Empirical Risk Minimization 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

Institute for Mathematical Stochastics
University of Göttingen

[1]joint work with Housen Li (University of Göttingen)

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 (cont')

Assumption (Standing for the whole talk)

- *The forward operator T is injective, compact and Hilbert-Schmidt*
- *The noise level σ is known, and we are interested in $\sigma \searrow 0$*

Note: σ can be pre-estimated (Rice '86; Hall et al. '90; Dette et al. '98).

Transformation to a standard sequence model: Let

- $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ the eigenvalues of T^*T and
- $e_1, e_2, \ldots$ the corresponding normalized eigenvectors.

Then $Y = Tf + \sigma\xi$ is equivalent to

$$Y_k = \sqrt{\lambda_k} \cdot f_k + \sigma\xi_k, \qquad k = 1, 2, \ldots,$$

with $Y_k := \langle \lambda_k^{-1/2} Te_k, Y \rangle$, $f_k := \langle f, e_k \rangle$, $\xi_k := \langle \lambda_k^{-1/2} Te_k, \xi \rangle \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$

Linear regularization methods

The maximum likelihood estimator for f is

$$\hat{f} := (T^*T)^{-1}T^*Y = \sum_{k=1}^{\infty} \lambda_k^{-\frac{1}{2}} Y_k e_k.$$

But: Compactness of $T \Rightarrow \lambda_k \searrow 0 \rightsquigarrow$ ill-posedness!

For regularization we consider filter based linear regularization methods:

$$\hat{f}_\alpha := q_\alpha(T^*T)T^*Y = \sum_{k=1}^{\infty} q_\alpha(\lambda_k)\sqrt{\lambda_k} Y_k e_k, \qquad \alpha \in \mathcal{A}$$

Definition (Filter, see e.g. (Engl et al. '96))

We call $q_\alpha : [0, \lambda_1] \to \mathbb{R}$ with $\alpha \in \mathcal{A} \subseteq \mathbb{R}_+$ a filter if there exist $C'_q, C''_q > 0$ such that for every $\alpha \in \mathcal{A}$ and every $\lambda \in [0, \lambda_1]$ it holds

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

Examples for linear regularization methods

Spectral cut-off regularization: $\frac{1}{\lambda}\mathbf{1}_{[\alpha,\infty)}(\lambda)$.

Can be interpreted as truncating the sum in the formula for the least squares estimator at $\lambda_k = \alpha$.

Tikhonov regularization: $q_\alpha(\lambda) = 1/(\lambda + \alpha)$.

By differentiation, it can be seen that in this case

$$\begin{aligned}\hat{f}_\alpha &= \arg\min_{f\in\mathcal{X}} \left[\|Tf\|^2 - \langle Tf, Y\rangle + \alpha\|f\|^2\right] \\ &= \arg\min_{f\in\mathcal{X}} \left[\|Tf - Y\|^2 + \alpha\|f\|^2\right] \quad \text{“formally”}\end{aligned}$$

Showalter regularization: $\frac{1-\exp\left(-\frac{\lambda}{\alpha}\right)}{\lambda}$.

$\hat{f}_\alpha$ corresponds to $u(1/\alpha)$ with the solution u of

$$\begin{cases} u'(t) &= -T^*Tu(t) + T^*Y, \quad t > 0 \\ u(0) &= 0 \end{cases}$$

Outline

Error measures

We will measure the error of an estimator $\hat{f}_\alpha$ w.r.t. the weak risk

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

or the strong risk

$$R_{\mathrm{s}}(\alpha, f) := \mathbb{E}\left[\|\hat{f}_\alpha - f\|_{\mathcal{X}}^2\right].$$

For the strong risk we have the error decomposition

$$\begin{aligned} R_{\mathrm{s}}(\alpha, f) &= \left\|\mathbb{E}\left[\hat{f}_\alpha\right] - f\right\|_{\mathcal{X}}^2 + \mathbb{E}\left[\left\|\hat{f}_\alpha - \mathbb{E}\left[\hat{f}_\alpha\right]\right\|_{\mathcal{X}}^2\right] \\ &= \|r_\alpha(T^*T) f\|_{\mathcal{X}}^2 + \sigma^2 \mathrm{tr}\left(q_\alpha(T^*T)^2 T^*T\right) \end{aligned}$$

where $r_\alpha(\lambda) = 1 - \lambda q_\alpha(\lambda)$.

Error analysis

$$R_{\mathrm{s}}(\alpha, f) = \|r_\alpha(T^*T) f\|_{\mathcal{X}}^2 + \sigma^2 \mathrm{tr}\left(q_\alpha(T^*T)^2 T^*T\right)$$

To derive rates of converges, both terms are estimated separately:

Bias $\|r_\alpha(T^*T) f\|_{\mathcal{X}}^2$ will be estimated exploiting (relative) smoothness of f

Variance $\mathrm{tr}\left(q_\alpha(T^*T)^2 T^*T\right)$ can be bounded by

$$-\left(\frac{C_q'}{\alpha}\right)^2 \int_0^\alpha t \,\mathrm{d}\Sigma(t) - (C_q'')^2 \int_\alpha^\infty \frac{1}{t} \,\mathrm{d}\Sigma(t)$$

with the counting function Σ of the eigenvalues of T^*T:

$$\Sigma(\alpha) := \#\left\{k \in \mathbb{N} \;\middle|\; \lambda_k \geq \alpha\right\}.$$

Assumptions to control the bias

Assumption (Source condition, smoothness of f)

The unknown truth f lies in

$$\mathcal{W}_\phi := \left\{ f^* \in \mathcal{X} \;\middle|\; f^* = \phi(T^*T)w, \|\omega\|_{\mathcal{X}} \leq 1 \right\}.$$

with ϕ such that $\psi(\lambda) := \lambda \phi^{-1}\left(\sqrt{\lambda}\right)$ is convex

- Source conditions measures the smoothness of f relative to the smoothing behavior of T.

Note: for any $f \in \mathcal{X}$ there exists ϕ such that $f \in \mathcal{W}_\phi$ (Mathé & Hofmann '08).

Note: Assuming that ψ is convex is no restriction at all, can be ensured by 'weakening' ϕ.

Assumptions to control the bias (cont')

Assumption (Qualification condition, compatibility with q_α)

The function ϕ is a qualification of q_α, this is (*recall:* $r_\alpha(\lambda) = 1 - \lambda q_\alpha(\lambda)$)

$$\sup_{\lambda \in [0, \|T^*T\|]} \phi(\lambda) |r_\alpha(\lambda)| \leq C_\phi \phi(\alpha)$$

Possible qualifications: (in case that $\phi(\lambda) = \lambda^\nu$)

Spectral cut-off: all functions ϕ ($\nu > 0$)

Tikhonov: ϕ 'asymptotically (at 0) above' the identity ($\nu \in (0, 1]$)

Showalter: ϕ 'asymptotically (at 0) above' some monomial ($\nu > 0$)

Assumptions to control the variance

Assumption (Approximation by smooth surrogate)

