P(k) = \begin{bmatrix} 1 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

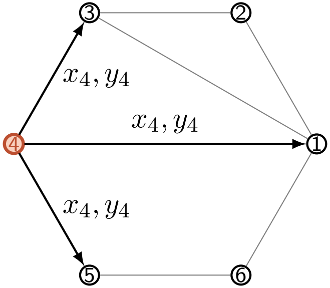


Ratio consensus

asynchronous communications with perfect channel feedback (Bénézit et al. 2010)

$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

$$P(k) = \begin{bmatrix} 1 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

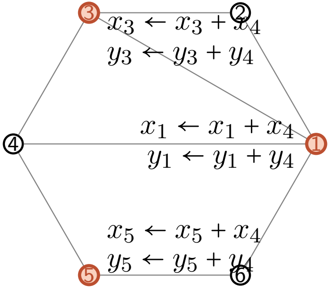


Ratio consensus

asynchronous communications with perfect channel feedback (Bénézit et al. 2010)

$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

$$P(k) = \begin{bmatrix} 1 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

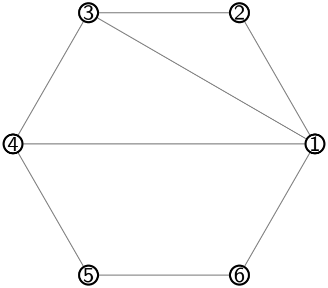


Ratio consensus

asynchronous communications with perfect channel feedback (Bénézit et al. 2010)

$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

$$P(k) = \begin{bmatrix} 1 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$



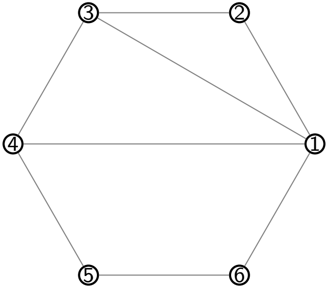
$$\begin{cases} x_i(k) \to \beta_i(k) \sum_j x_i(0) \\ \\ y_i(k) \to \beta_i(k) \sum_j y_i(0) \end{cases}$$

Ratio consensus

asynchronous communications with perfect channel feedback (Bénézit et al. 2010)

$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

$$P(k) = \begin{bmatrix} 1 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$



$$\begin{cases} x_i(k) \to \beta_i(k) \sum_j x_i(0) \\ \\ y_i(k) \to \beta_i(k) \sum_j y_i(0) \end{cases} \implies z_i(k) := \frac{x_i(k)}{y_i(k)} \to \frac{\sum_i x_i(0)}{\sum_i y_i(0)} = \frac{1}{N} \sum_i \theta_i$$

Robust ratio consensus

asynchronous commumincation ***without*** perfect channel feedback (Dominguez-Garcia et al. 2011)

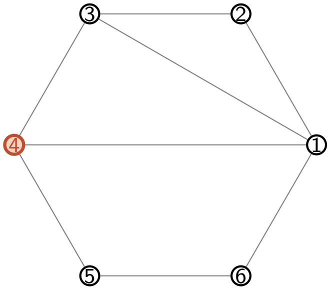
$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

Robust ratio consensus

asynchronous commumincation ***without*** perfect channel feedback (Dominguez-Garcia et al. 2011)

$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

$$P(k) = \begin{bmatrix} 1 & 0 & 0 & \mathbf{0} & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

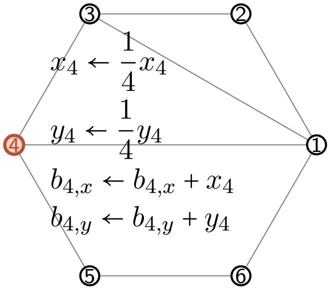


Robust ratio consensus

asynchronous commumincation ***without*** perfect channel feedback (Dominguez-Garcia et al. 2011)

$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

$$P(k) = \begin{bmatrix} 1 & 0 & 0 & \color{red}{0} & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$



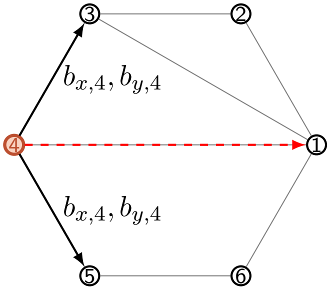
- $b_{i,x}$: total cumulative mass of x_i
- $\beta_{i,x}^{(j)}$: j's local estimate of $b_{i,x}$

