Learning to solve TV regularised problems with unrolled algorithms

Hamza Cherkaoui
Université Paris-Saclay, CEA, Inria
Gif-sur-Yvette, 91190, France
hamza.cherkaoui@cea.fr

Jeremias Sulam
Johns Hopkins University
jsulam1@jhu.edu

Thomas Moreau
Université Paris-Saclay, Inria, CEA,
Palaiseau, 91120, France
thomas.moreau@inria.fr

Abstract

Total Variation (TV) is a popular regularization strategy that promotes piece-wise constant signals by constraining the ℓ_1-norm of the first order derivative of the estimated signal. The resulting optimization problem is usually solved using iterative algorithms such as proximal gradient descent, primal-dual algorithms or ADMM. However, such methods can require a very large number of iterations to converge to a suitable solution. In this paper, we accelerate such iterative algorithms by unfolding proximal gradient descent solvers in order to learn their parameters for 1D TV regularized problems. While this could be done using the *synthesis* formulation, we demonstrate that this leads to slower performances. The main difficulty in applying such methods in the *analysis* formulation lies in proposing a way to compute the derivatives through the proximal operator. As our main contribution, we develop and characterize two approaches to do so, describe their benefits and limitations, and discuss the regime where they can actually improve over iterative procedures. We validate those findings with experiments on synthetic and real data.

1 Introduction

Ill-posed inverse problems appear naturally in signal and image processing and machine learning, requiring extra regularization techniques. Total Variation (TV) is a popular regularization strategy with a long history (Rudin et al., 1992), and has found a large number of applications in neuro-imaging (Fikret et al., 2013), medical imaging reconstruction (Tian et al., 2011), among myriad applications (Rodríguez, 2013; Darbon and Sigelle, 2006). TV promotes piece-wise constant estimates by penalizing the ℓ_1-norm of the first order derivative of the estimated signal, and it provides a simple, yet efficient regularization technique.

TV-regularized problems are typically convex, and so a wide variety of algorithms are in principle applicable. Since the ℓ_1 norm in the TV term is non-smooth, Proximal Gradient Descent (PGD) is the most popular choice (Rockafellar, 1976). Yet, the computation for the corresponding proximal operator (denoted prox-TV) represents a major difficulty in this case as it does not have a closed-form analytic solution. For 1D problems, it is possible to rely on dynamic programming to compute prox-TV, such as the taut string algorithm (Davies and Kovac, 2001; Condat, 2013a). Another alternative consists in computing the proximal operator with iterative first order algorithm (Chambolle, 2004; Beck and Teboulle, 2009; Boyd et al., 2011; Condat, 2013b). Other algorithms to solve TV-regularized

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

problems rely on primal dual algorithms (Chambolle and Pock, 2011; Condat, 2013b) or Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011). These algorithms typically use one sequence of estimates for each term in the objective and try to make them as close as possible while minimizing the associated term. While these algorithms are efficient for denoising problems – where one is mainly concerned with good reconstruction – they can result in estimate that are not very well regularized if the two sequences are not close enough.

When on fixed computational budget, iterative optimization methods can become impractical as they often require many iterations to give a satisfactory estimate. To accelerate the resolution of these problems with a finite (and small) number of iterations, one can resort to unrolled and learned optimization algorithms (see Monga et al. 2019 for a review). In their seminal work, Gregor and Le Cun (2010) proposed the Learned ISTA (LISTA), where the parameters of an unfolded Iterative Shrinkage-Thresholding Algorithm (ISTA) are learned with gradient descent and back-propagation. This allows to accelerate the approximate solution of a Lasso problem (Tibshirani, 1996), with a fixed number of iteration, for signals from a certain distribution. The core principle behind the success of this approach is that the network parameters can adaptively leverage the sensing matrix structure (Moreau and Bruna, 2017) as well as the input distribution (Giryes et al., 2018; Ablin et al., 2019). Many extensions of this original idea have been proposed to learn different algorithms (Sprechmann et al., 2012, 2013; Borgerding et al., 2017) or for different classes of problem (Xin et al., 2016; Giryes et al., 2018; Sulam et al., 2019). The motif in most of these adaptations is that all operations in the learned algorithms are either linear or separable, thus resulting in sub-differentials that are easy to compute and implement via back-propagation. Algorithm unrolling is also used in the context of bi-level optimization problems such as hyper-parameter selection. Here, the unrolled architecture provides a way to compute the derivative of the inner optimization problem solution compared to another variable such as the regularisation parameter using back-propagation (Bertrand et al., 2020).

The focus of this paper is to apply algorithm unrolling to TV-regularized problems in the 1D case. While one could indeed apply the LISTA approach directly to the *synthesis* formulation of these problems, we show in this paper that using such formulation leads to slower iterative or learned algorithms compared to their *analysis* counterparts. The extension of learnable algorithms to the analysis formulation is not trivial, as the inner proximal operator does not have an analytical or separable expression. We propose two architectures that can learn TV-solvers in their analysis form directly based on PGD. The first architecture uses an exact algorithm to compute the prox-TV and we derive the formulation of its weak Jacobian in order to learn the network's parameters. Our second method rely on a nested LISTA network in order to approximate the prox-TV itself in a differentiable way. This latter approach can be linked to inexact proximal gradient methods (Schmidt et al., 2011; Machart et al., 2012). These results are backed with numerical experiments on synthetic and real data. Concurrently to our work, Lecouat et al. (2020) also proposed an approach to differentiate the solution of TV-regularized problems. While their work can be applied in the context of 2D signals, they rely on smoothing the regularization term using Moreau-Yosida regularization, which results in smoother estimates from theirs learned networks. In contrast, our work allows to compute sharper signals but can only be applied to 1D signals.

The rest of the paper is organized as follows. In Section 2, we describe the different formulations for TV-regularized problems and their complexity. We also recall central ideas of algorithm unfolding. Section 3 introduces our two approaches for learnable network architectures based on PGD. Finally, the two proposed methods are evaluated on real and synthetic data in Section 4.

Notations For a vector $x \in \mathbb{R}^k$, we denote $\|x\|_q$ its ℓ_q-norm. For a matrix $A \in \mathbb{R}^{m\times k}$, we denote $\|A\|_2$ its ℓ_2-norm, which corresponds to its largest singular value and $A^\dagger$ denotes its pseudo-inverse. For an ordered subset of indices $\mathcal{S} \subset \{1, \dots, k\}$, $x_\mathcal{S}$ denote the vector in $\mathbb{R}^{|\mathcal{S}|}$ with element $(x_\mathcal{S})_t = x_{i_t}$ for $i_t \in \mathcal{S}$. For a matrix $A \in \mathbb{R}^{m\times k}$, $A_{:,\mathcal{S}}$ denotes the sub-matrix $[A_{:,i_1}, \dots A_{:,i_{|\mathcal{S}|}}]$ composed with the columns $A_{:,i_t}$ of index $i_t \in \mathcal{S}$ of A. For the rest of the paper, we refer to the operators $D \in \mathbb{R}^{k-1\times k}, \widetilde{D} \in \mathbb{R}^{k\times k}, L \in \mathbb{R}^{k\times k}$ and $R \in \mathbb{R}^{k\times k}$ as:

$$D = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix} \quad \widetilde{D} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ -1 & 1 & \ddots & \vdots \\ & \ddots & \ddots & 0 \\ & 0 & -1 & 1 \end{bmatrix} \quad L = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 1 & \dots & 1 & 1 \end{bmatrix} \quad R = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{bmatrix}$$

2 Solving TV-regularized problems

We begin by detailing the TV-regularized problem that will be the main focus of our work. Consider a latent vector $u \in \mathbb{R}^k$, a design matrix $A \in \mathbb{R}^{m \times k}$ and the corresponding observation $x \in \mathbb{R}^m$. The original formulation of the TV-regularized regression problem is referred to as the *analysis* formulation (Rudin et al., 1992). For a given regularization parameter $\lambda > 0$, it reads

$$\min_{u \in \mathbb{R}^k} P(u) = \frac{1}{2}\|x - Au\|_2^2 + \lambda \|u\|_{TV}, \tag{1}$$

where $\|u\|_{TV} = \|Du\|_1$, and $D \in \mathbb{R}^{k-1 \times k}$ stands for the first order finite difference operator, as defined above. The problem in (1) can be seen as a special case of a Generalized Lasso problem (Tibshirani and Taylor, 2011); one in which the analysis operator is D. Note that problem P is convex, but the TV-norm is non-smooth. In these cases, a practical alternative is the PGD, which iterates between a gradient descent step and the prox-TV. This algorithm's iterates read

$$u^{(t+1)} = \text{prox}_{\frac{\lambda}{\rho}\|\cdot\|_{TV}}\left(u^{(t)} - \frac{1}{\rho}A^\top(Au^{(t)} - x)\right) \quad , \tag{2}$$

where $\rho = \|A\|_2^2$ and the prox-TV is defined as

$$\text{prox}_{\mu\|\cdot\|_{TV}}(y) = \arg\min_{u \in \mathbb{R}^k} F_y(u) = \frac{1}{2}\|y - u\|_2^2 + \mu\|u\|_{TV}. \tag{3}$$

Problem (3) does not have a closed-form solution, and one needs to resort to iterative techniques to compute it. In our case, as the problem is 1D, the prox-TV problem can be addressed with a dynamic programming approach, such as the taut-string algorithm (Condat, 2013a). This scales as $O(k)$ in all practical situations and is thus much more efficient than other optimization based iterative algorithms (Rockafellar, 1976; Chambolle, 2004; Condat, 2013b) for which each iteration is $O(k^2)$ at best.

With a generic matrix $A \in \mathbb{R}^{m \times k}$, the PGD algorithm is known to have a sublinear convergence rate (Combettes and Bauschke, 2011). More precisely, for any initialization $u^{(0)}$ and solution u^*, the iterates satisfy

$$P(u^{(t)}) - P(u^*) \leq \frac{\rho}{2t}\|u^{(0)} - u^*\|_2^2, \tag{4}$$

where u^* is a solution of the problem in (1). Note that the constant ρ can have a significant effect. Indeed, it is clear from (4) that doubling ρ leads to consider doubling the number of iterations.

