SWAGGER: Sparsity Within and Across Groups for General Estimation and Recovery

Charles Saunders and Vivek K Goyal

Abstract—Penalty functions or regularization terms that promote structured solutions to optimization problems are of great interest in many fields. Proposed in this work is a nonconvex structured sparsity penalty that promotes one-sparsity within arbitrary overlapping groups in a vector. This allows one to enforce mutual exclusivity between components within solutions to optimization problems. We show multiple example use cases (including a total variation variant), demonstrate synergy between it and other regularizers, and propose an algorithm to efficiently solve problems regularized or constrained by the proposed penalty.

Index Terms—sparsity, structured sparsity, group sparsity, regularization, inverse problems

I. INTRODUCTION

SPARSITY-promoting regularization is an integral component of modern-day signal processing, optimization and machine learning. The most prevalent method, LASSO [1], uses the $\ell_1$ norm to promote solutions to optimization problems that have few nonzero coefficients. More recently, within some application domains there has been significant interest in structured sparsity. For instance, group sparsity aims to set entire groups of components within a vector entirely to zero and allow the components in other groups to take on any value [2]. This is often achieved by regularizing with the $\ell_{2,1}$ norm over groups, called group LASSO (G-LASSO) [3]. Conversely, there is interest in sparsity penalties that promote intra-group sparsity, that is, ensuring groups within a vector are themselves sparse. Such penalties have been used to great success in areas such as deep learning [4], computer vision [5], [6], and medicine [7], [8]. Previous work in this area includes elitist [9], [10], or exclusive LASSO (E-LASSO) [11] formulations that are convex and result in sparse, disjoint groups, and the nonconvex ‘sparsity within and across groups’ (SWAG) [12] plus extensions [13] that allow for overlapping groups of components.

We propose here a more general, nonconvex structured sparsity penalty that allows for sparsity within and across overlapping groups for general estimation and recovery (SWAGGER). The SWAGGER formulation encodes mutual exclusivity between pairs of components, or a transform of the components, using an easily constructed sparsity structure matrix. This results in one-sparse groups with minimal bias in the nonzero entries, where bias is typically an unwanted byproduct of using convex sparsity penalties such as elitist LASSO. We demonstrate the utility of SWAGGER in a number of settings, including a novel total variation-style denoising in one and two dimensions and modeling occlusions plus allowing mutually exclusive discretizations in imaging problems. Additionally, we introduce a novel algorithm to efficiently solve problems regularized or constrained by this proposed penalty.

II. PROPOSED PENALTY

We introduce a general structured sparsity penalty

$$R(x) = \Phi(Bx)^{T}S\Phi(Bx),$$

(1)

for $x \in \mathbb{R}^N$. For certain choices of the matrix $B \in \mathbb{R}^{M \times N}$, where $M$ is any integer, function $\Phi : \mathbb{R}^M \rightarrow \mathbb{R}^M$, and matrix $S \in \mathbb{R}^{M \times M}$, this penalizes deviation from a selected form of structured sparsity. Possible prototypes for $B$ include the identity matrix, finite difference matrices, or a transform to a different basis; see Section VII. The function $\Phi$ introduces a nonlinearity that helps to bound the penalty from below. We constrain $\Phi$ in Prop. 1 below, and examples are discussed in Section VII. The matrix $S$ is key, as it encodes the structural sparsity properties. We define $S$ as:

$$S_{i,j} = S_{j,i} = \mu, \quad (2)$$

where $\mu \in [0,1]$ is the strength of exclusivity between a pair of components, $\Phi(Bx)_i$ and $\Phi(Bx)_j$. When $\mu = 0$, there is no exclusivity between the two components. When $\mu > 0$, it indicates that the two components should not be nonzero simultaneously (with the strength of exclusivity increasing with $\mu$). We bound $\mu$ in the [0,1] range for convenience and note that any scaling can be absorbed into a multiplicative regularization parameter. Given this interpretation, the diagonal of $S$ will be zero and hence $S$ will typically be indefinite, resulting in nonconvexity. Examples for $S$ are given in Section IV.

In order for the penalty function to be useful in regularizing minimization problems, it must be bounded from below. Prop. 1 provides simple sufficient conditions for this (proof in Appendix A).

Proposition 1: In order for (1) to be bounded from below, it is sufficient to require:

1) $S_{i,j} \geq 0 \quad \forall (i,j) \in \{1,2,\ldots,M\}^2$; and
2) $\Phi(x) \geq 0 \text{ elementwise} \quad \forall x$.

