Linesearch-enhanced forward-backward methods for inexact nonconvex scenarios

Marco Prato

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dipartimento di Scienze Fisiche, Informatiche e Matematiche
Universitá di Modena e Reggio Emilia

OASIS Optimization Algorithms and Software for Inverse problemS

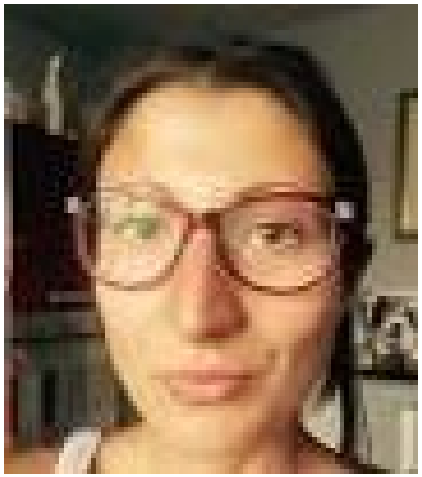
Optimization Algorithms and Software for Inverse problemS
www.oasis.unimore.it

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

STILE Workshop - Bologna, 16 June 2025

Joint work with:

Silvia Bonettini



Giorgia Franchini

Danilo Pezzi

Main paper:

- Bonettini S., Franchini G. Pezzi D., Prato M. 2025, Linesearch-Enhanced Forward–Backward Methods for Inexact Nonconvex Scenarios. *SIAM Journal on Imaging Sciences* **18**(2), 1314–1343

Bayesian/variational approach to image restoration

Direct discrete model

$$\underset{\text{data}}{b} = \underset{\text{linear model}}{Hu_{true}} + \underset{\text{noise}}{\nu}$$

$$\left(b, \nu \in \mathbb{R}^m, H \in \mathbb{R}^{m \times n}, u_{true} \in \mathbb{R}^n\right)$$

Goal:

Try to recover u_{true} by knowing b and H.

Variational formulation

$$u_{true} \simeq \tilde{u} \in \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} E(u, x; b), \quad E(u, x; b) = d(Hu, x; b) + \mathcal{R}(u, x)$$

- $d(Hu, x; b)$ expresses the data discrepancy
- $\mathcal{R}(u, x)$ is a regularization term, enforcing some desired property on u_{true};
- $x \in \mathbb{R}^p$ represents all the hyperparameters of the discrepancy and regularization functions

Supervised learning of the hyperparameters

Given a dataset $\mathcal{D} = \{(b^{(i)}, u_{true}^{(i)})\}_{i=1}^{S}$, where $b^{(i)} = Hu_{true}^{(i)} + \nu^{(i)}$, choose a suitable energy functional $E(u, x; b^{(i)})$ whose minimizers, for a proper array of parameters x, are supposed to be close to $u_{true}^{(i)}$.

Goal:

Find the optimal set of parameters x

Bilevel optimization problem

$$\min_{x \in \mathbb{R}^p} \quad \sum_{i=1}^{S} \mathcal{L}(\tilde{u}^{(i)}(x), u_{true}^{(i)}) + R(x)$$

$$\text{s.t.} \quad \tilde{u}^{(i)}(x) = \operatorname*{argmin}_{u \in \mathbb{R}^n} E(u, x; b^{(i)}), \quad i = 1, \ldots, S,$$

- $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is the loss function
- $R(x)$ is a regularization term on the hyperparameters.

Bilevel optimization problem

$$\min_{x\in\mathbb{R}^p} \overbrace{\sum_{i=1}^{S} \mathcal{L}(\tilde{u}^{(i)}(x), u_{true}^{(i)})}^{f_0(x)} + \overbrace{R(x)}^{f_1(x)} \quad \text{s.t.} \quad \tilde{u}^{(i)}(x) = \operatorname*{argmin}_{u\in\mathbb{R}^n} E(u, x; b^{(i)})$$