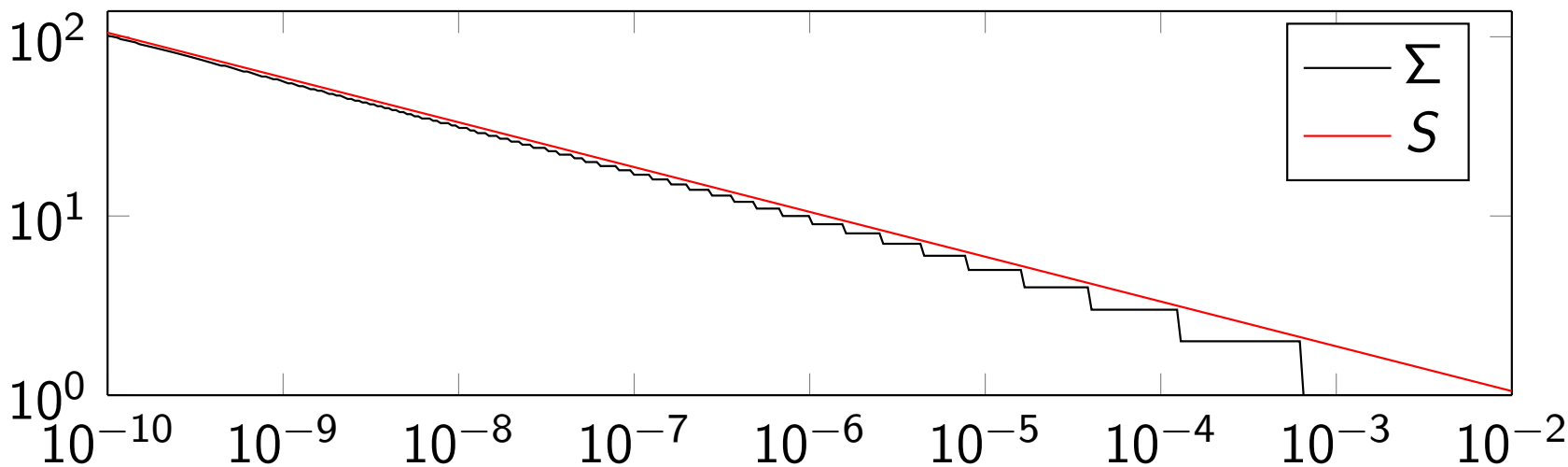
There exists $S \in C^2$ *approximating* $\Sigma(\alpha) := \#\left\{k \in \mathbb{N} \;\middle|\; \lambda_k \geq \alpha\right\}$ *with*

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

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

(3) $\lim_{\alpha \nearrow \infty} S(\alpha) = \lim_{\alpha \nearrow \infty} S'(\alpha) = 0$ *(behavior above* λ_1*)*

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



A priori convergence rates

Theorem (Bissantz, Hohage, Munk, Ruymgaart '07)

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

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

$$\sup_{f \in \mathcal{W}_\phi} R_{\mathrm{s}}(\alpha_*, f) \lesssim \phi(\alpha_*)^2 = \sigma^2 \frac{S(\alpha_*)}{\alpha_*} \qquad \text{as } \sigma \searrow 0.$$

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

$$\sup_{f \in \mathcal{W}_\phi} R_{\mathrm{w}}(\alpha_*, f) \lesssim \alpha_* \phi(\alpha_*)^2 = \sigma^2 S(\alpha_*) \qquad \text{as } \sigma \searrow 0.$$

Mildly ill-posed situation: Example

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

Corollary (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*

$$\sup_{f \in \mathcal{W}_b} R_{\mathrm{s}}(\alpha^*, f) \lesssim (\sigma^2)^{\frac{b}{a+b+1}}.$$

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

$$\sup_{f \in \mathcal{W}_b} R_{\mathrm{w}}(\alpha^*, f) \lesssim (\sigma^2)^{\frac{a+b}{a+b+1}}.$$

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

Outline

A-posteriori parameter choice

Above results rely on knowledge of the smoothness of f.

In practice, such information is not available! $\rightsquigarrow$ Adaptation?

Literature:

- discrepancy principle (Morozov '66; Davies & Anderssen '86; Lukas '95; Blanchard et al. '18);
- generalized cross validation (GCV) (Wahba '77; Golub et al. '79; Lukas '93);
- Lepskiĭ-type (Lepskiĭ '91; Mathé '06; Mathé & Pereverzev '06; Werner & Hohage '12).
- many more, see (Bauer & Lukas '11) for a survey.

Parameter choice by risk minimization

The optimal $\alpha \in \mathcal{A}$ is given by the strong oracle

$$\alpha_{o,\mathrm{s}} \in \operatorname*{arg\,min}_{\alpha \in \mathcal{A}} R_{\mathrm{s}}(\alpha, f) = \operatorname*{arg\,min}_{\alpha \in \mathcal{A}} \mathbb{E}\left[\| \hat{f}_\alpha - f \|^2\right]$$

which clearly depends on f.

We can try to mimic $\alpha_{o,\mathrm{s}}$ by the weak oracle

$$\alpha_{o,\mathrm{w}} \in \operatorname*{arg\,min}_{\alpha \in \mathcal{A}} R_{\mathrm{w}}(\alpha, f) = \operatorname*{arg\,min}_{\alpha \in \mathcal{A}} \mathbb{E}\left[\| T\hat{f}_\alpha - Tf \|^2\right]$$

which depends on Tf (also unknown).

Intuition: Replace Tf by Y.

Empirical prediction risk minimization

We consider the parameter choice rule α_{URE} given by

$$\alpha_{\mathrm{URE}} \in \operatorname*{arg\,min}_{\alpha \in \mathcal{A}} \left(\| T\hat{f}_\alpha \|^2 - 2\langle T\hat{f}_\alpha, Y \rangle + 2\sigma^2 \mathrm{tr}\left(s_\alpha\left(T^*T\right)\right)\right).$$

where $s_\alpha(\lambda) := \lambda q_\alpha(\lambda)$.

Note:

$$\begin{aligned}
&\mathbb{E}\left[\| T\hat{f}_\alpha \|^2 - 2\langle T\hat{f}_\alpha, Y \rangle \right] \\
&= \underbrace{\sum_{k=1}^{\infty} \lambda_k (1 - s_\alpha(\lambda_k))^2 f_k^2 + \sigma^2 \sum_{k=1}^{\infty} s_\alpha(\lambda_k)^2}_{R_{\mathrm{w}}(\alpha, f)} \underbrace{- 2\sigma^2 \sum_{k=1}^{\infty} s_\alpha(\lambda_k)}_{2\sigma^2 \mathrm{tr}(s_\alpha(T^*T))} - \sum_{k=1}^{\infty} \lambda_k f_k^2,
\end{aligned}$$

i.e.

$$\mathbb{E}\left[\| T\hat{f}_\alpha \|^2 - 2\langle T\hat{f}_\alpha, Y \rangle + 2\sigma^2 \mathrm{tr}\left(s_\alpha\left(T^*T\right)\right)\right] = R_{\mathrm{w}}\left(\alpha, f\right) + C$$

with a constant C independent of α.

Literature review on α_{URE}

- goes back to ideas of (Mallow '73) and (Stein '81) for 'direct' regression
- closely related to GCV (Li '86; Wahba '90; Efron '01)
- popular & attractive in practice (Lukas '98; Vogel '02; Bauer & Lukas '11)
- minimax order optimal w.r.t. the weak risk $R_{\mathrm{w}}(\alpha, f)$ (Li '86; Vogel '86; Lukas '93; Kneip '94)
- distributional behavior studied in (Lucka et al. '18)
- minimax order optimal w.r.t. the strong risk $R_{\mathrm{s}}(\alpha, f)$ if the problem is mildly ill-posed and spectral cut-off regularization is used (Chernousova & Golubev '14)

Does this also work for other filter-based regularization methods?