The definition for $S$ in (2) meets condition 1) in Prop. 1. Except for a discussion of extensions in Section VII we assume $\Phi(x)$ is the absolute value $|x|$, which satisfies condition 2).

III. PROPERTIES OF THE SWAGGER PENALTY

When $\Phi(\cdot) = | \cdot |$, the Hessian of the penalty is given by

$$\nabla^2_x R(x) = 2XSX,$$
where $X = \text{diag}(\text{sign}(x))$. The sparsity structure matrix, $S$, is symmetric and has zeros along its main diagonal. $S$ must be indefinite as it has at least one positive eigenvalue (see [14]) and the sum of the eigenvalues is zero (since the trace of a symmetric matrix is equal to the sum of its eigenvalues and $\text{Tr}(S) = 0$). The Hessian has the same structure and can therefore not be positive semidefinite outside of the trivial case where $x = 0$. Thus, the penalty is always nonconvex for practical purposes. More analysis of the eigenvalues of matrices of this kind can be found in [15].

We can compare SWAGGER with the convex E-LASSO penalty, which uses the $\ell_1$-squared norm to achieve intra-group sparsity. In the case of a one-sparse group (see Section IV-A), we can introduce a modified sparsity matrix, $S_\eta = I^T - \eta I$, where scalar parameter $\eta \in [0, 1]$ gives us E-LASSO for $\eta = 0$ and SWAGGER for $\eta = 1$:

$$|x|^T S_\eta |x| = \|x\|^2 - \eta \|x\|^2_2.$$

When $\eta = 1$, we see the introduction of the $-\|x\|^2$ term which both de-biases the penalty and also ensures the penalty equals zero when $x$ is one-sparse. For a more general $S$, the closest convex penalty would be to replace $S$ with $S - cI$, where $c$ takes on the value of the smallest eigenvalue of $S$.

Maintaining convexity of a full cost function: There has recently been interest in so-called ‘convex nonconvex’ sparsity regularization where parameters of a nonconvex penalty are set to values that maintain convexity of the full cost function with the data fidelity, e.g., [16]. This is achievable with the SWAGGER penalty, also. Consider the problem of minimizing a cost function of the form $\frac{1}{2} \|Ax - y\|_2^2 + \lambda \|x\|^T S |x|$. One can replace $S$ with $S - cI$, where $c = \lambda_{\min}(S) + \frac{1}{2\lambda} \lambda_{\min}(A^T A)$, to achieve nonconvex regularization whilst maintaining the overall convexity of the problem (see derivation in Appendix E). This is only useful when the smallest eigenvalue of $A^T A$ is not too small and the regularization strength parameter $\lambda$ is not too large. The work presented outside of this section is only focused on the use of the unmodified $S$, but further analysis of the intermediate convexity paradigm may be of interest.

IV. NOTABLE CASES FOR $S$

A. Canonical One-sparsity

Define $\Phi(\cdot) = \|\cdot\|$ and $B = I$. In this case, we obtain

$$R(x) = |x|^T S |x|. \quad (3)$$

This is the ‘sparsity within and across overlapping groups’ formulation which is introduced in [13]. If we choose $S = I^T - I$, where $I$ is a column vector of 1’s of appropriate dimension, then we see that

$$R(x) = |x|^T S |x| = |x|^T (I^T - I) |x| = (|x|^T 1)^2 - |x|^T |x| = \|x\|^2 - \|x\|^2_2, \quad (4)$$

which is equal to zero for any one-sparse $x$. Hence, this can be used to promote one-sparse solutions to optimization problems. See Fig. [1](a) for a graphical depiction.

B. Intra-group Sparsity

Define $\Phi(\cdot) = \|\cdot\|$ and $B = I$. Consider a vector $x \in \mathbb{R}^g$ where $g$ is a number of groups and $n$ is the number of components in each group. Then, we can promote one-sparsity within each group using $S \in \mathbb{R}^{g \times g}$ defined in blocks:

$$S = \begin{bmatrix} S & 0 & \cdots & 0 \\ 0 & \bar{S} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \bar{S} \end{bmatrix},$$

where $\bar{S} = I^T - I_g$ (see Fig. [1](b)). This is now a block-separable problem, equivalent to applying the penalty defined in Section IV-A to different sections of $x$, i.e., $\sum_{k=1}^g |x_k|^T S |x_k|$, where $k$ indexes blocks of $n$ components. One use case for this formulation is in modeling occlusions in imaging problems, which we address with an example in Section VII-C.