Under some assumptions, f_0 is smooth and

$$\nabla f_0(x) = -\sum_{i=1}^{S} \nabla^2_{xu} E(\tilde{u}^{(i)}(x), x; b^{(i)}) \nabla^2_{uu} E(\tilde{u}^{(i)}(x), x; b^{(i)})^{-1} \nabla_u \mathcal{L}(\tilde{u}^{(i)}(x), u_{true}^{(i)})$$

- Function and gradient evaluation require the exact minimizer of $E(u, x; b^{(i)})$
- Function and gradient can be approximated within any given tolerance ϵ
  1. Compute $\bar{u}^{(i)}(x) \simeq \tilde{u}^{(i)}(x)$ such that (iterative optimization method)

     $$\|\nabla_u E(\bar{u}^{(i)}(x), x; b^{(i)})\| \leq \epsilon$$

     and set

     $$\begin{aligned} \bar{f} &= \textstyle\sum_{i=1}^{S} \mathcal{L}(\bar{u}^{(i)}(x), u_{true}^{(i)}) \\ g &= -\textstyle\sum_{i=1}^{S} \nabla^2_{xu} E(\bar{u}^{(i)}(x), x; b^{(i)}) \nabla^2_{uu} E(\bar{u}^{(i)}(x), x; b^{(i)})^{-1} \nabla_u \mathcal{L}(\bar{u}^{(i)}(x), u_{true}^{(i)}) \end{aligned}$$

  2. If $E(\cdot, x; b)$ is strongly convex and its second derivatives are Lipschitz continuous

     $$|\bar{f} - f_0(x)| \leq C_f \epsilon \qquad \text{and} \qquad \|g - \nabla f_0(x)\| \leq C_g \epsilon$$

Reference problem and main assumptions

$$\min_{x \in \mathbb{R}^p} \quad \sum_{i=1}^{S} \mathcal{L}(\tilde{u}^{(i)}(x), u_{true}^{(i)}) + R(x)$$

$$\min_{x \in \mathbb{R}^p} f(x) \equiv f_0(x) + f_1(x)$$

Assumption: the objective function f as the sum of a smooth, possibly nonconvex term and a convex, possibly nonsmooth one

f_0 is smooth

we have the **gradient**

$$\nabla f_0(x)$$

f_1 is closed and convex

we have the **proximity** (or resolvent) operator:

$$\text{prox}_{f_1}(z) = \underset{x \in \mathbb{R}^p}{\text{argmin}}\, f_1(x) + \frac{1}{2}\|x - z\|^2$$

Forward-backward method with inexact gradient evaluation

- Forward-backward iteration with inexact gradient evaluation:

$$\begin{aligned} g^{(k)} &\simeq \nabla f_0(x^{(k)}) \text{ such that } \|g^{(k)} - \nabla f_0(x^{(k)})\| \leq \epsilon_k \\ y^{(k)} &= \operatorname{prox}_{\alpha_k f_1}(x^{(k)} - \alpha_k g^{(k)}) \\ d^{(k)} &= y^{(k)} - x^{(k)} \\ x^{(k+1)} &= x^{(k)} + \lambda_k d^{(k)} \end{aligned}$$

- Existing approaches: prefixed tolerance sequence $\{\epsilon_k\}_{k\in\mathbb{N}}$ with a prescribed decay rate, restrictions on the steplength parameters (e.g. $\alpha_k, \lambda_k < 2/L$)

  Example:

$$\epsilon_k \sim k^{-2}, \qquad \lambda_k = 1, \qquad \alpha_k = 1/L$$

  + some adaptive tricks if L is not available ([Pedregosa 2016])

- Forward-backward iteration with inexact gradient evaluation:

$$\begin{aligned} g^{(k)} &\simeq \nabla f_0(x^{(k)}) \text{ such that } \|g^{(k)} - \nabla f_0(x^{(k)})\| \leq \epsilon_k \\ y^{(k)} &= \operatorname{prox}_{\alpha_k f_1}(x^{(k)} - \alpha_k g^{(k)}) \\ d^{(k)} &= y^{(k)} - x^{(k)} \\ x^{(k+1)} &= x^{(k)} + \lambda_k d^{(k)} \end{aligned}$$

