Unbiased Risk Estimation as Parameter Choice Rule for Filter-based Regularization Methods

Frank Werner[1]

Statistical Inverse Problems in Biophysics Group
Max Planck Institute for Biophysical Chemistry, Göttingen

and

Felix Bernstein Institute for Mathematical Statistics in the Biosciences
University of Göttingen

Chemnitz Symposium on Inverse Problems 2017 (on Tour in Rio)

[1]joint work with Housen Li

Outline

Outline

Statistical inverse problems

Setting: $\mathcal{X}, \mathcal{Y}$ Hilbert spaces, $T : \mathcal{X} \to \mathcal{Y}$ bounded, linear

Task: Recover unknown $f \in \mathcal{X}$ from noisy measurements

$$Y = Tf + \sigma\xi$$

Noise: ξ is a standard Gaussian white noise process, $\sigma > 0$ noise level

The model has to be understood in a weak sense:

$$Y_g := \langle Tf, g\rangle_{\mathcal{Y}} + \sigma \langle \xi, g\rangle \qquad \text{for all } g \in \mathcal{Y}$$

with $\langle \xi, g\rangle \sim \mathcal{N}\left(0, \|g\|_{\mathcal{Y}}^2\right)$ and $\mathbb{E}\left[\langle \xi, g_1\rangle \langle \xi, g_2\rangle\right] = \langle g_1, g_2\rangle_{\mathcal{Y}}$.

Statistical inverse problems

Assumptions:

- T is injective and Hilbert-Schmidt ($\sum \sigma_k^2 < \infty$, σ_k singular values)
- σ is known exactly

As the problem is ill-posed, regularization is needed. Consider filter-based regularization schemes

$$\hat{f}_\alpha := q_\alpha(T^*T)T^*Y, \qquad \alpha > 0.$$

Aim:

A posteriori choice of α such that rate of convergence (as $\sigma \searrow 0$) is order optimal (no loss of log-factors)

Note: Heuristic parameter choice rules might work here as well, as the Bakushinskiĭ veto does not hold in our setting (Becker '11).

Outline

The discrepancy principle

- For deterministic data: $\alpha_{\mathrm{DP}} = \max\left\{\alpha > 0 \;\middle|\; \left\|T\hat{f}_\alpha - Y\right\|_{\mathcal{Y}} \leq \tau\sigma\right\}$
- But here: $Y \notin \mathcal{Y}$! Either pre-smoothing ($Y \rightsquigarrow Z := T^*Y \in \mathcal{X}$) ...
- ... or discretization: $Y \in \mathbb{R}^n$, $\xi \sim \mathcal{N}_n(0, I_n)$ and choose

$$\alpha_{\mathrm{DP}} = \max\left\{\alpha > 0 \;\middle|\; \left\|T\hat{f}_\alpha - Y\right\|_2 \leq \tau\sigma\sqrt{n}\right\}$$

Pros:

- Easy to implement
- Works for all q_α
- Order-optimal convergence rates

Cons:

- How to choose $\tau \geq 1$?
- Only discretized meaningful
- Early saturation

Davies & Anderssen '86, Lukas '95, Blanchard, Hoffmann & Reiß '16

The quasi-optimality criterion

- Neubauer '08 ($r_\alpha(\lambda) = 1 - \lambda q_\alpha(\lambda)$): $\alpha_{\mathrm{QO}} = \operatorname*{argmin}_{\alpha>0} \left\| r_\alpha(T^*T)\hat{f}_\alpha \right\|_{\mathcal{X}}$
- But for spectral cut-off $r_\alpha(T^*T)\hat{f}_\alpha = 0$ for all $\alpha > 0$
- Alternative formulation for Tikhonov regularization if candidates $\alpha_1 < ... < \alpha_m$ are given:

$$n_{\mathrm{QO}} = \operatorname*{argmin}_{1\leq n\leq m-1} \left\| \hat{f}_{\alpha_n} - \hat{f}_{\alpha_{n+1}} \right\|_{\mathcal{X}}, \qquad \alpha_{\mathrm{QO}} := \alpha_{n_{\mathrm{QO}}}.$$

Pros:

- Easy to implement, very fast
- No knowledge of σ necessary
- Order-optimal convergence rates in mildly ill-posed situations

Cons:

- Only for special q_α
- Additional assumptions on noise and/or f necessary
- Performance unclear in severely ill-posed situations

Bauer & Kindermann '08, Bauer & Reiß '08, Bauer & Kindermann '09

The Lepskiĭ-type balancing principle

- For given α, the standard deviation of $\hat{f}_\alpha$ can be bounded by

$$\mathsf{std}\,(\alpha) := \sigma \sqrt{\operatorname{Tr}\left(q_{\alpha_k}\left(T^*T\right)^2 T^*T\right)}$$

- If candidates $\alpha_1 < ... < \alpha_m$ are given:

$$n_{\mathrm{LEP}} = \max\left\{j \;\middle|\; \left\|\hat{f}_{\alpha_j} - \hat{f}_{\alpha_k}\right\|_{\mathcal{X}} \leq 4\kappa \mathsf{std}\,(\alpha_k) \text{ for all } 1 \leq k \leq j\right\}$$

and $\alpha_{\mathrm{LEP}} = \alpha_{n_{\mathrm{LEP}}}$

Pros:

- Works for all q_α
- Robust in practice
- convergence rates (mildly / severely ill-posed)

Cons:

- Computationally expansive
- $\kappa \geq 1$ depends on decay of σ_k
- loss of log factor compared to order-optimal rate

Bauer & Pereverzev '05, Mathé '06, Mathé & Pereverzev '06

Unbiased risk estimation

- Dating back to ideas of Mallows '73 and Stein '81 let

$$\hat{r}(\alpha, Y) := \left\| T\hat{f}_\alpha \right\|_{\mathcal{Y}}^2 - 2\left\langle T\hat{f}_\alpha, Y \right\rangle + 2\sigma^2 \mathrm{Tr}\left(T^* T q_\alpha\left(T^* T\right)\right)$$

  and choose $\alpha_{\mathrm{URE}} = \mathrm{argmin}_{\alpha>0}\, \hat{r}(\alpha, Y)$
- Note that $\mathbb{E}\left[\hat{r}(\alpha, Y)\right] = \mathbb{E}\left[\left\| T\hat{f}_\alpha - Tf \right\|_{\mathcal{Y}}^2\right] - c$ with c independent of α (**U**nbiased **R**isk **E**stimation)

For spectral cut-off and in mildly ill-posed situations, this gives order optimal rates (Chernousova & Golubev '14). Besides this, only optimality in the image space is known (Li '87, Lukas '93, Kneip '94). Distributional behavior of α_{URE}: Lucka et al '17.

In general: Pros? Cons? Convergence rates? Order optimality?

Outline

Assumptions

Filter

$$\alpha \left|q_\alpha(\lambda)\right| \leq C'_q \qquad \text{and} \qquad \lambda \left|q_\alpha(\lambda)\right| \leq C''_q.$$

Source condition

$$\mathcal{W}_\phi := \left\{ f \in \mathcal{X} \;\middle|\; f = \phi(T^*T)w, \|\omega\|_{\mathcal{X}} \leq C \right\}.$$

Note: for any $f \in \mathcal{X}$ there exists ϕ such that $f \in \mathcal{W}_\phi$.

Qualification condition

The function ϕ is a qualification of q_α if

$$\sup_{\lambda \in [0, \|T^*T\|} \phi(\lambda) \left|1 - \lambda q_\alpha(\lambda)\right| \leq C_\phi \phi(\alpha)$$

Assumptions

Let

$$\Sigma(x) := \#\left\{k \mid \sigma_k^2 \geq x\right\}$$

be the counting function of the singular values of T

Approximation by smooth surrogate

There exists $S \in C^2$, $\alpha_1 \in (0, \|T^*T\|]$ and $C_S \in (0,2)$ such that

(1) $\lim_{\alpha \searrow 0} S(\alpha)/\Sigma(\alpha) = 1$ (approximation)

(2) $S' < 0$ (decreasing)

(3) $\lim_{\alpha \nearrow \infty} S(\alpha) = \lim_{\alpha \nearrow \infty} S'(\alpha) = 0$ (behavior above σ_1^2)

(4) $\lim_{\alpha \searrow 0} \alpha S(\alpha) = 0$ (Hilbert-Schmidt)

(5) $\alpha S'(\alpha)$ is integrable on $(0, \alpha_1]$

(6) $\frac{S''(\alpha)}{-S'(\alpha)} \leq \frac{C_S}{\alpha}$ on $(0, \alpha_1]$

A priori convergence rates

Bissantz, Hohage, Munk, Ruymgaart '07

Let $\alpha_* \phi(\alpha_*)^2 = \sigma^2 S(\alpha_*)$.

