Learning regularizers 
bilevel optimization 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

- \( \tilde{s}_\ell \in \mathbb{R}^n \) speech signals, \( s_\ell = Q_\Delta(\tilde{s}_\ell) \) quantized signals
- Goal: Recover the \( \tilde{s}_\ell \) from the \( s_\ell \)
Learning to dequantize speech signals

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

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

(Look for signal with sparse DCT but respecting quantization bounds.)
Learning to dequantize speech signals

- [Brauer, L., Gerkmann 16]: Take $\hat{s}_\ell = DCT^{-1}(x)$ where $x$ solves
  \[
  \min_x \|x\|_1 \quad \text{s.t.} \quad \|DCT^{-1}(x) - s_\ell\|_\infty \leq \Delta/2 \quad (*)
  \]
  (Look for signal with sparse DCT but respecting quantization bounds.)

- [Brauer, Zhao, L., Fingscheidt, 19]: Improve method by learning better linear map than DCT.
  \[
  \min_K \sum_\ell \|Kx^N_\ell - \bar{s}_\ell\|_2^2 \quad \text{s.t.} \quad x^N_\ell \quad N\text{-the iterate of Chambolle-Pock for } (*)
  \]
Learning to dequantize speech signals

- [Brauer, L., Gerkmann 16]: Take \( \hat{s}_\ell = DCT^{-1}(x) \) where \( x \) solves
  \[
  \min_x \|x\|_1 \quad \text{s.t.} \quad \|DCT^{-1}(x) - s_\ell\|_\infty \leq \Delta/2 \quad (*)
  \]
  (Look for signal with sparse DCT but respecting quantization bounds.)

- [Brauer, Zhao, L., Fingscheidt, 19]: Improve method by learning better linear map than DCT.
  \[
  \min K \sum_{\ell} \|Kx_N^\ell - \bar{s}_\ell\|_2^2 \quad \text{s.t.} \quad x_N^\ell \quad N\text{-the iterate of Chambolle-Pock for (*)}
  \]

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

![Graphs showing SNR vs. number of layers for different methods]
Bilevel learning and unrolling

A toy model

Expressivity

Results
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, classically \( ||f||^2 \) or \( ||\nabla f||^2 \), also total variation \( ||\nabla f||_1, \sum_i \phi(\langle f, v_i \rangle) \) for some \( \phi \),…


- 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.
Learning regularizers

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

$$
\min_{\mathcal{R}} \frac{1}{m} \sum_{i=1}^{m} \ell(\hat{f}_i, f_i^\dagger)
$$

s.t. $\hat{f}_i \in \arg\min_f \mathcal{D}(Af, g_i^\delta) + \alpha \mathcal{R}(f)$.

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

$$ \min_R \frac{1}{m} \sum_{i=1}^{m} \ell(S(g^\delta_i), f^\dagger_i) $$

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

$$ \min_R \frac{1}{m} \sum_{i=1}^{m} \ell(A_N(g^\delta_i), f^\dagger_i) $$

Need to “differentiate through iterations”

If $\mathcal{D}$ is least squares may use

$$ f^{n+1} = \text{prox}_{\tau\alpha_{R}}(f^n - \tau A^*(Af^n - g^\delta)) $$

differentiation may be possible by automatic differentiation
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?
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?

⇝ Build tractable toy model and analyze everything explicitly!
Outline

Bilevel learning and unrolling

A toy model

Expressivity

Results
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 \( g^\delta \))
  - \( y \): “labels” (was: ground truth \( f^\dagger \))
  - Goal: Predict \( y \) from \( x \) (was: reconstruct \( f^\dagger \) from \( g^\delta \))
- Model consists of:
  - **Data**: Distributions for \( x \) and \( y \)
  - **Lower level problem**: similarity \( D \) and model for regularizer \( R \)
  - An **algorithm** to unroll
  - **Upper level problem**: loss function \( \ell \)
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

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

Model for the noise: Normally distributed with

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$$
The toy lower level problem and algorithm to unroll

