Learning regularizers
bilevel opitimization or unrolling?

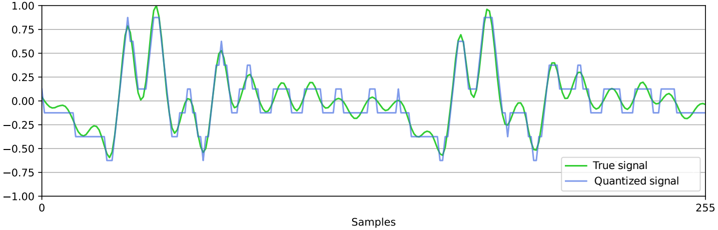
Dirk Lorenz
(joint work with Timo de Wolff, Christoph Brauer, Niklas Breustedt)

Center for Industrial Mathematics, University of Bremen

Workshop on Deep Learning, Image Analysis, Inverse Problems, and Optimisation, Lyon, November 2023

Learning to dequantize speech signals

- $\bar{s}_\ell \in \mathbb{R}^n$ speech signals, $s_\ell = Q_\Delta(\bar{s}_\ell)$ quantized signals
- Goal: Recover the $\bar{s}_\ell$ from the s_ℓ



Center for Industrial Mathematics

Learning to dequantize speech signals

- [Brauer, L., Gerkmann 16]: Take $\hat{s}_\ell = DCT^{-1}(x)$ where x solves

$$\min_x \|x\|_1 \quad \text{s.t.} \quad \|DCT^{-1}(x) - s_\ell\|_\infty \leq \Delta/2 \qquad (*)$$

(Look for signal with sparse DCT but respecting quantization bounds.)

Center for Industrial Mathematics

Learning to dequantize speech signals

- [Brauer, L., Gerkmann 16]: Take $\hat{s}_\ell = DCT^{-1}(x)$ where x solves

$$\min_x \|x\|_1 \quad \text{s.t.} \quad \|DCT^{-1}(x) - s_\ell\|_\infty \leq \Delta/2 \qquad (*)$$

(Look for signal with sparse DCT but respecting quantization bounds.)

- [Brauer, Zhao, L., Fingscheidt, 19]: Improve method by learning better linear map than DCT.

$$\min_K \sum_\ell \|Kx_\ell^N - \bar{s}_\ell\|_2^2 \text{ s.t. } x_\ell^N \ N\text{-the iterate of Chambolle-Pock for } (*)$$

Center for Industrial Mathematics

Learning to dequantize speech signals

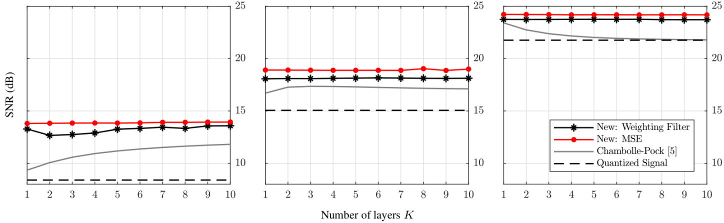
- [Brauer, L., Gerkmann 16]: Take $\hat{s}_\ell = DCT^{-1}(x)$ where x solves

$$\min_x \|x\|_1 \quad \text{s.t.} \quad \|DCT^{-1}(x) - s_\ell\|_\infty \leq \Delta/2 \qquad (*)$$

  (Look for signal with sparse DCT but respecting quantization bounds.)
- [Brauer, Zhao, L., Fingscheidt, 19]: Improve method by learning better linear map than DCT.

  $$\min_K \sum_\ell \|Kx_\ell^N - \bar{s}_\ell\|_2^2 \text{ s.t. } x_\ell^N \text{ } N\text{-the iterate of Chambolle-Pock for } (*)$$

- Results for different depths (with learned stepsizes as well):