2.1 Synthesis formulation

An alternative formulation for TV-regularized problems relies on removing the analysis operator D from the ℓ_1-norm and translating it into a synthesis expression (Elad et al., 2007). Removing D from the non-smooth term simplifies the expression of the proximal operator by making it separable, as in the Lasso. The operator D is not directly invertible but keeping the first value of the vector u allows for perfect reconstruction. This motivates the definition of the operator $\widetilde{D} \in \mathbb{R}^{k \times k}$, and its inverse $L \in \mathbb{R}^{k \times k}$, as defined previously. Naturally, L is the discrete integration operator. Considering the change of variable $z = \widetilde{D}u$, and using the operator $R \in \mathbb{R}^{k \times k}$, the problem in (1) is equivalent to

$$\min_{z \in \mathbb{R}^k} S(z) = \frac{1}{2}\|x - ALz\|_2^2 + \lambda\|Rz\|_1. \tag{5}$$

Note that for any $z \in \mathbb{R}^k$, $S(z) = P(Lz)$. There is thus an exact equivalence between solutions from the synthesis and the analysis formulation, and the solution for the analysis can be obtained with $u^* = Lz^*$. The benefit of this formulation is that the problem above now reduces to a Lasso problem (Tibshirani, 1996). In this case, the PGD algorithm is reduced to the ISTA with a closed-form proximal operator (the soft-thresholding). Note that this simple formulation is only possible in 1D where the first order derivative space is unconstrained. In larger dimensions, the derivative must be constrained to verify the Fubini's formula that enforces the symmetry of integration over dimensions. While it is also possible to derive synthesis formulation in higher dimension (Elad et al., 2007), this does not lead to simplistic proximal operator.

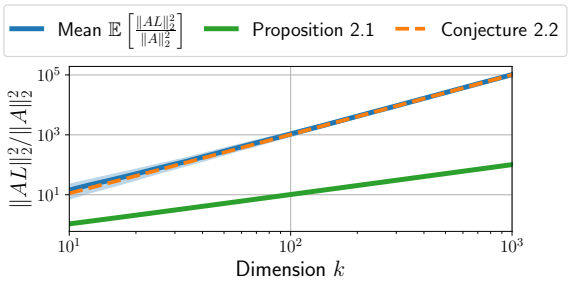


Figure 1: Evolution of $\mathbb{E}\left[\frac{\|AL\|_2^2}{\|A\|_2^2}\right]$ *w.r.t* the dimension k for random matrices A with *i.i.d* normally distributed entries. In light blue is the confidence interval [0.1, 0.9] computed with the quantiles. We observe that it scales as $O(k^2)$ and that our conjectured bound seems tight.

For this synthesis formulation, with a generic matrix $A \in \mathbb{R}^{m\times k}$, the PGD algorithm has also a sublinear convergence rate (Beck and Teboulle, 2009) such that

$$P(u^{(t)}) - P(u^*) \leq \frac{2\widetilde{\rho}}{t}\|u^{(0)} - u^*\|_2^2, \tag{6}$$

with $\widetilde{\rho} = \|AL\|_2^2$ (see Subsection F.1 for full derivation). While the rate of this algorithm is the same as in the analysis formulation – in $O(\frac{1}{t})$ – the constant $\widetilde{\rho}$ related to the operator norm differs. We now present two results that will characterize the value of $\widetilde{\rho}$.

Proposition 2.1. *[Lower bound for the ratio $\frac{\|AL\|_2^2}{\|A\|_2^2}$ expectation] Let A be a random matrix in $\mathbb{R}^{m\times k}$ with* i.i.d *normally distributed entries. The expectation of $\|AL\|_2^2/\|A\|_2^2$ is asymptotically lower bounded when k tends to ∞ by*

$$\mathbb{E}\left[\frac{\|AL\|_2^2}{\|A\|_2^2}\right] \geq \frac{2k+1}{4\pi^2} + o(1)$$

The full proof can be found in Subsection F.3. The lower bound is constructed by using $A^TA \succeq \|A\|_2^2 u_1 u_1^\top$ for a unit vector u_1 and computing explicitely the expectation for rank one matrices. To assess the tightness of this bound, we evaluated numerically $\mathbb{E}\left[\frac{\|AL\|_2^2}{\|A\|_2^2}\right]$ on a set of 1000 matrices sampled with *i.i.d* normally distributed entries. The results are displayed *w.r.t* the dimension k in Figure 1. It is clear that the lower bound from Proposition 2.1 is not tight. This is expected as we consider only the leading eigenvector of A to derive it in the proof. The following conjecture gives a tighter bound.

Conjecture 2.2 (Expectation for the ratio $\frac{\|AL\|_2^2}{\|A\|_2^2}$)**.** *Under the same conditions as in Proposition 2.1, the expectation of $\|AL\|_2^2/\|A\|_2^2$ is given by*

$$\mathbb{E}\left[\frac{\|AL\|_2^2}{\|A\|_2^2}\right] = \frac{(2k+1)^2}{16\pi^2} + o(1) \ .$$

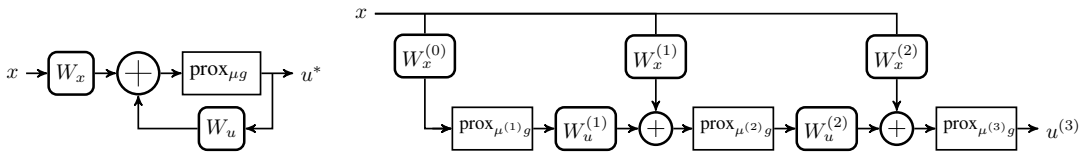
We believe this conjecture can potentially be proven with analogous developments as those in Proposition 2.1, but integrating over all dimensions. However, a main difficulty lies in the fact that integration over all eigenvectors have to be carried out jointly as they are not independent. This is subject of current ongoing work.

Finally, we can expect that $\widetilde{\rho}/\rho$ scales as $\Theta(k^2)$. This leads to the observation that $\frac{\widetilde{\rho}}{2} \gg \rho$ in large enough dimension. As a result, the analysis formulation should be much more efficient in terms of iterations than the synthesis formulation – as long as the prox-TVcan be dealt with efficiently.

2.2 Unrolled iterative algorithms

As shown by Gregor and Le Cun (2010), ISTA is equivalent to a recurrent neural network (RNN) with a particular structure. This observation can be generalized to PGD algorithms for any penalized least squares problem of the form

$$u^*(x) = \arg\min_u \mathcal{L}(x,u) = \frac{1}{2}\|x - Bu\|_2^2 + \lambda g(u) \ , \tag{7}$$



(a) **PGD** - Recurrent Neural Network (b) **LPGD** - Unfolded network for Learned PGD with $T = 3$

Figure 2: **Algorithm Unrolling** - Neural network representation of iterative algorithms. The parameters $\Theta^{(t)} = \{W_x^{(t)}, W_u^{(t)}, \mu^{(t)}\}$ can be learned by minimizing the loss (10) to approximate good solution of (7) on average.

where g is proper and convex, as depicted in Figure 2a. By unrolling this architecture with T layers, we obtain a network $\phi_{\Theta^{(T)}}(x) = u^{(T)}$ – illustrated in Figure 2b – with parameters $\Theta^{(T)} = \{W_x^{(t)}, W_u^{(t)}, \mu^{(t)}\}_{t=1}^T$, defined by the following recursion

$$u^{(0)} = B^\dagger x \; ; \qquad u^{(t)} = \text{prox}_{\mu^{(t)} g}(W_x^{(t)} x + W_u^{(t)} u^{(t-1)}) \; . \tag{8}$$

As underlined by (4), a good estimate $u^{(0)}$ is crucial in order to have a fast convergence toward $u^*(x)$. However, this chosen initialization is mitigated by the first layer of the network which learns to set a good initial guess for $u^{(1)}$. For a network with T layers, one recovers exactly the T-th iteration of PGD if the weights are chosen constant equal to

$$W_x^{(t)} = \frac{1}{\rho} B^\top, \qquad W_u^{(t)} = (\text{Id} - \frac{1}{\rho} B^\top B) \; , \qquad \mu^{(t)} = \frac{\lambda}{\rho}, \qquad \text{with } \rho = \|B\|_2^2 \; . \tag{9}$$

In practice, this choice of parameters are used as initialization for a posterior training stage. In many practical applications, one is interested in minimizing the loss (7) for a fixed B and a particular distribution over the space of x, $\mathcal{P}$. As a result, the goal of this training stage is to find parameters $\Theta^{(T)}$ that minimize the risk, or expected loss, $\mathbb{E}[\mathcal{L}(x, \phi_{\Theta^{(T)}}(x))]$ over $\mathcal{P}$. Since one does not have access to this distribution, and following an empirical risk minimization approach with a given training set $\{x_1, \dots x_N\}$ (assumed sampled *i.i.d* from $\mathcal{P}$), the network is trained by minimizing

$$\min_{\Theta^{(T)}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(x_i, \phi_{\Theta^{(T)}}(x_i)) \; . \tag{10}$$

Note that when $T \to +\infty$, the presented initialization in (9) gives a global minimizer of the loss for all x_i, as the network converges to exact PGD. When T is fixed, however, the output of the network is not a minimizer of (7) in general. Minimizing this empirical risk can therefore find a weight configuration that reduces the sub-optimality of the network relative to (7) over the input distribution used to train the network. In such a way, the network learns an algorithm to approximate the solution of (7) for a particular class or distributions of signals. It is important to note here that while this procedure can accelerate the resolution the problem, the learned algorithm will only be valid for inputs x_i coming from the same input distribution $\mathcal{P}$ as the training samples. The algorithm might not converge for samples which are too different from the training set, unlike the iterative algorithm which is guaranteed to converge for any sample.

This network architecture design can be directly applied to TV regularised problems if the synthesis formulation (5) is used. Indeed, in this case PGD reduces to the ISTA algorithm, with $B = AL$ and $\text{prox}_{\mu g} = \text{ST}(\cdot, \mu)$ becomes simply a soft-thresholding operator (which is only applied on the coordinates $\{2, \dots k\}$, following the definition of R). However, as discussed in Proposition 2.1, the conditioning of the synthesis problem makes the estimation of the solution slow, increasing the number of network layers needed to get a good estimate of the solution. In the next section, we will extend these learning-based ideas directly to the analysis formulation by deriving a way to obtain exact and approximate expressions for the sub-differential of the non-separable prox-TV.

