Doubly Sparse Explicitly Conditioned Transform Learning

It is of interest a modified version of the explicitly conditioned unstructured transform presented in Problem 4 in [1], following a similar idea to that of the authors in [5]. The proposed factorization, $W = T\Omega$, represents the distinctive feature of what will be referred to as the *doubly sparse explicitly conditioned transform learning* problem.

1.1 Problem Formulation

Let $Y \in \mathbb{R}^{n\times N}$ and $X \in \mathbb{R}^{n\times N}$ denote the data matrix and its corresponding representation, where n is the dimension of each of the N signals in the given set. Considering the factorization $W = T\Omega$ and the sparsity constraint imposed on the learnt component, a change of variable is introduced in the explicitly conditioned transform learning Problem 4 in [1]:

$$\begin{aligned} \min_{T\Omega\in\mathbb{R}^{n\times n},\, X\in\mathbb{R}^{n\times N}} \quad & \|T\Omega Y - X\|_F^2 \\ \text{s.t.} \quad & \|X_i\|_0 \le r, \quad \forall i, \\ & \|T\|_0 \le R, \quad \kappa(T\Omega) \le \rho, \quad \|T\Omega\|_F = \tau, \end{aligned} \tag{1.1}$$

where R represents the target sparsity of matrix T.

The decision variable $T\Omega$ can be reduced to T, as Ω represents a fixed, conventionally known matrix. Following the same consideration, it is introduced the variable $\tilde{Y} = \Omega Y$, as the learnt transform acts only over the spectral components in the domain induced by Ω. The norm-preserving property of the orthogonal canonical transform also simplifies the constraints imposed on the condition number and the Frobenius norm, not affecting the singular spectrum of the overall transform. The resulting formulation for the *doubly sparse explicitly conditioned transform learning* problem follows:

$$
\begin{aligned}
\min_{T\in\mathbb{R}^{n\times n},\, X\in\mathbb{R}^{n\times N}} \quad & \|T\tilde{Y} - X\|_F^2 \\
\text{s.t.} \quad & \|X_i\|_0 \le r, \quad \forall i, \\
& \|T\|_0 \le R, \quad \kappa(T) \le \rho, \quad \|T\|_F = \tau.
\end{aligned} \tag{1.2}
$$

This problem is the same as Problem 4 in [1], with an additional constraint on the learnt transform matrix. The exact factorial order $T\Omega$ has been favored over its equivalent form, ΩT, for its convenience, as potential algorithms will compute the multiplication ΩY exactly once at the beginning, leaving it unchanged throughout their iterations.

An optimal solution for the transform update step can not be derived under both the sparsity and conditioning constraint. Hence, an ℓ_1-norm relaxation is introduced and an additional positive-definiteness constraint is imposed on the matrix T, thereby transforming the conditioning constraint on the singular value spectrum into a restriction on the eigenvalue spectrum:

$$
\begin{aligned}
\min_{T\in\mathbb{R}^{n\times n},\, X\in\mathbb{R}^{n\times N}} \quad & \|T\tilde{Y} - X\|_F^2 - \lambda\|T\|_1 \\
\text{s.t.} \quad & \|X_i\|_0 \le r, \quad \forall i, \\
& T \succ 0, \quad \kappa(T) \le \rho, \quad \|T\|_F = \tau.
\end{aligned} \tag{1.3}
$$

1.2 Proposed Algorithm

The sparse representation X^{k+1} is computed in the same manner as in the work of the aforementioned authors, by zeroing out the smallest absolute value components in each column X_i^{k+1}, retaining the first r largest ones.

The proposed algorithm for the transform update step is a projected variant of the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) introduced in [6], which is a proximal gradient descent method enhanced with Nesterov acceleration. Specifically, an unconstrained gradient step is applied to the smooth component of the objective function, followed by a proximal operation on the decision variable T, i.e. the soft-thresholding operator promoting the non-smooth convex ℓ_1 regularization term, and finally by a projection onto the feasible set.