Center for Industrial Mathematics

Outline

Center for
Industrial Mathematics

Learning regularizers

- Variational regularization of inverse problem $Af = g^\delta$

$$\min_f \mathcal{D}(Af, g^\delta) + \alpha \mathcal{R}(f)$$

- $\mathcal{D}$: similarity measure, often $\|Af - g^\delta\|^2$
  $\mathcal{R}$: regularizer, clasically $\|f\|^2$ or $\|\nabla f\|^2$, also total variation $\|\nabla f\|_1$, $\sum_i \phi(\langle f, v_i \rangle)$ for some ϕ,...
- General regularization theory available [Burger, Osher 2004], [Burger, Resmerita 2005], [Scherzer et al. 2009]
- Need to solve the optimization problem!
- Need to choose $\mathcal{D}$ and $\mathcal{R}$...
- $\mathcal{D}$ can be motivated by noise characteristic, generally least squares often good, despite of noise characteristic.
- Influence of $\mathcal{R}$ much bigger in practice, much less clear how to choose.

Center for Industrial Mathematics

Learning regularizers

- Idea: Having paired data $f_i^\dagger$ and g_i^δ (with $g_i^\delta = Af_i + \text{noise}$), $i = 1 \ldots, m$, learn regularizer $\mathcal{R}$ by *empirical risk minimization*

$$\min_{\mathcal{R}} \frac{1}{m} \sum_{i=1}^{m} \ell(\hat{f}_i, f_i^\dagger)$$
$$\text{s.t.} \quad \hat{f}_i \in \arg\min_f \mathcal{D}(Af, g_i^\delta) + \alpha \mathcal{R}(f).$$

- ℓ: Loss, typically $\ell(\hat{f}, f^\dagger) = \|\hat{f} - f^\dagger\|^2$
- Needs a model for $\mathcal{R}$ to optimize over!
- $\rightsquigarrow$ *Bi-level optimization problem*, generally very hard to solve...
- Upper and lower level problems
- [Tappen et al., 2007, Peyré, Fadili 2011, Pock et al. 2013, de los Reyes et al. 2017]

Use unrolling

- If lower level problem has unique solution, consider solution map $S(g^\delta) = \hat{f}$, and obtain

$$\min_R \frac{1}{m} \sum_{i=1}^{m} \ell(S(g_i^\delta), f_i^\dagger)$$

- Optimization needs derivative of solution operator S
- Circumventing this: *Unroll* an optimization algorithm

$$A_N(g^\delta) = \text{output of } N\text{th iteration of convergent algorithm}$$

and consider

$$\min_R \frac{1}{m} \sum_{i=1}^{m} \ell(A_N(g_i^\delta), f_i^\dagger)$$

- Need to "differentiate through iterations"
- If $\mathcal{D}$ is least squares may use

$$f^{n+1} = \operatorname{prox}_{\tau\alpha\mathcal{R}}(f^n - \tau A^*(Af^n - g^\delta)).$$

differentiation may be possible by automatic differentiation

Center for Industrial Mathematics

Unrolling vs bi-level

- Which one is better?
- Does unrolling approach converge to bi-level approach for $N \to \infty$?
- Why did the deeper unrolling not increase quality in the example of speech dequantization?

Center for Industrial Mathematics

Unrolling vs bi-level

- Which one is better?
- Does unrolling approach converge to bi-level approach for $N \to \infty$?
- Why did the deeper unrolling not increase quality in the example of speech dequantization?

$\rightsquigarrow$ Build tractable toy model and analyze everything explicitly!

Center for Industrial Mathematics

Outline

A tractable model

- Goal: develop a tractable toy model which can be analyzed explicitly
- Switch to notation from learning theory:
  - x: "objects" (was: noisy data g^δ)
  - y: "labels" (was: ground truth $f^\dagger$)
  - Goal: Predict y from x (was: reconstruct $f^\dagger$ from g^δ)
- Model consists of:
  - **Data:** Distributions for x and y
  - **Lower level problem:** similarity $\mathcal{D}$ and model for regularizer $\mathcal{R}$
  - An **algorithm** to unroll
  - **Upper level problem:** loss function ℓ

Center for Industrial Mathematics

The toy data

- Consider a denoising problem in $\mathbb{R}^n$, i.e.