3 Back-propagating through TV proximal operator

Our two approaches to define learnable networks based on PGD for TV-regularised problems in the analysis formulation differ on the computation of the prox-TV and its derivatives. Our first approach

consists in directly computing the weak derivatives of the exact proximal operator while the second one uses a differentiable approximation.

3.1 Derivative of prox-TV

While there is no analytic solution to the prox-TV, it can be computed exactly (numerically) for 1D problems using the taut-string algorithm (Condat, 2013a). This operator can thus be applied at each layer of the network, reproducing the architecture described in Figure 2b. We define the LPGD-Taut network $\phi_{\Theta^{(T)}}(x)$ with the following recursion formula

$$\phi_{\Theta^{(T)}}(x) = \text{prox}_{\mu^{(T)}\|\cdot\|_{TV}}\left(W_x^{(T)}x + W_u^{(T)}\phi_{\Theta^{(T-1)}}(x)\right) \tag{11}$$

To be able to learn the parameters through gradient descent, one needs to compute the derivatives of (10) *w.r.t* the parameters $\Theta^{(T)}$. Denoting $h = W_x^{(t)}x + W_u^{(t)}\phi_{\Theta^{(t-1)}}(x)$ and $u = \text{prox}_{\mu^{(t)}\|\cdot\|_{TV}}(h)$, the application of the chain rule (as implemented efficiently by automatic differentiation) results in

$$\frac{\partial \mathcal{L}}{\partial h} = J_x(h, \mu^{(t)})^\top \frac{\partial \mathcal{L}}{\partial u}\,, \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial \mu^{(t)}} = J_\mu(h, \mu^{(t)})^\top \frac{\partial \mathcal{L}}{\partial u}\,, \tag{12}$$

where $J_x(h,\mu) \in \mathbb{R}^{k\times k}$ and $J_\mu(h,\mu) \in \mathbb{R}^{k\times 1}$ denotes the weak Jacobian of the output of the proximal operator u with respect to the first and second input respectively. We now give the analytic formulation of these weak Jacobians in the following proposition.

Proposition 3.1. *[Weak Jacobian of prox-TV] Let $x \in \mathbb{R}^k$ and $u = prox_{\mu\|\cdot\|_{TV}}(x)$, and denote by $\mathcal{S}$ the support of $z = \widetilde{D}u$. Then, the weak Jacobian J_x and J_μ of the prox-TV relative to x and μ can be computed as*

$$J_x(x,\mu) = L_{:,\mathcal{S}}(L_{:,\mathcal{S}}^\top L_{:,\mathcal{S}})^{-1}L_{:,\mathcal{S}}^\top \quad \textit{and} \quad J_\mu(x,\mu) = -L_{:,\mathcal{S}}(L_{:,\mathcal{S}}^\top L_{:,\mathcal{S}})^{-1}\,\text{sign}(Du)_{\mathcal{S}}$$

The proof of this proposition can be found in Subsection G.1. Note that the dependency in the inputs is only through $\mathcal{S}$ and $\text{sign}(Du)$, where u is a short-hand for $\text{prox}_{\mu\|\cdot\|_{TV}}(x)$. As a result, computing these weak Jacobians can be done efficiently by simply storing $\text{sign}(Du)$ as a mask, as it would be done for a RELU or the soft-thresholding activations, and requiring just $2(k-1)$ bits. With these expressions, it is thus possible to compute gradient relatively to all parameters in the network, and employ them via back-propagation.

3.2 Unrolled prox-TV

As an alternative to the previous approach, we propose to use the LISTA network to approximate the prox-TV (3). The prox-TV can be reformulated with a synthesis approach resulting in a Lasso *i.e.*

$$z^* = \arg\min_z \frac{1}{2}\|h - Lz\|_2^2 + \mu\|Rz\|_1 \tag{13}$$

The proximal operator solution can then be retrieved with $\text{prox}_{\mu\|\cdot\|_{TV}}(h) = Lz^*$. This problem can be solved using ISTA, and approximated efficiently with a LISTA network Gregor and Le Cun (2010). For the resulting architecture – dubbed LPGD-LISTA – $\text{prox}_{\mu\|\cdot\|_{TV}}(h)$ is replaced by a nested LISTA network with a fixed number of layers T_{in} defined recursively with $z^{(0)} = Dh$ and

$$z^{(\ell+1)} = \text{ST}\left(W_z^{(\ell,t)}z^{(\ell)} + W_h^{(\ell,t)}\Phi_{\Theta^{(t)}}, \frac{\mu^{(\ell,t)}}{\rho}\right)\,. \tag{14}$$

Here, $W_z^{(\ell,t)}, W_h^{(\ell,t)}, \mu^{(\ell,t)}$ are the weights of the nested LISTA network for layer ℓ. They are initialized with weights chosen as in (9) to ensure that the initial state approximates the prox-TV. Note that the weigths of each of these inner layers are also learned through back-propagation during training.

The choice of this architecture provides a differentiable (approximate) proximal operator. Indeed, the LISTA network is composed only of linear and soft-thresholding layers – standard tools for deep-learning libraries. The gradient of the network's parameters can thus be computed using classic automatic differentiation. Moreover, if the inner network is not trained, the gradient computed with this method will converge toward the gradient computed using Proposition 3.1 as T_{in} goes to ∞ (see Proposition G.2). Thus, in this untrained setting with infinitely many inner layers, the network is equivalent to LPGD-Taut as the output of the layer also converges toward the exact proximal operator.

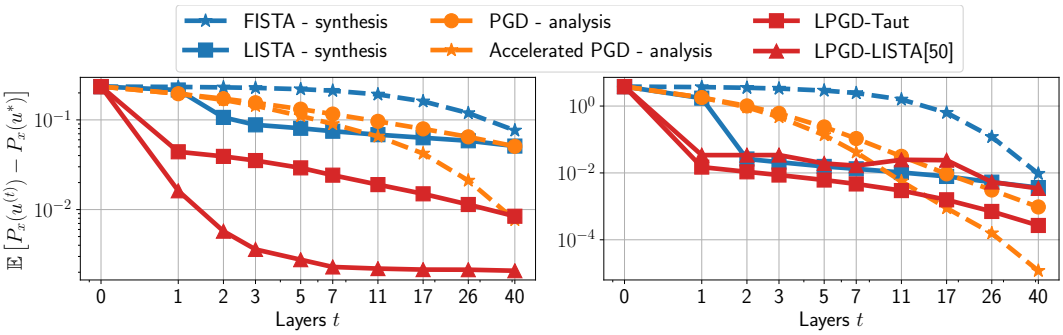


Figure 3: **Performance comparison** for different regularisation levels (*left*) $\lambda = 0.1$, (*right*) $\lambda = 0.8$. We see that synthesis formulations are outperformed by the analysis counter part. Both our methods are able to accelerate the resolution of (20), at least in the first iterations.

Connections to inexact PGD A drawback of approximating the prox-TV via an iterative procedure is, precisely, that it is not exact. This optimization error results from a trade-off between computational cost and convergence rate. Using results from Machart et al. (2012), one can compute the scaling of T and T_{in} to reach an error level of δ with an untrained network. Proposition G.3 shows that without learning, T should scale as $O(\frac{1}{t})$ and T_{in} should be larger than $O(\ln(\frac{1}{\delta}))$. This scaling gives potential guidelines to set these parameters, as one can expect that learning the parameters of the network would reduce these requirement.

4 Experiments

All experiments are performed in Python using `PyTorch` (Paszke et al., 2019). We used the implementation[1] of Barbero and Sra (2018) to compute TV proximal operator using taut-string algorithm. The code to reproduce the figures is available online[2].

In all experiments, we initialize $u_0 = A^\dagger x$. Moreover, we employed a normalized λ_{reg} as a penalty parameter: we first compute the value of $\lambda_{\max}$ (which is the minimal value for which $z = 0$ is solution of (5)) and we refer to λ as the ratio so that $\lambda_{reg} = \lambda\lambda_{\max}$, with $\lambda \in [0, 1]$ (see Appendix D). As the computational complexity of all compared algorithms is the same except for the proximal operator, we compare them in term of iterations.

4.1 Simulation

We generate $n = 2000$ times series and used half for training and other half for testing and comparing the different algorithms. We train all the network's parameters jointly – those to approximate the gradient for each iteration along with those to define the inner proximal operator. The full training process is described in Appendix A. We set the length of the source signals $(u_i)_{i=1}^n \in \mathbb{R}^{n\times k}$ to $k = 8$ with a support of $|S| = 2$ non-zero coefficients (larger dimensions will be showcased in the real data application). We generate $A \in \mathbb{R}^{m\times k}$ as a Gaussian matrix with $m = 5$, obtaining then $(u_i)_{i=1}^n \in \mathbb{R}^{n\times p}$. Moreover, we add Gaussian noise to measurements $x_i = Au_i$ with a signal to noise ratio (SNR) of 1.0.

We compare our proposed methods, LPGD-Taut network and the LPGD-LISTA with $T_{in} = 50$ inner layers to PGD and Accelerated PGD with the analysis formulation. For completeness, we also add the FISTA algorithm for the synthesis formulation in order to illustrate Proposition 2.1 along with its learned version.

Figure 3 presents the risk (or expected function value, P) of each algorithm as a function of the number of layers or, equivalently, iterations. For the learned algorithms, the curves in t display the performances of a network with t layer trained specifically. We observe that all the synthesis formulation algorithms are slower than their analysis counterparts, empirically validating Proposition 2.1.

[1] Available at https://github.com/albarji/proxTV

[2] Available at https://github.com/hcherkaoui/carpet.