Additional assumptions for the analysis

Assumption (on the filter q_α)

(a) $\alpha \mapsto \{q_\alpha(\lambda_k)\}_{k=1}^\infty$ *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.*

Note: 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).

Additional assumptions for the analysis (cont')

Assumption (on the decay of the eigenvalues of T)

(d) *There exist* $\alpha_1 \in (0, \lambda_1]$ *and* $C_S > 0$ *such that*

$$\frac{1}{\alpha} \int_0^\alpha S(t)\, \mathrm{d}t \leq C_s S(\alpha) \qquad \text{for all} \qquad \alpha \in (0, \alpha_1].$$

(e) *There exists a constant* $C_q > c_q^{-2}$ *such that*

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

Note: Conditions implying (d) are also assumed in (Bissantz et al. '07) to prove minimax optimality under a priori choice of α

Note: (e) is needed to bound generalized moments; the constant C comes from a concentration result by (Kneip '94)

Oracle inequality

$$\gamma_\sigma := \frac{R_{\mathrm{w}}(\alpha_{o,\mathrm{w}}, f)}{\sigma^2}$$

Theorem (Li & W. '18)

There are positive constants C_1, C_2 and C_3, independent of f and σ, such that

$$\mathbb{E}\left[\|\hat{f}_{\alpha_{\mathrm{URE}}} - f\|^2\right] \leq \psi^{-1}\Big(\sigma^2\big(2\gamma_\sigma + C_1\big)\Big) + \sigma^2 C_3\left(\frac{\gamma_\sigma + \sqrt{\gamma_\sigma}}{S^{-1}\left(2C_q\gamma_\sigma\right)} + C_2\right)$$

Gives a comparison of the strong risk under α_{URE} with the weak risk R_{w} under the oracle $\alpha_{o,\mathrm{w}}$.

Note: $\mathbb{E}\left[\|\hat{f}_{\alpha_{\mathrm{URE}}} - f\|^2\right] \neq \mathbb{E}\left[R_{\mathrm{s}}\left(\alpha_{\mathrm{URE}}, f\right)\right]$.

Convergence rates

Theorem (Li & W. '18)

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* $C_1, C_2, C_3 > 0$ *independent of* σ *such that*

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

as $\sigma \searrow 0$.

If there is $C_4 > 0$ such that $S(C_4 x) \geq C_2 C_q S(x)$, then this equals the a priori rate

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

Order optimality in mildly ill-posed situations

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

Corollary (Oracle inequality)

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

Corollary (Convergence rate)

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

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

which is minimax order-optimal (no loss of log-factors!).

Outline

Overview

In the following we will compare the following parameter choice rules for different mildly and exponentially ill-posed problems:

- Discrepancy principle
- Quasi-optimality criterion
- Lepskiĭ-type balancing principle
- Empirical risk minimization

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

The discrepancy principle

- For deterministic data: $\alpha_{\mathrm{DP}} = \max\left\{\alpha > 0 \mid \left\| T\hat{f}_\alpha - Y \right\|_{\mathcal{Y}} \leq \tau\sigma\right\}$
- 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

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

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 '18)

The quasi-optimality criterion

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

$$n_{\mathrm{QO}} = \arg\min_{1\le n\le m-1} \left\| \hat{f}_{\alpha_n} - \hat{f}_{\alpha_{n+1}} \right\|_{\mathcal{X}}, \qquad \alpha_{\mathrm{QO}} := \alpha_{n_{\mathrm{QO}}}.$$

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

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

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

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

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 λ_k
- loss of log factor compared to order-optimal rate

(Bauer & Pereverzev '05; Mathé '06; Mathé & Pereverzev '06)

Unbiased risk estimation

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

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 (2018). Empirical risk minimization as parameter choice rule for general linear regularization methods. Submitted, *arXiv:* 1703.07809.

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 eigenvalues λ_k of T^*T satisfy $\lambda_k \asymp k^{-4}$.

We choose

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

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

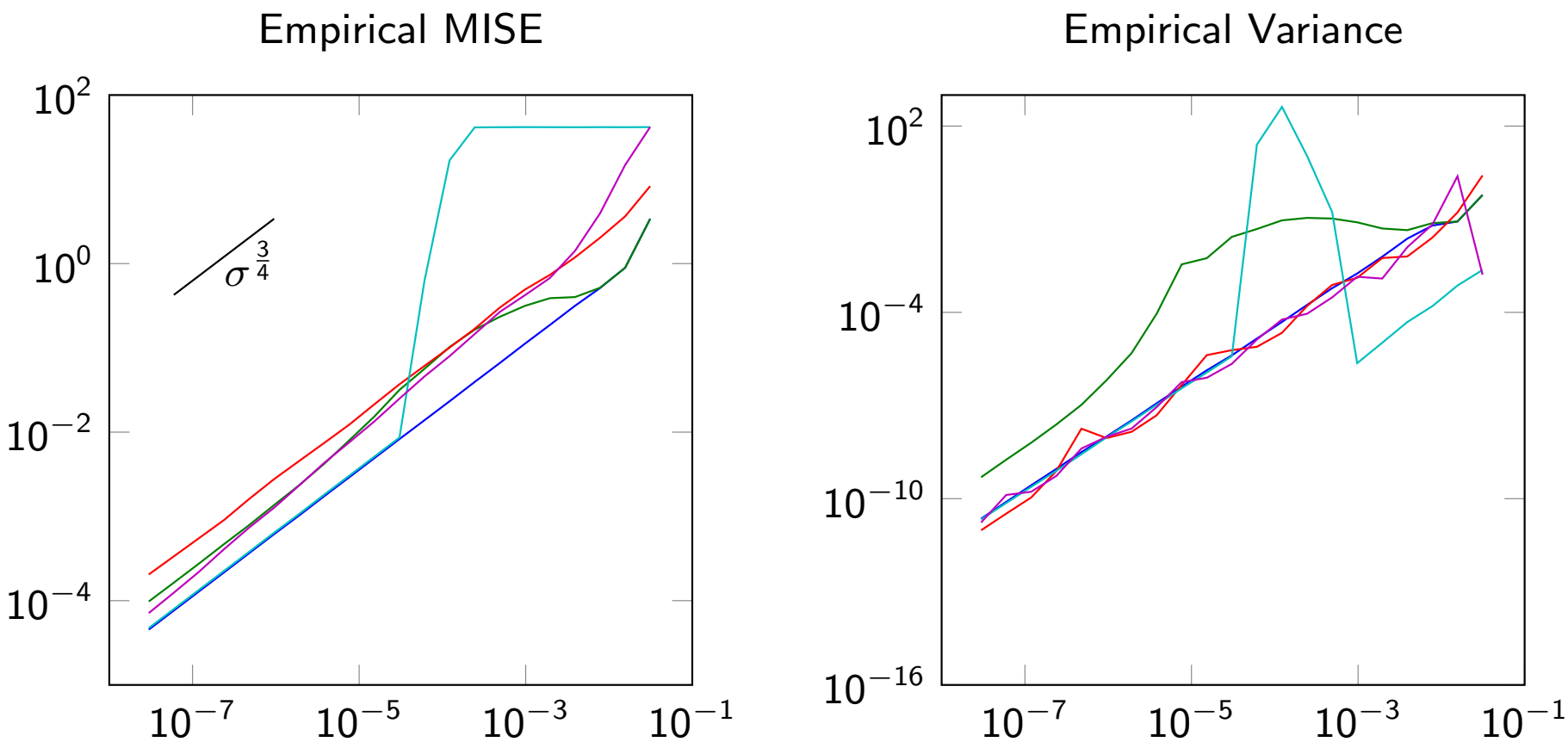


Figure: Empirical MISE and variance of $\|\hat{f} - f\|_2^2$ over 10^4 repetitions: α_o (——), α_{DP} (——), α_{QO} (——), α_{LEP} (——), α_{URE} (——).

