Regularization independent of the noise level: an analysis of quasi-optimality

Markus Reiß*

Institute of Mathematics
Humboldt University Berlin
www.math.hu-berlin.de/∼mreiss

Conference *Applied Inverse Problems*
Vienna, 21 July 2009

*joint with F. Bauer

Outline

Idea

Prototypical inverse problem

$$\boxed{y = Ax + \xi}$$

with operator $A : H \to H$ self-adjoint, positive and compact;
A has singular basis (u_k, λ_k) with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \downarrow 0$.

Truncated SVD estimators:

$$\boxed{\hat{x}^{(n)} := \sum_{k=1}^{\ell(n)} \lambda_k^{-1} \langle y, u_k \rangle u_k}$$

with cut-off function $\ell : \mathbb{N} \to \mathbb{N}$ ($\ell(n+1) > \ell(n)$).

Quasi-optimality criterion for parameter n:

Look where $\hat{x}^{(n)}$ stabilizes:

$$\boxed{n^* := \operatorname{argmin}_n \|\hat{x}^{(n+1)} - \hat{x}^{(n)}\|}$$

Idea

Prototypical inverse problem

$$\boxed{y = Ax + \xi}$$

with operator $A : H \to H$ self-adjoint, positive and compact;
A has singular basis (u_k, λ_k) with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \downarrow 0$.

Truncated SVD estimators:

$$\boxed{\hat{x}^{(n)} := \sum_{k=1}^{\ell(n)} \lambda_k^{-1} \langle y, u_k \rangle u_k}$$

with cut-off function $\ell : \mathbb{N} \to \mathbb{N}$ ($\ell(n+1) > \ell(n)$).

Quasi-optimality criterion for parameter n:

Look where $\hat{x}^{(n)}$ stabilizes:

$$\boxed{n^* := \operatorname{argmin}_n \|\hat{x}^{(n+1)} - \hat{x}^{(n)}\|}$$

Idea

Prototypical inverse problem

$$\boxed{y = Ax + \xi}$$

with operator $A : H \to H$ self-adjoint, positive and compact;
A has singular basis (u_k, λ_k) with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \downarrow 0$.

Truncated SVD estimators:

$$\boxed{\hat{x}^{(n)} := \sum_{k=1}^{\ell(n)} \lambda_k^{-1} \langle y, u_k \rangle u_k}$$

with cut-off function $\ell : \mathbb{N} \to \mathbb{N}$ ($\ell(n+1) > \ell(n)$).

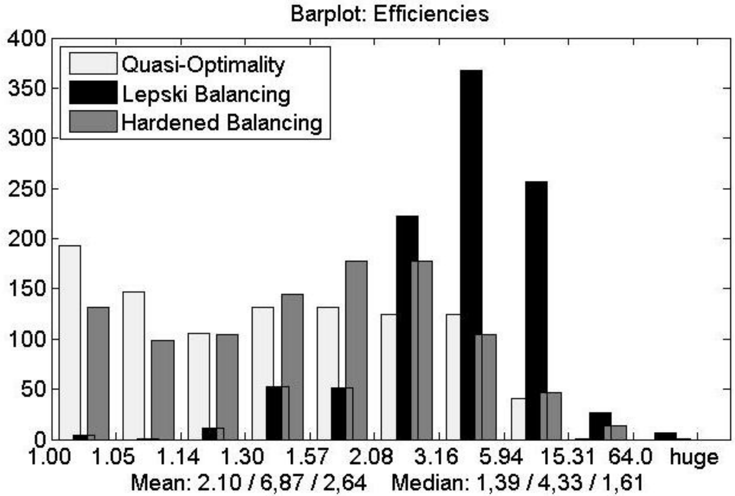
Quasi-optimality criterion for parameter n:

Look where $\hat{x}^{(n)}$ stabilizes:

$$\boxed{n^* := \operatorname{argmin}_n \|\hat{x}^{(n+1)} - \hat{x}^{(n)}\|}$$

Practical performance

Efficiency: $\frac{\|x^{(n*)}-x\|}{\min_n \|x^{(n)}-x\|}$



(Inverse problem in option calibration for mathematical finance)