- **Our approach**: adaptive choices for ϵ_k, α_k and line–search for λ_k

Our adaptive accuracy criterion

At each iteration k, compute $g^{(k)} \in \mathbb{R}^n$ such that

$$\|\nabla f_0(x^{(k)}) - g^{(k)}\| \leq \underbrace{C\|x^{(k)} - x^{(k-1)}\|}_{\epsilon_k},$$

for some $C \geq 0$.

$$\min_{x \in \mathbb{R}^p} f(x) \equiv f_0(x)$$

- Standard gradient method:

$$\begin{aligned} g^{(k)} &= \nabla f_0(x^{(k)}) \\ d^{(k)} &= -g^{(k)} \\ x^{(k+1)} &= x^{(k)} + \lambda_k d^{(k)} \end{aligned}$$

- Armijo line–search rule: find $\lambda_k > 0$ such that

$$f(x^{(k+1)}) \leq f(x^{(k)}) + \sigma \lambda_k \Delta_k$$

where $\sigma \in (0, 1)$ and

$$\Delta_k = \langle g^{(k)}, d^{(k)} \rangle < 0$$

$$\min_{x \in \mathbb{R}^p} f(x) \equiv f_0(x) + \iota_\Omega(x)$$

- Standard gradient projection method:

$$\begin{aligned} g^{(k)} &= \nabla f_0(x^{(k)}) \\ y^{(k)} &= P_\Omega(x^{(k)} - \alpha_k g^{(k)}) \\ d^{(k)} &= y^{(k)} - x^{(k)} \\ x^{(k+1)} &= x^{(k)} + \lambda_k d^{(k)} \end{aligned}$$

- Armijo line–search rule: find $\lambda_k \in (0, 1]$ such that

$$f(x^{(k+1)}) \leq f(x^{(k)}) + \sigma \lambda_k \Delta_k$$

where $\sigma \in (0, 1)$ and

$$\Delta_k = \langle g^{(k)}, d^{(k)} \rangle + \frac{1}{2\alpha_k} \|d^{(k)}\|^2 \leq -\frac{1}{2\alpha_k} \|d^{(k)}\|^2 < 0$$

$$\min_{x \in \mathbb{R}^p} f(x) \equiv f_0(x) + f_1(x)$$

- Forward-backward method with exact gradient:

$$\begin{aligned} g^{(k)} &= \nabla f_0(x^{(k)}) \\ y^{(k)} &= \mathrm{prox}_{\alpha_k f_1}(x^{(k)} - \alpha_k g^{(k)}) \\ d^{(k)} &= y^{(k)} - x^{(k)} \\ x^{(k+1)} &= x^{(k)} + \lambda_k d^{(k)} \end{aligned}$$

- Generalized Armijo line–search rule ([Bonettini et al 2016]): find $\lambda_k \in (0, 1]$ such that

$$f(x^{(k+1)}) \leq f(x^{(k)}) + \sigma \lambda_k \Delta_k$$

where $\sigma \in (0, 1)$ and

$$\Delta_k = f_1(y^{(k)}) - f_1(x^{(k)}) + \langle g^{(k)}, d^{(k)} \rangle + \frac{1}{2\alpha_k} \|d^{(k)}\|^2 \leq -\frac{1}{2\alpha_k} \|d^{(k)}\|^2 < 0$$

$$\min_{x\in\mathbb{R}^p} f(x) \equiv f_0(x) + f_1(x)$$

- Forward-backward method with inexact gradient:

$$\begin{aligned} g^{(k)} &\simeq \nabla f_0(x^{(k)}) \text{ such that } \|g^{(k)} - \nabla f_0(x^{(k)})\| \leq \underbrace{C\|x^{(k)} - x^{(k-1)}\|}_{\epsilon_k} \\ y^{(k)} &= \operatorname{prox}_{\alpha_k f_1}(x^{(k)} - \alpha_k g^{(k)}) \\ d^{(k)} &= y^{(k)} - x^{(k)} \\ x^{(k+1)} &= x^{(k)} + \lambda_k d^{(k)} \end{aligned}$$

