siam 2026 Conference on Optimization

Why the noise model matters

A performance gap in learned regularization

Sebastian Banert, Christoph Brauer, Dirk Lorenz, Lionel Tondji

DLR Institute of Lightweight Systems

Motivation: Learning to regularize in a supervised fashion

Institute of Lightweight Systems

Task

Recover ground truth $x^\dagger \in \mathbb{R}^n$
from noisy observation $y \in \mathbb{R}^m$

use $\mathcal{R}$

Bilevel problem (empirical risk)

$$\text{minimize}_{\mathcal{R}} \quad \frac{1}{m}\sum_{i=1}^{m} \|\hat{x}(y_i) - x_i^\dagger\|^2$$
$$\text{s.t.} \quad \hat{x}(y_i) \in \operatorname{argmin}_x \ \mathcal{D}(Ax, y_i) + \mathcal{R}(x)$$

via

via

Variational problem

$$\text{minimize}_x \quad \underbrace{\mathcal{D}(Ax, y)}_{\text{Discrepancy term}} + \underbrace{\mathcal{R}(x)}_{\text{Regularizer}}$$

learn $\mathcal{R}$ from

Paired training data

$$\{(x_1^\dagger, y_1), \ldots, (x_m^\dagger, y_m)\}$$

Motivation: Learning to regularize in a supervised fashion

Institute of Lightweight Systems

Task

Recover ground truth $x^\dagger \in \mathbb{R}^n$
from noisy observation $y \in \mathbb{R}^m$

use $\mathcal{R}$

Bilevel problem (true risk)

$$\text{minimize}_{\mathcal{R}} \quad \mathbb{E}\, \|\hat{x}(y) - x^\dagger\|^2$$
$$\text{s.t.} \quad \hat{x}(y) \in \operatorname{argmin}_x \; \mathcal{D}(Ax, y) + \mathcal{R}(x)$$

via

via

Variational problem

$$\text{minimize}_x \quad \underbrace{\mathcal{D}(Ax, y)}_{\text{Discrepancy term}} + \underbrace{\mathcal{R}(x)}_{\text{Regularizer}}$$

learn $\mathcal{R}$ from

Uncorrelated random variables

$x^\dagger$ with $\mathbb{E}(x^\dagger) = \mu_{x^\dagger}$, $\operatorname{Cov}(x^\dagger) = \Sigma_{x^\dagger}$
ε with $\mathbb{E}(\varepsilon) = 0$, $\operatorname{Cov}(\varepsilon) = \Sigma_\varepsilon$
$y = Ax + \varepsilon$

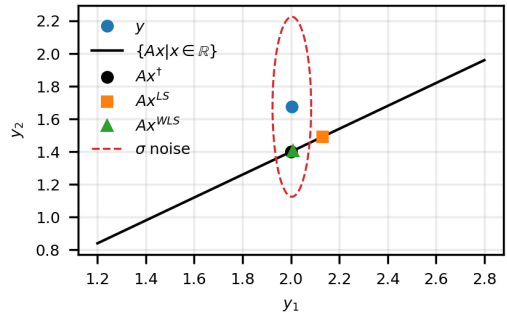
Outline

- Recap of noise weighting
- Six different variational regularization formulations:

  Tikhonov, quadratic, Lavrentiev – with and without noise weighting + LMMSE estimator
- Investigation of possible performance gaps between methods
- Numerical experiments

Noise weighting

- $\mathcal{D}(Ax, y) = \|Ax - y\|^2$ implicitly assumes $\varepsilon = y - Ax \sim \mathcal{N}(0, \sigma^2 I_M)$
- If $\varepsilon \sim \mathcal{N}(0, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2))$ then the ML discrepancy term is $\|\mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)^{-1}(Ax - y)\|^2$
- If ε has correlated components and $\varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon)$ then the ML discrepancy term is $\|\Sigma_\varepsilon^{-1}(Ax - y)\|^2$
- In the following, we consider $\|\Omega(Ax - y)\|^2 = \|Ax - y\|_\Omega^2$ for arbitrary symmetric and positive semidefinite Ω



Interpretation: Stronger penalty for residuals in components with low uncertainty / noise whitening