C. Local Neighborhood Sparsity

Sparsity can be promoted in overlapping local neighborhoods in an ordered sequence of indexes by restricting the nonzeros to a band around the diagonal. This gives a symmetric Toeplitz $S$ (see Fig. [1](c)). For instance, an $S$ with neighbor distance $n = 1$ is given by

$$S = \begin{bmatrix} 0 & 1 & 0 & \cdots \\ 1 & 0 & 1 & \ddots \\ 0 & 1 & 0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$  

For a general neighbor distance $n$, $S$ is given by

$$S_{i,j} = \begin{cases} \mu, & |i - j| \leq n, \ i \neq j; \\ 0, & \text{otherwise}, \end{cases}$$

where again $\mu \in [0, 1]$ defines a strength of exclusivity. This imposes local-$n$-neighborhood one-sparsity, which ensures nonzeros in a solution occur with a minimum separation distance of $n$ indices.
V. CHOICES FOR $\Phi(\cdot)$

The typical choice for $\Phi(\cdot)$ is the absolute value $|\cdot|$. This
is the main focus of this paper and admits fast algorithms to
result in one-sparse groups with a prescribed structure, with
minimal bias. However, we note that other choices can enjoy
interesting properties. For instance, using $\Phi(x) = \max(0,x)$
or $\Phi(x) = \max(0,-x)$ acts on positive or negative values
only, respectively, and meets the condition outlined in Prop.[1]

Thus, combining two constraint terms,

$$\hat{x} = \arg\min_{x} f(x)$$

$$\text{s.t.} \quad \max(0,x)S\max(0,x) = 0 \quad (5)$$

$$\max(0,-x)S\max(0,-x) = 0$$

can result in a solution where only positive correlation is
penalized, i.e., a group may contain one positive and negative
component, rather than being one-sparse. In [4], exclusive
sparsity regularization is used to reduce redundancy in the
learned kernels in convolutional neural networks. The $\Phi(\cdot) =
\max(0,\cdot)$ formulation could function as a relaxed approach to
this, allowing more degrees of freedom in the filter kernels.

In a case where there can be a large disparity between the
size of components within groups in the solution vector, one
may wish to impose a nonlinear scaling and use $\Phi(x) = |x|^\kappa$
where $\kappa < 1$. This reduces the effect of large but possibly
incorrect components on other, potentially correct components.
This can be seen in use in the two-dimensional local neighbor-
hood total variation example in Section VII-B.

VI. ALGORITHMS

First, we note that when $\Phi(\cdot) = |\cdot|$ and $B = I$, the penalty
can be expressed

$$R(x) = |x|^T S |x| = \|x\|^2 - |x|^T (I I^T - S - I) |x|. \quad (6)$$

An efficient algorithm to evaluate the proximal operator of the
$\ell^2_1$ norm is available (see Appendix B) [17]. Hence, we can
target to efficiently solve any problem of the form:

$$\hat{x} = \arg\min_{x} f(x)$$

$$\text{s.t.} \quad |x|^T S |x| = 0$$

using a proximal subgradient method (Algorithm 1). This
simply requires us to evaluate the (sub)gradient of $f(x) -
\|x\|^2 − |x|^T (I I^T - S - I) |x|$ and the proximal operator of $\|x\|^2$.
In practice, we find the convergence speed of this algorithm is
not negatively affected by the use of a subgradient of $−|x|^T (I I^T - S - I) |x|$ as this term will always aim to push the
components of $x$ away from the discontinuity in the gradient
at $x = 0$. Furthermore, in cases with no overlapping groups,
this term equals zero.

As this penalty is nonconvex, it is prudent to start with an
initial estimate of $x$ that minimizes the data fidelity. Given
this, Algorithm 1 outlines the proximal subgradient scheme
used to approximately solve the problem in (7) by introducing
a Lagrange multiplier $\lambda$.

In Algorithm 1 $P_{\alpha \lambda_k}$ is the proximal operator associated
with the $\ell^2_1$ term, $\|\cdot\|_1$, and $\alpha$ is a step size that can be
fixed or computed dynamically. There exists an accelerated
proximal gradient algorithm that is guaranteed to converge to
a stationary point even for nonconvex problems [18], which
could be used here also. The algorithm for this is outlined in
Appendix C.

When $B$ is not the identity, we can instead solve a problem of the
form:

$$\hat{x} = \arg\min_{x} f(x)$$

$$\text{s.t.} \quad Bx = z$$

$$|x|^T S |x| = 0 \quad (8)$$

using the alternating direction method of multipliers (ADMM)
[19], which also uses Algorithm 1 or an accelerated variation
thereof at each iteration.