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

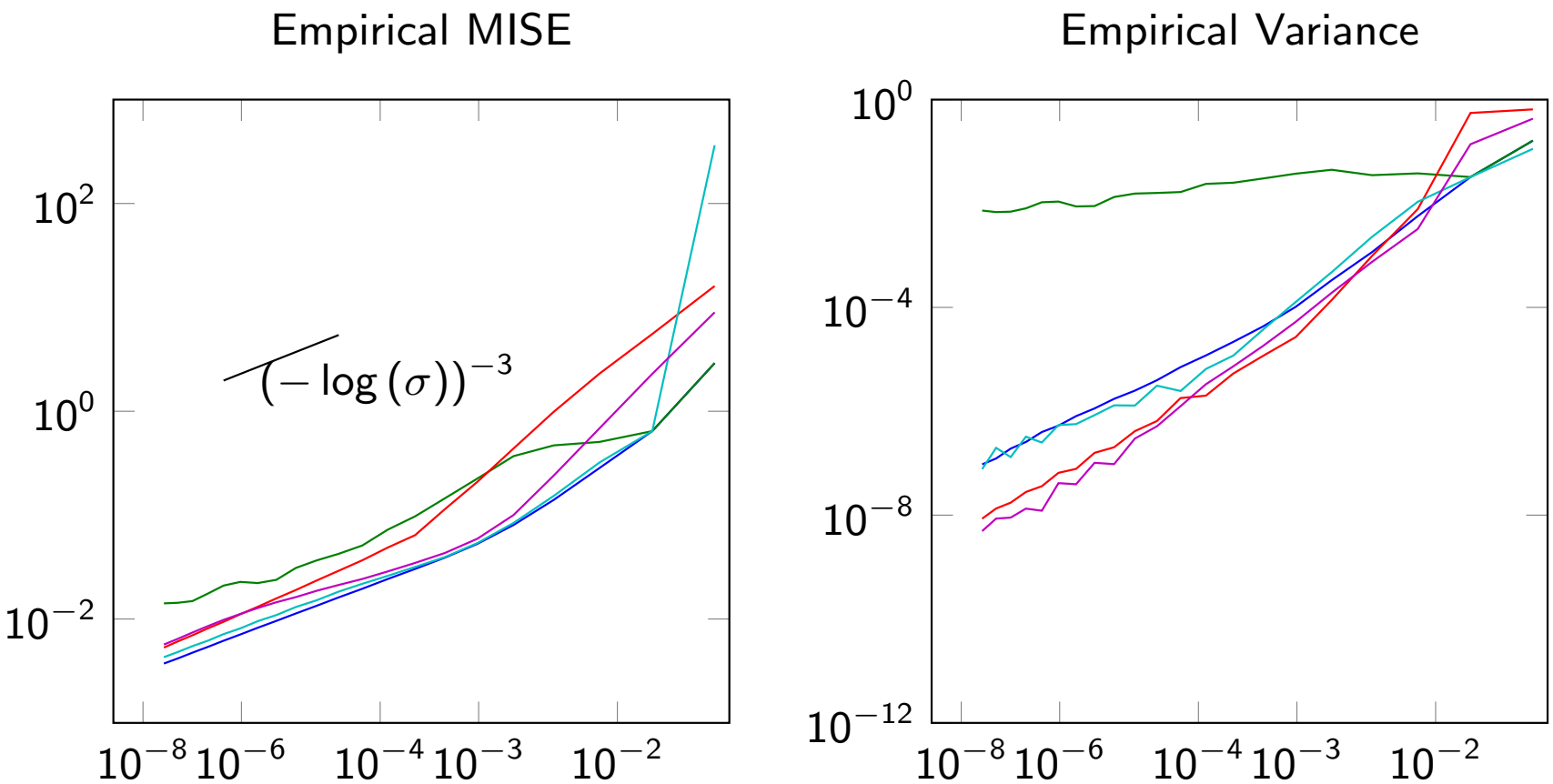


Figure: Empirical MISE and variance of $\|\hat{f} - f\|_2^2$ over 10^4 repetitions: α_o (——), α_{DP} (——), α_{QO} (——), α_{LEP} (——), α_{URE} (——).