- Line–search without errors aims to make $\{f(x^{(k)})\}_{k\in\mathbb{N}}$ monotone decreasing ($d^{(k)}$ is a descent direction for f).
- **Our approach**: make the sequence $\{f(x^{(k)}) + \gamma\|x^{(k)} - x^{(k-1)}\|^2\}_{k\in\mathbb{N}}$ monotone decreasing: find $\lambda_k \in (0,1]$ such that

$$f(x^{(k+1)}) + \gamma\|x^{(k+1)} - x^{(k)}\|^2 \leq f(x^{(k)}) + \gamma\|x^{(k)} - x^{(k-1)}\|^2 + \sigma\lambda_k\Delta_k$$

where $\sigma \in (0,1)$ and

$$\Delta_k = f_1(y^{(k)}) - f_1(x^{(k)}) + \langle g^{(k)}, d^{(k)}\rangle + \frac{1}{2\alpha_k}\|d^{(k)}\|^2 \leq -\frac{1}{2\alpha_k}\|d^{(k)}\|^2 < 0$$

Bilevel optimization problem

$$\min_{x\in\mathbb{R}^p} \overbrace{\sum_{i=1}^{S} \mathcal{L}(\tilde{u}^{(i)}(x), u_{true}^{(i)})}^{f_0(x)} + \overbrace{R(x)}^{f_1(x)} \quad \text{s.t.} \quad \tilde{u}^{(i)}(x) = \operatorname*{argmin}_{u\in\mathbb{R}^n} E(u, x; b^{(i)})$$

Under some assumptions, f_0 is smooth and

$$\nabla f_0(x) = -\sum_{i=1}^{S} \nabla^2_{xu} E(\tilde{u}^{(i)}(x), x; b^{(i)}) \nabla^2_{uu} E(\tilde{u}^{(i)}(x), x; b^{(i)})^{-1} \nabla_u \mathcal{L}(\tilde{u}^{(i)}(x), u_{true}^{(i)})$$

- Function and gradient evaluation require the exact minimizer of $E(u, x; b^{(i)})$
- Function and gradient can be approximated within any given tolerance ϵ
  1. Compute $\bar{u}^{(i)}(x) \simeq \tilde{u}^{(i)}(x)$ such that (iterative optimization method)

     $$\|\nabla_u E(\bar{u}^{(i)}(x), x; b^{(i)})\| \leq \epsilon$$

     and set

     $$\begin{aligned} \bar{f} &= \textstyle\sum_{i=1}^{S} \mathcal{L}(\bar{u}^{(i)}(x), u_{true}^{(i)}) \\ g &= -\textstyle\sum_{i=1}^{S} \nabla^2_{xu} E(\bar{u}^{(i)}(x), x; b^{(i)}) \nabla^2_{uu} E(\bar{u}^{(i)}(x), x; b^{(i)})^{-1} \nabla_u \mathcal{L}(\bar{u}^{(i)}(x), u_{true}^{(i)}) \end{aligned}$$

  2. If $E(\cdot, x; b)$ is strongly convex and its second derivatives are Lipschitz continuous

     $$|\bar{f} - f_0(x)| \leq C_f \epsilon \qquad \text{and} \qquad \|g - \nabla f_0(x)\| \leq C_g \epsilon$$

Line–search with inexact gradient and function evaluations

- **Last step**: compute $f^{(k)} \simeq f(x^{(k)})$ and make the sequence $\{f^{(k)} + \gamma\|x^{(k)} - x^{(k-1)}\|^2\}_{k\in\mathbb{N}}$ monotone decreasing

Line–search

Choose $\gamma > 0$ at the beginning and assume α_k be given.