The model

$$\boxed{y = Ax + \xi \text{ resp. } y_k = \lambda_k x_k + \xi_k \text{ resp. } \tilde{x}_k = x_k + \lambda_k^{-1}\xi_k}$$

Statistical noise ξ, here: $\xi_k \sim N(0, \varepsilon_k^2)$ independent.
Assume $\sigma_k := \varepsilon_k \lambda_k^{-1} \to \infty$ (*ill-posed* problem):

$$\boxed{\tilde{x}_k = x_k + \sigma_k \eta_k, \quad \eta_k \sim N(0,1)}$$

Bias-variance decomposition of mean squared error (MSE):

$$\mathbb{E}[\|x^{(n)} - x\|^2] = \mathbb{E}\left[\sum_{k=1}^{\ell(n)} (\tilde{x}_k - x_k)^2 + \sum_{k=\ell(n)+1}^{\infty} x_k^2\right]$$

$$= \underbrace{\sum_{k=1}^{\ell(n)} \sigma_k^2}_{\text{Variance } s(n)} + \underbrace{\sum_{k=\ell(n)+1}^{\infty} x_k^2}_{\text{Bias}^2 \beta(n)}$$

The model

$$\boxed{y = Ax + \xi \text{ resp. } y_k = \lambda_k x_k + \xi_k \text{ resp. } \tilde{x}_k = x_k + \lambda_k^{-1}\xi_k}$$

Statistical noise ξ, here: $\xi_k \sim N(0, \varepsilon_k^2)$ independent.
Assume $\sigma_k := \varepsilon_k \lambda_k^{-1} \to \infty$ (*ill-posed* problem):

$$\boxed{\tilde{x}_k = x_k + \sigma_k \eta_k, \quad \eta_k \sim N(0,1)}$$

Bias-variance decomposition of mean squared error (MSE):

$$\mathbb{E}[\|x^{(n)} - x\|^2] = \mathbb{E}\left[\sum_{k=1}^{\ell(n)} (\tilde{x}_k - x_k)^2 + \sum_{k=\ell(n)+1}^{\infty} x_k^2\right]$$

$$= \underbrace{\sum_{k=1}^{\ell(n)} \sigma_k^2}_{\text{Variance } s(n)} + \underbrace{\sum_{k=\ell(n)+1}^{\infty} x_k^2}_{\text{Bias}^2 \beta(n)}$$

Estimating the error?

$$\boxed{\mathbb{E}[\|x^{(n)} - x\|^2] = s(n) + \beta(n)}$$

On the other hand, **in the mean**:

$$\boxed{\mathbb{E}[\|x^{(n+1)} - x^{(n)}\|^2] = s(n+1) - s(n) + \beta(n) - \beta(n+1)}$$

Choose the cut-off function $\ell(\bullet)$ for geometrically growing variance:

$$\exists 1 < c_s \leqslant C_s : \; c_s s(n) < s(n+1) < C_s s(n)$$

(typical: $\ell(n) = \lfloor q^n \rfloor$ for white noise and polynomial d.o.i.)
Hence: $s(n+1) - s(n) \asymp s(n)$ and we have

$$\mathbb{E}[\|x^{(n+1)} - x^{(n)}\|^2] \asymp \mathbb{E}[\|x^{(n)} - x\|^2] \text{ if } \beta(n) - \beta(n+1) \asymp \beta(n).$$

Estimating the error?

$$\boxed{\mathbb{E}[\|x^{(n)} - x\|^2] = s(n) + \beta(n)}$$

On the other hand, **in the mean**:

$$\boxed{\mathbb{E}[\|x^{(n+1)} - x^{(n)}\|^2] = s(n+1) - s(n) + \beta(n) - \beta(n+1)}$$

Choose the cut-off function $\ell(\bullet)$ for geometrically growing variance:

$$\exists 1 < c_s \leqslant C_s : \; c_s s(n) < s(n+1) < C_s s(n)$$

(typical: $\ell(n) = \lfloor q^n \rfloor$ for white noise and polynomial d.o.i.)
Hence: $s(n+1) - s(n) \asymp s(n)$ and we have

$$\mathbb{E}[\|x^{(n+1)} - x^{(n)}\|^2] \asymp \mathbb{E}[\|x^{(n)} - x\|^2] \text{ if } \beta(n) - \beta(n+1) \asymp \beta(n).$$

Average case scenario

What functions x satisfy $\beta(n) - \beta(n+1) \asymp \beta(n)$?
Typical setting: $x_k^2 \asymp k^{-2\alpha}$ for some $\alpha > 1/2$ suffices.

