Automatic Differentiation of Proximal Algorithms in Bilevel Learning for Imaging

Patricio Guerrero

Signal Processing Systems
TU Delft

Edinburgh, June 2026

bilevel optimization setting

formulation and challenges

We considerer the generalized bilevel problem

$$\min_{\theta} L(\theta) := \ell(s^\star(\theta))$$
$$\text{subject to } s^\star(\theta) = \arg\min_{s} f(s, \theta)$$

where

- $s^\star(\theta)$ is an estimated signal with parameters θ
- ℓ is a smooth upper level loss function on the estimated speech
- f is a *nonsmooth*, convex on s, lower level loss function
- s or θ is high-dimensional

- The upper level loss is smooth, we can apply a first-order method

$$\partial L(\theta_k) = \nabla \ell\left(s^\star(\theta_k)\right) \partial s^\star(\theta_k)$$
$$\theta_{k+1} = \theta_k - \beta \nabla L(\theta_k)$$

and compute $\partial L(\theta_k)$ with (reverse) automatic differentiation (AD)

- The upper level loss is smooth, we can apply a first-order method

$$\partial L(\theta_k) = \nabla \ell \left(s^\star(\theta_k)\right) \partial s^\star(\theta_k)$$
$$\theta_{k+1} = \theta_k - \beta \nabla L(\theta_k)$$

  and compute $\partial L(\theta_k)$ with (reverse) automatic differentiation (AD)

- $s^\star(\theta_k)$ may not have a close-form solution, we consider $s^\star(\theta)$ as the solution of the fixed-point equation

$$s(\theta) = F(s(\theta), \theta)$$

  and approximate it with the iteration

$$s_{j+1}(\theta) = F(s_j(\theta), \theta)$$

conservative derivatives and AD

How to define / compute $\partial L(\theta)$?

Conservative derivatives (Bolte, Pauwels, 2021)

For a locally Lipschitz function $F\colon \mathbb{R}^n \to \mathbb{R}^m$, a set-valued map $D_F\colon \mathbb{R}^n \to \mathbb{R}^{m\times n}$ is a *conservative derivative* of F if, for every absolutely continuous $\gamma\colon [0,1] \to \mathbb{R}^n$,

$$\frac{d}{dt}F(\gamma(t)) = V\,\dot\gamma(t), \quad \text{for all } V \in D_F(\gamma(t)), \quad \text{for almost all } t \in [0,1].$$

How to define / compute $\partial L(\theta)$?

Conservative derivatives (Bolte, Pauwels, 2021)

For a locally Lipschitz function $F\colon \mathbb{R}^n \to \mathbb{R}^m$, a set-valued map $D_F\colon \mathbb{R}^n \to \mathbb{R}^{m\times n}$ is a *conservative derivative* of F if, for every absolutely continuous $\gamma\colon [0,1] \to \mathbb{R}^n$,

$$\frac{d}{dt}F(\gamma(t)) = V\,\dot{\gamma}(t), \quad \text{for all } V \in D_F(\gamma(t)), \quad \text{for almost all } t \in [0,1].$$

Then the chain-rule becomes available again

$$D_L(\theta) \coloneqq \nabla \ell\,(s(\theta))\; D_s(\theta)$$

we will focus now on the term $D_s(\theta)$

How to compute the error between two (conservative derivative) sets ?

Excess, Beer et al., 1993

Given two bounded sets of matrices $\mathcal{A}$ and $\mathcal{B}$. The *excess* of $\mathcal{A}$ over $\mathcal{B}$ is

$$e(\mathcal{A}, \mathcal{B}) := \sup_{A \in \mathcal{A}} \inf_{B \in \mathcal{B}} \|A - B\|$$

- not symmetric
- verifies triangle inequality

We will assume the following

Assumption 1, (Grazzi et al., 2024)

1. The step map $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz, *definable*, piece-wise Lipschitz smooth
2. For all (s, θ)
$$\|D_{F,s}(s, \theta)\|_{\text{sup}} \leq q < 1$$

We will assume the following

Assumption 1, (Grazzi et al., 2024)

1. The step map $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz, *definable*, piece-wise Lipschitz smooth
2. For all (s, θ)
$$\|D_{F,s}(s,\theta)\|_{\text{sup}} \leq q < 1$$

- 2. implies $D_{F,s}(\cdot, \theta)$ is a contraction, then there is a fixed point $s^\star$ of $F(\cdot, \theta)$. We will denote D_s^{fix} the conservative derivate on $s^\star$