Robust ratio consensus

asynchronous commumincation ***without*** perfect channel feedback (Dominguez-Garcia et al. 2011)

$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

$$P(k) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$



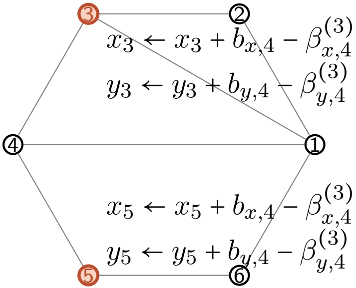
- $b_{i,x}$: total cumulative mass of x_i
- $\beta_{i,x}^{(j)}$: j's local estimate of $b_{i,x}$

Robust ratio consensus

asynchronous commumincation ***without*** perfect channel feedback (Dominguez-Garcia et al. 2011)

$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

$$P(k) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$



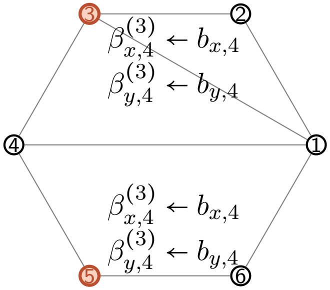
- $b_{i,x}$: total cumulative mass of x_i
- $\beta^{(j)}_{i,x}$: j's local estimate of $b_{i,x}$

Robust ratio consensus

asynchronous commumincation ***without*** perfect channel feedback (Dominguez-Garcia et al. 2011)

$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$

$$P(k) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

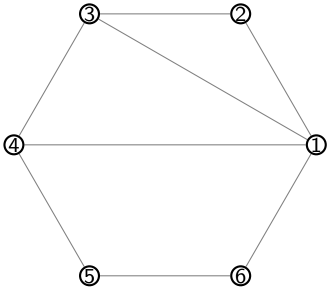


- $b_{i,x}$: total cumulative mass of x_i
- $\beta_{i,x}^{(j)}$: j's local estimate of $b_{i,x}$

Robust ratio consensus

asynchronous commumincation ***without*** perfect channel feedback (Dominguez-Garcia et al. 2011)

$$\begin{cases} \boldsymbol{x}(k+1) = P(k)\boldsymbol{x}(k) \\ x_i(0) = \theta_i \\ \boldsymbol{y}(k+1) = P(k)\boldsymbol{y}(k) \\ y_i(0) = 1 \end{cases}$$



$$P(k) = \begin{bmatrix} 1 & 0 & 0 & \mathbf{0} & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$z_i(k) = \frac{x_i(k)}{y_i(k)} \to \frac{1}{N} \sum_j \theta_i$$

- $b_{i,x}$: total cumulative mass of x_i
- $\beta_{i,x}^{(j)}$: j's local estimate of $b_{i,x}$

SURE in our distributed regression settings

Recap: KL + average consensus

$$\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu} \tag{48}$$

$$\widehat{f}_E(x) := \begin{bmatrix} \phi_1(x) & \cdots & \phi_E(x) \end{bmatrix} H\boldsymbol{y} \qquad H := \left(\frac{G^T G}{M} + \lambda \frac{\sigma_\nu^2}{M} \Lambda_E^{-1} \right)^{-1} \frac{G^T}{M} \tag{49}$$

$$\frac{G^T G}{M} = \frac{1}{M} \sum_{m=1}^{M} G_m^T G_m \qquad \frac{G^T \boldsymbol{y}}{M} = \frac{1}{M} \sum_{m=1}^{M} G_m^T y_m \tag{50}$$

Recap: KL + average consensus

$$\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu} \tag{48}$$

$$\widehat{f}_E(x) := \begin{bmatrix} \phi_1(x) & \cdots & \phi_E(x) \end{bmatrix} H \boldsymbol{y} \qquad H := \left(\frac{G^T G}{M} + \lambda \frac{\sigma_\nu^2}{M} \Lambda_E^{-1} \right)^{-1} \frac{G^T}{M} \tag{49}$$

$$\frac{G^T G}{M} = \frac{1}{M} \sum_{m=1}^{M} G_m^T G_m \qquad \frac{G^T \boldsymbol{y}}{M} = \frac{1}{M} \sum_{m=1}^{M} G_m^T y_m \tag{50}$$