A far more general average case analysis is possible:
Consider independent coefficients $x_k \sim N(0, \gamma_k^2)$:
(denote by $\tilde{\mathbb{E}}$ the joint mean w.r.t. to noise and prior law)

$$b(n) := \tilde{\mathbb{E}}[\beta(n)] = \tilde{\mathbb{E}}\Big[\sum_{k=\ell(n)+1}^{\infty} x_k^2\Big] = \sum_{k=\ell(n)+1}^{\infty} \gamma_k^2$$

Assumption on geometric decay of average bias:

$$\exists 1 < c_b \leqslant C_b : \; c_b b(n+1) < b(n) < C_b b(n+1)$$

Typical setting: $\gamma_k^2 \asymp k^{-2\alpha}$ for some $\alpha > 1/2$ suffices.

Average case scenario

What functions x satisfy $\beta(n) - \beta(n+1) \asymp \beta(n)$?
Typical setting: $x_k^2 \asymp k^{-2\alpha}$ for some $\alpha > 1/2$ suffices.

A far more general average case analysis is possible:
Consider independent coefficients $x_k \sim N(0, \gamma_k^2)$:
(denote by $\tilde{\mathbb{E}}$ the joint mean w.r.t. to noise and prior law)

$$b(n) := \tilde{\mathbb{E}}[\beta(n)] = \tilde{\mathbb{E}}\Big[\sum_{k=\ell(n)+1}^{\infty} x_k^2\Big] = \sum_{k=\ell(n)+1}^{\infty} \gamma_k^2$$

Assumption on geometric decay of average bias:

$$\exists 1 < c_b \leqslant C_b : \; c_b b(n+1) < b(n) < C_b b(n+1)$$

Typical setting: $\gamma_k^2 \asymp k^{-2\alpha}$ for some $\alpha > 1/2$ suffices.

Main result

Theorem

The quasi-optimality index n^* is almost surely well-defined:

$$\boxed{n^* := \operatorname{argmin}_n \|\hat{x}^{(n+1)} - \hat{x}^{(n)}\|}$$

and the data-driven estimator satisfies the *oracle inequality*

$$\boxed{\tilde{\mathbb{E}}\left[\|\hat{x}^{(n*)} - x\|^{\alpha}\right]^{1/\alpha} \leqslant K \min_{n \geqslant 1} \tilde{\mathbb{E}}\left[\|\hat{x}^{(n)} - x\|^{2}\right]^{1/2}}$$

with a constant $K = K(c_s, c_b, C_s, C_b, r, \alpha) > 0$ for all

$$0 < \alpha < \frac{r}{\max(\frac{\log(C_b)}{\log(c_b)}, \frac{\log(C_s)}{\log(c_s)})}.$$

What moments α do we attain?

$$\boxed{\tilde{\mathbb{E}}\left[\|\hat{x}^{(n*)} - x\|^{\alpha}\right]^{1/\alpha} \leqslant K \min_{n \geqslant 1} \tilde{\mathbb{E}}\left[\|\hat{x}^{(n)} - x\|^{2}\right]^{1/2}}$$

for all $0 < \alpha < r/\max(\frac{\log(C_b)}{\log(c_b)}, \frac{\log(C_s)}{\log(c_s)})$ with

$$r := \inf_{n \geqslant 1} \frac{|(s(n+1) - s(n)) + (b(n) - b(n+1))|}{\max_{\ell(n) < k \leqslant \ell(n+1)}(\sigma_k^2 + \gamma_k^2)} \geqslant 1.$$

Example

ξ white noise, $\lambda_k \asymp k^{-\nu}$, $\ell(n) = \lfloor q^n \rfloor$, $\gamma_k \asymp k^{-\mu}$
(hence: $c_s = C_s$, $c_b = C_b$)
Then: $r \geqslant (\ell(2) - \ell(1)) \min((2\nu + 1)^{-1}, (2\mu + 1)^{-1})$
We attain the higher moments α, the larger $q > 1$.

