Variational Justification of Cycle Spinning for Wavelet-Based Solutions of Inverse Problems

Ulugbek S. Kamilov, Student Member, IEEE, Emrah Bostan, Student Member, IEEE, and Michael Unser, Fellow, IEEE

Abstract—Cycle spinning is a widely used approach for improving the performance of wavelet-based methods that solve linear inverse problems. Extensive numerical experiments have shown that it significantly improves the quality of the recovered signal without increasing the computational cost. In this letter, we provide the first theoretical convergence result for cycle spinning for solving general linear inverse problems. We prove that the sequence of reconstructed signals is guaranteed to converge to the minimizer of some global cost function that incorporates all wavelet shifts.

Index Terms—Cycle spinning, linear inverse problems, wavelet regularization.

I. INTRODUCTION

T he problem of estimating an unknown signal from noisy linear observations is fundamental in signal processing. The estimation task is often formulated as the linear inverse problem

\[ y = Hx + n, \]  

(1)

where the goal is to reconstruct the unknown signal \( x \in \mathbb{R}^N \) from the noisy measurements \( y \in \mathbb{R}^M \). The matrix \( H \in \mathbb{R}^{M \times N} \) models the response of the acquisition device; the vector \( n \) represents the measurement noise, which is often assumed to be i.i.d. Gaussian. When Problem (1) is ill-posed, the standard approach is to rely on the estimator

\[ \hat{x} = \arg \min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2} \| y - Hx \|^2 + \lambda \phi(x) \right\}, \]  

(2)

where the function \( \phi \) is a regularizer that promotes solutions with desirable properties, and where \( \lambda > 0 \) controls the degree of regularization. In the wavelet-based framework, regularization is achieved by favoring solutions that have a “sparse” wavelet expansion. A popular approach is to use the non-smooth convex function \( \phi(x) = \| Wx \|_1 \), where \( W \in \mathbb{R}^{N \times N} \) represents a wavelet transform [1]–[3]. Although (2) generally does not admit a closed-form solution with non-quadratic regularizers, it can still be computed efficiently using the iterative shrinkage/thresholding algorithm (ISTA) [1]–[3]. Based on the definition of our forward model, the algorithm can be expressed as

\[ z^t = \hat{x}^{t-1} - \gamma \lambda T(\lambda \hat{x}^{t-1} - y) \]  

(3a)

\[ \hat{x}^t = W^T T(\lambda Wz^t; \gamma \lambda), \]  

(3b)

where \( \gamma > 0 \) is a step-size that can be determined a priori to ensure convergence of the algorithm [4]. Iterations in (3) combine gradient-descent steps (3a) with pointwise soft-thresholding operations (3b), of the form

\[ T(w; \lambda) = (|w| - \lambda)_+ \text{sgn}(w). \]  

(4)

Because of its simplicity, ISTA has become the method of choice for finding sparse solutions. More recently, several accelerated versions of ISTA have been developed that provide state-of-the-art rates of convergence [4]–[6].

The theory of wavelet-regularized reconstruction is often formulated with orthogonal wavelet transforms. However, in order to make regularized wavelet-based methods truly competitive, one needs to make the transform shift-invariant. The concept was first introduced in wavelet-based denoising under the name of cycle spinning [7]–[9]. The recursive approach to cycle spinning was presented and analyzed in [10]. Cycle spinning was then applied to more-general linear inverse problems by Figueiredo and Nowak [1]. Currently, it is used in the majority of wavelet-based reconstruction algorithms to obtain higher-quality solutions with less-blocky artifacts [11]–[14]. However, it is rarely accounted for in the accompanying theory.

Let the matrix \( W_k \) denote an orthogonal wavelet transform with \( k \)-th shift applied to all the basis functions in \( W \). We consider the \( K \) different shifts \( W_1, \ldots, W_K \) that are required to get a shift-invariant version of \( W \). Then, a simple way to implement cycle spinning, without increasing the memory usage, is to consider

\[ z^t = \hat{x}^{t-1} - \gamma \lambda T(\lambda \hat{x}^{t-1} - y) \]  

(5a)

\[ \hat{x}^t = W_k^T T(W_k x^t; \gamma \lambda), \]  