Types of regularization

$$\hat{x}_\theta^{\text{Method}}(y) = W^{\text{Method}} y + b^{\text{Method}}$$

Institute of Lightweight Systems

Method	Minimization problem	Map	Parameters θ
Tikhonov w/ noise weight	$\frac{1}{2}\|Ax-y\|_\Omega^2 + \frac{1}{2}\|R(x-x_0)\|^2$	$\hat{x}_\theta(y) = (A^T\Omega A + R^TR)^{-1}(A^T\Omega y + R^TRx_0)$ $W = (A^T\Omega A + R^TR)^{-1}A^T\Omega$ $b = (A^T\Omega A + R^TR)^{-1}R^TRx_0$	$\Omega \in \mathbb{S}^m_{\succ 0}$ $R \in \mathbb{R}^{n\times n}$ $x_0 \in \mathbb{R}^n$
Tikhonov w/o noise weight	$\frac{1}{2}\|Ax-y\|^2 + \frac{1}{2}\|R(x-x_0)\|^2$	$\hat{x}_\theta(y) = (A^TA + R^TR)^{-1}(A^Ty + R^TRx_0)$ $W = (A^TA + R^TR)^{-1}A^T$ $b = (A^TA + R^TR)^{-1}R^TRx_0$	$R \in \mathbb{R}^{n\times n}$ $x_0 \in \mathbb{R}^n$
Quadratic w/ noise weight	$\frac{1}{2}\|Ax-y\|_\Omega^2 + \frac{1}{2}\langle x-x_0, M(x-x_0)\rangle$	$\hat{x}_\theta(y) = (A^T\Omega A + M)^{-1}(A^T\Omega y + Mx_0)$ $W = (A^T\Omega A + M)^{-1}A^T\Omega$ $b = (A^T\Omega A + M)^{-1}Mx_0$	$\Omega \in \mathbb{S}^m_{\succ 0}$ $M \in \mathbb{S}^n$ $x_0 \in \mathbb{R}^n$
Quadratic w/o noise weight	$\frac{1}{2}\|Ax-y\|^2 + \frac{1}{2}\langle x-x_0, M(x-x_0)\rangle$	$\hat{x}_\theta(y) = (A^TA + M)^{-1}(A^Ty + Mx_0)$ $W = (A^TA + M)^{-1}A^T$ $b = (A^TA + M)^{-1}Mx_0$	$M \in \mathbb{S}^n$ $x_0 \in \mathbb{R}^n$
Lavrentiev w/ noise weight	-	$\hat{x}_\theta(y) = (A^T\Omega A + M)^{-1}(A^T\Omega y + Mx_0)$ $W = (A^T\Omega A + M)^{-1}A^T\Omega$ $b = (A^T\Omega A + M)^{-1}Mx_0$	$\Omega \in \mathbb{S}^m_{\succ 0}$ $M \in \mathbb{R}^{n\times n}$ $x_0 \in \mathbb{R}^n$
Lavrentiev w/o noise weight	-	$\hat{x}_\theta(y) = (A^TA + M)^{-1}(A^Ty + Mx_0)$ $W = (A^TA + M)^{-1}A^T$ $b = (A^TA + M)^{-1}Mx_0$	$M \in \mathbb{R}^{n\times n}$ $x_0 \in \mathbb{R}^n$

First estimate of optimal risks

- Generalization of all methods: $x_\theta^{\mathrm{Aff}}(y) = W^{\mathrm{Aff}} y + b^{\mathrm{Aff}}$ with parameters W^{Aff} and b^{Aff}
- Optimal risk of each method: $R_{\mathrm{Method}} = \inf_\theta \ \mathbb{E}_{x^\dagger,\varepsilon} \ \|\hat{x}_\theta^{\mathrm{Method}}(Ax^\dagger + b) - x^\dagger\|^2$
- Then, by construction:

$$R_{\mathrm{Aff}} \leq \begin{cases} R_{\mathrm{Lav}} \leq R_{\mathrm{Quad}} \leq R_{\mathrm{Tikh}} \\ R_{\mathrm{Lav}(\Omega)} \leq R_{\mathrm{Quad}(\Omega)} \leq R_{\mathrm{Tikh}(\Omega)} \end{cases}$$

LMMSE vs. Tikh(Ω)