Recall that $s^\star(\theta)$ is approximated as the FP iteration

$$s_{j+1}(\theta) = F(s_j(\theta), \theta).$$

Then D_s is recursively computed by

$$D_{s_{j+1}}(\theta) = D_F(s_j(\theta), \theta)\, D_{s_j}(\theta) \tag{ID}$$

Recall that $s^\star(\theta)$ is approximated as the FP iteration

$$s_{j+1}(\theta) = F(s_j(\theta), \theta).$$

Then D_s is recursively computed by

$$D_{s_{j+1}}(\theta) = D_F(s_j(\theta), \theta)\, D_{s_j}(\theta) \tag{ID}$$

Theorem (Grazzi et al., 2024)

Under Assumption 1, let $J \in \mathbb{N}$, then (ID) satisfies

$$e\left(D_{s_J}(\theta), D_s^{\text{fix}}(\theta)\right) \leq \frac{B_\theta}{1-q} q^J + \bar{s}_0\, C(\theta) J q^{J-1}$$

where $q < 1$, $\bar{s}_0 := \|s_0(\theta) - s^\star(\theta)\|$, $B_\theta := \|D_{F,\theta}(s(\theta,\theta))\|_{\sup}$

truncated automatic differentiation

- What if s or θ are high-dimensional ?

Then reverse AD is memory expensive. If we unroll (ID) for J iterations, we can prove (closely to Shaban et al., 2019 in the smooth setting)

$$D_{s_J}(\theta) = \sum_{j=0}^{J} \mathcal{B}_j \mathcal{A}_{j+1} \cdots \mathcal{A}_J$$

with

$$\begin{aligned} \mathcal{A}_j &= D_{F,s}(s(\theta), \theta) \\ \mathcal{B}_j &= D_{F,\theta}(s(\theta), \theta). \end{aligned}$$

- J could be too large to handle $\dim(s) + \dim(\theta)$
  We consider T-*truncated* reverse AD with $T \ll J$

$$D_{s_{J,T}}(\theta) = \sum_{j=J-T+1}^{J} \mathcal{B}_j \mathcal{A}_{j+1} \cdots \mathcal{A}_J$$

  and analyse the introduced error

- We can write

$$D_{s_J}(\theta) = E_{s_T}(\theta) + D_{s_{J,T}}(\theta)$$

  with E_{s_T} the early $J - T$ sensitivities of s w.r.t. θ

Can we bound the excess ?

Can we bound the excess ?

- If the step map F is the proximal gradient descent (pGD) map

$$F(s, \theta) = \mathrm{prox}_{\gamma g}\left(s - \gamma \nabla_s f(s(\theta, \alpha))\right), \quad \gamma > 0 \tag{1}$$

  for a convex, lower-semicontinuous function g, and a β-Lipschitz smooth function f

- This F solves the *regularized* lower-level problem

$$\min_s f(s, \theta) + g(s) \tag{2}$$

Theorem (truncated AD excess bound for pGD)

Under Assumption 1, with F the pGD step map, for $\gamma \leq 1/\beta$, we have

$$e\left(D_{s_{J,T}}(\theta), D_{s_J}(\theta)\right) \leq \frac{B_\theta}{\gamma\alpha}\left(1 - \gamma\alpha\right)^T$$

where

$$B_\theta = \max_{j \in \{0, \cdots, J-T\}} \|\mathcal{B}_j\|_{\text{sup}} \tag{3}$$

- The proof is inspired from Shaban et al., 2019 covering the smooth case

Proof: We have

$$\begin{aligned} e\left(D_{s_{J,K}}(\theta), D_{s_J}(\theta)\right) &= e\left(D_{s_{J,K}}(\theta), E_{s_T}(\theta) + D_{s_{J,T}}(\theta)\right) \\ &\leq e\left(D_{s_{J,K}}(\theta), D_{s_{J,T}}(\theta)\right) + \|E_{s_T}\|_{\sup} \\ &\leq \|E_{s_T}\|_{\sup} \end{aligned}$$

Then, we also have

$$\begin{aligned} E_{s_T} &= \sum_{j=0}^{J-K} \mathcal{B}_j \mathcal{A}_{j+1} \cdots \mathcal{A}_J \\ &= \sum_{j=0}^{J-T} \mathcal{B}_j \mathcal{A}_{j+1} \cdots \mathcal{A}_{J-T} \ \mathcal{A}_{J-T+1} \cdots \mathcal{A}_J \\ &\leq B_\theta \frac{1}{\gamma\alpha} (1-\gamma\alpha)^T \end{aligned}$$

because it is known that $\|D_{\mathrm{prox}_{\gamma\psi}}\|_{\sup} \leq 1$

Corollary

Under Assumption 1, let $J, T \in \mathbb{N}$, $T \ll J$, we have

$$e\left(D_{s_{J,T}}(\theta), D_s^{\mathrm{fix}}(\theta)\right) \leq \frac{B_\theta}{1-q} q^J + \bar{s}_0 C(\theta) J q^{J-1} + B_\theta \frac{1}{\gamma\alpha}(1-\gamma\alpha)^T$$