(i) If ϕ is a qualification of q_α, then

$$\sup_{f \in \mathcal{W}_\phi} \mathbb{E}\left[\left\|\hat{f}_{\alpha_*} - f\right\|_{\mathcal{X}}^2\right] \lesssim \phi(\alpha_*)^2 = \sigma^2 \frac{S(\alpha_*)}{\alpha_*} \qquad \text{as } \sigma \searrow 0.$$

(ii) If $\lambda \mapsto \sqrt{\lambda}\phi(\lambda)$ is a qualification of the filter q_α, then

$$\sup_{f \in \mathcal{W}_\phi} \mathbb{E}\left[\left\|T\hat{f}_{\alpha_*} - Tf\right\|_{\mathcal{Y}}^2\right] \lesssim \alpha_* \phi(\alpha_*)^2 = \sigma^2 S(\alpha_*) \qquad \text{as } \sigma \searrow 0.$$

Mildly ill-posed situation: Example

Assume $\sigma_k^2 \asymp k^{-a}$, $\mathcal{W}_b := \left\{ f \in \mathcal{X} : \sum_{k=1}^{\infty} k^b f_k^2 \leq 1 \right\}$ with $a > 1, b > 0$:

Bissantz, Hohage, Munk, Ruymgaart '07

Let $\alpha_* \asymp (\sigma^2)^{a/(a+b+1)}$.

- If $\phi(\lambda) = \lambda^{b/2a}$ is a qualification of q_α, then

$$\sup_{f \in \mathcal{W}_b} \mathbb{E}\left[\left\| \hat{f}_{\alpha_*} - f \right\|_{\mathcal{X}}^2 \right] \lesssim (\sigma^2)^{\frac{b}{a+b+1}}.$$

- If $\phi(\lambda) = \lambda^{b/2a+1/2}$ is a qualification of q_α, then

$$\sup_{f \in \mathcal{W}_b} \mathbb{E}\left[\left\| T\hat{f}_{\alpha_*} - Tf \right\|_{\mathcal{Y}}^2 \right] \lesssim (\sigma^2)^{\frac{a+b}{a+b+1}}.$$

These rates are order optimal over $\mathcal{W}_b$.

Unbiased risk estimation vs. the oracle

Recall that

$$\hat{r}(\alpha, Y) := \left\| T\hat{f}_\alpha \right\|_{\mathcal{Y}}^2 - 2\left\langle T\hat{f}_\alpha, Y \right\rangle + 2\sigma^2 \mathrm{Tr}\left(T^*Tq_\alpha\left(T^*T\right)\right)$$

is an unbiased estimator for

$$r(\alpha, f) := \mathbb{E}\left[\left\| T\hat{f}_\alpha - Tf \right\|_{\mathcal{Y}}^2\right].$$

In the following, we will compare

$$\alpha_{URE} = \underset{\alpha>0}{\mathrm{argmin}}\, \hat{r}(\alpha, Y) \quad \text{and} \quad \alpha_o = \underset{\alpha>0}{\mathrm{argmin}}\, r(\alpha, f).$$

Additional assumptions

(a) $\alpha \mapsto \{q_\alpha(\sigma_k^2)\}_{k=1}^\infty$ is strictly monotone and continuous as $\mathbb{R} \to \ell^2$.
(b) As $\alpha \searrow 0$, $\alpha q_\alpha(\alpha) \geq c_q > 0$.
(c) For $\alpha > 0$, the function $\lambda \mapsto \lambda q_\alpha(\lambda)$ is non-decreasing.

Satisfied by Tikhonov, spectral cut-off, Landweber, iterated Tikhonov and Showalter regularization, under proper parametrization. E.g. Tikhonov with re-parametrization $\alpha \mapsto \sqrt{\alpha}$ $(q_\alpha(\lambda) = 1/(\sqrt{\alpha} + \lambda))$ violates (b).