(5b)

where \( k_t = 1 + (t - 1 \mod K) \) is an iteration-dependent circular shift of the signal. Although Iteration (5) has nearly the same computational cost as (3), it yields results of significantly higher quality, as illustrated in Fig. 2. Also note that (5) is different from the original formulation of cycle spinning in [7], where the thresholded wavelet-coefficients corresponding to different shifts are simply averaged. The scheme in (5) is conceptually closer to recursive cycle spinning that was developed for signal denoising in [10].
II. MAIN RESULT

The apparent limitation of the cycle-spinning algorithm (5) lies in its greedy nature. At each iteration, the algorithm simply makes a locally optimal step towards the minimization of

$$F_k(x) = \frac{1}{2} \| y - Hx \|_2^2 + \lambda \| W_k x \|_1,$$

(6)

instead of using the information available from all possible shifts. We are not aware of any prior analysis of the convergence properties of such a scheme. Our main theorem below establishes the first result.

**Theorem 1.** Let $f(x) = (1/K) \sum_{k=1}^K F_k(x)$, where

$$F_k(x) = \frac{1}{2} \| y - Hx \|_2^2 + \lambda \| W_k x \|_1.$$

(7)

Assume that the feasible set $\mathcal{X} \subseteq \mathbb{R}^N$ is nonempty, convex, bounded, and closed. Set $\gamma_1 = 1/(L\sqrt{t})$, where $L$ is any constant such that $L > \lambda_{\text{max}}(H^T H)$. Let $\tilde{x}^t$ be an arbitrary vector in $\mathcal{X}$, and $(\tilde{x}^t)_{t \in \mathbb{N}}$ be the sequence generated by (5). Then,

$$\lim_{t \to \infty} f(\tilde{x}^t) = f^*,$$

(8)

where $f^* = \min_{x \in \mathcal{X}} f(x)$.

The proof is provided in the Appendix.

When $K = 1$, cycle spinning reduces to the standard ISTA, which is known to converge [4]. What our theorem proves is that, by iteratively cycling through $K$ orthogonal wavelets, we are minimizing a cost function that penalizes $\ell_1$-norm of all the shifts simultaneously, with the advantage that the underlying regularizer is truly shift-invariant.

III. EXPERIMENTS

We illustrate the theorem with two simple examples. In the first example, we consider the estimation of a piecewise constant signal of length $N = 128$ corrupted by AWGN corresponding to an input signal-to-noise ratio (SNR) of 30 dB. An interesting property of such signals is that they can be sparsified with the finite-difference operator, which justifies the use of TV regularization [6], [15]. Since the TV regularizer corresponds to an $\ell_1$-penalty applied to the finite differences of the signal, our theorem indicates that it can also be minimized with cycle spinning when $W$ corresponds to the Haar-wavelet basis with one level of decomposition and a zero weight in the lowpass. In Fig. 1, we plot the per-iteration gap $(f(\tilde{x}^t) - f^*)$, where $\tilde{x}^t$ is computed with cycle spinning and $f$ is the TV-regularized least-squares cost. We set $\lambda = 0.65$ and, following our analysis, we set the step-size to $\gamma_1 = 1/(4\sqrt{t})$. As expected, we observe that, as $t \to \infty$, we have that $(f(\tilde{x}^t) - f^*) \to 0$. Moreover, we note that, for large $t$, the slope of $(f(\tilde{x}^t) - f^*)$ in log-log domain tends to $-1/2$, which indicates the asymptotic rate of convergence $O(1/\sqrt{t})$.

In the second example, we consider an image-deblurring problem where the Cameraman image of size $256 \times 256$ is blurred with a $7 \times 7$ Gaussian kernel of standard deviation 2 with the addition of AWGN of variance $\sigma^2 = 10^{-5}$. In Fig. 2, we present the result of the reconstruction with three different methods: standard Haar-domain $\ell_1$-regularization, anisotropic TV [6], and cycle spinning with 1D Haar-basis functions applied horizontally and vertically to imitate TV. The regularization parameters for the standard wavelet approach and TV were optimized for the least-error performance. The regularization parameter of cycle spinning was set by rescaling the regularization parameter of TV according to $\lambda_{\text{CS}} = \sqrt{2K}\lambda_{\text{TV}}$, with $K = 4^1$. Therefore, we expect cycle spinning to again match the TV solution. It is clear from Fig. 2 that cycle spinning outperforms the standard wavelet regularization (improvement of at least 2 dB). As expected, the solution obtained by cycle spinning exactly matches that of TV both visually and in terms of SNR.