1. Compute any value $f^{(k)} \simeq f(x^{(k)})$ such that
$$|f^{(k)} - f(x^{(k)})| \leq C_f\|x^{(k)} - x^{(k-1)}\|^2$$
for some $C_f \geq 0$.
2. Backtracking: compute $x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$, with $\lambda_k \in (0, 1]$ such that
$$f^{(k+1)} + \gamma\|x^{(k+1)} - x^{(k)}\|^2 \leq f^{(k)} + \gamma\|x^{(k)} - x^{(k-1)}\|^2 + \sigma\lambda_k\Delta_k$$
where $\sigma \in (0, 1)$ and
$$|f^{(k+1)} - f(x^{(k+1)})| \leq C_f\|x^{(k+1)} - x^{(k)}\|^2$$
$$\Delta_k = f_1(y^{(k)}) - f_1(x^{(k)}) + \frac{1}{2\alpha_k}\|d^{(k)}\|^2 + \langle g^{(k)}, d^{(k)}\rangle \leq -\frac{1}{2\alpha_k}\|d^{(k)}\|^2 < 0$$

- The line–search is well posed
- The sequence $\{f^{(k)} + \gamma\|x^{(k)} - x^{(k-1)}\|^2\}_{k\in\mathbb{N}}$ is monotone decreasing ([Bonettini et al 2025])

Back to bilevel problems

Given a dataset $\mathcal{D} = \{(b^{(i)}, u_{true}^{(i)})\}_{i=1}^{S}$, solve

$$\min_{x \in \mathbb{R}^p} \quad \sum_{i=1}^{S} \mathcal{L}(\tilde{u}^{(i)}(x), u_{true}^{(i)}) + R(x) \quad \text{s.t.} \quad \tilde{u}^{(i)}(x) = \operatorname*{argmin}_{u \in \mathbb{R}^n} E(u, x; b^{(i)})$$

Remark

At each iterate k, compute $u^{(i,k)} \simeq \tilde{u}^{(i)}(x^{(k)})$ satisfying

$$\|\nabla_u E(u^{(i,k)}; x^{(k)}, b^{(i)})\| \leq \rho \min\{\|x^{(k)} - x^{(k-1)}\|^2, \|x^{(k)} - x^{(k-1)}\|\}$$

Then, compute $f^{(k)}, g^{(k)}$ as

$$\begin{aligned} f^{(k)} &= \textstyle\sum_{i=1}^{S} \mathcal{L}(u^{(i,k)}, u_{true}^{(i)}) \\ g^{(k)} &= -\textstyle\sum_{i=1}^{S} \nabla_{xu}^2 E(u^{(i,k)}, x^{(k)}, b^{(i)}) \nabla_{uu}^2 E(u^{(i,k)}, x^{(k)}\, b^{(i)})^{-1} \nabla_u \mathcal{L}(u^{(i,k)}, u_{true}^{(i)}) \end{aligned}$$

We have

$$\|g^{(k)} - \nabla f_0(x^{(k)})\| \leq C_g \|x^{(k)} - x^{(k-1)}\|, \text{ for some } C_g > 0$$

$$|f^{(k)} - f(x^{(k)})| \leq C_f \|x^{(k)} - x^{(k-1)}\|^2 \text{ for some } C_f \geq 0$$

NB:

- the algorithm implementation does not require the knowledge of C_g, C_f, but only to provide the forcing term $\rho \geq 0$.
- the linear system for computing the inexact gradient can be solved inexactly, within the same tolerance.

inexact Proximal Gradient method with Line–search (iPGL)

$$\begin{aligned} \tilde{y}^{(k)} &= \operatorname{prox}_{\alpha_k f_1}(x^{(k)} - \alpha_k g^{(k)}) \\ d^{(k)} &= \tilde{y}^{(k)} - x^{(k)} \\ x^{(k+1)} &= x^{(k)} + \lambda_k d^{(k)} \end{aligned}$$

where

$$\begin{aligned} \|g^{(k)} - \nabla f_0(x^{(k)})\| &\leq C_g \|x^{(k)} - x^{(k-1)}\| \\ f^{(k+1)} + \gamma {\lambda_k}^2 \|d^{(k)}\|^2 &\leq f^{(k)} + \gamma \|x^{(k)} - x^{(k-1)}\|^2 + \sigma \lambda_k \Delta_k \end{aligned}$$

with $|f^{(k)} - f(x^{(k)})| \leq C_f \|x^{(k)} - x^{(k-1)}\|^2$.