A severely ill-posed situation - backwards heat equation

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

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

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

A severely ill-posed situation - Tikhonov regularization

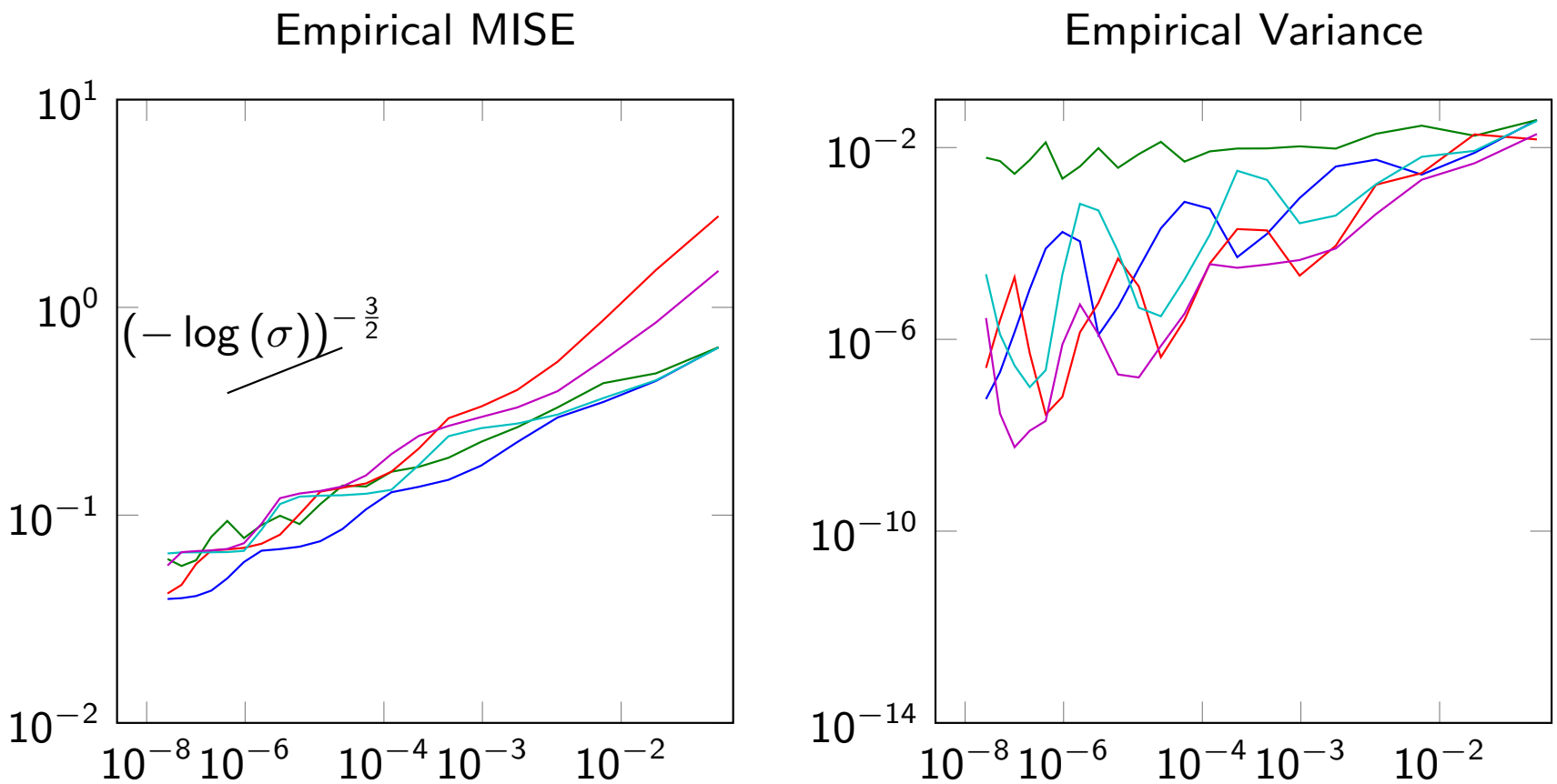


Figure: Empirical MISE and variance of $\|\hat{f} - f\|_2^2$ over 10^4 repetitions: α_o (——), α_{DP} (——), α_{QO} (——), α_{LEP} (——), α_{URE} (——).

Efficiency simulations

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

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

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

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

Efficiency simulations - results

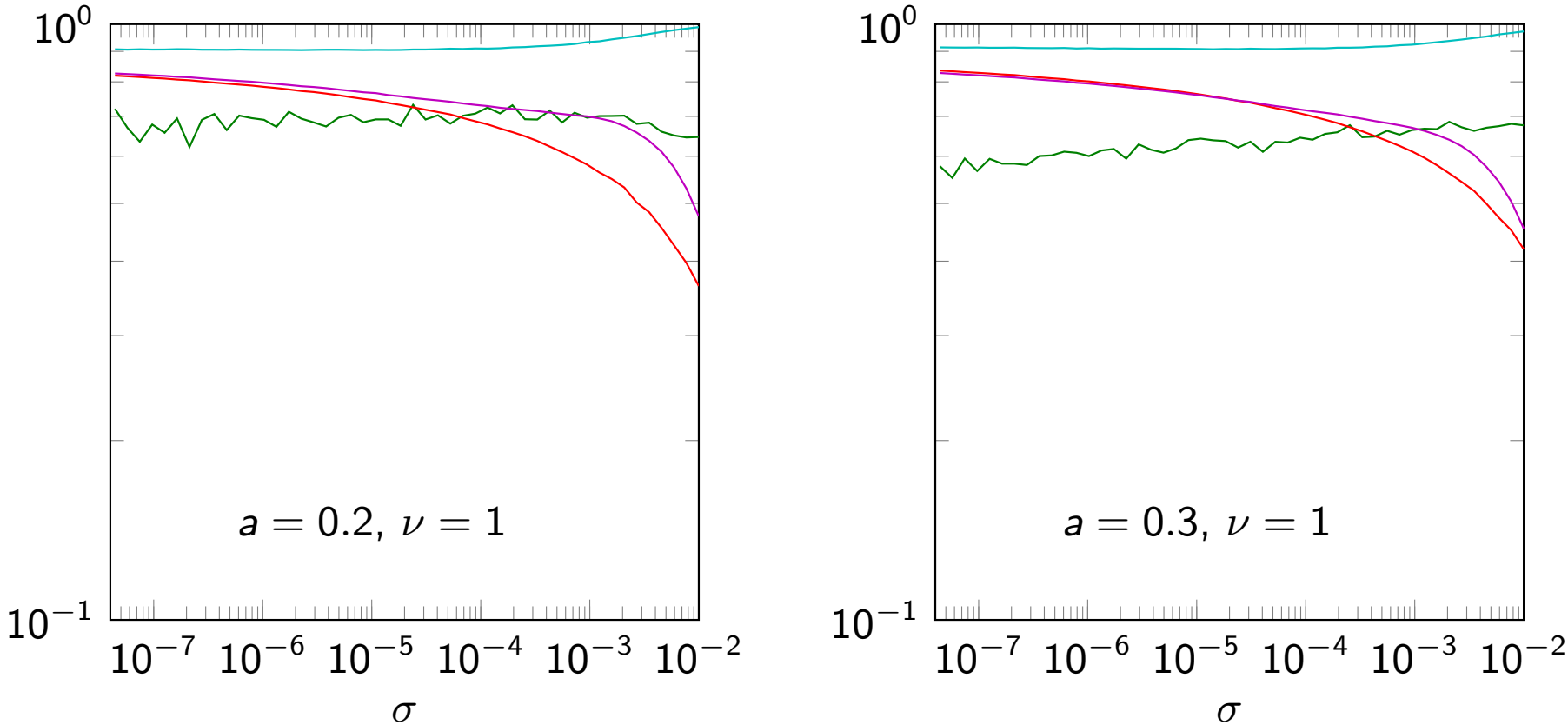


Figure: R_{QO} (——), R_{DP} (——), R_{LEP} (——), and R_{URE} (——)