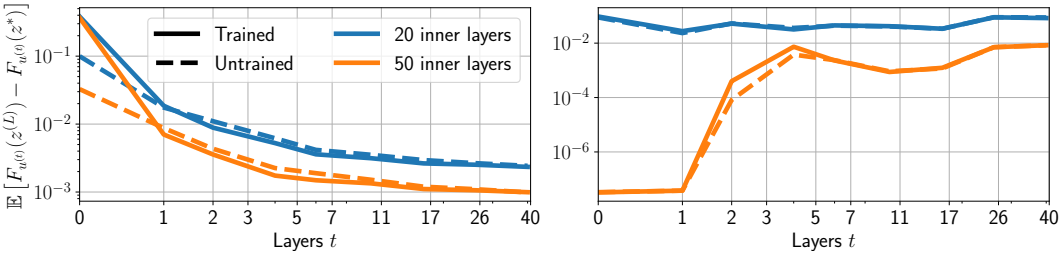


Figure 4: **Proximal operator error comparison** for different regularisation levels (*left*) $\lambda = 0.1$, (*right*) $\lambda = 0.8$. We see that learn the trained unrolled prox-TV barely improve the performance. More interestingly, in a high sparsity context, after a certain point, the error sharply increase.

Moreover, both of the proposed methods accelerate the resolution of (20) in a low iteration regime. However, when the regularization parameter is high ($\lambda = 0.8$), we observe that the performance of the LPGD-LISTA tends to plateau. It is possible that such a high level of sparsity require more than 50 layers for the inner network (which computes the prox-TV). According to Section 3.2, the error associated with this proximity step hinders the global convergence, making the loss function decrease slowly. Increasing the number of inner layers would alleviate this issue, though at the expense of increased computational burden for both training and runtime. For LPGD-Taut, while the Taut-string algorithm ensures that the recovered support is exact for the proximal step, the overall support can be badly estimated in the first iterations. This can lead to un-informative gradients as they greatly depend on the support of the solution in this case, and explain the reduced performances of the network in the high sparsity setting.

Inexact prox-TV With the same data $(x_i)_{i=1}^n \in \mathbb{R}^{n \times m}$, we empirically investigate the error of the prox-TV $\epsilon_k^{(t)} = F_{u^{(t)}}(z^{(t)}) - F_{u^{(t)}}(z^*)$ and evaluate it for c with different number of layers ($T \in [20, 50]$). We also investigate the case where the parameter of the nested LISTA in LPGD-LISTA are trained compared to their initialization in untrained version.

Figure 4 depicts the error ϵ_k for each layer. We see that learning the parameters of the unrolled prox-TV in LPGD-LISTA barely improves the performance. More interestingly, we observe that in a high sparsity setting the error sharply increases after a certain number of layers. This is likely cause by the high sparsity of the estimates, the small numbers of iterations of the inner network (between 20 and 50) are insufficient to obtain an accurate solution to the proximal operator. This is in accordance with inexact PGD theory which predict that such algorithm has no exact convergence guarantees (Schmidt et al., 2011).

4.2 fMRI data deconvolution

Functional magnetic resonance imaging (fMRI) is a non-invasive method for recording the brain activity by dynamically measuring blood oxygenation level-dependent (BOLD) contrast, denoted here x. The latter reflects the local changes in the deoxyhemoglobin concentration in the brain Ogawa et al. (1992) and thus indirectly measures neural activity through the neurovascular coupling. This coupling is usually modelled as a linear and time-invariant system and characterized by its impulse response, the so-called haemodynamic response function (HRF), denoted here h. Recent developments propose to estimate either the neural activity signal independently (Fikret et al., 2013; Cherkaoui et al., 2019b) or jointly with the HRF (Cherkaoui et al., 2019a; Farouj et al., 2019). Estimating the neural activity signal with a fixed HRF is akin to a deconvolution problem regularized with TV-norm,

$$\min_{u \in \mathbb{R}^k} P(u) = \frac{1}{2}\|h * u - x\|_2^2 + \lambda \|u\|_{TV} \tag{15}$$

To demonstrate the usefulness of our approach with real data, where the training set has not the exact same distribution than the testing set, we compare the LPGD-Taut to Accelerated PGD for the analysis formulation on this deconvolution problem. We choose two subjects from the UK Bio Bank (UKBB) dataset (Sudlow et al., 2015), perform the usual fMRI processing and reduce the dimension of the problem to retain only 8000 time-series of 250 time-frames, corresponding to a record of 3 minute 03 seconds. The full preprocessing pipeline is described in Appendix B. We train

the LPGD taut-string network solver on the first subject and Figure 5 reports the performance of the two algorithms on the second subject for $\lambda = 0.1$. The performance is reported relatively to the number of iteration as the computational complexity of each iteration or layer for both methods is equivalent. It is clear that LPGD-Taut converges faster than the Accelerated PGD even on real data. In particular, acceleration is higher when the regularization parameter λ is smaller. As mentioned previously, this acceleration is likely to be caused by the better learning capacity of the network in a low sparsity context. The same experiment is repeated for $\lambda = 0.8$ in Figure C.1.

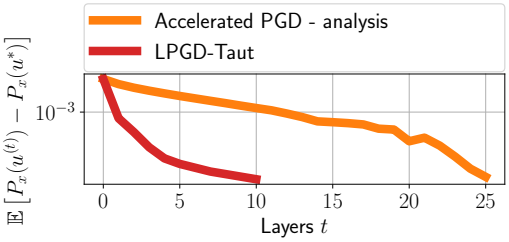


Figure 5: **Performance comparison** ($\lambda = 0.1$) between our analytic prox-TV derivative method and the PGD in the analysis formulation for the HRF deconvolution problem with fMRI data. Our proposed method outperform the FISTA algorithm in the analysis formulation.

5 Conclusion

This paper studies the optimization of TV-regularised problems via learned PGD. We demonstrated, both analytically and numerically, that it is better to address these problems in their original analysis formulation rather than resort to the simpler (alas slower) synthesis version. We then proposed two different algorithms that allow for the efficient computation and derivation of the required prox-TV, exactly or approximately. Our experiments on synthetic and real data demonstrate that our learned networks for prox-TV provide a significant advantage in convergence speed.

Finally, we believe that the principles presented in this paper could be generalized and deployed in other optimization problems, involving not just the TV-norm but more general analysis-type priors. In particular, this paper only apply for 1D TV problems because the equivalence between Lasso and TV is not exact in higher dimension. In this case, we believe exploiting a dual formulation (Chambolle, 2004) for the problem could allow us to derive similar learnable algorithms.

Broader Impact

This work attempts to shed some understanding into empirical phenomena in signal processing – in our case, piecewise constant approximations. As such, it is our hope that this work encourages fellow researchers to invest in the study and development of principled machine learning tools. Besides these, we do not foresee any other immediate societal consequences.

Acknowledgement

We gratefully acknowledge discussions with Pierre Ablin, whose suggestions helped us completing some parts of the proofs. H. Cherkaoui is supported by a CEA PhD scholarship. J. Sulam is partially supported by NSF Grant 2007649.

References

P. Ablin, T. Moreau, M. Massias, and A. Gramfort. Learning step sizes for unfolded sparse coding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13100–13110, Vancouver, BC, Canada, 2019.

F. Alfaro-Almagro, M. Jenkinson, N. K. Bangerter, J. L. R. Andersson, L. Griffanti, G. Douaud, S. N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, D. Vidaurre, M. Webster, P. McCarthy, C. Rorden, A. Daducci, D. C. Alexander, H. Zhang, I. Dragonu, P. M. Matthews, K. L. Miller, and S. M. Smith. Image Processing and Quality Control for the first 10,000 Brain Imaging Datasets from UK Biobank. *NeuroImage*, 166:400–424, 2018.

À. Barbero and S. Sra. Modular proximal optimization for multidimensional total-variation regularization. *The Journal of Machine Learning Research*, 19(1):2232–2313, Jan. 2018.

A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Q. Bertrand, Q. Klopfenstein, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon. Implicit differentiation of Lasso-type models for hyperparameter optimization. In *International Conference on Machine Learning (ICML)*, volume 2002.08943, pages 3199–3210, online, Apr. 2020.

M. Borgerding, P. Schniter, and S. Rangan. AMP-Inspired Deep Networks for Sparse Linear Inverse Problems. *IEEE Transactions on Signal Processing*, 65(16):4293–4308, 2017.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

A. Chambolle. An Algorithm for Total Variation Minimization and Applications. *Journal of Mathematical Imaging and Vision*, 20(1/2):89–97, Jan. 2004.

A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, May 2011.

P. L. Chebyshev. *Théorie Des Mécanismes Connus Sous Le Nom de Parallélogrammes*. Imprimerie de l'Académie impériale des sciences, 1853.

H. Cherkaoui, T. Moreau, A. Halimi, and P. Ciuciu. Sparsity-based Semi-Blind Deconvolution of Neural Activation Signal in fMRI. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019a.

H. Cherkaoui, T. Moreau, A. Halimi, and P. Ciuciu. fMRI BOLD signal decomposition using a multivariate low-rank model. In *European Signal Processing Conference (EUSIPCO)*, Coruña, Spain, 2019b.

P. L. Combettes and H. H. Bauschke. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

L. Condat. A Direct Algorithm for 1D Total Variation Denoising. *IEEE Signal Processing Letters*, 20(11):1054–1057, 2013a.

L. Condat. A Primal–Dual Splitting Method for Convex Optimization Involving Lipschitzian, Proximable and Linear Composite Terms. *Journal of Optimization Theory and Applications*, 158 (2):460–479, Aug. 2013b.

J. Darbon and M. Sigelle. Image Restoration with Discrete Constrained Total Variation Part I: Fast and Exact Optimization. *Journal of Mathematical Imaging and Vision*, 26(3):261–276, Dec. 2006.

P. L. Davies and A. Kovac. Local Extremes, Runs, Strings and Multiresolution. *The Annals of Statistics*, 29(1):1–65, Feb. 2001.

C. A. Deledalle, S. Vaiter, J. Fadili, and G. Peyré. Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection. *SIAM Journal on Imaging Sciences*, 7(4):2448–2487, 2014.

M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23(3):947–968, June 2007.

Y. Farouj, F. I. Karahanoglu, and D. V. D. Ville. Bold Signal Deconvolution Under Uncertain HÆModynamics: A Semi-Blind Approach. In *IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, pages 1792–1796, Venice, Italy, Apr. 2019. IEEE.

I. K. Fikret, C. Caballero-gaudes, F. Lazeyras, and V. D. V. Dimitri. Total activation: fMRI deconvolution through spatio-temporal regularization. *NeuroImage*, 73:121–134, 2013.