Additional features:

- Add a new parameter, a s.p.d. scaling matrix D_k which determines a different metric at each iterate:

$$\text{replace } \|x\| \text{ with } \|x\|_{D_k} = x^T D_k x$$

- Approximate proximal operator evaluation ([Salzo&Villa 2012], [Bonettini et al 2017, 2023]).

Theorem

Let $\{x^{(k)}\}_{k\in\mathbb{N}}$ be the sequence generated by iPGL where

- $\alpha_k \in [\alpha_{min}, \alpha_{max}]$
- ∇f_0 is Lipschitz

Then, the following holds true.

1. Any limit point of the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ is a stationary point of f and $\lim_{k\to\infty} f^{(k)} = f(x^*)$.
2. If, in addition, the function defined as

$$\mathcal{F}(x, s) = f(x) + \frac{1}{2}s^2,$$

is a KL function and the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ is bounded, then $\{x^{(k)}\}_{k\in\mathbb{N}}$ converges to a stationary point x^* of f.

NB: α_k does not play any role in the convergence proof. It can be used as an acceleration strategy to improve practical performances (Barzilai-Borwein, Ritz values [Fletcher 2012]).

Let the test set $\{(b^{(i)}, u_{true}^{(i)})\}_{i=1}^{20}$ be given, where

- $u_{true}^{(i)} \in \mathbb{R}^{128^2}$ is a natural image from BSDS representing the ground truth
- $b^{(i)}$ is obtained by blurring the ground truth with a 5×5 kernel then adding Gaussian noise

Compute the optimal parameters x_1, x_2, x_3, x_4 by solving

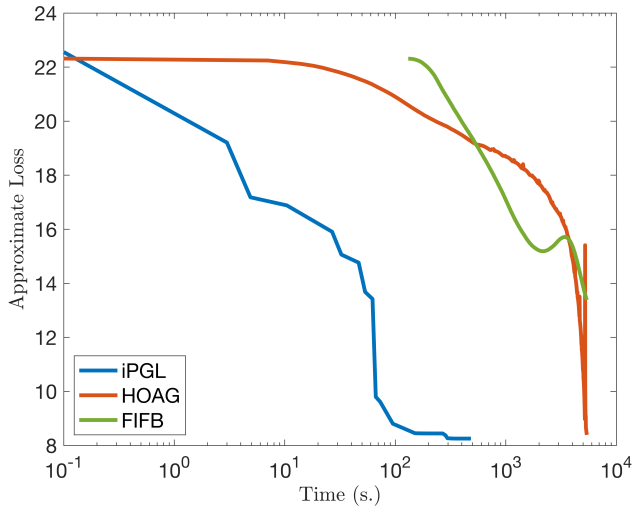
$$\min_{x \in \mathbb{R}^4} \quad f(x) \equiv \sum_{i=1}^{20} \frac{1}{2} \|\tilde{u}^{(i)}(x) - u_{true}^{(i)}\|^2 + \iota_{\mathcal{X}}(x)$$

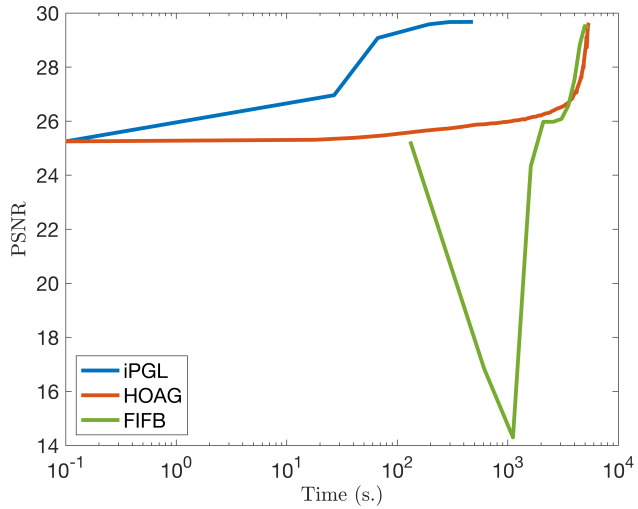
$$\text{s.t.} \quad \tilde{u}^{(i)}(x) = \operatorname*{argmin}_{u \in \mathbb{R}^{128^2}} E(u, x; b^{(i)}) \equiv \tfrac{1}{2} \|H(x_2, x_3, x_4)u - b^{(i)}\|^2 + x_1 \|\mathcal{D}u\|_{1,\varepsilon}$$
$$i = 1, \dots, 20$$