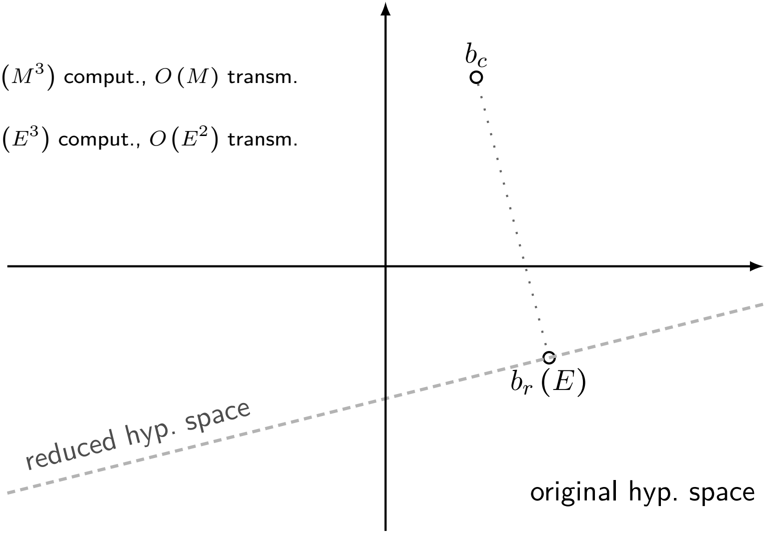
Questions:

- how shall we tune E?
- how shall we tune λ?

Tuning of E (not in this presentation)

b_c: $O\left(M^3\right)$ comput., $O\left(M\right)$ transm.

b_r: $O\left(E^3\right)$ comput., $O\left(E^2\right)$ transm.



Tuning of λ: a SURE-based approach

Recall:

$$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2 + 2\text{tr}\left(S\Sigma\right) \text{ is an unbiased estimator of the risk } \mathbb{E}\left[\|\boldsymbol{y}^* - \widehat{\boldsymbol{y}}\|_2^2\right] \text{ with } \widehat{\boldsymbol{y}} = S\boldsymbol{y} \tag{51}$$

Tuning of λ: a SURE-based approach

Recall:

$$\left\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\right\|^{2}+2\operatorname{tr}\left(S\Sigma\right) \text{ is an unbiased estimator of the risk } \mathbb{E}\left[\left\|\boldsymbol{y}^{*}-\widehat{\boldsymbol{y}}\right\|_{2}^{2}\right] \text{ with } \widehat{\boldsymbol{y}}=S\boldsymbol{y} \tag{51}$$

In our case:

$$S=\frac{G^{T}G}{M}\left(\frac{G^{T}G}{M}+\frac{\gamma\sigma_{\nu}^{2}}{M}\Lambda_{E}^{-1}\right)^{-1} \tag{52}$$

$$\Sigma=\sigma_{\nu}^{2}\frac{G^{T}G}{M^{2}}. \tag{53}$$

Tuning of λ: a SURE-based approach

Recall:

$$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2 + 2\mathrm{tr}\left(S\Sigma\right) \text{ is an unbiased estimator of the risk } \mathbb{E}\left[\|\boldsymbol{y}^* - \widehat{\boldsymbol{y}}\|_2^2\right] \text{ with } \widehat{\boldsymbol{y}} = S\boldsymbol{y} \tag{51}$$

In our case:

$$S = \frac{G^T G}{M}\left(\frac{G^T G}{M} + \frac{\gamma\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1} \tag{52}$$

$$\Sigma = \sigma_\nu^2 \frac{G^T G}{M^2}. \tag{53}$$

Important consideration: the original process was $\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu}$, but this SURE approach considers only $G\boldsymbol{a} + \boldsymbol{\nu}$!