What moments α do we attain?

$$\boxed{\tilde{\mathbb{E}}\left[\|\hat{x}^{(n*)} - x\|^{\alpha}\right]^{1/\alpha} \leqslant K \min_{n\geqslant 1} \tilde{\mathbb{E}}\left[\|\hat{x}^{(n)} - x\|^{2}\right]^{1/2}}$$

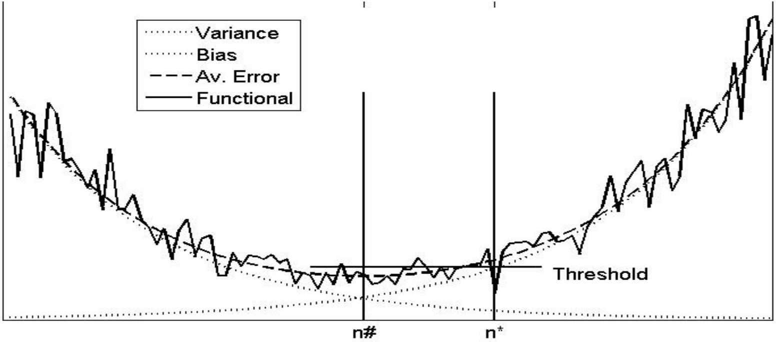
for all $0 < \alpha < r/\max(\frac{\log(C_b)}{\log(c_b)}, \frac{\log(C_s)}{\log(c_s)})$ with

$$r := \inf_{n\geqslant 1} \frac{|(s(n+1) - s(n)) + (b(n) - b(n+1))|}{\max_{\ell(n)<k\leqslant \ell(n+1)}(\sigma_k^2 + \gamma_k^2)} \geqslant 1.$$

Example

ξ white noise, $\lambda_k \asymp k^{-\nu}$, $\ell(n) = \lfloor q^n \rfloor$, $\gamma_k \asymp k^{-\mu}$
(hence: $c_s = C_s$, $c_b = C_b$)
Then: $r \geqslant (\ell(2) - \ell(1)) \min((2\nu + 1)^{-1}, (2\mu + 1)^{-1})$
We attain the higher moments α, the larger $q > 1$.

Idea of proof



1. Compare n^* with index $n^\#$ such that $s(n^\# + 1) > b(n^\# + 1)$, but $s(n^\#) \leqslant b(n^\#)$.
2. Derive deviation bounds for generalized χ^2-distribution at $+\infty$ and(!) 0.
3. Bound uniformly the deviations from the mean errors $s(n)$ resp. $b(n)$.

Discussion and summary

- Heuristic quasi-optimality criterion works well in practice.
- Counterexamples à la Bakushinskii show that it does not always work.
- The quasi-optimality criterion estimates the MSE for geometric $s(n)$ and $b(n)$ without using the noise level.
- Average case analysis with geometrically decaying weights in the coefficients.
- We obtain a general oracle-type inequality.
- The sparser the subsampling $\ell(\bullet)$, the higher the moments we achieve.
- Generalisation to other methods than TSVD seem feasible.

Thank you very much for your attention!

Discussion and summary

- Heuristic quasi-optimality criterion works well in practice.
- Counterexamples à la Bakushinskii show that it does not always work.
- The quasi-optimality criterion estimates the MSE for geometric $s(n)$ and $b(n)$ without using the noise level.
- Average case analysis with geometrically decaying weights in the coefficients.
- We obtain a general oracle-type inequality.
- The sparser the subsampling $\ell(\bullet)$, the higher the moments we achieve.
- Generalisation to other methods than TSVD seem feasible.

Thank you very much for your attention!