- **Simple quadratic problem**
  \[
  \hat{y} = \arg\min_z \frac{1}{2} \|z - x\|^2_2 + \frac{1}{2} \|Rz\|^2_2
  \]
  with
  \[
  R \in \mathbb{R}^{k \times n}.
  \]

- **Bilevel**: Explicit solution
  \[
  \hat{y} = (I + R^T R)^{-1} x.
  \]

- **Unrolling**: Unroll gradient descent with stepsize $\omega$ and $z^0 = 0$:
  \[
  \hat{y} = z^N = z^{N-1} - \omega((I + R^T R)z^{N-1} - x)
  \]
  \[
  = \omega \sum_{j=0}^{N-1} (I - \omega(I + R^T R))^j x
  \]
  \[
  = (I + R^T R)^{-1} (I - (I - \omega(I + R^T R))^N) x.
  \]
Toy upper level problem

- Loss

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

- Minimize true (population) risk:

\[ \mathcal{E} = \mathbb{E}_{\varepsilon \sim \mathcal{N}} \frac{1}{2} \|\hat{y} - y\|_2^2. \]
Toy upper level problem

- **Loss**
  \[ \ell(\hat{y}, y) = \frac{1}{2} \| \hat{y} - y \|^2 \]

- **Minimize true (population) risk:**
  \[ \mathcal{E} = \mathbb{E}_{\epsilon \sim \mathcal{N}} \mathbb{E}_{y \sim \mathcal{D}} \frac{1}{2} \| \hat{y} - y \|^2. \]

- **Risk of a denoiser \( T_R \)**
  \[ \mathcal{E}(T_R) = \mathbb{E}_{\epsilon \sim \mathcal{N}} \mathbb{E}_{y \sim \mathcal{D}} \frac{1}{2} \| T_R(y + \epsilon) - y \|^2. \]

Recall:

- **Bilevel:** \( T_R = (I + R^T R)^{-1} \)
- **Unrolling:** \( T_R = \omega \sum_{j=0}^{N-1} (I - \omega (I + R^T R))^j \)

Both linear maps!
Total model

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

where

**Bilevel:** \( \mathbf{T}_R = (\mathbf{I} + \mathbf{R}^T \mathbf{R})^{-1} \)

**Unrolling:** \( \mathbf{T}_R = \omega \sum_{j=0}^{N-1} (\mathbf{I} - \omega(\mathbf{I} + \mathbf{R}^T \mathbf{R}))^j \)
Lemma

If data $y$ and noise $\varepsilon$ are independent, we have for linear $T$

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

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

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

- **Bilevel:** \( \mathbf{T}_R = (\mathbf{I} + \mathbf{R}^T \mathbf{R})^{-1} \)
- **Unrolling:** \( \mathbf{T}_R = \omega \sum_{j=0}^{N-1} (\mathbf{I} - \omega (\mathbf{I} + \mathbf{R}^T \mathbf{R}))^j \)

- For unrolling: Could also minimize over stepsize \( \omega \)!
- 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

Bilevel learning and unrolling

A toy model

Expressivity

Results
Theorem

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

$$A_k = \{ T \in \mathbb{R}^{n \times n} \mid \dim(\text{Eig}(T, 1)) \geq n - k, \quad T^T = T, \quad 0 \prec T \preceq I \}$$

Proof.

- Spectral calculus: $T = f(R^T R), \quad f(s) = 1/(1 + s) \rightsquigarrow 0 \prec T \preceq I$
- $\text{rank}(R^T R) \leq k$ implies $\dim(\text{Eig}(T, 1)) \geq n - k$
Expressivity of unrolling

Theorem

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

1. \( N \) even:

\[
\mathcal{B}_{N,k,\omega} = \left\{ U \in \mathbb{R}^{n \times n} \mid U = U^T, \text{dim}(\text{Eig}(U, 1 - (1 - \omega)^N)) \geq n - k, \right. \\
\left. U \preceq [1 - (1 - \omega)^N]I \right\}
\]