Tuning of λ: a SURE-based approach

Recall:

$$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2 + 2\mathrm{tr}\left(S\Sigma\right) \text{ is an unbiased estimator of the risk } \mathbb{E}\left[\|\boldsymbol{y}^* - \widehat{\boldsymbol{y}}\|_2^2\right] \text{ with } \widehat{\boldsymbol{y}} = S\boldsymbol{y} \tag{51}$$

In our case:

$$S = \frac{G^T G}{M}\left(\frac{G^T G}{M} + \frac{\gamma\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1} \tag{52}$$

$$\Sigma = \sigma_\nu^2 \frac{G^T G}{M^2}. \tag{53}$$

Important consideration: the original process was $\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu}$, but this SURE approach considers only $G\boldsymbol{a} + \boldsymbol{\nu}$! However, for large M, $G^T Z\boldsymbol{b}$ vanishes $\implies$ SURE score above is an asymptotically unbiased estimator of the actual risk

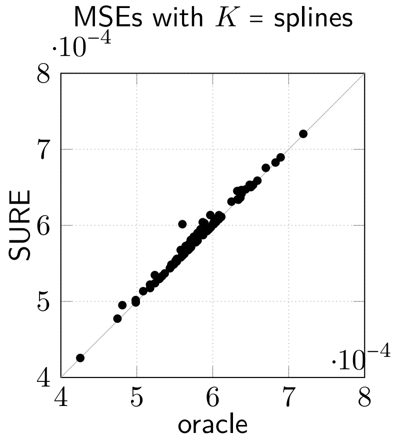
Does this work? Analysis on synthetic data

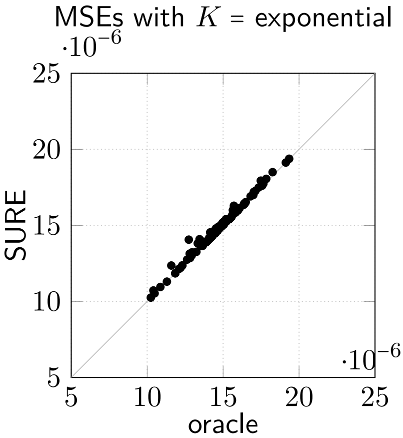
Does this work? Analysis on synthetic data

- $M = 10000$, 1000 Monte-Carlo runs, K = splines or exponential
- $\Lambda = 50$ potential λs, log-spaced in $\left[10^{-3}, 10^{3}\right]$
- $\exists$ "oracle" that knows f and thus what is the best λ

Does this work? Analysis on synthetic data

- $M = 10000$, 1000 Monte-Carlo runs, K = splines or exponential
- $\Lambda = 50$ potential λs, log-spaced in $\left[10^{-3}, 10^{3}\right]$
- $\exists$ "oracle" that knows f and thus what is the best λ





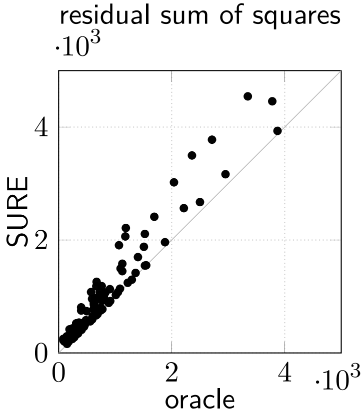
Does this work? Analysis on field data

Does this work? Analysis on field data

- "Colorado rain" UCI dataset
- $K\left(x, x'\right) = \exp\left(-10 \left\|x - x'\right\|_2^2\right)$ fixed, Λ as before
- 1000 Monte-Carlo runs, each with 2 random months of data as training set and 1 random month as test

Does this work? Analysis on field data

- "Colorado rain" UCI dataset
- $K(x, x') = \exp\left(-10 \|x - x'\|_2^2\right)$ fixed, Λ as before
- 1000 Monte-Carlo runs, each with 2 random months of data as training set and 1 random month as test



Some brief concluding remarks

Some brief concluding remarks

- SURE approaches can be valid alternatives to other model selection strategies
- seem to be suitable for distributed average-consensus based settings

Some brief concluding remarks

- SURE approaches can be valid alternatives to other model selection strategies
- seem to be suitable for distributed average-consensus based settings

What now?

- time-varying estimation
- generalizations for other distributed estimation settings & big data
  (not discussed in this presentation)

Purposes of this seminar

1. discuss about a useful tool
2. connect with you