Efficiency simulations - results

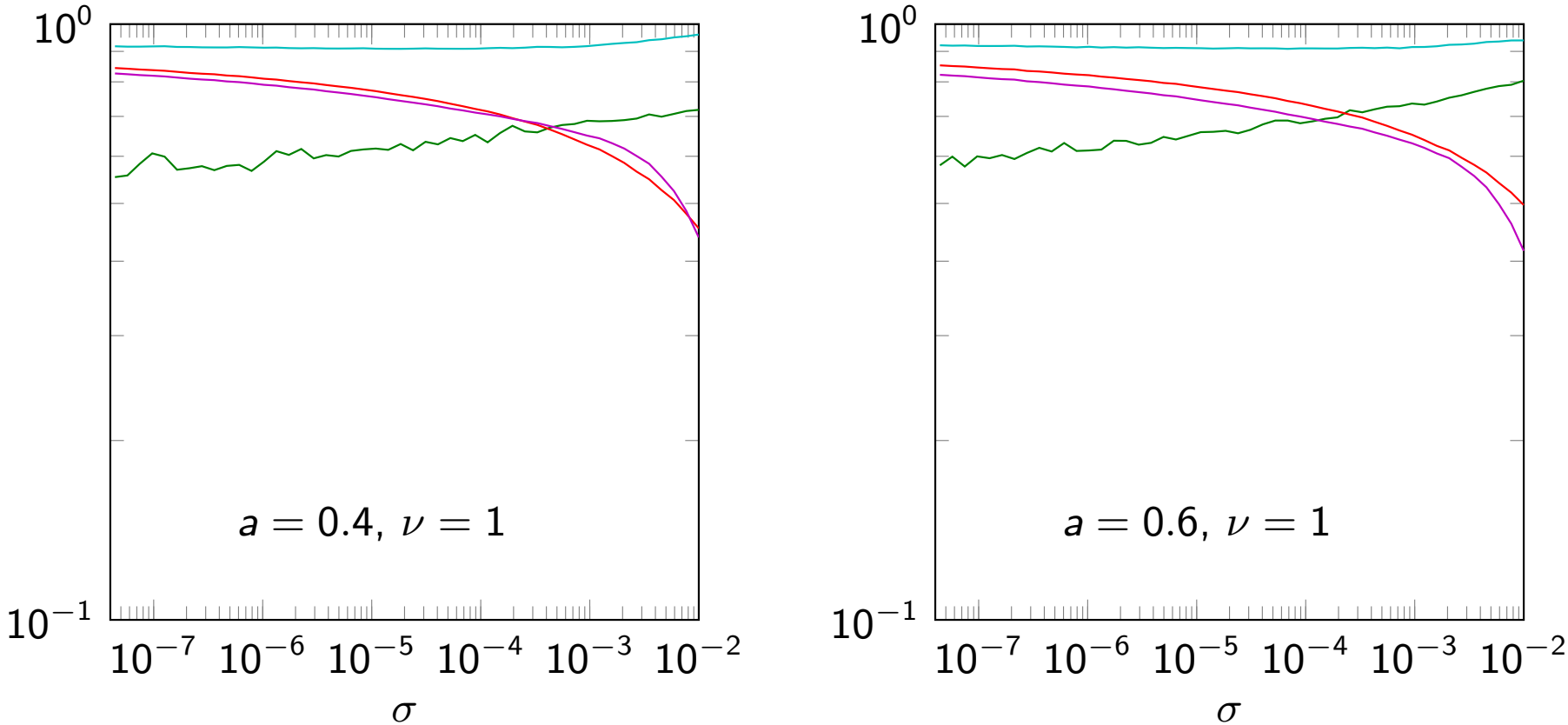


Figure: R_{QO} (——), R_{DP} (——), R_{LEP} (——), and R_{URE} (——)

Efficiency simulations - results

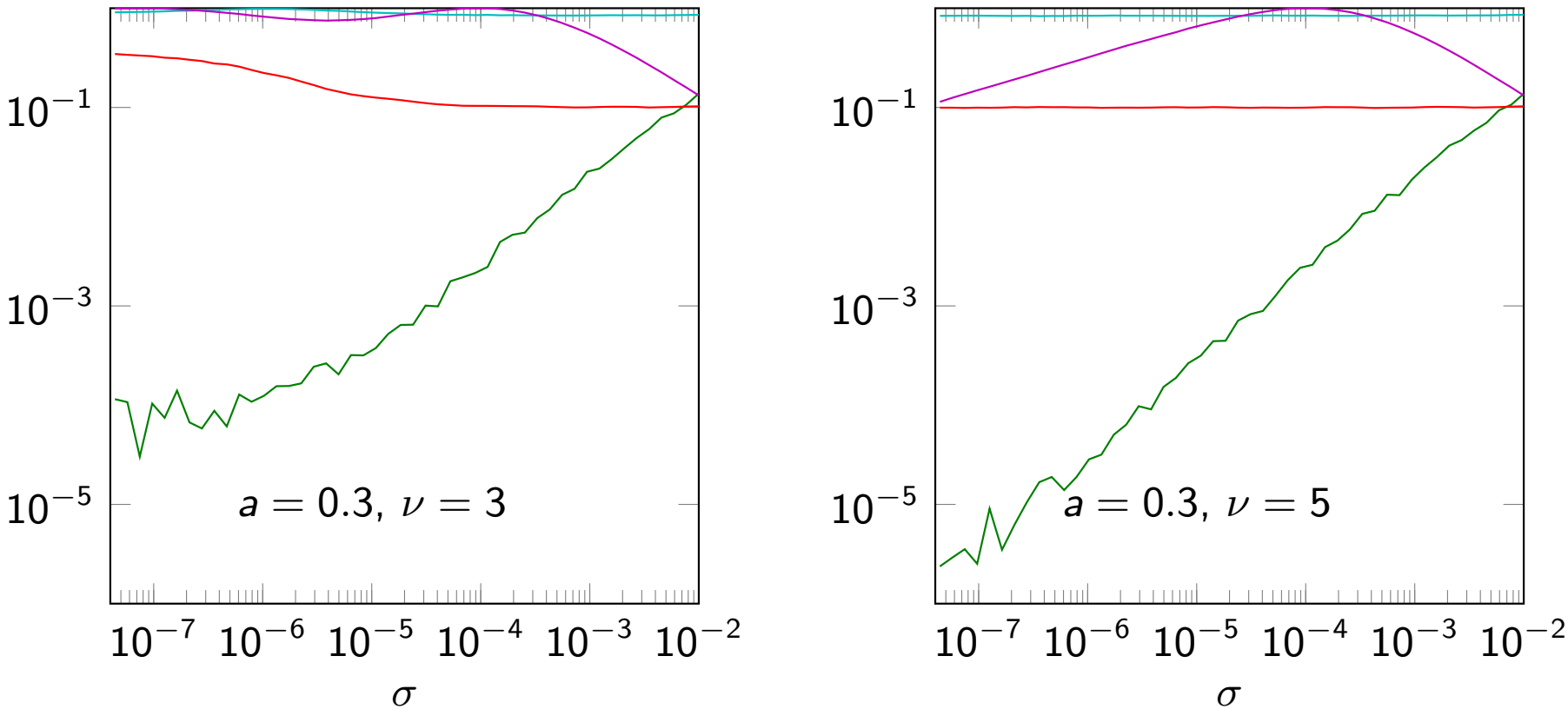


Figure: R_{QO} (——), R_{DP} (——), R_{LEP} (——), and R_{URE} (——)

Outline

Recall: Oracle inequality

$$\gamma_\sigma := \frac{R_{\mathrm{w}}(\alpha_{o,\mathrm{w}}, f)}{\sigma^2}$$

Theorem (Li & W. '18)

There are positive constants C_1, C_2 and C_3, independent of f and σ, such that

$$\mathbb{E}\left[\|\hat{f}_{\alpha_{\mathrm{URE}}} - f\|^2\right] \leq \psi^{-1}\Big(\sigma^2\big(2\gamma_\sigma + C_1\big)\Big) + \sigma^2 C_3\left(\frac{\gamma_\sigma + \sqrt{\gamma_\sigma}}{S^{-1}\left(2C_q\gamma_\sigma\right)} + C_2\right)$$

We will now develop a general methodology for proving such oracle inequalities based on three ingredients:

1. Bounds for generalized moments of $R_{\mathrm{w}}(\alpha_{\mathrm{URE}}, f)/\sigma^2$
2. Controlling the bias
3. A comparison for the occurring variance terms

Step 1: Bounds for generalized moments

Theorem (Kneip '94)

There are constants C'_ξ, C''_ξ such that for all $x \geq 0$ and all $f \in \mathcal{X}$

$$\mathbb{P}\left\{\frac{R_{\mathrm{w}}(\alpha_{\mathrm{URE}}, f)}{\sigma^2} - \gamma_\sigma \geq x\right\} \leq C'_\xi \exp\left(-C''_\xi \min\left\{\sqrt{x}, \frac{x}{\sqrt{\gamma_\sigma}}\right\}\right).$$

Bounds for generalized moments follow as a Corollary:

$$\begin{aligned}
\mathbb{E}\left[\Psi\left(\frac{R_{\mathrm{w}}(\alpha_{\mathrm{URE}}, f)}{\sigma^2}\right)\right] &= \int_0^\infty \Psi'(x)\, \mathbb{P}\left\{\frac{R_{\mathrm{w}}(\alpha_{\mathrm{URE}}, f)}{\sigma^2} \geq x\right\} \mathrm{d}x \\
&\leq \int_0^{2\gamma_\sigma} \Psi'(x)\, \mathrm{d}x + C'_\xi \int_{2\gamma_\sigma}^\infty \Psi'(x) \exp\left(-C''_\xi \sqrt{x - \gamma_\sigma}\right) \mathrm{d}x \\
&\leq \Psi(2\gamma_\sigma) + C'_\xi \int_0^\infty \Psi'(x) \exp\left(-C''_\xi \sqrt{\frac{x}{2}}\right) \mathrm{d}x \\
&= \Psi(2\gamma_\sigma) + C_\Psi.
\end{aligned}$$

Step 2: Controlling the bias

Recall the assumptions:

- $f = \phi(T^*T)\, w$, $\|w\|_{\mathcal{X}} \leq 1$, i.e. $f_k \leq \phi(\lambda_k)$
- $\psi(\lambda) = \lambda \phi^{-1}\left(\sqrt{\lambda}\right)$ is convex, $c\psi(x) \leq \psi(cx)$ for $c \in [0,1]$
- No qualification condition so far!