(d) $\psi(\lambda) := \lambda \phi^{-1}\left(\sqrt{\lambda}\right)$ is convex
(e) There exists a constant $C_q > c_q^{-2}$ such that

$$\int_1^\infty \Psi'(C_q x) \exp\left(-C\sqrt{\frac{x}{2}}\right) \mathrm{d}x < \infty \qquad \text{with } \Psi(x) := \frac{x}{\left(S^{-1}(x)\right)^2}.$$

for some explicitly known $C > 0$.

(d) can always be satisfied by weakening ϕ, (e) restricts the decay of the singular values

Oracle inequality

Li & W. '16

There are positive constants C_i, $i = 1, \ldots, 6$, such that for all $f \in \mathcal{W}_\phi$ it holds

$$\mathbb{E}\left[\left\|\hat{f}_{\alpha_{\mathrm{URE}}} - f\right\|_{\mathcal{X}}^2\right] \leq C_1 \psi^{-1}\left(C_2 r(\alpha_o, f) + C_3 \sigma^2\right) + C_4 \sigma^2 + C_5 \frac{r(\alpha_o, f) + \sigma\sqrt{r(\alpha_o, f)}}{S^{-1}\left(C_6 \frac{r(\alpha_o, f)}{\sigma^2}\right)}$$

as $\sigma \searrow 0$.

Gives a comparison of the strong risk under α_{URE} with the weak risk under the oracle α_o.

Convergence rates

Li & W. '16

If also $\lambda \mapsto \sqrt{\lambda}\phi(\lambda)$ is a qualification of the filter q_α, then for $\alpha_* \phi(\alpha_*)^2 = \sigma^2 S(\alpha_*)$ there are $C_1, C_2, C_3 > 0$ such that

$$\sup_{f \in \mathcal{W}_\phi} \mathbb{E}\left[\left\|\hat{f}_{\alpha_{\mathrm{URE}}} - f\right\|_{\mathcal{X}}^2\right] \leq C_1 \sigma^2 \frac{S(\alpha_*)}{\alpha_*} + C_2 \frac{\sigma^2 S(\alpha_*)}{S^{-1}\left(C_3 S(\alpha_*)\right)}$$

as $\sigma \searrow 0$.

If there is $C_4 > 0$ such that $S(C_4 x) \geq C_3 S(x)$, then this equals the a priori rate

$$\sup_{f \in \mathcal{W}_\phi} \mathbb{E}\left[\left\|\hat{f}_{\alpha_{\mathrm{URE}}} - f\right\|_{\mathcal{X}}^2\right] \lesssim \phi(\alpha_*)^2 = \sigma^2 \frac{S(\alpha_*)}{\alpha_*}.$$

Order optimality in mildly ill-posed situations

Assume $\sigma_k^2 \asymp k^{-a}$, $\mathcal{W}_b := \left\{ f \in \mathcal{X} : \sum_{k=1}^{\infty} k^b f_k^2 \leq 1 \right\}$ with $a > 1, b > 0$:

Oracle inequality

For all $f \in \mathcal{W}$:

$$\mathbb{E}\left[\left\| \hat{f}_{\alpha_{\mathrm{URE}}} - f \right\|_{\mathcal{X}}^2 \right] \lesssim r(\alpha_o, f)^{\frac{b}{a+b}} + \sigma^{-2a} r(\alpha_o, f)^{1+a} + \sigma^{1-2a} r(\alpha_o, f)^{\frac{1+2a}{2}}.$$

Convergence rate

Thus, if $\lambda \mapsto \lambda^{b/2a+1/2}$ is a qualification of q_α, then

$$\sup_{f \in \mathcal{W}_b} \mathbb{E}\left[\left\| \hat{f}_{\alpha_{\mathrm{URE}}} - f \right\|_{\mathcal{X}}^2 \right] \lesssim \sigma^{\frac{2b}{a+b+1}}$$

which is order-optimal.

Unbiased risk estimation - pros and cons

$$\alpha_{\mathrm{URE}} = \operatorname*{argmin}_{\alpha>0} \left[\left\| T\hat{f}_\alpha \right\|_{\mathcal{Y}}^2 - 2\left\langle T\hat{f}_\alpha, Y \right\rangle + 2\sigma^2 \mathrm{Tr}\left(T^*Tq_\alpha\left(T^*T\right)\right)\right]$$

Pros:

- Works for many q_α
- order-optimal convergence rates in mildly ill-posed situations
- no loss of log factor
- no tuning parameter

Cons:

- Computationally expansive
- Early saturation
- performance in severely ill-posed situations unclear

H. Li and F. Werner (2017). Empirical risk minimization as parameter choice rule for general linear regularization methods. Submitted, *arXiv:* 1703.07809.