R. Giryes, Y. C. Eldar, A. M. Bronstein, and G. Sapiro. Tradeoffs between Convergence Speed and Reconstruction Accuracy in Inverse Problems. *IEEE Transaction on Signal Processing*, 66(7): 1676–1690, 2018.

K. Gregor and Y. Le Cun. Learning Fast Approximations of Sparse Coding. In *International Conference on Machine Learning (ICML)*, pages 399–406, 2010.

B. Lecouat, J. Ponce, and J. Mairal. Designing and Learning Trainable Priors with Non-Cooperative Games. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, June 2020.

P. Machart, S. Anthoine, and L. Baldassarre. Optimal Computational Trade-Off of Inexact Proximal Methods. *preprint ArXiv*, 1210.5034, 2012.

V. Monga, Y. Li, and Y. C. Eldar. Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing. *preprint ArXiv*, 1912.10557, Dec. 2019.

T. Moreau and J. Bruna. Understanding Neural Sparse Coding with Matrix Factorization. In *International Conference on Learning Representation (ICLR)*, Toulon, France, 2017.

S. Ogawa, D. W. Tank, R. Menon, J. M. Ellermann, S. G. Kim, H. Merkle, and K. Ugurbil. Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955, July 1992.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*, page 12, Vancouver, BC, Canada, 2019.

R. T. Rockafellar. Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

P. Rodríguez. Total Variation Regularization Algorithms for Images Corrupted with Different Noise Models: A Review. *Journal of Electrical and Computer Engineering*, 2013:1–18, 2013.

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, Nov. 1992.

M. Schmidt, N. Le Roux, and F. R. Bach. Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1458–1466, Grenada, Spain, 2011.

J. W. Silverstein. On the eigenvectors of large dimensional sample covariance matrices. *Journal of Multivariate Analysis*, 30(1):1–16, July 1989.

P. Sprechmann, A. M. Bronstein, and G. Sapiro. Learning Efficient Structured Sparse Models. In *International Conference on Machine Learning (ICML)*, pages 615–622, Edinburgh, Great Britain, 2012.

P. Sprechmann, R. Litman, and T. Yakar. Efficient Supervised Sparse Analysis and Synthesis Operators. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 908–916, South Lake Tahoe, United States, 2013.

C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779, Mar. 2015.

J. Sulam, A. Aberdam, A. Beck, and M. Elad. On Multi-Layer Basis Pursuit, Efficient Algorithms and Convolutional Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2019.

Z. Tian, X. Jia, K. Yuan, T. Pan, and S. B. Jiang. Low-dose CT reconstruction via edge-preserving total variation regularization. *Physics in Medicine and Biology*, 56(18):5949–5967, Sept. 2011.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 58(1):267–288, 1996.

R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39 (3):1335–1371, June 2011.

B. Xin, Y. Wang, W. Gao, and D. Wipf. Maximal Sparsity with Deep Networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4340–4348, 2016.

A Network training process strategy

Here, we give a more detailed description of the training procedure used in Section 4.

Optimization algorithm for training In our experiments, all networks are trained using Gradient Descent (GD) with back-tracking line search. The gradients are computed using automatic differentiation in Pytorch (Paszke et al., 2019) for most layers and the weak Jacobian proposed in Subsection 3.1 for the back-propagation through the prox-TV. The learning is stopped once a step-size of $\eta_{limit} = 10^{-20}$ is reached in the back-tracking step. For LPGD-LISTA, the weights of the inner LISTA computing the prox-TV are trained jointly with the parameters of the outer unrolled PGD.

Weight initialization All layers for an unrolled algorithm are initialized using the values of weights in (9) that ensure the output of the layer with T layers corresponds to the output of T iterations of the original algorithm. To further stabilize the training, we use a layer-wise approach. When training a network with $T_1 + T_2$ layers after having trained a network with T_1 layers, the first T_1 layers in the new network are initialized with the weights of the one trained previously, and the remaining layers are initialized using weights value from (9). This ensures that the initial value of the loss for the new network is smaller than the one from the shallower one if the unrolled algorithm is monotonous (as it is the case for PGD).

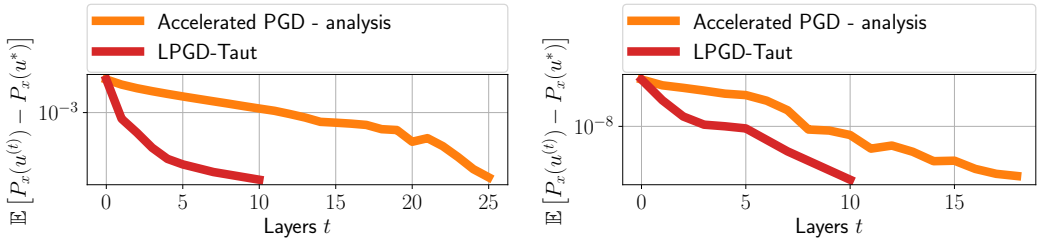
B Real fMRI data acquisition parameters and preprocessing strategy

In this section, we complete the description of the resting-state fMRI (rs-fMRI) data used for the experiment of Fig. 5. For this experiment, we investigate the 6 min long rs-fMRI acquisition (TR=0.735 s) from the UK Bio Bank dataset (Sudlow et al., 2015). The following pre-processing steps were applied on the images: motion correction, grand-mean intensity normalisation, high-pass temporal filtering, Echo planar imaging unwarping, Gradient Distortion Correction unwarping and structured artefacts removal by Independant Components Analysis. More details on the processing pipeline can found in Alfaro-Almagro et al. (2018).

On top of this preprocessing, we perform a standard fMRI preprocessing proposed in the python package Nilearn[3]. This standard pipeline includes to detrend the data, standardize it and filter high and low frequencies to reduce the presence of noise.

C Real fMRI data experiment addition results

Here, we provide an extra experiment for Subsection 4.2 with $\lambda = 0.8\lambda_{\max}$ and recall the previous one with $\lambda = 0.1\lambda_{\max}$ to help performance comparison in different regularization regime.



(a) $\lambda = 0.1\lambda_{\max}$ (b) $\lambda = 0.8\lambda_{\max}$

Figure C.1: **Performance comparison** between LPGD-Taut and iterative PGD for the analysis formulation for the HRF deconvolution problem with fMRI data. Our proposed method outperform the FISTA algorithm in the analysis formulation. We notice a slight degradation of the acceleration in this high sparsity context.

[3] `https://nilearn.github.io`

We can see that the performance drop in the higher sparsity setting compared to the performances with $\lambda = 0.1\lambda_{\max}$ but LPGD-Taut still outperforms iterative algorithms for this task on real data.

D Computing $\lambda_{\max}$ for the TV regularized problem

The definition of $\lambda_{\max}$ is the smallest value for the regularisation parameter λ such that the solution of the TV-regularized problem is constant. This corresponds to the definition of $\lambda_{\max}$ in the Lasso, which is the smallest regularisation parameter such that 0 is solution. We here derive its analytic value which is used to rescale all experiments. This is important to define an equiregularisation set for the training and testing samples, to have a coherent and generalizable training.

Proposition D.1. *The value of $\lambda_{\max}$ for the TV-regularized problem is*

$$\lambda_{\max} = \|A^\top(Ac\mathbf{1} - x)\|_\infty$$

where $c = \dfrac{\sum_{i=1}^p S_i x_i}{\sum_{i=1}^p S_i^2}$ *and* $S_i = \sum_{j=1}^k A_{i,j}$.

Proof. We first derive the constant solution of the ℓ_2-regression problem associated to (1). For $c \in \mathbb{R}$, we consider a constant vector $c\mathbf{1}$. The best constant solution for the ℓ_2-regression problem is obtained by solving

$$\min_{c\in\mathbb{R}} f_x(c) = \frac{1}{2}\|x - cA\mathbf{1}\|_2^2 \ . \tag{16}$$

The first order optimality condition in c reads

$$\nabla f_x(c) = \sum_{i=1}^n (\sum_{j=1}^k A_{i,j})(c\sum_{j=1}^k A_{i,j} - x_i) = \sum_{i=1}^n S_i(cS_i - x_i) = 0 \ , \tag{17}$$

and thus $c = \dfrac{\sum_{i=1}^p S_i x_i}{\sum_{i=1}^p S_i^2}$.

Then, we look at the conditions on λ to ensure that the constant solution $c\mathbf{1}$ is solution of the regularized problem. The first order conditions on the regularized problem reads

$$0 \in \partial P_x(c\mathbf{1}) = A^\top(Ac\mathbf{1} - x) + \lambda\partial\|Dc\mathbf{1}\|_1 \tag{18}$$

Next, we develop the previous equality:

$$\forall j \in \{2, \dots k\}, \quad A_j^\top(Ac\mathbf{1} - x) \in \lambda\partial(\|Dc\mathbf{1}\|_1)_j = [-\lambda, \lambda] \quad \text{since } Dc\mathbf{1} = 0 \tag{19}$$

Thus, the constrains are all satisfied for $\lambda \geq \lambda_{\max}$, with $\lambda_{max} = \|A^\top(Ac\mathbf{1} - x)\|_\infty$ and as c is solution for the unregularized problem reduced to a constant, $c\mathbf{1}$ is solution of the TV-regularized problem for all $\lambda \geq \lambda_{\max}$.

□

E Dual formulation

In this work, we devote our effort in the analysis formulation depicted in (1). In this section, we propose to investigate the dual formulation corresponding to (1) in order to rationalize our choice to focus on approaches that solve the prox-TV with an iterative method.

Dual derivation First, we derive the dual of the analysis formulation for the prox-TV.