$$x = y + \varepsilon \in \mathbb{R}^n$$

  Problem: Given pairs of noisy x and clean y, learn a denoiser

- Model for clean data y: $y \sim \mathcal{D}$ characterized by

$$y = \lambda \mathbf{1}, \quad \mathbb{E}(\lambda) = \mu, \quad \mathrm{Var}(\lambda) = \theta^2.$$

- Model for the noise: Normally distributed with

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$$

Center for Industrial Mathematics

The toy lower level problem and algorithm to unroll

- Simple quadratic problem

$$\hat{y} = \arg\min_{z} \frac{1}{2}\|z - x\|_2^2 + \frac{1}{2}\|\mathbf{R}z\|_2^2$$

with

$$\mathbf{R} \in \mathbb{R}^{k \times n}.$$

- **Bilevel:** Explicit solution

$$\hat{y} = (\mathbf{I} + \mathbf{R}^T\mathbf{R})^{-1}x.$$

- **Unrolling:** Unroll gradient descent with stepsize ω and $z^0 = 0$:

$$\begin{aligned}
\hat{y} = z^N = z^{N-1} - \omega((\mathbf{I} + \mathbf{R}^T\mathbf{R})z^{N-1} - x) \\
= \omega \sum_{j=0}^{N-1} (I - \omega(\mathbf{I} + \mathbf{R}^T\mathbf{R}))^j x \\
= (\mathbf{I} + \mathbf{R}^T\mathbf{R})^{-1}(\mathbf{I} - (\mathbf{I} - \omega(I + \mathbf{R}^T\mathbf{R}))^N)x.
\end{aligned}$$

Toy upper level problem

- Loss

$$\ell(\hat{y}, y) = \frac{1}{2}\|\hat{y} - y\|_2^2$$

- Minimize true (population) risk:

$$\mathcal{E} = \mathop{\mathbb{E}}_{\substack{\varepsilon\sim\mathcal{N} \\ y\sim\mathcal{D}}} \frac{1}{2}\|\hat{y} - y\|_2^2.$$

Center for Industrial Mathematics

Toy upper level problem

- Loss

$$\ell(\hat{y}, y) = \frac{1}{2}\|\hat{y} - y\|_2^2$$

- Minimize true (population) risk:

$$\mathcal{E} = \mathop{\mathbb{E}}_{\substack{\varepsilon\sim\mathcal{N} \\ y\sim\mathcal{D}}} \frac{1}{2}\|\hat{y} - y\|_2^2.$$

- Risk of a denoiser $\mathbf{T_R}$

$$\mathcal{E}(\mathbf{T_R}) = \mathop{\mathbb{E}}_{\substack{\varepsilon\sim\mathcal{N} \\ y\sim\mathcal{D}}} \frac{1}{2}\|\mathbf{T_R}(y+\varepsilon) - y\|_2^2$$

Recall:

$$\text{Bilevel:} \quad \mathbf{T_R} = (\mathbf{I} + \mathbf{R}^T\mathbf{R})^{-1}$$

$$\text{Unrolling:} \quad \mathbf{T_R} = \omega \sum_{j=0}^{N-1} (\mathbf{I} - \omega(\mathbf{I} + \mathbf{R}^T\mathbf{R}))^j$$

Both linear maps!

Total model

$$\min_{\mathbf{R}\in\mathbb{R}^{k\times n}} \mathop{\mathbb{E}}_{\substack{\varepsilon\sim\mathcal{N}\\ y\sim\mathcal{D}}} \frac{1}{2}\|\mathbf{T}_{\mathbf{R}}(y+\varepsilon)-y\|_2^2$$

where

$$\textbf{Bilevel:}\quad \mathbf{T}_{\mathbf{R}} = (\mathbf{I}+\mathbf{R}^T\mathbf{R})^{-1}$$

$$\textbf{Unrolling:}\quad \mathbf{T}_{\mathbf{R}} = \omega\sum_{j=0}^{N-1}(\mathbf{I}-\omega(\mathbf{I}+\mathbf{R}^T\mathbf{R}))^j$$