Institute of Lightweight Systems

Theorem (LMMSE Estimator)

$$\underset{W,b}{\text{minimize}} \ \mathbb{E}_{x^\dagger,\varepsilon}\|\hat{x}_\theta^{Aff}(Ax^\dagger+\varepsilon)-x^\dagger\|^2$$

is solved by

$$W = \Sigma_{x^\dagger}A^T(A\Sigma_{x^\dagger}A^T+\Sigma_\varepsilon)^{-1} \quad \textit{and}$$
$$b = (I-W^{Aff}A)\mu_{x^\dagger}.$$

Theorem (Optimal Tikh(Ω))

$$\underset{\Omega,R,x_0}{\text{minimize}} \ \mathbb{E}_{x^\dagger,\varepsilon}\|\hat{x}_\theta^{Tikh(\Omega)}(Ax^\dagger+\varepsilon)-x^\dagger\|^2$$

is solved by

$$\Omega = \Sigma_\varepsilon^{-1}, \quad R^TR = \Sigma_{x^\dagger}^{-1}, \quad x_0 = \mu_{x^\dagger}$$

and it holds that $\hat{x}_{\theta^*}^{Tikh} = \hat{x}_{\theta^*}^{Aff}$.

$$\implies R_{\text{Aff}} = R_{\text{Lav}(\Omega)} = R_{\text{Quad}(\Omega)} = R_{\text{Tikh}(\Omega)} \overset{<?}{\leq} R_{\text{Lav}} \overset{<?}{\leq} R_{\text{Quad}} \overset{<?}{\leq} R_{\text{Tikh}}$$

Lav vs. Quad

Institute of Lightweight Systems

Theorem (Optimal Lav)

$$\underset{M,x_0}{\text{minimize}}\ \mathbb{E}_{x^\dagger,\varepsilon}\|\hat{x}_\theta^{Lav}(Ax^\dagger+\varepsilon)-x^\dagger\|^2$$

is solved by

$$M = A^T\Sigma_\varepsilon A(A^TA)^{-1}\Sigma_{x^\dagger}^{-1} \quad \textit{and}$$
$$x_0 \in \mu_{x^\dagger} + \ker(M).$$

Theorem (Optimal Quad)

$$\underset{M,x_0}{\text{minimize}}\ \mathbb{E}_{x^\dagger,\varepsilon}\|\hat{x}_\theta^{Quad}(Ax^\dagger+\varepsilon)-x^\dagger\|^2$$

is solved by

$$M = N^{-1} - A^TA \quad \textit{and} \quad x_0 \in \mu_{x^\dagger} + \ker(M)$$

with $B = A^T(A\Sigma_{x^\dagger}A^T+\Sigma_\varepsilon)A$ and N being the unique symmetric solution of $A^TA\Sigma_{x^\dagger}+\Sigma_{x^\dagger}A^TA = NB+BN$.

$$\overset{\text{Possibly}}{\Longrightarrow} R_{\text{Aff}} = R_{\text{Lav}(\Omega)} = R_{\text{Quad}(\Omega)} = R_{\text{Tikh}(\Omega)} \overset{<?}{\leq} R_{\text{Lav}} \overset{M^{\text{Lav}}\text{ not sym.}}{<} R_{\text{Quad}} \overset{M^{\text{Quad}}\text{ not pos. def.}}{<} R_{\text{Tikh}}$$

Theorem

The following are equivalent:

1. *The best linear map for Lavrentiev regularization*

$$W^{\mathrm{Lav}} = (A^T A + A^T \Sigma_\varepsilon A (A^T A)^{-1} \Sigma_{x^\dagger}^{-1})^{-1} A^T$$

is equal to the linear map from the LMMSE estimator

$$W^{\mathrm{LMMSE}} = \Sigma_{x^\dagger} A^T (A \Sigma_{x^\dagger} A^T + \Sigma_\varepsilon)^{-1}.$$

2. *The noise covariance Σ_ε leaves the kernel of A^T invariant, i.e., exactly if $A^T \Sigma_\varepsilon P_{\ker(A^T)} = 0$.*

Examples with $A^T \Sigma_\varepsilon P_{\ker(A^T)} = 0$:

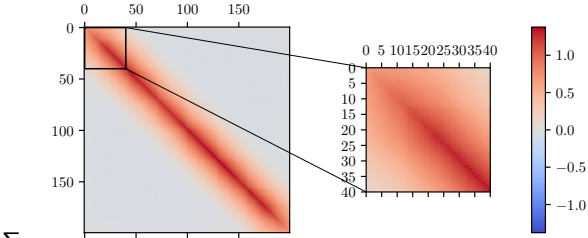
- $\ker(A^T) = \{0\}$
- $\ker(A) = \{0\}$ (here fulfilled for invertible $A \in \mathbb{R}^{n \times n}$)
- $\Sigma_\varepsilon = \sigma^2 I$
- $\Sigma_\varepsilon = \sigma^2 I + \tau^2 A A^T$ (e.g., when $y = A(x + \varepsilon_x) + \varepsilon_y$)

$$\overset{\text{Possibly}}{\Longrightarrow} R_{\mathrm{Aff}} = R_{\mathrm{Lav}(\Omega)} = R_{\mathrm{Quad}(\Omega)} = R_{\mathrm{Tikh}(\Omega)} < R_{\mathrm{Lav}} < R_{\mathrm{Quad}} < R_{\mathrm{Tikh}}$$

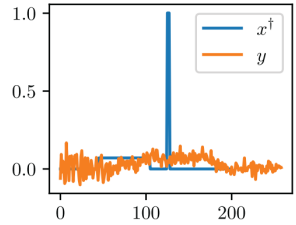
Deconvolution of plateau functions under structured noise

Institute of Lightweight Systems

- 50.000 discretized versions $x_i^\dagger$ of plateau functions $x_i(t)=\sum_{j=1}^{k}\left(a_j^2+0.01\right)\chi_{[c_j-b_j,\ c_j+b_j]}(t)$ with $n=200$ for training, 20.000 for testing
- $y_i = Ax_i^\dagger + \varepsilon_i \in \mathbb{R}^{259}$ with conv. matrix A, $\varepsilon_i \sim \mathcal{N}(0, \mathrm{diag}(\sigma_1^2,\ldots,\sigma_m^2))$ and decaying variance from $\sigma_1^2 = 10^{-2}$ to $\sigma_m^2 = 5\cdot 10^{-4}$



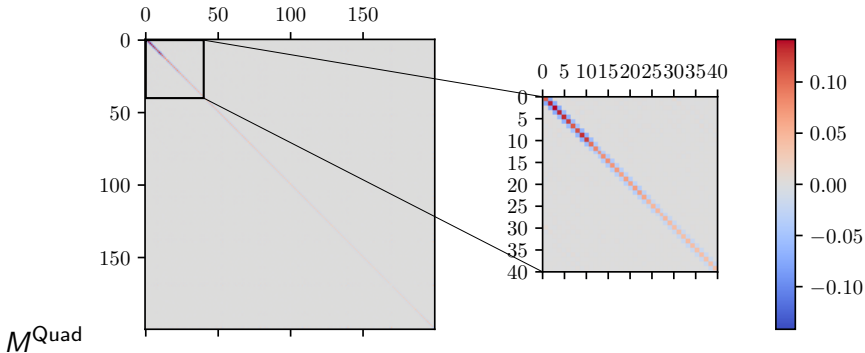
$\Sigma_{x^\dagger}$



Risks of theoretically optimal maps on test data:

R_{Aff}	23.12
R_{Lav}	23.23
R_{Quad}	23.50

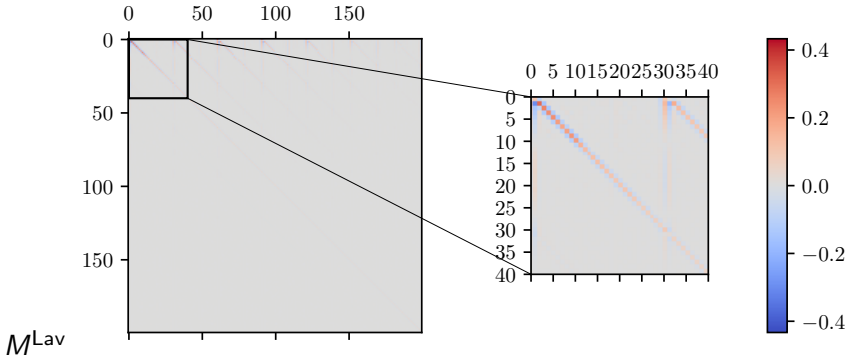
Deconvolution of plateau functions under structured noise