![Fig. 1. Estimation of a sparse signal from noisy measurements.](image1)

![Fig. 2. Reconstruction of Cameraman from blurry and noisy measurements.](image2)

The goal is for cycle spinning to be minimizing exactly the same cost function as TV. Thus, the factor $K$ is due to the number of shifts, while the factor $\sqrt{2}$ is due to the normalization of the Haar wavelets.
IV. CONCLUSION AND FUTURE WORK

We have established the convergence result of the popular cycle-spinning technique for solving linear inverse problems. We have proved that the algorithm converges to the minimizer of a regularized least-squares cost function where the regularizer is the $\ell_1$-norm of translation-invariant wavelet coefficients in the analysis form [16].

One can imagine numerous possible extensions of our results. The analysis presented in this letter was deterministic; an interesting avenue for future research would be to see if it also holds in the stochastic setting, where $k_t$ would be generated randomly at each iteration. Our analysis indicates that the rate of convergence of cycle spinning is no worse than $O(1/\sqrt{t})$. A possible direction of research is to search for a faster convergence rate by using various standard acceleration techniques for ISTA algorithms [4].

V. APPENDIX

A. Useful Facts from Convex Analysis

Before embarking on the actual proof of Theorem 1, it is convenient to summarize a few facts that will be used next.

A subgradient of a convex function $f$ at $x$ is any vector $g$ that satisfies the inequality $f(y) \geq f(x) + \langle g, y - x \rangle$, for all $y$. When $f$ is differentiable, the only possible choice for $g$ is $\nabla f(x)$. The set of subgradients of $f$ at $x$ is the subdifferential of $f$ at $x$, denoted $\partial f(x)$. The condition that $g$ be a subgradient of $f$ at $x$ can then be written $g \in \partial f(x)$.

The proximal operator is defined as

$$ x = \text{prox}_{\gamma r}(z) = \arg\min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2} \|x - z\|^2 + \gamma r(x) \right\}, \quad (9) $$

where $\gamma > 0$ and $r$ is a convex continuous function. The proximal operator is characterized by the following inclusion, for all $x, z \in \mathbb{R}^N$:

$$ x = \text{prox}_{\gamma r}(z) \Leftrightarrow z - x \in \gamma \partial r(x). \quad (10) $$

We say that an operator $T : \mathcal{X} \to \mathcal{X}$ is nonexpanding if, for all $x, z \in \mathcal{X}$, it satisfies $\|T^T x - T^T z\| \leq \|x - z\|$. Note that the proximal operator is nonexpanding.

Next, we present the Browder-Göhde-Kirk’s fixed-point theorem. It is a standard theorem in convex analysis (see [17, Theorem 4.19]).

**Theorem 2:** Let $\mathcal{X}$ be a nonempty bounded closed convex subset of $\mathbb{R}^N$ and let $T : \mathcal{X} \to \mathcal{X}$ be a nonexpansive operator. Then, the operator $T$ has a nonempty set of fixed points or that $T$ has a fixed point.

We now state a technical lemma that will be used in our proof.

**D. Lemma 1**

For all $t > 1$, and for any $x^* \in \mathcal{X}$, we have that

$$ F_{k_t}(\hat{x}^t) = F_{k_t}(x^*) \leq \frac{1}{2\gamma_t} \left( \|\hat{x}^t - x^*\|^2 - \|\hat{x}^t - x^*\|^2 \right) + 6\gamma_t G^2. \quad (13) $$

**Proof:** The optimality condition of (12) implies that there must exist a vector $g^t \in \partial R_{k_t}(\hat{x}^t)$ such that

$$ \hat{x}^t = \hat{x}^t - \gamma_t \left( \nabla d(\hat{x}^t - 1) + g^t \right). \quad (14) $$

Then, we can write

$$ \|\hat{x}^t - x^*\|^2 = \|\hat{x}^t - x^* - (\nabla d(\hat{x}^t - 1) + g^t) - x^*\|^2 $$

$$ = \|\hat{x}^t - x^*\|^2 - 2\gamma_t \nabla d(\hat{x}^t - 1) + g^t, \hat{x}^t - x^*\|^2 + \gamma_t \|\nabla d(\hat{x}^t - 1) + g^t\|^2. \quad (15) $$