1.2.1 Unconstrained Gradient Step

The first stage in the FISTA scheme reduces to applying a gradient step on the smooth function $f(T) = \|T\tilde{Y} - X\|_F^2$ in the decision variable T. The descent direction is determined by the first-order derivative of the objective function:

$$
\nabla f(T) = (T\tilde{Y} - X)\tilde{Y}. \tag{1.4}
$$

The Nesterov acceleration method follows:

$$T_k = T_k + \beta_k (T_k - T_{k-1}) \tag{1.5}$$

$$T_{k+1} = T_k - \alpha_k \nabla f(T_k). \tag{1.6}$$

Authors of [6] provide an explicit formula for the acceleration parameter, β_k, with constant stepsize:

$$\beta_k = \frac{t_k - 1}{t_{k+1}}, \tag{1.7}$$

where

$$t_1 = 1, \tag{1.8}$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}. \tag{1.9}$$

The proposed strategy for choosing the step size is an alternation between α_k^{LONG} and α_k^{SHORT} introduced by Jonathan Barzilai and Jonathan M. Borwein in [7]:

$$\alpha_k^{\text{LONG}} = \frac{\langle \Delta T, \Delta T \rangle}{\langle \Delta T, \Delta G \rangle}, \tag{1.10}$$

$$\alpha_k^{\text{SHORT}} = \frac{\langle \Delta T, \Delta G \rangle}{\langle \Delta G, \Delta G \rangle}, \tag{1.11}$$

where

$$\Delta T = T_k - T_{k-1}, \tag{1.12}$$

$$\Delta G = \nabla f(T_k) - \nabla f(T_{k-1}). \tag{1.13}$$

Since both the Nesterov acceleration scheme and the computation of the α_{BB} parameter require at least one prior iteration, a classical gradient descent step is performed at $k = 1$, with α_1 determined by *exact line search*:

$$\alpha_1 = \frac{1}{4} \frac{\|\nabla f(T)\|_F^2}{\|\nabla f(T) Y\|_F^2}. \tag{1.14}$$

1.2.2 Proximal Step

The next stage of the algorithm involves applying the soft-thresholding operator to the decision variable T, as it corresponds to the proximal mapping associated with the ℓ_1-norm penalty:

$$T_{k+1} = \mathcal{S}_{\lambda}(T_{k+1}), \tag{1.15}$$

where

$$\mathcal{S}_{\lambda}(T_{k+1}) = \text{sign}(T_{k+1}) \odot \max(|T_{k+1}| - \lambda, 0). \tag{1.16}$$

1.2.3 Spectrum Projection Step

The projection onto the feasible set reduces to finding the transform that is closest to the one obtained in the previous steps while satisfying the hard constraints, which can be formulated as a *matrix nearness* problem:

$$\begin{aligned} \min_{\hat{T} \in \mathbb{R}^{n \times n}} \quad & \|\hat{T} - T\|_F^2 \\ \text{s.t.} \quad & \hat{T} \succ 0, \quad \kappa(\hat{T}) \le \rho, \quad \|\hat{T}\|_F = \tau. \end{aligned} \tag{1.17}$$

As $\hat{T}$ lies in the interior of the positive semi-definite cone, the singular values coincide with the eigenvalues:

$$\begin{aligned} \min_{\hat{T} \in \mathbb{R}^{n \times n}} \quad & \|\hat{T} - T\|_F^2 \\ \text{s.t.} \quad & \hat{T} = \hat{T}^\top, \quad \lambda_{min}(\hat{T}) \ge c, \\ & \frac{\lambda_{max}(\hat{T})}{\lambda_{min}(\hat{T})} \le \rho, \quad \|\hat{T}\|_F = \tau, \end{aligned} \tag{1.18}$$

where c is a small positive tolerance.