Also, the conservative derivatives of the upper-level loss $D_L(\theta)$ verifies

$$e\left(\nabla \ell\left(s(\theta)\right) D_{s_{J,T}}, D_L(\theta)\right) \leq \|\nabla \ell\left(s(\theta)\right)\| e\left(D_{s_{J,T}}(\theta), D_s^{\mathrm{fix}}(\theta)\right)$$

- What about a generalized nonsmooth step map F ? The only (non obvious) assumption on F is to be nonexpansive :

$$\|D_{F,s}\|_{\sup} \leq 1$$

numerical experiments

- For an image s, we consider the reconstruction lower-level problem as

$$\min_{s} \tfrac{1}{2}\|As - b\|^2 + \theta\, g(Ls) + h(s), \quad \theta_i > 0,$$

where g, h are convex, lower-semicontinuous, A, L linear operators
- and the learning upper-level problem as simply

$$L(\theta) := \tfrac{1}{2}\|s(\theta) - s_{\text{true}}\|^2$$

where s_{true} is some available ground-truth image.

- We consider TGV regularization

$$\mathrm{tgv}_{\theta_1}(s_0) = \min_{s_1} \|\nabla s_0 - s_1\|_{2,1} + \theta_1 \|E s_1\|_{2,1}, \quad \theta_1 > 0. \tag{4}$$

with

$$L = \begin{pmatrix} \nabla & -I \\ 0 & E \end{pmatrix}, \quad g\colon (s_0, s_1) \mapsto \begin{pmatrix} \|s_0\|_{2,1} \\ \|s_1\|_{2,1} \end{pmatrix}$$

and $h(s) = \imath_{\mathbb{R}_+}(s_0)$

- We solved with pDG and Condat-Vu[1] for the proximity operator

$$\mathrm{prox}_{\gamma\theta g}(b) = \arg\min_s \tfrac{1}{2}\|s - b\|^2 + \gamma\theta g(s)$$

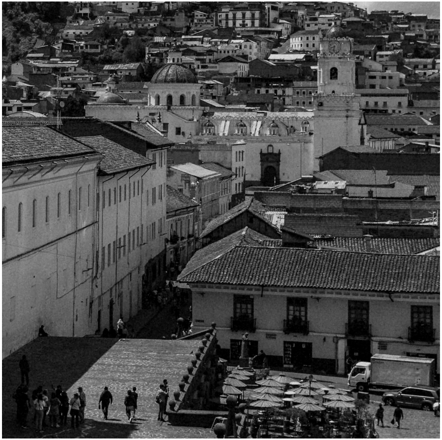
is now a denoising problem for a noisy image b regularized with $g + h$

[1] PG, Bellens, Dewulf, arXiv:2412.10034. 2024.