Proposition E.1. *[Dual re-parametrization for the analysis formulation TV problem (1)]*

Considering the primal analysis problem (with operator and variables defined as previously)

$$P_x(u) = \frac{1}{2}\|x - Au\|_2^2 + \lambda\|Du\|_1 \tag{20}$$

Then, the dual formulation reads:

$$p = -\min_{v} \frac{1}{2}\|{A^\dagger}^\top D^\top v\|_2^2 - v^\top D A^\dagger x \tag{21}$$

$$\text{s.t. } \|v\|_\infty \leq \lambda \tag{22}$$

Proof. Defining, f and g, such as $f(u) = \frac{1}{2}\|x - Au\|_2^2$ and $g(u) = \lambda\|u\|_1$ and by denoting p the minimum of (20) *w.r.t* u, the problem reads:

$$p = \min_{u} \quad f(u) + g(Du) \tag{23}$$

With the Fenchel-Rockafellar duality theorem, we derive the dual re-parametrization:

$$p = -\min_{v} \; f^*(-D^\top v) + g^*(v) \tag{24}$$

Note, that in this case we have the equality with p since the problem (1) is μ-strongly convex, with $\mu = \frac{1}{2}$.

We have $g^*(v) = -\min_u \; g(u) - v^\top u$. With a component-wise minimization, we obtain $g^*(v)_i = \delta_{|v_i| \leq \lambda}$ with δ being the convex indicator. Thus, we deduce that $g^*(v) = \delta_{\|v\|_\infty \leq \lambda}$.

Then, we have $f^*(v) = -\min_u \; f(u) - v^\top u$. By cancelling the gradient we obtain: $f^*(v) = \frac{1}{2}\|{A^\dagger}^\top v\|_2^2 + v^\top A^\dagger x$

This allows use to conclude the demonstration. Note that, if we set $A = \mathrm{Id}$, we obtain the same problem as (Barbero and Sra, 2018; Chambolle, 2004).

□

Performance comparison We propose to compare the performance of different iterative solvers to assess their performance.

We generate $n = 1000$ times series to compare the performance between the different algorithms. We set the length of the source signals $(u_i)_{i=1}^n \in \mathbb{R}^{n \times k}$ to $k = 40$ with a support of $|S| = 4$ non-zero coefficients. We generate $A \in \mathbb{R}^{m \times k}$ as a Gaussian matrix with $m = 40$, obtaining then $(u_i)_{i=1}^n \in \mathbb{R}^{n \times p}$. Moreover, we add Gaussian noise to measurements $x_i = Au_i$ with a signal to noise ratio (SNR) of 1.0.

We select the PGD and its accelerated version with the synthesis primal formulation (5) ("Synthesis primal A/PGD"), the PGD and its accelerated version with the analysis primal formulation ("Analysis primal A/PGD"). We consider also the PGD and its accelerated version (Chambolle, 2004), for the analysis dual formulation ("Analysis dual A/PGD") and finally we add the primal/dual algorithm (Condat, 2013b) for the analysis primal formulation ("Analysis primal dual GD").

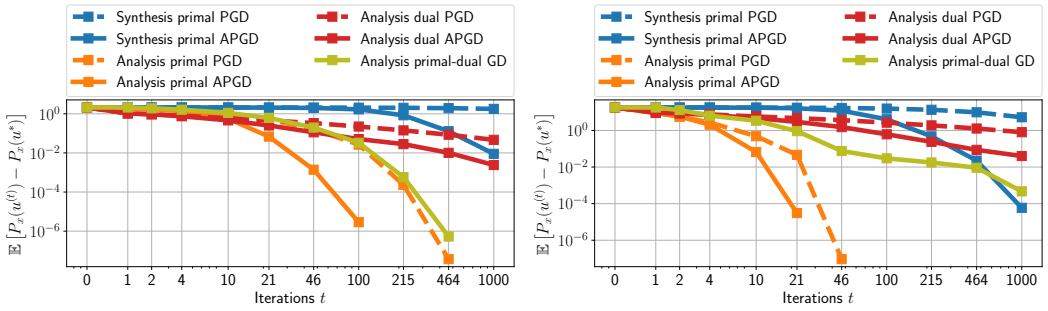
Figure E.1 a proposes performance comparison for an exhaustive selection of the algorithm used to solve (1). We see that the analysis primal formulation proposes the best performance for each regularization parameter. We notice that the Condat (2013b) provides good performance too. Finally, the synthesis primal formulation along with the analysis dual formulation provides the slowest performance. Those results reinforces our choice to focus on the PGD of the analysis primal formulation.

F Proof for Section 2

F.1 Convergence rate of PGD for the synthesis formulation (6)

Proof. The convergence rate of ISTA for the synthesis formulation reads

$$S(z^{(t)}) - S(z^*) \leq \frac{\widetilde{\rho}}{2t}\|z^{(0)} - z^*\|_2^2 \ . \tag{25}$$



(a) $\lambda = 0.1\lambda_{\max}$ (b) $\lambda = 0.8\lambda_{\max}$

Figure E.1: **Performance comparison** between the iterative solver for the synthesis and analysis formulation with the corresponding primal, dual or primal-dual re-parametrization. We notice that the primal analysis formulation provides the best performance in term of iterations. We also observe that the higher the regularization parameter, the faster the performance for each algorithm.

We use $S(z^{(t)}) = P(Lz^{(t)}) = P(u^{(t)})$ to get the correct left-hand side term. For the right hand side, we use $z^{(0)} = \widetilde{D}u^{(0)}$, and $z^* = \widetilde{D}u^*$, which gives $\|z^{(0)} - z^*\|_2^2 = \|\widetilde{D}(u^{(0)} - u^*)\|_2^2 \leq 4\|u^{(0)} - u^*\|_2^2$. The last majoration comes from the fact that $\|\widetilde{D}\|_2^2 \leq 4$, as shown per Lemma F.1. This yields

$$P(u^{(t)}) - P(u^*) \leq \frac{2\widetilde{\rho}}{t}\|u^{(0)} - u^*\|_2^2 \ . \tag{26}$$

□

F.2 Computing the spectrum of L

Lemma F.1. *[Singular values of L] The singular values of $L \in \mathbb{R}^{k\times k}$ are given by*

$$\sigma_l = \frac{1}{2\cos(\frac{\pi l}{2k+1})}, \qquad \forall l \in \{1, \dots, k\} \ .$$

Thus, $\|L\|_2 = \frac{2k+1}{\pi} + o(1)$.

Proof. As L is invertible, so is $L^\top L$. To compute the singular values σ_l of L, we will compute the eigenvalues μ_l of $(L^\top L)^{-1}$ and use the relation

$$\sigma_l = \frac{1}{\sqrt{\mu_l}} \tag{27}$$

With simple computations, we obtain

$$M_k = (L^\top L)^{-1} = L^{-1}(L^\top)^{-1} = \widetilde{D}\widetilde{D}^\top = \begin{bmatrix} 1 & -1 & 0 & \dots & \\ -1 & 2 & -1 & 0 & \dots \\ & \ddots & \ddots & \ddots & \\ & 0 & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \tag{28}$$

This matrix is tri-diagonal with a quasi-toepliz structure. Its characteristic polynomial $P_k(\mu)$ is given by:

$$P_k(\mu) = |\mu\,\mathrm{Id} - M_k| = \begin{vmatrix} \mu-1 & 1 & 0 & \dots & \\ 1 & \mu-2 & 1 & 0 & \dots \\ & \ddots & \ddots & \ddots & \\ & 0 & 1 & \mu-2 & 1 \\ & & 0 & 1 & \mu-2 \end{vmatrix} \tag{29}$$

$$= (\mu-1)Q_{k-1}(\mu) - Q_{k-2}(\mu) \tag{30}$$

where (30) is obtained by developing the determinant relatively to the first line and $Q_k(\mu)$ is the characteristic polynomial of matrix $\widetilde{M_k}$ equal to M_k except for the the top left coefficient which is replaced by 2

$$\widetilde{M_k} = \begin{bmatrix} 2 & -1 & 0 & \dots & \\ -1 & 2 & -1 & 0 & \dots \\ & \ddots & \ddots & \ddots & \\ & 0 & -1 & 2 & -1 \\ & & 0 & -1 & 2 \end{bmatrix} \tag{31}$$

Using the same development as (30), one can show that Q_k verifies the recurrence relationship

$$Q_k(\mu) = (\mu - 2)Q_{k-1}(\mu) - Q_{k-2}(\mu); \qquad Q_1(\mu) = 2 - \mu, \quad Q_0(\mu) = 1 \ . \tag{32}$$

Using this with (30) yields

$$P_k(\mu) = Q_k(\mu) + Q_{k-1}(\mu) \tag{33}$$

With the change of variable $\nu = \frac{\mu - 2}{2}$ and denoting $U_k(\nu) = Q_k(2 + 2\nu)$, the recursion becomes

$$U_k(\nu) = 2\nu U_{k-1}(\nu) - U_{k-2}(\nu); \qquad U_1(\nu) = 2\nu, \quad U_0(\mu) = 1 \ . \tag{34}$$

This recursion defines the Chebyshev polynomials of the second kind (Chebyshev, 1853) which verifies the following relation

$$\forall \theta \in [0, 2\pi], \quad U_k(\cos(\theta))\sin(\theta) = \sin((k+1)\theta) \ . \tag{35}$$

Translating this relationship to Q_k gives

$$\forall \theta \in [0, 2\pi], \quad Q_k(2 + 2\cos(\theta))\sin(\theta) = \sin((k+1)\theta) \ . \tag{36}$$

Using this in (33) shows that for $\theta \in [0, 2\pi[$ P_k verify

$$P_k(2 + 2\cos(\theta)\sin(\theta) = \sin((k+1)\theta) + \sin(k\theta) \ . \tag{37}$$

The equation

$$\sin((k+1)\theta) + \sin(k\theta) = 0 \ , \tag{38}$$

has l solution in $[0, 2\pi[$ that are given by $\theta_l = \frac{2\pi l}{2k+1}$ for $l \in \{1, \dots n\}$. As for all l, $\sin(\theta_l) \neq 0$, the values $\mu_l = 2 + 2\cos(\theta_l) = 4\cos^2(\frac{\pi l}{2k+1})$ are the roots of P_k and therefor the eigenvalues of M_k. Using (27) yields the expected value for σ_l.