For fixed α, the bias $\|r_\alpha(T^*T) f\|_{\mathcal{X}} = \sum_{k=1}^{\infty} r_\alpha(\lambda_k)^2 f_k^2$ can be controlled as follows:

$$\psi\left(\sum_{k=1}^{n} r_\alpha(\lambda_k)^2 f_k^2\right) \leq \psi\left(\sum_{k=1}^{n} f_k^2 \frac{r_\alpha(\lambda_k)^2}{1+\sum_{i=1}^{n} r_\alpha(\lambda_i)^2}\right)\left(1+\sum_{i=1}^{n} r_\alpha(\lambda_i)^2\right)$$

$$\leq \sum_{k=1}^{n} \psi\left(f_k^2\right) r_\alpha(\lambda_k)^2 \leq \sum_{k=1}^{n} f_k^2 \phi^{-1}(f_k)\, r_\alpha(\lambda_k)^2 \leq \sum_{k=1}^{\infty} \lambda_k f_k^2 r_\alpha(\lambda_k)^2$$

This shows $\|r_\alpha(T^*T) f\|_{\mathcal{X}}^2 \leq \psi^{-1}\left(\|r_\alpha(T^*T)\, Tf\|_{\mathcal{Y}}^2\right)$.

Step 3: Comparison of the variance terms

Recall the assumptions:

- The counting function Σ is well-approximated by S
- We have $\int_0^\alpha S(t)\,\mathrm{d}t \leq C_S \alpha S(\alpha)$
- $\alpha q_\alpha(\alpha) \geq c_q$ as $\alpha \searrow 0$

For fixed α, the term $\operatorname{tr}\left(q_\alpha(T^*T)^2\, T^*T\right) = -\int_0^\infty t q_\alpha(t)^2\,\mathrm{d}\Sigma(t)$ can be controlled as follows:

$$\begin{aligned}
-\int_0^\infty t q_\alpha(t)^2\,\mathrm{d}\Sigma(t) &\leq -\left(\frac{C_q'}{\alpha}\right)^2 \int_0^\alpha t\,\mathrm{d}\Sigma(t) - \left(C_q''\right)^2 \int_\alpha^\infty \frac{1}{t}\,\mathrm{d}\Sigma(t) \\
&\leq \frac{C}{\alpha^2}\int_0^\alpha \Sigma(t)\,\mathrm{d}t \\
&\leq \frac{C'}{\alpha^2}\int_0^\alpha S(t)\,\mathrm{d}t \\
&\leq C' C_S \frac{S(\alpha)}{\alpha}
\end{aligned}$$

Analogously one derives $\operatorname{tr}\left(q_\alpha(T^*T)^4\,(T^*T)^2\right) \leq C'' C_S \frac{S(\alpha)}{\alpha^2}$.

Step 3: Comparison of the variance terms (cont')

But furthermore

$$\begin{aligned} S(\alpha) \lesssim \Sigma(\alpha) &= -\int_{\alpha}^{\infty} \mathrm{d}\Sigma(t) \\ &\leq -\int_{\alpha}^{\infty} \alpha^2 q_\alpha(\alpha)^2 \, \mathrm{d}\Sigma(t) \\ &\leq -\int_{0}^{\infty} t^2 q_\alpha(t)^2 \, \mathrm{d}\Sigma(t) \\ &= \operatorname{tr}\left(\left(q_\alpha(T^*T)\, T^*T \right)^2 \right). \end{aligned}$$

This implies

$$\operatorname{tr}\left(q_\alpha(T^*T)^2\, T^*T \right) \lesssim \Psi_1\left(\operatorname{tr}\left(\left(q_\alpha(T^*T)\, T^*T \right)^2 \right) \right),$$
$$\operatorname{tr}\left(q_\alpha(T^*T)^4\, T^*T \right) \lesssim \Psi_2\left(\operatorname{tr}\left(\left(q_\alpha(T^*T)\, T^*T \right)^2 \right) \right),$$

with $\Psi_1(x) = x/S^{-1}(x)$ and $\Psi_2(x) = x/(S^{-1}(x))^2$.

Golubev's lemma

Lemma (Golubev '10)

Given $c_k : \mathcal{A} \to \mathbb{R}$, $k = 1, 2, \ldots$, *with* $\mathcal{A} \subseteq \mathbb{R}_+$, *define*

$$\zeta(\alpha) := \sum_{k=1}^{\infty} c_k(\alpha)(\xi_k^2 - 1) \quad \text{and} \quad \kappa(\alpha)^2 := \sum_{k=1}^{\infty} c_k(\alpha)^2$$

for $\xi_k \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.
If κ *is well-defined, continuous and strictly monotone, then for some* $C > 0$

$$\mathbb{E}\left[\sup_{\alpha \in \mathcal{A}} \left(\zeta(\alpha) - x\kappa(\alpha)^2\right)_+\right] \leq \frac{C}{x} \qquad \text{for all } x > 0.$$

Proof of the oracle inequality

Bias-variance decomposition:

$$\begin{aligned}\mathbb{E}\left[\|\hat{f}_{\alpha_{\mathrm{URE}}} - f\|^2\right] &\leq 2\mathbb{E}\left[\|r_{\alpha_{\mathrm{URE}}}(T^*T)f\|_{\mathcal{X}}^2\right] + 2\sigma^2\mathbb{E}\left[\|q_{\alpha_{\mathrm{URE}}}(T^*T)T^*\xi\|^2\right] \\ &= 2\mathbb{E}\left[\|r_{\alpha_{\mathrm{URE}}}(T^*T)f\|_{\mathcal{X}}^2\right] + 2\sigma^2\mathbb{E}\left[\sum_{k=1}^{\infty}\lambda_k q_{\alpha_{\mathrm{URE}}}(\lambda_k)^2\xi_k^2\right].\end{aligned}$$

Bias term:

$$\begin{aligned}\mathbb{E}\left[\|r_\alpha(T^*T)f\|_{\mathcal{X}}^2\right] &\leq \mathbb{E}\left[\psi^{-1}\left(\|r_\alpha(T^*T)Tf\|_{\mathcal{Y}}^2\right)\right] && \text{[controlling the bias]} \\ &\leq \mathbb{E}\left[\psi^{-1}\left(R_{\mathrm{w}}(\alpha_{\mathrm{URE}}, f)\right)\right] \\ &\leq \psi^{-1}\left(\mathbb{E}\left[R_{\mathrm{w}}(\alpha_{\mathrm{URE}}, f)\right]\right) && \text{[Jensen's inequality]} \\ &\leq \psi^{-1}\left(2R_{\mathrm{w}}(\alpha_{o,\mathrm{w}}, f) + C_1\sigma^2\right) && \text{[moment bounds]}\end{aligned}$$