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

\[
\mathcal{B}_{N,k,\omega} = \left\{ U \in \mathbb{R}^{n \times n} \mid U = U^T, \text{dim}(\text{Eig}(U, 1 - (1 - \omega)^N)) \geq n - k, \right. \\
\left. c_{N,\omega} I \preceq U \right\}
\]

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

- Spectral calculus \( T = f(R^T R) \) with
  \[
f(s) = \omega \sum_{j=0}^{N-1} (1 - \omega(1 + s))^j = \frac{1-(1-\omega(1+s))^N}{1+s}
\]

- \( R^T R \) has eigenvalue 0 “at least \( n - k \) times” \( \Rightarrow \) \( T \) 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
Using expressivity results we can calculate optimal risks

\[
\min_{R \in \mathbb{R}^{k \times n}} \mathbb{E}_{\varepsilon \sim \mathcal{N}} \frac{1}{2} \| T(y + \varepsilon) - y \|_2^2
\]
explicitly

For unrolling we can even learn (i.e. optimize) over stepsize \( \omega \)

Quite some mess of case distinctions and not very informative
Some results

Theorem

1. Best linear denoiser for our toy model is

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

2. Best bilevel denoiser does not exist, but

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

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 \( \omega \).

\( \Rightarrow \) Calculate best risks numerically and consider risk ratios
Some results

Theorem

1. Best linear denoiser for our toy model is
   \[ T^* = \frac{\mu^2 + \theta^2}{n(\mu^2 + \theta^2) + \sigma^2} \mathbf{1}_{n \times n} \]  
   with

2. Best bilevel denoiser does not exist,
   \[ \inf_{T=(I+R^T R)^{-1}} \mathcal{E}(T) = \begin{cases} \frac{\sigma^2}{2} (n - k) : k < n \\ \frac{\sigma^2}{2} \frac{n}{n(\mu^2 + \theta^2)} : n \end{cases} \]

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 $\omega$.

→ Calculate best risks numerically and consider risk ratios
Also analyzed slightly more general data model:

For $j = 1, \ldots, n$:

\[ y_j \sim \mathcal{D} \quad \text{i.i.d.} \]

\[ \mathbb{E}(y_j) = \mu, \quad \text{Var}(y_j) = \theta^2 \]

Different bias-variance decomposition

\[ \mathbb{E}_{\varepsilon \sim \mathcal{N}, y \sim \mathcal{D}} \frac{1}{2} \| T(y + \varepsilon) - y \|_2^2 = \frac{\mu^2}{2} \| (T - I)1 \|_2^2 + \frac{\theta^2}{2} \| T - I \|_F^2 + \frac{\sigma^2}{2} \| T \|_F^2 \]

\[ =: \mathcal{E}_{\text{i.i.d}}(T) \]

Related best risks also pretty messy…
Outline

Bilevel learning and unrolling

A toy model

Expressivity

Results
Risk ratios with best linear denoiser

\( n = 500, \mu = 1, \sigma = 0.9: \)

\[ \frac{\xi_{\text{linear}}}{\xi_{\text{bilevel}}} \]

\[ k \]

\[ 0 \quad 100 \quad 200 \quad 300 \quad 400 \quad 500 \]

\[ 0 \quad 0.2 \quad 0.4 \quad 0.6 \quad 0.8 \quad 1 \]

\[ \theta \]

- 0.00
- 0.16
- 0.32
- 0.48
- 0.64
- 0.8
- case
- const.
- i.i.d.
Risk ratios between bilevel and unrolling

\[ n = 500, \mu = 1: \]

\[ \Rightarrow \text{unrolling generally better than bilevel} \]
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 $\alpha$. 
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
Double check: Do results fit theory?

Optimal risks according to theory with parameters as in experiment:

![Graph 1: MSE on Test Data](image1)

![Graph 2: Best unrolling risk from Theorem 1.3](image2)
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. $\mathbb{R} = r \in \mathbb{R}$:
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

Bilevel learning and unrolling

A toy model

Expressivity

Results