The singular value of L is thus obtain for $l = k$ and we get

$$\|L\|_2 = \sigma_k = \frac{1}{2\cos(\frac{\pi k}{2k+1})} = \frac{1}{2\cos(\frac{\pi}{2}(1 - \frac{1}{2k+1}))} \ , \tag{39}$$

$$= \frac{1}{2\sin(\frac{\pi}{2}\frac{1}{2k+1})} = \frac{2k+1}{\pi} + o(1) \ . \tag{40}$$

Where the last approximation comes from $\frac{1}{sin(x)} = 1/x + o(1)$ when x is close to 0. □

F.3 Proof for Proposition 2.1

Proposition 2.1. *[Lower bound for the ratio $\frac{\|AL\|_2^2}{\|A\|_2^2}$ expectation] Let A be a random matrix in $\mathbb{R}^{m \times k}$ with* i.i.d *normally distributed entries. The expectation of $\|AL\|_2^2 / \|A\|_2^2$ is asymptotically lower bounded when k tends to ∞ by*

$$\mathbb{E}\left[\frac{\|AL\|_2^2}{\|A\|_2^2}\right] \geq \frac{2k+1}{4\pi^2} + o(1)$$

Proof. Finding the norm of AL can be written as

$$\|AL\|_2^2 = \max_{x \in \mathbb{R}^k} xL^\top A^\top ALx; \quad \textit{s.t. } \|x\|_2 = 1 \tag{41}$$

From Lemma F.1, we can write $L = W^\top \Sigma V$ with V, W two unitary matrices and Σ a diagonal matrix with $\Sigma_{l,l} = \sigma_l$ for all $l \in \{1, .., k\}$.

First, we consider the case where $A^\top A$ is a rank one matrix with $A^\top A = \|A\|_2^2 u_1 u_1^\top$, with vector u_1 uniformly sampled from the ℓ_2-ball and fixed $\|A\|_2$. In this case, as W is unitary, $w_1 = W u_1$ is also a vector uniformly sampled from the sphere. Also as V is unitary, it is possible to re-parametrize (41) by $y = Vx$ such that

$$\max_{y \in \mathbb{R}^k} \|A\|_2^2 y^\top \Sigma w_1 w_1^\top \Sigma y; \quad \textit{s.t. } \|y\|_2 = 1 \tag{42}$$

This problem can be maximized by taking $y = \frac{\Sigma u_1}{\|\Sigma u_1\|_2}$, which gives

$$\|AL\|_2^2 = \|A\|_2^2 \|\Sigma w_1\|_2^2 \tag{43}$$

Then, we compute the expectation of $\|\Sigma w_1\|_2^2$ with respect with w_1, a random vector sampled in the ℓ_2 unit ball,

$$\mathbb{E}_{w_1}[\|\Sigma w_1\|_2^2] = \sum_{l=1}^{k} \sigma_l^2 \mathbb{E}[u_{1,i}^2] = \sum_{l=1}^{k} \frac{1}{4\cos^2 \frac{\pi l}{2k+1}} \frac{1}{k} = \frac{1}{2\pi} \sum_{l=1}^{k} \frac{\pi}{2k} \frac{1}{\cos^2 \frac{\pi l}{2k+1}} \ . \tag{44}$$

Here, we made use of the fact that for a random vector u_1 on the sphere in dimension k, $\mathbb{E}[u_{1,i}] = \frac{1}{k}$ In the last part of the equation, we recognize a Riemann sum for the interval $[0, \frac{\pi}{2}[$. However, $x \mapsto \frac{1}{\cos^2(x)}$ is not integrable on this interval. As the function is positive and monotone, we can still use the integral to highlight the asymptotic behavior of the series. For k large enough, we consider the integral

$$\int_0^{\frac{\pi}{2} - \frac{\pi}{2k+1}} \frac{1}{\cos^2(x)} dx = \left[\frac{sin(x)}{\cos(x)}\right]_0^{\frac{\pi}{2} - \frac{\pi}{2k+1}} = \frac{\cos \frac{\pi}{2k+1}}{\sin \frac{\pi}{2k+1}} = \frac{2k+1}{\pi} + o(1) \tag{45}$$

Thus, for k large enough, we obtain

$$\mathbb{E}_{w_1}\left[\|\Sigma w_1\|_2^2\right] = \frac{1}{2\pi}\left(\frac{2k+1}{\pi} + o(1)\right) \tag{46}$$

Thus, we get

$$\mathbb{E}\left[\frac{\|AL\|_2^2}{\|A\|_2^2}\right] = \left(\frac{k + \frac{1}{2}}{\pi^2} + o(1)\right) \tag{47}$$

This concludes the case where $A^\top A$ is of rank-1 with uniformly distributed eigenvector.

In the case where $A^\top A$ is larger rank, it is lower bounded by $\|A\|_2^2 u_1 u_1^\top$ where u_1 is its eigenvector associated to its largest eigenvalue, since it is *psd*. Since $A^\top A$ is a Whishart matrix, its eigenvectors are uniformly distributed on the sphere (Silverstein, 1989). We can thus use the same lower bound as previously for the whole matrix.

□

G Proof for Section 3

G.1 Proof for Proposition 3.1

Proposition 3.1. *[Weak Jacobian of prox-TV] Let $x \in \mathbb{R}^k$ and $u = prox_{\mu\|\cdot\|_{TV}}(x)$, and denote by $\mathcal{S}$ the support of $z = \widetilde{D}u$. Then, the weak Jacobian J_x and J_μ of the prox-TV relative to x and μ can be computed as*

$$J_x(x,\mu) = L_{:,\mathcal{S}}(L_{:,\mathcal{S}}^\top L_{:,\mathcal{S}})^{-1} L_{:,\mathcal{S}}^\top \quad \textit{and} \quad J_\mu(x,\mu) = -L_{:,\mathcal{S}}(L_{:,\mathcal{S}}^\top L_{:,\mathcal{S}})^{-1} \operatorname{sign}(Du)_{\mathcal{S}}$$

First, we recall Lemma G.1 to weakly derive the soft-thresholding.

Lemma G.1 (Weak derivative of the soft-thresholding; Deledalle et al. 2014)**.** *The soft-thresholding operator ST : $\mathbb{R} \times \mathbb{R}_+ \mapsto \mathbb{R}$ defined by ST$(t,\tau) = \operatorname{sign}(t)(|t| - \tau)_+$ is weakly differentiable with weak derivatives*

$$\frac{\partial \operatorname{ST}}{\partial t}(t,\tau) = \mathbb{1}_{\{|t|>\tau\}} \ , \qquad \textit{and} \qquad \frac{\partial \operatorname{ST}}{\partial \tau}(t,\tau) = -\operatorname{sign}(t) \cdot \mathbb{1}_{\{|t|>\tau\}} \ ,$$

where

$$\mathbb{1}_{\{|t|>\tau\}} = \begin{cases} 1, & \textit{if } |t| > \tau, \\ 0, & \textit{otherwise.} \end{cases}$$

A very important remark here is to notice that if one denote $z = \mathrm{ST}(t, \tau)$, one can rewrite these weak derivatives as

$$\frac{\partial \mathrm{ST}}{\partial t}(t,\tau) = \mathbb{1}_{\{|z|>0\}} \ , \qquad \text{and} \qquad \frac{\partial \mathrm{ST}}{\partial \tau}(t,\tau) = -\operatorname{sign}(z) \cdot \mathbb{1}_{\{|z|>0\}} \ . \tag{48}$$

Indeed, when $|t| > \tau$, $|z| = |t| - \tau > 0$ and conversely, $|z| = 0$ when $|t| < \tau$. Moreover, when $|t| > \tau$, we have $\operatorname{sign}(t) = \operatorname{sign}(z)$ and thus the two expressions for $\frac{\partial \mathrm{ST}}{\partial \tau}$ match.

Using this Lemma G.1, we now give the proof of Proposition 3.1.

Proof. The proof is inspired from the proof from Bertrand et al. (2020, Proposition 1). We denote $u(x,\mu) = \operatorname{prox}_{\mu\|\cdot\|_{TV}}(x)$, hence $u(x,\mu)$ is defined by

$$u(x,\mu) = \arg\min_{\hat{u}} \frac{1}{2}\|x - \hat{u}\|_2^2 + \mu\|\hat{u}\|_{TV} \tag{49}$$

Equivalently, as we have seen previously in (5), using the change of variable $\hat{u} = L\hat{z}$ and minimizing over $\hat{z}$ gives

$$\min_{\hat{z}} \frac{1}{2}\|x - L\hat{z}\|_2 + \mu\|R\hat{z}\|_1 \ . \tag{50}$$

We denote by $z(x,\mu)$ the minimizer of the previous equation. Thus, the solution $u(x,\mu)$ of the original problem (49) can be recovered using $u(,\mu) = Lz(x,\mu)$. Iterative PGD can be used to solve (50) and $z(x, mu)$ is a fixed point of the iterative procedure. That is to say the solution z verifies

$$\begin{cases} z_1(x,\mu) = & z_1(x,\mu) - \frac{1}{\rho}(L^\top(Lz(x,\mu) - x))_1 \ , \\ z_i(x,\mu) = & \mathrm{ST}\left(z_i(x,\mu) - \frac{1}{\rho}(L^\top(Lz(x,\mu) - x))_i, \frac{\mu}{\rho}\right) \quad \text{for } i = 2 \dots k \ . \end{cases} \tag{51}$$

Using the result from Lemma G.1, we can differentiate (51) and obtain the following equation for the weak Jacobian $\widehat{J}_x(x,\mu) = \frac{\partial z}{\partial x}(x,\mu)$ of $z(x,\mu)$ relative to x

$$\widehat{J}_x(x,\mu) = \begin{pmatrix} 1 \\ \mathbb{1}_{\{|z_2(x,\mu)|>0\}} \\ \vdots \\ \mathbb{1}_{\{|z_k(x,\mu)|>0\}} \end{pmatrix} \odot \left[(\mathrm{Id} - \frac{1}{\rho}L^\top L)\widehat{J}_x(x,\mu) + \frac{1}{\rho}L^\top \mathrm{Id} \right] \ . \tag{52}$$

Identifying the non-zero coefficient in the indicator vectors yields

$$\begin{cases} \widehat{J}_{x,\mathcal{S}^c}(x,\mu) &= 0 \\ \widehat{J}_{x,\mathcal{S}}(x,\mu) &= (\mathrm{Id} - \frac{1}{\rho}L_{:,\mathcal{S}}^\top L_{:,\mathcal{S}})\widehat{J}_{x,\mathcal{S}}(x,\mu) + \frac{1}{\rho}L_{:,\mathcal{S}}^\top \ . \end{cases} \tag{53}$$

As, L is invertible, so is $L_{:,\mathcal{S}}^\top L_{:,\mathcal{S}}$ for any support $\mathcal{S}$ and solving the second equations yields the following

$$\widehat{J}_{x,\mathcal{S}} = (L_{:,\mathcal{S}}^\top L_{:,\mathcal{S}})^{-1} L_{:,\mathcal{S}}^\top \tag{54}$$

Using $u = Lz$ and the chain rules yields the expecting result for the weak Jacobian relative to x, noticing that as $\hat{J}_{x,\mathcal{S}^c} = 0$, $L\hat{J}_x = L_{:,\mathcal{S}}\hat{J}_{x,\mathcal{S}}$.