$\rightarrow$ **not positive semidefinite (smallest EV** ≈ -0.19**)**

Deconvolution of plateau functions under structured noise

Institute of Lightweight Systems



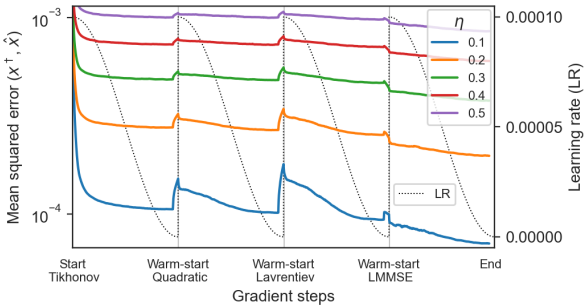
M^{Lav}

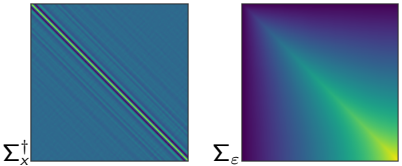
$\rightarrow$ **not symmetric**

Dereverberation of speech signals under wind noise

Institute of Lightweight Systems

- Speech data from IEEE-Harvard Corpus ($n = 500$, 21.147 frames $x_i^\dagger$ for training, 4.601 for testing)
- $y_i = Ax_i^\dagger + \eta w_i \in \mathbb{R}^{2n-1}$ with reverberation matrix A, different noise levels η and wind noise w_i
- $w_i \approx$ perturbation $\circ$ modulation $\circ$ low-pass $\circ$ Brownian
- Theoretically optimal maps + learned via gradient descent





Learned R_{Aff}	$\min_{W,b}$	$\frac{1}{mn}\sum_{i=1}^m \lVert Wy_i+b-x_i^\dagger \rVert^2$
Learned R_{Lav}	$\min_{M,x_0}$	$\frac{1}{mn}\sum_{i=1}^m \lVert (A^TA+M)^{-1}(A^Ty_i+x_0)-x_i^\dagger \rVert^2$
Learned R_{Quad}	$\min_{L,x_0}$	$\frac{1}{mn}\sum_{i=1}^m \lVert (A^TA+\frac{1}{2}[L+L^T])^{-1}(A^Ty_i+x_0)-x_i^\dagger \rVert^2$
Learned R_{Tikh}	$\min_{R,x_0}$	$\frac{1}{mn}\sum_{i=1}^m \lVert (A^TA+R^TR)^{-1}(A^Ty_i+x_0)-x_i^\dagger \rVert^2$

Results on Training data

Noise level η	0.1	0.2	0.3	0.4	0.5
Optimal R_{Aff}	7.05e-05	1.98e-04	3.80e-04	6.03e-04	8.56e-04
Optimal R_{Lav}	7.28e-05	2.05e-04	3.93e-04	6.24e-04	8.84e-04
Optimal R_{Quad}	9.16e-05	2.31e-04	4.25e-04	6.62e-04	9.29e-04
Learned R_{Aff}	7.10e-05	1.99e-04	3.79e-04	6.03e-04	8.55e-04
Learned R_{Lav}	9.38e-05	2.54e-04	4.67e-04	7.11e-04	9.80e-04
Learned R_{Quad}	1.02e-04	2.68e-04	4.82e-04	7.29e-04	1.00e-03
Learned R_{Tikh}	1.06e-04	2.76e-04	4.86e-04	7.32e-04	1.01e-03

Results on Test data

Noise level η	0.1	0.2	0.3	0.4	0.5
Optimal R_{Aff}	7.51e-05	2.11e-04	4.05e-04	6.41e-04	9.06e-04
Optimal R_{Lav}	7.35e-05	2.05e-04	3.93e-04	6.22e-04	8.81e-04
Optimal R_{Quad}	8.92e-05	2.25e-04	4.17e-04	6.48e-04	9.08e-04
Learned R_{Aff}	7.50e-05	2.10e-04	4.01e-04	6.36e-04	9.05e-04
Learned R_{Lav}	9.48e-05	2.53e-04	4.66e-04	7.11e-04	9.86e-04
Learned R_{Quad}	1.01e-04	2.62e-04	4.72e-04	7.15e-04	9.88e-04
Learned R_{Tikh}	1.04e-04	2.69e-04	4.76e-04	7.18e-04	9.91e-04

Summary

If the noise is not too simple and you don't learn the noise weight, you leave something on the table.

Paper: Banert, S., Brauer, C., Lorenz, D., Tondji, L. (2026). Why the noise model matters: a performance gap in learned regularization. Inverse Problems, 42(2), 025005.

Questions?