clean image

noisy image

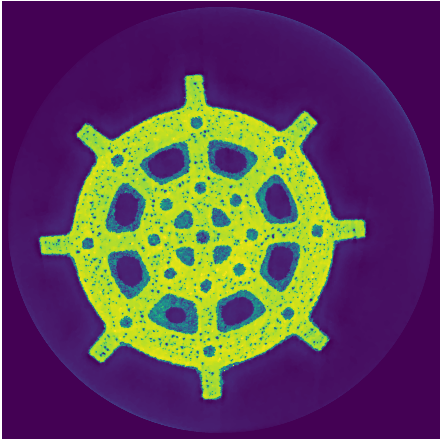
tv



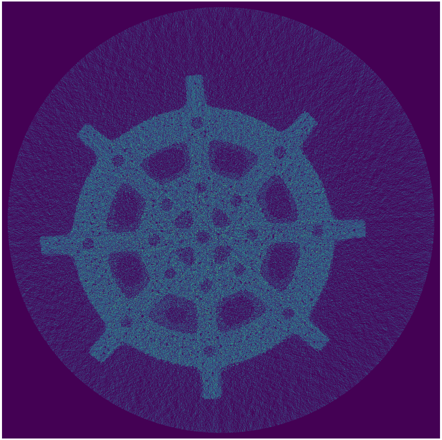
tgv

tv

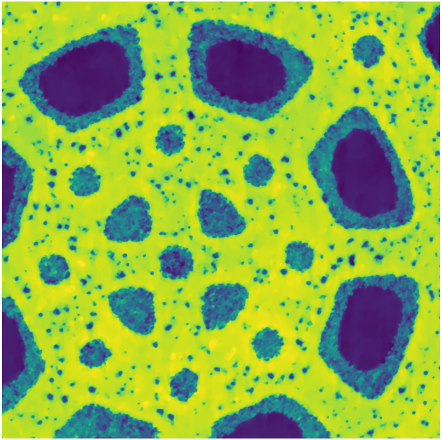
tgv



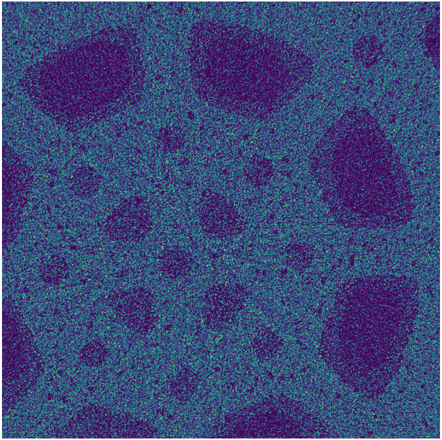
fully sampled



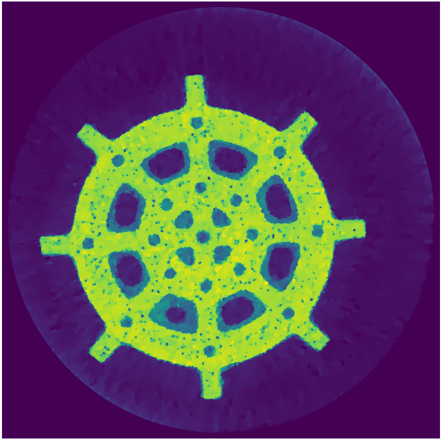
undersampled



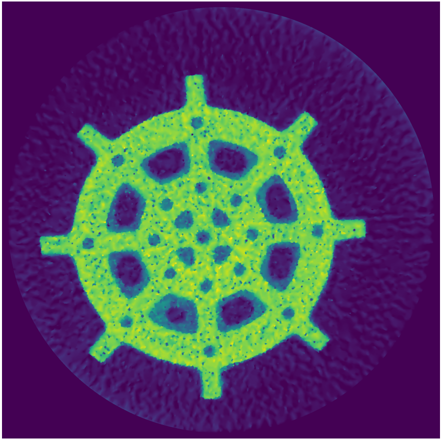
fully sampled



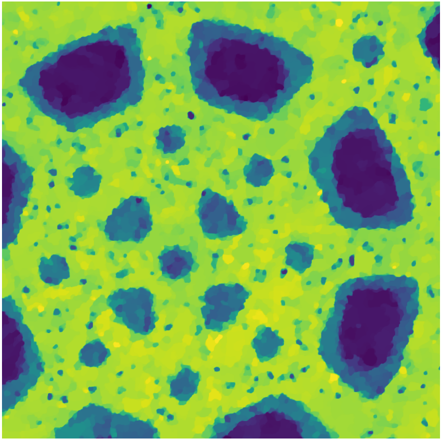
undersampled



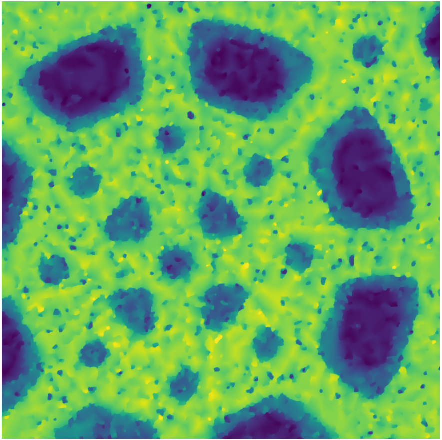
undersampled tv



undersampled tgv



undersampled tv



undersampled tgv

some conclusions

- We have presented some connexions between ideas of related backpropagation (Shaban) and nonsmooth AD (Grazzi)

- Extension when F is accelerated pGD (FISTA) requires more work, it seems not to work numerically, further analysis needed
- Extension when F is a primal-dual map also unclear
- Bounds are valid for sufficiently small step size γ, further analysis needed to consider adaptive step sizes *e.g.* by backtracking

thanks !