Outline

A mildly ill-posed situation - antiderivative

Let $T : \mathbf{L}^2([0,1]) \to \mathbf{L}^2([0,1])$ given by

$$(Tf)(x) = \int_0^1 \min\{x(1-y), y(1-x)\} f(y) \,\mathrm{d}y$$

As $(Tf)'' = -f$ the singular values σ_k of T satisfy $\sigma_k \asymp k^{-2}$.

We choose

$$f(x) = \begin{cases} x & \text{if } 0 \leq x \leq \frac{1}{2}, \\ 1-x & \text{if } \frac{1}{2} \leq x \leq 1. \end{cases}$$

Fourier coefficients: $f_k = \frac{(-1)^k - 1}{4\pi^3 k^2}$, so the optimal rate is $\mathcal{O}\left(\sigma^{\frac{3}{4}-\varepsilon}\right)$ for any $\varepsilon > 0$.

A mildly ill-posed situation - Tikhonov regularization

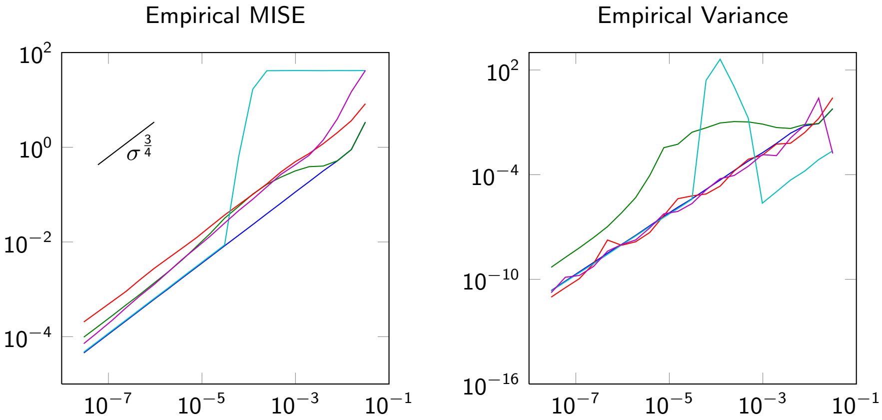

Figure: Empirical MISE and variance of $\|\hat{f} - f\|_2^2$ over 10^4 repetitions: α_o (blue line), α_{DP} (magenta line), α_{QO} (cyan line), α_{LEP} (red line), α_{URE} (green line).

A severely ill-posed situation - satellite gradiometry

Let $R > 1$ and $S \subset \mathbb{R}^2$ the unit sphere. Given $g = \frac{\partial^2 u}{\partial r^2}$ on RS find f in

$$\begin{cases} \Delta u = 0 & \text{in } \mathbb{R}^d \setminus B, \\ u = f & \text{on } S, \\ |u(x)| = \mathcal{O}\left(\|x\|_2^{-1}\right) & \text{as } \|x\|_2 \to \infty. \end{cases}$$

Corresponding $T : \mathbf{L}^2(S, \mu) \to \mathbf{L}^2(RS, \mu)$ has singular values $\sigma_k = |k|(|k|+1) R^{-|k|-2}$.

We choose

$$f(x) = \frac{\pi}{2} - |x|, \qquad x \in [-\pi, \pi]$$

Optimal rate of convergence is $\mathcal{O}\left((-\log(\sigma))^{-3+\varepsilon}\right)$ for any $\varepsilon > 0$.