By using the triangle inequality, we can bound the last term as

$$ \|\nabla d(\hat{x}^t - 1) + g^t\|^2 \leq 4G^2. \quad (16) $$
To bound the second term we proceed in two steps. We first write that

\[ \langle \nabla d(\tilde{x}^{t-1}), \tilde{x}^{t-1} - x^* \rangle \geq d(\tilde{x}^{t-1}) - d(x^*) \]

\[ \geq d(\tilde{x}^t) - \langle \nabla d(\tilde{x}^t), \tilde{x}^t - \tilde{x}^{t-1} \rangle - d(x^*) \]

\[ = d(\tilde{x}^t) - d(x^*) - \langle \nabla d(\tilde{x}^t), -\gamma_t (\nabla d(\tilde{x}^{t-1}) + g^t) \rangle \]

\[ \geq d(\tilde{x}^t) - d(x^*) - 2\gamma_t G^2, \]  

(17)

where in (a) and (b) we used the convexity of \( d \), in (c) we used the Cauchy-Schwarz inequality and the bound on the gradient. Then, in a similar fashion, we can also write that

\[ \langle g^t, \tilde{x}^{t-1} - x^* \rangle = \langle g^t, \tilde{x}^t - x^* \rangle - \langle g^t, \tilde{x}^t - \tilde{x}^{t-1} \rangle \]

\[ \geq R_{k_t}(\tilde{x}^t) - R_{k_t}(x^*) - \langle g^t, -\gamma_t (\nabla d(\tilde{x}^{t-1}) + g^t) \rangle \]

\[ \geq R_{k_t}(\tilde{x}^t) - R_{k_t}(x^*) - 2\gamma_t G^2, \]  

(18)

where in (a) we used the convexity of \( R_{k_t} \), in (b) we used the Cauchy-Schwarz inequality followed with a bound on the gradient.

By plugging (16), (17), and (18) into (15), by reorganizing the terms, and by using the definition \( F_{k_t}(x) = d(x) + R_{k_t}(x) \), we obtain the claim.

**Lemma 2:** With \((\tilde{x}^t)_{t\in\mathbb{N}}\) in \( \mathcal{X} \) and \( \tilde{x} \in \mathcal{X} \), let \( \tilde{x}^t \to \tilde{x} \). Then,

\[ \lim_{n \to \infty} \left\{ \frac{1}{nK} \sum_{t=1}^{nK} F_{k_t}(\tilde{x}^t) \right\} = f(\tilde{x}). \]  

(19)

**Proof:** We introduce the shorthand notation \( \delta_t = F_{k_t}(\tilde{x}^t) - F_{k_t}(\tilde{x}) \). The convergence of \( \tilde{x}^t \) and the continuity of \( F_{k_t} \) imply that, for a given \( \epsilon > 0 \), there exists an \( m \) such that, for all \( t > mK \), \( |\delta_t| = F_{k_t}(\tilde{x}^t) - F_{k_t}(\tilde{x}) < \epsilon/2 \). Then,

\[ \frac{1}{nK} \sum_{t=1}^{nK} F_{k_t}(\tilde{x}^t) - f(\tilde{x}) = \frac{1}{nK} \sum_{t=1}^{nK} (F_{k_t}(\tilde{x}^t) - F_{k_t}(\tilde{x})) \]

\[ \leq \frac{1}{nK} \sum_{t=1}^{mK} \delta_t + \frac{1}{nK} \sum_{t=mK+1}^{nK} \delta_t \]

\[ \leq \frac{mK}{nK} \left( \max_{t \in [1, mK]} |\delta_t| \right) + \frac{1 - mK}{nK} \left( \max_{t \in [mK+1, nK]} |\delta_t| \right) \]

\[ \leq \frac{\max_{t \in [1, mK]} |\delta_t|}{n} + \frac{1 - mK}{nK} \frac{\max_{t \in [mK+1, nK]} |\delta_t|}{n} \].