Center for Industrial Mathematics

Bias-variance decomposition of upper level

Lemma

If data y and noise ε are independent, we have for linear $\mathbf{T}$

$$\mathop{\mathbb{E}}_{\substack{\varepsilon\sim\mathcal{N}\\ y\sim\mathcal{D}}} \frac{1}{2}\|\mathbf{T}(y+\varepsilon)-y\|_2^2 = \mathop{\mathbb{E}}_{y\sim\mathcal{D}} \frac{1}{2}\|(\mathbf{T}-\mathbf{I})y\|_2^2 + \mathop{\mathbb{E}}_{\varepsilon\sim\mathcal{N}} \frac{1}{2}\|\mathbf{T}\varepsilon\|_2^2.$$

For $y=\lambda\mathbf{1}, \quad \mathbb{E}(\lambda)=\mu, \quad \mathrm{Var}(\lambda)=\theta^2$ *and* $\varepsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})$ *get*

$$\mathop{\mathbb{E}}_{\substack{\varepsilon\sim\mathcal{N}\\ y\sim\mathcal{D}}} \frac{1}{2}\|\mathbf{T}(y+\varepsilon)-y\|_2^2 = \frac{\mu^2+\theta^2}{2}\|(\mathbf{T}-\mathbf{I})\mathbf{1}\|_2^2 + \frac{\sigma^2}{2}\|\mathbf{T}\|_F^2.$$

Center for Industrial Mathematics

Total model (once again)

$$\min_{\mathbf{R}\in\mathbb{R}^{k\times n}} \frac{\mu^2+\theta^2}{2}\|(\mathbf{T_R}-\mathbf{I})\mathbf{1}\|_2^2 + \frac{\sigma^2}{2}\|\mathbf{T_R}\|_F^2$$

where

$$\textbf{Bilevel:} \quad \mathbf{T_R} = (\mathbf{I}+\mathbf{R}^T\mathbf{R})^{-1}$$

$$\textbf{Unrolling:} \quad \mathbf{T_R} = \omega\sum_{j=0}^{N-1}(\mathbf{I}-\omega(\mathbf{I}+\mathbf{R}^T\mathbf{R}))^j$$

- For unrolling: Could also minimize over stepsize ω!
- Dependence on k (# rows of $\mathbf{R}$)?
- Very nonlinear in $\mathbf{R}$.
- First study *expressivity*, i.e. characterize the set of possible $\mathbf{T_R}$

Outline

Center for
Industrial Mathematics

Expressivity of bilevel learning

Theorem

The set of possible unrolling denoisers $\mathbf{T} = (\mathbf{I} + \mathbf{R}^T\mathbf{R})^{-1}$ *for* $\mathbf{R} \in \mathbb{R}^{k \times n}$ *is*

$$\mathcal{A}_k = \left\{ \mathbf{T} \in \mathbb{R}^{n \times n} \mid \dim(\mathrm{Eig}(\mathbf{T}, 1)) \geq n - k, \quad \mathbf{T}^T = \mathbf{T}, \quad 0 \prec \mathbf{T} \preccurlyeq \mathbf{I} \right\}$$

Proof.

- Spectral calculus: $T = f(R^TR)$, $f(s) = 1/(1+s) \rightsquigarrow 0 \prec \mathbf{T} \preccurlyeq \mathbf{I}$
- $rank(R^TR) \leq k$ implies $\dim(\mathrm{Eig}(\mathbf{T}, 1)) \geq n - k$

□

Center for Industrial Mathematics

Expressivity of unrolling

Theorem

The set of possible bilevel denoisers $\mathbf{T} = \omega \sum_{j=0}^{N-1} (\mathbf{I} - \omega(\mathbf{I} + \mathbf{R}^T\mathbf{R}))^j$ *for* $\mathbf{R} \in \mathbb{R}^{k\times n}$ *is*