- x_1 is the weight of the regularization term in the energy functional
- x_2, x_3, x_4 are the blurring matrix parameters
- $\mathcal{X} = \{x \in \mathbb{R}^4 : 10^{-9} \leq x_1 \leq 10 \wedge x_2 + x_3 + x_4 = 1\}$
- $\mathcal{D}$ is the discrete gradient and $\|\cdot\|_{1,\varepsilon}$ is a C^2 convex approximation of the 1-norm

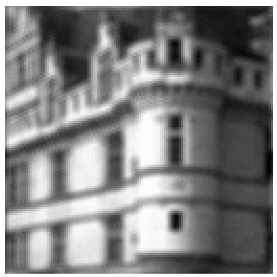
Numerical assessment

- SGP ([Bonettini et al 2009]) used for inner minimization
- performance comparison with HOAG ([Pedregosa 2016]) and FIFB ([Suonperä&Valkonen 2024])
- quality of the results is evaluated on a test set of 10 further images

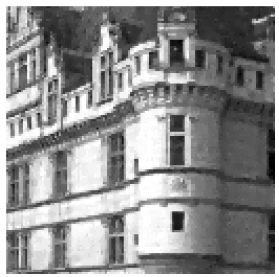




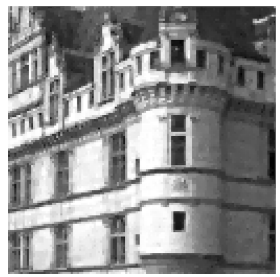
ground truth



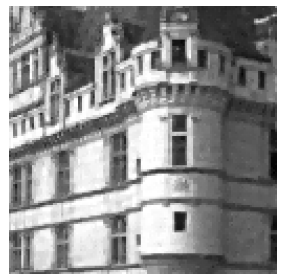
noisy blurred



iPGL



HOAG



FIFB

Conclusions and perspectives

Proposal:

- Forward-backward method allowing inexactness in
  - gradient evaluation
  - function evaluation
  - proximity operator evaluation
- line-search with inexact function values
- convergence analysis in the KL framework complying with the implementation
- solves bilevel problems
- may include inertial and variable metric techniques

Future work:

- consider different (unrolling–like) inexactness criteria
- bounded, nondiminishing errors
- application to CT images (?)

Bibliography

- J. Barzilai and J. M. Borwein 1988, Two-point step size gradient methods, IMA Journal on Numerical Analysis 8, 141–148
- S. Bonettini, G. Franchini, D. Pezzi, M. Prato 2025, Linesearch-enhanced forward-backward methods for inexact nonconvex scenarios. SIAM Journal on Imaging Sciences 18(2), 1314–1343
- S. Bonettini, P. Ochs, M. Prato, S. Rebegoldi 2023, An abstract convergence framework with application to inertial inexact forward-backward methods. Computational Optimization and Applications 84(2), 319–362
- S. Bonettini, I. Loris, F. Porta, M. Prato, S. Rebegoldi 2017, On the convergence of a linesearch based proximal-gradient method for nonconvex optimization. Inverse Problems 33, 055005
- S. Bonettini, R. Zanella, L. Zanni 2009, A scaled gradient projection method for constrained image deblurring. Inverse Problems 25, 015002
- R. Fletcher 2012, A limited memory steepest descent method. Mathematical Programming 135, 413–436
- F. Pedregosa 2016, Hyperparameter optimization with approximate gradient. Proceedings of Machine Learning Research 48, 737–746
- S. Salzo and S. Villa 2012, Inexact and accelerated proximal point algorithms. Journal of Convex Analysis 19, 1167–1192
- E. Suonperä and T. Valkonen 2024, Linearly convergent bilevel optimization with single-step inner methods. Computational Optimization and Applications 87, 571–610