Proof of the oracle inequality (cont')

Variance term:

$$\mathbb{E}\left[\sum_{k=1}^{\infty}\lambda_k q_{\alpha_{\mathrm{URE}}}(\lambda_k)^2\xi_k^2\right] = \mathbb{E}\left[\sum_{k=1}^{\infty}\lambda_k q_{\alpha_{\mathrm{URE}}}(\lambda_k)^2\right] + \mathbb{E}\left[\sum_{k=1}^{\infty}\lambda_k q_{\alpha_{\mathrm{URE}}}(\lambda_k)^2(\xi_k^2-1)\right]$$

$$= \underbrace{\mathbb{E}\left[\mathrm{tr}\left(q_{\alpha_{\mathrm{URE}}}\left(T^*T\right)T^*T\right)\right]}_{\text{Term A}} + \underbrace{\mathbb{E}\left[\sum_{k=1}^{\infty}\lambda_k q_{\alpha_{\mathrm{URE}}}(\lambda_k)^2(\xi_k^2-1)\right]}_{\text{Term B}}$$

Term A:

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{tr}\left(q_{\alpha_{\mathrm{URE}}}\left(T^*T\right)^2T^*T\right)\right] \\
\lesssim&\mathbb{E}\left[\Psi_1\left(\mathrm{tr}\left(\left(q_{\alpha_{\mathrm{URE}}}\left(T^*T\right)T^*T\right)^2\right)\right)\right] && \text{[variance comparison]} \\
\leq&\mathbb{E}\left[\Psi_1\left(\frac{\mathbb{R}_{\mathrm{w}}\left(\alpha_{\mathrm{URE}},f\right)}{\sigma^2}\right)\right] \\
\lesssim&\Psi_1\left(2\gamma_\sigma\right)+C_{\Psi_1} && \text{[moment bounds]} \\
=&\frac{\gamma_\sigma}{S^{-1}\left(\gamma_\sigma\right)}+C_{\Psi_1}
\end{aligned}$$

Proof of the oracle inequality (cont')

Term B: Apply Golubev's lemma on ordered processes with $\zeta(\alpha) := \sum_{k=1}^{\infty} \lambda_k q_\alpha(\lambda_k)^2(\xi_k^2 - 1)$, $\kappa(\alpha)^2 := \sum_{k=1}^{\infty} \lambda_k^2 q_\alpha(\lambda_k)^4$:

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{\infty} \lambda_k q_{\alpha_{\mathrm{URE}}}(\lambda_k)^2(\xi_k^2 - 1)\right] &\le \mathbb{E}\left[\left(\zeta(\alpha_{\mathrm{URE}}) - x\kappa(\alpha_{\mathrm{URE}})^2\right)_+\right] + \mathbb{E}\left[x\kappa(\alpha_{\mathrm{URE}})^2\right] \\
&\le \mathbb{E}\left[\sup_{\alpha\in\mathcal{A}}\left(\zeta(\alpha) - x\kappa(\alpha)^2\right)_+\right] + \mathbb{E}\left[x\kappa(\alpha_{\mathrm{URE}})^2\right] \\
&\le \frac{C}{x} + x\mathbb{E}\left[\sum_{k=1}^{\infty} \lambda_k^2 q_{\alpha_{\mathrm{URE}}}(\lambda_k)^4\right] \qquad \text{for all } x > 0.
\end{aligned}$$

Minimizing over $x > 0$ gives

$$\mathbb{E}\left[\sum_{k=1}^{\infty} \lambda_k q_{\alpha_{\mathrm{URE}}}(\lambda_k)^2(\xi_k^2 - 1)\right] \lesssim \left(\mathbb{E}\left[\sum_{k=1}^{\infty} \lambda_k^2 q_{\alpha_{\mathrm{URE}}}(\lambda_k)^4\right]\right)^{1/2}$$

Proof of the oracle inequality (cont')

Continue with Term B:

$$
\begin{aligned}
&\left(\mathbb{E}\left[\sum_{k=1}^{\infty} \lambda_k^2 q_{\alpha_{\mathrm{URE}}}(\lambda_k)^4\right]\right)^{1/2} \\
\lesssim &\left(\mathbb{E}\left[\Psi_2\left(\operatorname{tr}\left(\left(q_{\alpha_{\mathrm{URE}}}\left(T^* T\right) T^* T\right)^2\right)\right)\right]\right)^{1/2} && \text{[variance comparison]} \\
\leq &\left(\mathbb{E}\left[\Psi_2\left(\frac{\mathbb{R}_{\mathrm{w}}\left(\alpha_{\mathrm{URE}}, f\right)}{\sigma^2}\right)\right]\right)^{1/2} \\
\lesssim &\left(\Psi_2\left(2\gamma_\sigma\right) + C_{\Psi_2}\right)^{1/2} && \text{[moment bounds]} \\
= &\frac{\sqrt{\gamma_\sigma}}{S^{-1}\left(\gamma_\sigma\right)} + \sqrt{C_{\Psi_2}}
\end{aligned}
$$

Recall: Convergence rates

Theorem (Li & W. '18)

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* $C_1, C_2, C_3 > 0$ *independent of* σ *such that*

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

as $\sigma \searrow 0$.

Proof of the convergence rates theorem

By a priori results with $\alpha_* \phi(\alpha_*)^2 = \sigma^2 S(\alpha_*)$ under the posed qualification condition:

$$\gamma_\sigma = \frac{R_{\mathrm{w}}(\alpha_{o,\mathrm{w}}, f)}{\sigma^2} \leq \frac{R_{\mathrm{w}}(\alpha_*, f)}{\sigma^2} \lesssim S(\alpha_*).$$

Thus the oracle inequality gives:

$$\begin{aligned}
\sup_{f \in \mathcal{W}_\phi} \mathbb{E}\left[\|\hat{f}_{\alpha_{\mathrm{URE}}} - f\|^2\right] &\lesssim \sup_{f \in \mathcal{W}_\phi} \psi^{-1}\left(\sigma^2\left(2\gamma_\sigma + C_1\right)\right) + \sigma^2 \frac{\gamma_\sigma + \sqrt{\gamma_\sigma}}{S^{-1}(2C_q\gamma_\sigma)} \\
&\lesssim \psi^{-1}\left(\sigma^2\left(2S(\alpha_*) + C_1\right)\right) + \sigma^2 \frac{S(\alpha_*) + \sqrt{S(\alpha_*)}}{S^{-1}(2C_q S(\alpha_*))} \\
&\lesssim \psi^{-1}\left(\alpha_* \phi(\alpha_*)^2\right) + \frac{\sigma^2 S(\alpha_*)}{S^{-1}(C_2 C_q S(\alpha_*))} \\
&= \phi(\alpha_*)^2 + \frac{\sigma^2 S(\alpha_*)}{S^{-1}(C_2 C_q S(\alpha_*))} \\
&= \sigma^2 \frac{S(\alpha_*)}{\alpha_*} + \frac{\sigma^2 S(\alpha_*)}{S^{-1}(C_2 C_q S(\alpha_*))}
\end{aligned}$$

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
- Proofs:
  - general methodology to prove (strong-to-weak) oracle inequalities

Thank you for your attention!