1. *N even:*

$$\mathcal{B}_{N,k,\omega} = \left\{ \mathbf{U} \in \mathbb{R}^{n\times n} \;\middle|\; \begin{array}{l} \mathbf{U} = \mathbf{U}^\top,\ \dim(\mathrm{Eig}(\mathbf{U}, 1-(1-\omega)^N)) \geq n-k, \\ \mathbf{U} \preccurlyeq [1-(1-\omega)^N]\mathbf{I} \end{array} \right\}$$

2. *N odd: There exists a constant* $c_{N,\omega}$ *such that*

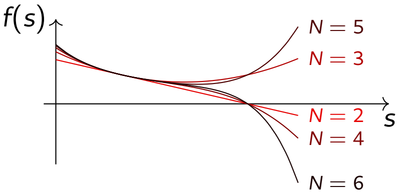
$$\mathcal{B}_{N,k,\omega} = \left\{ \mathbf{U} \in \mathbb{R}^{n\times n} \;\middle|\; \begin{array}{l} \mathbf{U} = \mathbf{U}^\top,\ \dim(\mathrm{Eig}(\mathbf{U}, 1-(1-\omega)^N)) \geq n-k, \\ c_{N,\omega}\mathbf{I} \preccurlyeq \mathbf{U} \end{array} \right\}.$$

Roughly $\omega\left(\frac{1}{2} + \frac{1}{N+1}\right) \leq c_{N,\omega} \leq \omega\left(\frac{1}{2} + \frac{1}{N}\left(\frac{1+\log(N)/2}{2-\frac{\log(N)}{N}}\right)\right)$

Center for Industrial Mathematics

Expressivity of unrolling - proof

- Spectral calculus $T = f(R^T R)$ with
  $$f(s) = \omega \sum_{j=0}^{N-1} (1 - \omega(1+s))^j = \tfrac{1-(1-\omega(1+s))^N}{1+s}$$
- $R^T R$ has eigenvalue 0 "at least $n-k$ times" $\rightsquigarrow$ T has eigenvalue $1-(1-\omega)^N$ "at least $n-k$ times"
- Upper and lower bounds on f imply eigenvalue bounds for T



- N even: $f(s) \leq f(0) = 1-(1-\omega)^N$, unbounded from below
- N odd: f unbounded from above, single global minimum $c_{N,\omega}$ with no explicit expression

Center for Industrial Mathematics

Best risks

- Using expressivity results we can calculate optimal risks

$$\min_{\mathbf{R}\in\mathbb{R}^{k\times n}} \mathop{\mathbb{E}}_{\substack{\varepsilon\sim\mathcal{N}\\ y\sim\mathcal{D}}} \frac{1}{2}\|\mathbf{T}(y+\varepsilon)-y\|_2^2$$

  explicitly
- For unrolling we can even learn (i.e. optimize) over stepsize ω
- Quite some mess of case distinctions and not very informative

Center for Industrial Mathematics

Some results

Theorem

1. *Best linear denoiser for our toy model is*

$$\mathbf{T}^* = \frac{\mu^2 + \theta^2}{n(\mu^2 + \theta^2) + \sigma^2} \mathbf{1}_{n \times n} \quad \textit{with} \quad \mathcal{E}(\mathbf{T}^*) = \frac{\sigma^2}{2} \frac{n(\mu^2 + \theta^2)}{n(\mu^2 + \theta^2) + \sigma^2}$$

2. *Best bilevel denoiser* *does not exist*, *but*

$$\inf_{\mathbf{T} = (\mathbf{I} + \mathbf{R}^T \mathbf{R})^{-1}} \mathcal{E}(\mathbf{T}) = \begin{cases} \frac{\sigma^2}{2}(n-k) & : \quad k < n \\ \frac{\sigma^2}{2} \frac{n(\mu^2+\theta^2)}{n(\mu^2+\theta^2)+\sigma^2} & : \quad n = k \end{cases}$$