(20)

Now, considering \( \bar{n} \geq \sqrt{2m/\epsilon} \max_{t \in [1, mK]} |\delta_t| \) realizing that the second term is bounded by \( \epsilon/2 \), we conclude that, for a given \( \epsilon > 0 \), there exists \( \bar{n} \) such that, for all \( n > \bar{n} \),

\[ \frac{1}{nK} \sum_{t=1}^{nK} F_{k_t}(\tilde{x}^t) - f(\tilde{x}) < \epsilon. \]  

(21)

**E. Proof of Theorem 1**

Let \( x^* \) denote a minimizer of \( f \). We introduce the shorthand notation \( \Delta_t = |\tilde{x}^{t-1} - x^*|^2 \). By following an approach similar to [19], we sum the bound in Lemma 1

\[ \sum_{t=1}^{nK} (F_{k_t}(\tilde{x}^t) - F_{k_t}(x^*)) = \sum_{t=1}^{nK} F_{k_t}(\tilde{x}^t) - nK f(x^*) \]

\[ \leq \frac{1}{2} \sum_{t=1}^{nK} \frac{1}{\gamma_t} (\Delta_t - \Delta_{t-1}) + 6G^2 \sum_{t=1}^{nK} \gamma_t \]

\[ \leq \frac{1}{2\gamma_0} \Delta_0 + \frac{1}{2} \sum_{t=1}^{nK} \left( \frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right) \Delta_t + 6G^2 \sum_{t=1}^{nK} \gamma_t \]

\[ \leq \frac{D^2}{2\gamma_0} + 6G^2 \sum_{t=1}^{nK} \gamma_t, \]  

(22)

where in (a) we used the the boundedness of \( \mathcal{X} \). Then, by choosing \( \gamma_t = 1/(L\sqrt{t}) \), by using the bound \( \sum_{t=1}^{nK} 1/\sqrt{t} \leq 2\sqrt{T} \), and by dividing the two sides of inequality by \( nK \), we obtain

\[ \frac{1}{nK} \sum_{t=1}^{nK} F_{k_t}(\tilde{x}^t) - f(x^*) \leq C \sqrt{1/n}, \]  

(23)

where the constant \( C \) is given by

\[ C = \frac{LD^2}{2\sqrt{K}} + \frac{12G^2}{L\sqrt{K}}. \]  

(24)

To complete the proof, we argue that the sequence \( \{\tilde{x}^t\}_{t\in\mathbb{N}} \) converges to a fixed point in \( \mathcal{X} \). On one hand, note that, due to the nonexpansiveness of \( \text{prox}_{\gamma_t R_{k_t}} \), and the Lipschitz property of \( d \), the operator \( T_0 : \mathcal{X} \to \mathcal{X} \),

\[ T_0 x = \text{prox}_{\gamma_t R_{k_t}} (x - \gamma_t d(x)) \]

(25)

is nonexpanding for any \( \gamma \in (0, 1/L) \). Therefore, the composition \( T = T_0 \cdots T_1 \) is also nonexpanding. Then, from Theorem 2, we know that there exists at least one fixed-point of \( T \) in \( \mathcal{X} \). On the other hand, for our choice \( \gamma_t = 1/(L\sqrt{T}) \), Theorem 3 implies that the subsequence generated via \( \tilde{x}^{K(t-1)} = T_{K(t-1)} \cdots T_0 \tilde{x} \) converges to the fixed-point \( \tilde{x} \) in \( T \). Since, \( \gamma_t \to 0 \), we have that \( \tilde{x}^{t-1} \to \tilde{x} \) and conclude that \( \tilde{x}^t \to \tilde{x} \). Finally, this allows us to show that

\[ 0 \leq f(\tilde{x}) - f(x^*) \]

\[ \leq \lim_{n \to \infty} \left\{ \frac{1}{nK} \sum_{t=1}^{K} F_{k_t}(\tilde{x}^t) \right\} - f^* \leq 0, \]  

(26)

where (a) comes from the optimality of \( x^* \), (b) from Lemma 2, and (c) from the upper bound (23).

**ACKNOWLEDGMENT**

The existence of this letter owes much to intense discussions with M. Nilchian. Also, P. Pad and J. P. Ward were extremely helpful during inspiring exchanges related to convex analysis. The authors are also grateful for A. Amini’s help and encouragement.
REFERENCES