Similarly, concerning, $\widehat{J}_\mu(x,\mu)$, we use the result from Lemma G.1 an differentiale (51) and obtain $\widehat{J}_\mu(x,\mu) = \frac{\partial z}{\partial \mu}(x,\mu)$ of $z(x,\mu)$ relative to μ

$$\widehat{J}_\mu(x,\mu) = \begin{pmatrix} 1 \\ \mathbb{1}_{\{|z_2(x,\mu)|>0\}} \\ \vdots \\ \mathbb{1}_{\{|z_k(x,\mu)|>0\}} \end{pmatrix} \odot \left[(\mathrm{Id} - \frac{1}{\rho}L^\top L)\widehat{J}_\mu(x,\mu) \right] + \frac{1}{\rho} \begin{pmatrix} 1 \\ -\operatorname{sign}(z_2(x,\mu))\mathbb{1}_{\{|z_2(x,\mu)|>0\}} \\ \vdots \\ -\operatorname{sign}(z_k(x,\mu))\mathbb{1}_{\{|z_k(x,\mu)|>0\}} \end{pmatrix} \ . \tag{55}$$

Identifying the non-zero coefficient in the indicator vectors yields

$$\begin{cases} \widehat{J}_{\mu,\mathcal{S}^c}(x,\mu) &= 0 \\ \widehat{J}_{\mu,\mathcal{S}}(x,\mu) &= \widehat{J}_{\mu,\mathcal{S}^c}(x,\mu) - \frac{1}{\rho} L_{:,\mathcal{S}}^\top L_{:,\mathcal{S}} \widehat{J}_{\mu,\mathcal{S}^c}(x,\mu) - \frac{1}{\rho}\operatorname{sign}(z_S(x,\mu)) \ . \end{cases} \tag{56}$$

As previous, solving the second equation yields the following

$$\widehat{J}_{\mu,\mathcal{S}} = -(L_{:,\mathcal{S}}^\top L_{:,\mathcal{S}})^{-1} \operatorname{sign}(z_S(x,\mu)) \tag{57}$$

Using $u = Lz$ and the chain rules yields the expecting result for the weak Jacobian relative to μ, noticing that as $\hat{J}_{\mu,\mathcal{S}^c} = 0$.

□

G.2 Convergence of the weak Jacobian

Proposition G.2. *Linear convergence of the weak Jacobian We consider the mapping* $z^{(T_i n)} : , \mu\mathbb{R}^k \times \mathbb{R}_+ \mapsto \mathbb{R}^k$ *defined where* $z^{(T_i n)}(x)$ *is defined by recursion*

$$z^{(t)}(x,\mu) = ST(z^{(t-1)}(x,\mu) - \frac{1}{\|L\|_2^2} L^\top (Lz^{(t-1)}(x,\mu) - x), \frac{\mu}{\|L\|_2^2} \ . \tag{58}$$

Then the weak $\mathcal{J}_x = L\frac{\partial z^{(T_i n)}}{\partial x}$ *and* $\mathcal{J}_\mu = L\frac{\partial z^{(T_i n)}}{\partial \mu}$ *of this mapping relative to the inputs* x *and* μ *converges linearly toward the weak Jacobian* J_x *and* J_μ *of* $prox_{\mu\|\cdot\|_{TV}}(x)$ *defined in Proposition 3.1.*

This mapping defined in (58) corresponds to the inner network in LPGD-LISTA when the weights of the network have not been learned.

Proof. As L is invertible, problem (50) is strongly convex and have a unique solution. We can thus apply the result from Bertrand et al. (2020, Proposition 2) which shows the linear convergence of the weak Jacobian $\hat{\mathcal{J}}_x = \frac{\partial z^{(T_i n)}}{\partial x}$ and $\hat{\mathcal{J}}_\mu = \frac{\partial z^{(T_i n)}}{\partial \mu}$ for ISTA toward $\hat{J}_x$ and $\hat{J}_\mu$ of the synthesis formulation of the prox. Using the linear relationship between the analysis and the synthesis formulations yields the expected result. □

G.3 Estimating T_{in} and T to achieve δ error

Using inexact proximal gradient descent results from Schmidt et al. (2011) and Machart et al. (2012), we compute the scaling of T_{in} and T to achieve a given error level $\delta > 0$.

Proposition G.3. *[Scaling of* T *and* T_{in} w.r.t *the error level* δ*] Let* δ *the error defined such as* $P_x(u^{(T)}) - P_x(u^*) \leq \delta$.
We suppose there exists some constants $C_0 \geq \|u^{(0)} - u^*\|_2$ *and* $C_1 \geq \max_\ell \|u^{(\ell)} - prox_{\frac{\mu}{\rho}}(u^{(\ell)})\|_2$
Then, T *the number of layers for the global network and* T_{in} *the inner number of layers for the prox-TV scale are given by*

$$T_{in} = \frac{\ln\frac{1}{\delta} + \ln 6\sqrt{2\rho}C_1}{\ln\frac{1}{1-\gamma}} \quad \textit{and} \quad T = \frac{2\rho C_0^2}{\delta}$$

with ρ *defined as in (2)*

Proof. As discussed by Machart et al. (2012), the global convergence rate of inexact PGD with T_{in} inner iteration is given by

$$P_x(u^{(T)}) - P_x(u^*) \leq \\ \frac{\rho}{2T}\left(\|u^{(0)} - u^*\|_2 + 3\sum_{\ell=1}^{T}\sqrt{\frac{2(1-\gamma)^{T_{in}}\|u^{(\ell-1)} - \operatorname{prox}_{\frac{\mu}{\rho}}(u^{(\ell-1)})\|_2^2}{\rho}}\right)^2 , \tag{59}$$

where γ is the condition number for L *i.e.* $\frac{\cos(\frac{\pi}{2k+1})}{\sin(\frac{\pi}{2k+1})}$.

We are looking for minimal parameters T and T_{in} such that the error bound in (59) is bellow a certain error level δ.

We consider the case where there exists some constants $C_0 \geq \|u^{(0)} - u^*\|_2$ and $C_1 \geq \max_\ell \|u^{(\ell)} - \text{prox}_{\frac{\mu}{\rho}}(u^{(\ell)})\|_2$ upper bounding how far the initialization can be compared to the result of the global problem and the sub-problems respectively.

We denote $\alpha_1 = 3\sqrt{\frac{2}{\rho}}C_1$. The right hand side of (59) can be upper bounded by as

$$\frac{\rho}{2T}\left(\|u^{(0)} - u^*\|_2 + 3\sum_{\ell=1}^{T}\sqrt{\frac{2(1-\gamma)^{T_{in}}\|u^{(\ell-1)} - \text{prox}_{\frac{\mu}{\rho}}(u^{(\ell-1)})\|_2^2}{\rho}}\right)^2 \leq \frac{\rho}{2T}\left(C_0 + \alpha_1 T(1-\gamma)^{T_{in}/2}\right)^2 \tag{60}$$

Then, we are looking for T, T_{in} such that this upper bound is lower than δ, *i.e.*

$$\frac{\rho}{2T}\left(C_0 + \alpha_1 T(1-\gamma)^{T_{in}/2}\right)^2 \leq \delta \tag{61}$$

$$\Leftrightarrow \left(C_0 + \alpha_1 T(1-\gamma)^{T_{in}/2}\right)^2 - \frac{2\delta}{\rho}T \leq 0 \tag{62}$$

$$\Leftrightarrow \left(C_0 + \alpha_1 T(1-\gamma)^{T_{in}/2} - \sqrt{\frac{2\delta}{\rho}}\sqrt{T}\right)\underbrace{\left(B + \alpha_1 T(1-\gamma)^{T_{in}/2} + \sqrt{\frac{2\delta}{\rho}}\sqrt{T}\right)}_{\geq 0} \leq 0 \tag{63}$$

$$\Leftrightarrow C_0 + \alpha_1 T(1-\gamma)^{T_{in}/2} - \sqrt{\frac{2\delta}{\rho}}\sqrt{T} \leq 0 \tag{64}$$

$$\tag{65}$$

Denoting $\alpha_2 = \sqrt{\frac{2\delta}{\rho}}$ and $X = \sqrt{T}$, we get the following function of X and T_{in}

$$f(X, T_{in}) = \alpha_1(1-\gamma)^{T_{in}/2}X^2 - \alpha_2 X + C_0 \tag{66}$$

The inequality $f(X, T_{in}) \leq 0$ has a solution if and only if $\alpha_2^2 - 4C_0\alpha_1(1-\gamma)^{T_{in}/2} \geq 0$ *i.e.*

$$T_{in} \geq 2\frac{\ln\frac{\alpha_2^2}{4\alpha_1 C_0}}{\ln 1-\gamma}$$

Taking the minimal value for T_{in} *i.e.* $T_{in} = 2\frac{\ln\frac{\alpha_2^2}{4\alpha_1 C_0}}{\ln 1-\gamma} = \frac{\ln\frac{1}{\delta} + \ln 6\sqrt{2\rho}C1}{\ln\frac{1}{1-\gamma}}$ yields

$$f(X, T_{in}) = \frac{\alpha_2^2}{4C_0}X^2 - \alpha_2 X + C_0 = \frac{\alpha_2^2}{4C_0}(X - \frac{2C_0}{\alpha_2})^2$$

for $X = \frac{2C_0}{\alpha_2} = \frac{\sqrt{2\rho}C_0}{\sqrt{\delta}}$ *i.e.* $T = \frac{2\rho C_0^2}{\delta}$. □