3. *Best unrolling denoisers exist but is it a mess of a formula...* *(results different for N even or odd and $k < n$ or $k = n$).* *For N either even or odd, best risk does not depend on N if optimized over stepsize ω.*

$\rightsquigarrow$ Calculate best risks numerically and consider risk ratios

Center for Industrial Mathematics

Some results

Theorem

1. *Best linear denoiser for our toy mod*

$$\mathbf{T}^* = \frac{\mu^2 + \theta^2}{n(\mu^2 + \theta^2) + \sigma^2}\mathbf{1}_{n\times n} \quad \textit{with}$$

2. *Best bilevel denoiser* *does not exist*,

$$\inf_{\mathbf{T}=(\mathbf{I}+\mathbf{R}^T\mathbf{R})^{-1}} \mathcal{E}(\mathbf{T}) = \begin{cases} \frac{\sigma^2}{2}(n \\ \frac{\sigma^2}{2}\frac{n(}{n(\mu^} \end{cases}$$

3. *Best unrolling denoisers exist but is it a mess of a formula…*
   (results different for N even or odd and $k < n$ or $k = n$).
   For N either even or odd, best risk does not depend on N if optimized over stepsize ω.

$\rightsquigarrow$ Calculate best risks numerically and consider risk ratios

i.i.d. model

- Also analyzed slightly more general data model:

$$\text{For } j = 1, \ldots, n: \quad y_j \sim \mathcal{D} \quad \text{i.i.d.}$$
$$\mathbb{E}(y_j) = \mu, \quad \mathrm{Var}(y_j) = \theta^2$$

- Different bias-variance decomposition

$$\mathop{\mathbb{E}}_{\substack{\varepsilon\sim\mathcal{N} \\ y\sim\mathcal{D}}} \frac{1}{2}\|\mathbf{T}(y+\varepsilon) - y\|_2^2 = \frac{\mu^2}{2}\|(\mathbf{T}-\mathbf{I})\mathbf{1}\|_2^2 + \frac{\theta^2}{2}\|\mathbf{T}-\mathbf{I}\|_F^2 + \frac{\sigma^2}{2}\|\mathbf{T}\|_F^2$$
$$=: \mathcal{E}_{\text{i.i.d}}(\mathbf{T})$$