A severely ill-posed situation - Tikhonov regularization

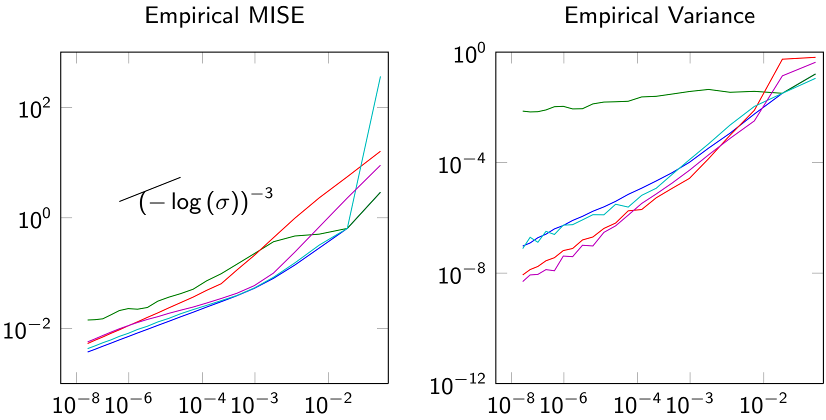

Figure: Empirical MISE and variance of $\|\hat{f} - f\|_2^2$ over 10^4 repetitions: α_o (——), α_{DP} (——), α_{QO} (——), α_{LEP} (——), α_{URE} (——).

A severely ill-posed situation - backwards heat equation

Let $\bar{t} > 0$. Given $g = u(\cdot, \bar{t})$ find f in

$$\begin{cases} \frac{\partial u}{\partial t}(x,t) = \frac{\partial^2 u}{\partial t^2}(x,t) & \text{in } (-\pi,\pi] \times (0,\bar{t}), \\ u(x,0) = f(x) & \text{on } [-\pi,\pi], \\ u(-\pi,t) = u(\pi,t) & \text{on } t \in (0,\bar{t}]. \end{cases}$$

Corresponding $T : \mathbf{L}^2([-\pi,\pi]) \to \mathbf{L}^2([-\pi,\pi])$ has singular values $\sigma_k = \exp\left(-k^2\bar{t}\right)$.

We choose

$$f(x) = \frac{\pi}{2} - |x|, \qquad x \in [-\pi,\pi]$$

Optimal rate of convergence is $\mathcal{O}\left((-\log(\sigma))^{-3/2+\varepsilon}\right)$ for any $\varepsilon > 0$.

A severely ill-posed situation - Tikhonov regularization

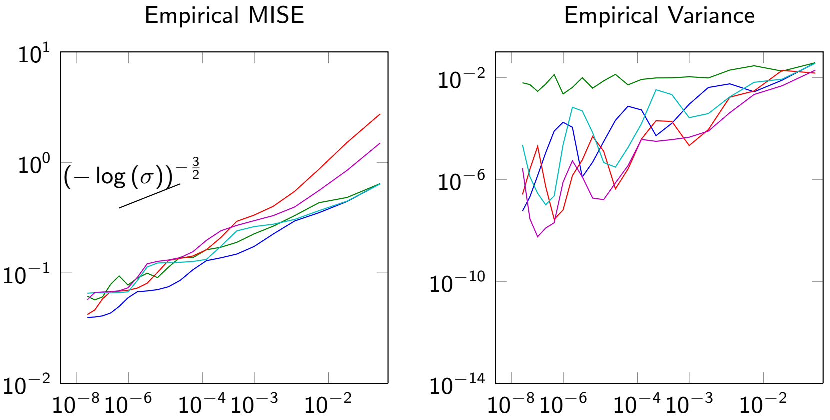

Figure: Empirical MISE and variance of $\|\hat{f} - f\|_2^2$ over 10^4 repetitions:
α_o (——), α_{DP} (——), α_{QO} (——), α_{LEP} (——), α_{URE} (——).

Efficiency simulations

Measure the efficiency of a parameter choice rule α_* by the fraction

$$R_* := \frac{\mathbb{E}\left[\left\|\hat{f}_{\alpha_o} - f\right\|_{\mathcal{X}}^2\right]}{\mathbb{E}\left[\left\|\hat{f}_{\alpha_*} - f\right\|_{\mathcal{X}}^2\right]}$$

Numerical approximations of these as functions of σ with different parameters $a, \nu > 0$ in the following setting:

- $\sigma_k = \exp(-ak)$
- $f_k = \pm k^{-\nu} \cdot \left(1 + \mathcal{N}\left(0, 0.1^2\right)\right)$
- $Y_k = \sigma_k \cdot f_k + \mathcal{N}\left(0, \sigma^2\right)$
- $k = 1, ..., 300$, 10^4 repetitions
- Tikhonov regularization