The first two constraints are enforced by clipping the eigenvalues of T below c and by projecting T onto its symmetric part via the symmetric-skew-symmetric decomposition. Consequently, the problem reduces to the eigenvalue vectors of T and $\hat{T}$, as the orthogonal matrices in the eigenvalue decomposition $Q \Lambda Q^\top$ remain unchanged:

$$\begin{aligned} \min_{diag(\hat{\Lambda})} \quad & \|diag(\hat{\Lambda}) - diag(\Lambda)\|_2^2 \\ \text{s.t.} \quad & l \le \hat{\lambda}_i \le \rho l, \quad \|\hat{\Lambda}\|_F = \tau. \end{aligned} \tag{1.19}$$

The problem essentially reduces to finding the optimal value of l that minimizes the cost function, for which a linear-time algorithm based on a geometric approach has been introduced in Algorithm 2.1 by the authors of [2]. The matrix $\hat{\Lambda}$ is thereafter normalized and scaled by τ, thereby satisfying the Frobenius norm constraint.

1.3 Conclusions

Although integrating the Barzilai-Borwein strategy for the α step size with FISTA may appear to be an aggressive approach, the numerical experiments show stability in most cases. The reason behind this choice is to favor the descent direction of the solution prior to its projection onto the feasible space, a structure encouraged by the proximal step and enforced by the spectrum projection step.

The sparsity constraint imposed by the penalty function effectively contracts the feasible space, and, by extension, the search space, having a positive impact on convergence for suitable choices of ρ. However, as the imposed condition number grows, extending the feasible set, the limited degrees of freedom implied by the sparsity constraint become a limiting factor. This is because the intersection of the feasible sets might not include the specific solution that tends towards Least Squares. **This aspect requires a more formal analysis**.

Another observation identifies the interdependence between parameters λ and τ. The Frobenius norm constraint impacts the magnitude of the entries of matrix T, upon which the soft-thresholding operator relies.

1.4 Alternatives

Burdakov *et al.* introduce a stabilized variant of the Barzilai–Borwein method in [3].

Another possible approach is to use the Variable Metric Proximal Gradient method introduced in [4].

Bibliography

[1] A. Pătrașcu, C. Rusu, and P. Irofti. "Learning Explicitly Conditioned Sparsifying Transforms." In: (2024). arXiv: 2403.03168.

[2] Jiao-fen Li, Wen Li, and Seak-Weng Vong. "Efficient algorithms for solving condition number-constrained matrix minimization problems." In: *Linear Algebra and its Applications* 607 (2020), pp. 190–230. ISSN: 0024-3795. DOI: https://doi.org/10.1016/j.laa.2020.08.007. URL: https://www.sciencedirect.com/science/article/pii/S0024379520303803.

[3] Oleg Burdakov, Yu-Hong Dai, and Na Huang. *Stabilized Barzilai-Borwein method*. 2019. arXiv: 1907.06409 [math.OC]. URL: https://arxiv.org/abs/1907.06409.

[4] Youngsuk Park, Sauptik Dhar, Stephen Boyd, and Mohak Shah. *Variable Metric Proximal Gradient Method with Diagonal Barzilai-Borwein Stepsize*. 2019. arXiv: 1910.07056 [math.OC]. URL: https://arxiv.org/abs/1910.07056.

[5] Saiprasad Ravishankar and Yoram Bresler. "Learning doubly sparse transforms for image representation." In: *2012 19th IEEE International Conference on Image Processing*. 2012, pp. 685–688. DOI: 10.1109/ICIP.2012.6466952.

[6] Amir Beck and Marc Teboulle. "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems." In: *SIAM J. Imaging Sci.* 2 (2009), pp. 183–202. URL: https://api.semanticscholar.org/CorpusID:3072879.

[7] JONATHAN BARZILAI and JONATHAN M. BORWEIN. "Two-Point Step Size Gradient Methods." In: *IMA Journal of Numerical Analysis* 8.1 (Jan. 1988), pp. 141–148. ISSN: 0272-4979. DOI: 10.1093/imanum/8.1.141. eprint: https://academic.oup.com/imajna/article-pdf/8/1/141/2402762/8-1-141.pdf. URL: https://doi.org/10.1093/imanum/8.1.141.