- Related best risks also pretty messy…

Center for Industrial Mathematics

Outline

Bilevel learning and unrolling

A toy model

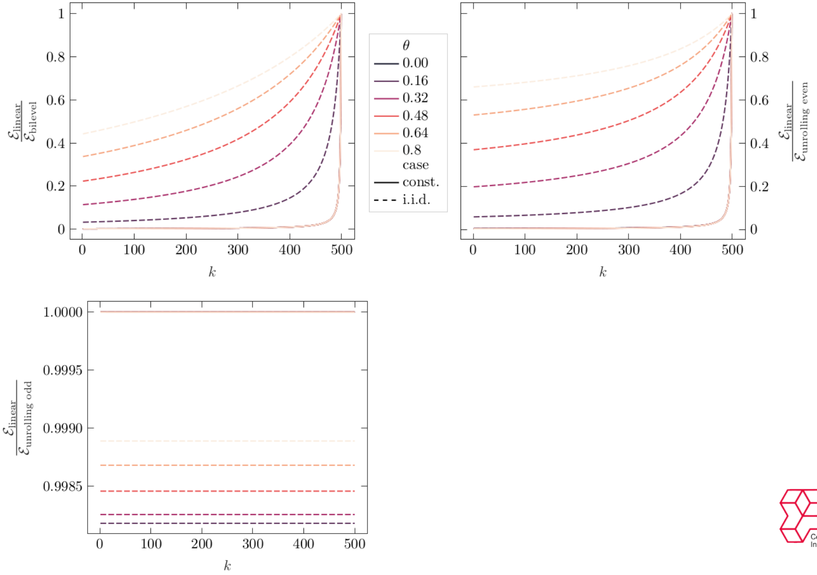
Expressivity

Results

Center for
Industrial Mathematics

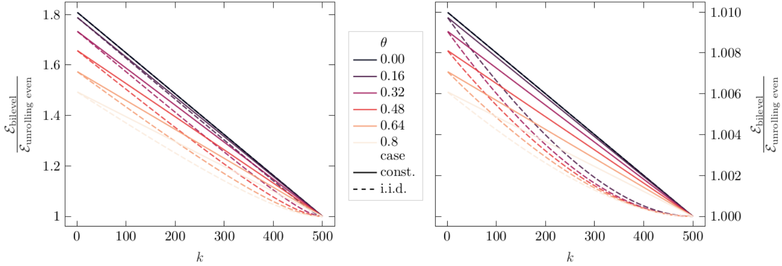
Risk ratios with best linear denoiser

$n = 500, \mu = 1, \sigma = 0.9$:



Risk ratios between bilevel and unrolling

$n = 500, \mu = 1$:



(a) $\sigma = 0.9$ (b) $\sigma = 0.1$

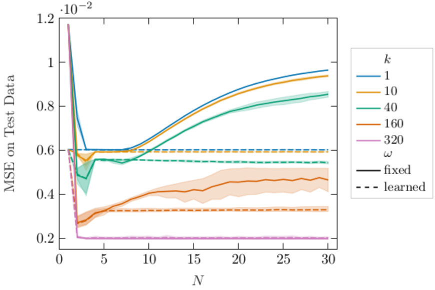
$\rightsquigarrow$ unrolling generally better than bilevel

Reality check: Numerical experiment

- Our model is very simple - how close to reality are the results?
- Experiment on speech data.
  - Data model:
    - y clean speech (part of IEEE speech corpus)
    - $x = y + \varepsilon$ with Gaussian noise, $n = 320$, $\sigma = 0.1$
  - Lower level problem and algorithm exactly like here.
  - Upper level problem: Empirical risk with least squares loss.
  - Numerical optimization with TensorFlow, standard optimization tricks applied (initialization, learning rates optimized...)
  - Also optimized over stepsize $\omega(\alpha) = \log(1 + \exp(\alpha))$ over α.

Center for Industrial Mathematics

Reality check, observations

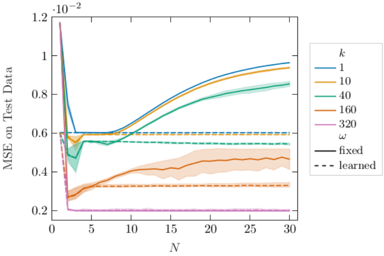


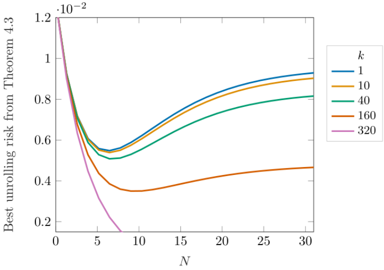
- With learned stepsize, MSE basically independent of depth N, as predicted
- Without learned stepsizes no dependence on parity, contrary to prediction
- Without learned stepsizes: Worse MSE for deeper unrolling

Center for Industrial Mathematics

Double check: Do results fit theory?

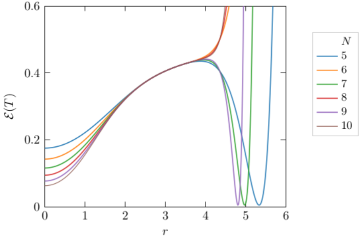
Optimal risks according to theory with parameters as in experiment:





Why don't we see the dependence on parity?

- Most theoretical findings confirmed.
- What about parity?
- Conjecture: Good denoisers for odd depth hidden in sharp local minima!
  In $k = n = 1$, i.e. $\mathbf{R} = r \in \mathbb{R}$:



Center for Industrial Mathematics

Thanks for listening!

Learning Variational Models with Unrolling and Bilevel Optimization
Christoph Brauer, Niklas Breustedt, Timo de Wolff, Dirk A. Lorenz
`https://arxiv.org/abs/2209.12651`

Center for Industrial Mathematics