Efficiency simulations - results

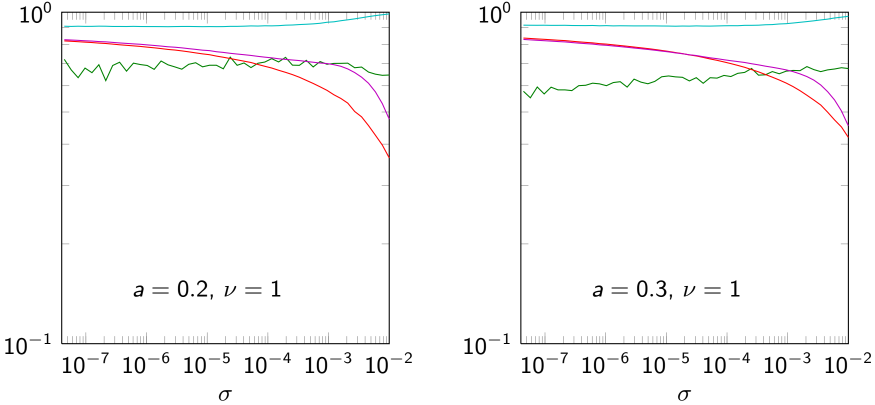

Figure: R_{QO} (——), R_{DP} (——), R_{LEP} (——), and R_{URE} (——)

Efficiency simulations - results

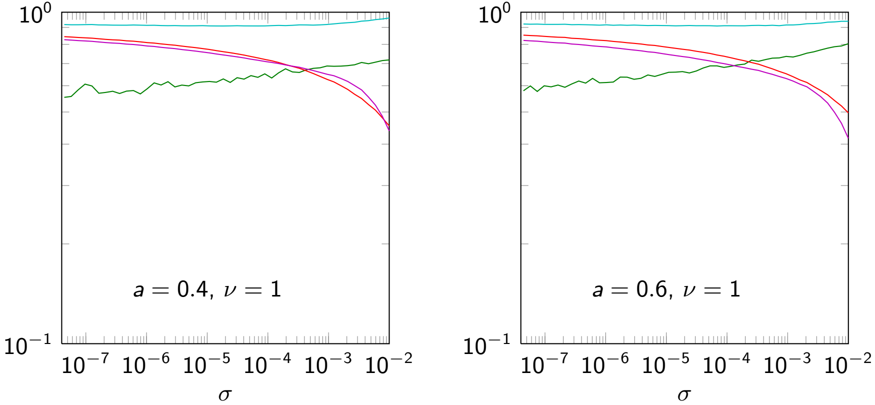

Figure: R_{QO} (——), R_{DP} (——), R_{LEP} (——), and R_{URE} (——)

Efficiency simulations - results

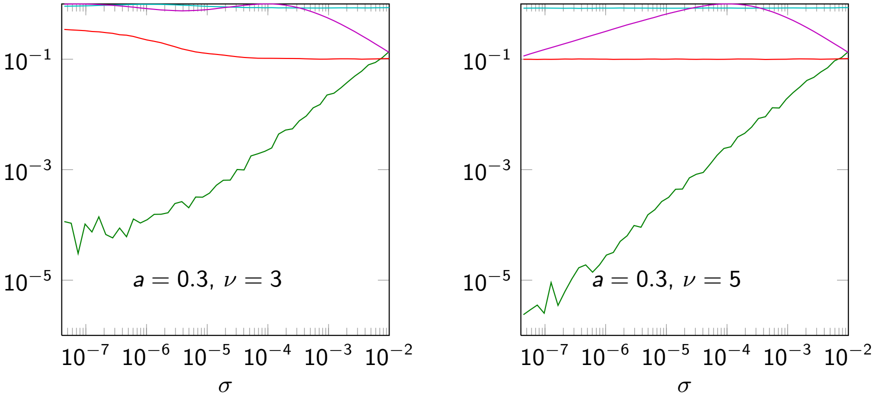

Figure: R_{QO} (——), R_{DP} (——), R_{LEP} (——), and R_{URE} (——)

Outline

Presented results

- Analysis of a parameter choice based on unbiased risk estimation:
  - oracle inequality
  - convergence rates
  - order optimality in mildly ill-posed situations
- Numerical comparison:
  - in this specific setting, quasi-optimality outperforms all other methods
  - unbiased risk estimation has higher variance (by design)
  - simulations suggest order optimality of quasi-optimality also in severely ill-posed situations, not clear for unbiased risk estimation

Thank you for your attention!