VII. EXAMPLE APPLICATIONS

A. Comparison with Other Methods

Three experiments were performed, where measurements $y$
were generated using $y = Ax$, with white Gaussian noise
added to achieve a signal-to-noise radio (SNR) of 25 dB. Matrix
$A \in \mathbb{R}^{25 \times 60}$ has random Gaussian entries realized for
each trial. The vector $x$ has entries $\pm U(0.5, 1.5)$ and is made
to fit a specific sparsity structure such that $|x|^T S |x| \approx 0$
for some $S$. Details of how this is achieved are presented in
Appendix D. In the first experiment, $S$ is fixed and enforces intra-
group one-sparsity in ten groups (see Section IV-B), where
$S \in \mathbb{R}^{5 \times 6}$. In the second experiment, a local neighborhood $S$
(see Section IV-C) is used with $n = 4$. In the third experiment,
a new binary $S$ is generated per trial with entries equal to zero
or one with $50\%$ probability. We compare the performance of
the SWAGGER penalty with other typical sparsity regularizers,
both structured and unstructured. The problems solved are of
the form $\hat{x} = \arg\min_{x} \frac{1}{2} \|Ax - y\|^2 + \lambda R(x)$, where $R(x)$ is:

1) LASSO: $\|x\|_1$

2) p-shrinkage [20], with $p = 0.5$

3) E-LASSO/\ell_{1,2}: $\sum_{i} \|x_i\|_1^2$

The E-LASSO penalty uses a sum of the $\ell_{1}$-squared term
over groups. We create one group per row in $S$, where each
group contains the nonzero components in each row (and the
diagonal), i.e., $g_i = \{ j \mid S_{ij} + \delta_{i,j} = 1 \}$. This penalty is not
designed for overlapping groups, so the results are poor in the
local neighborhood and random $S$ experiments. The pseudo-
inverse is used in the initialization for the nonconvex penalties,
$x^0 = (A^T A)^{-1} A^T y$.

Table I shows the results from the three experiments. The
metrics evaluated are the percentage of components correctly
Table I

Simulated results over 1000 trials for three different sparsity structures. Metrics are the percentage of nonzero components identified, the Jaccard index (0 - 1), and the mean-squared error (MSE) where the ground truth is nonzero.

| Experiment | Support correct (%) | Jaccard Index | MSE in support |
|------------|---------------------|---------------|----------------|
| **Group S** |                     |               |                |
| λ tuned for sparsity level |                     |               |                |
| SWAGGER    | 75.69               | 0.646         | 1.987          |
| E-LASSO    | 59.84               | 0.398         | 4.634          |
| p-shrinkage | 68.89               | 0.559         | 2.564          |
| LASSO      | 56.71               | 0.404         | 3.813          |
| λ tuned for Jaccard index |                     |               |                |
| SWAGGER    | 84.10               | 0.699         | 1.633          |
| E-LASSO    | 87.38               | 0.509         | 2.849          |
| p-shrinkage | 77.82               | 0.638         | 2.122          |
| LASSO      | 73.91               | 0.492         | 3.221          |
| **Local neighborhood S** |                     |               |                |
| λ tuned for sparsity level |                     |               |                |
| SWAGGER    | 79.89               | 0.701         | 1.653          |
| E-LASSO    | 60.37               | 0.241         | 5.432          |
| p-shrinkage | 68.02               | 0.545         | 2.671          |
| LASSO      | 56.46               | 0.402         | 3.813          |
| λ tuned for Jaccard index |                     |               |                |
| SWAGGER    | 85.18               | 0.754         | 1.328          |
| E-LASSO    | 82.18               | 0.306         | 5.278          |
| p-shrinkage | 76.92               | 0.625         | 2.187          |
| LASSO      | 72.10               | 0.479         | 3.318          |
| **Random S** |                     |               |                |
| λ tuned for sparsity level |                     |               |                |
| SWAGGER    | 95.32               | 0.927         | 0.171          |
| E-LASSO    | 56.20               | 0.147         | 3.151          |
| p-shrinkage | 89.68               | 0.845         | 0.461          |
| LASSO      | 70.90               | 0.563         | 1.708          |
| λ tuned for Jaccard index |                     |               |                |
| SWAGGER    | 95.84               | 0.932         | 0.151          |
| E-LASSO    | 88.64               | 0.206         | 3.173          |
| p-shrinkage | 93.15               | 0.888         | 0.327          |
| LASSO      | 88.26               | 0.661         | 1.173          |

Identified as nonzero, the Jaccard index (the intersection of the true support and estimated support divided by the union), and the mean-squared error (MSE) for the components that are nonzero in the ground truth. We see that SWAGGER outperforms the other algorithms in almost every case. It is unsurprising that the unstructured sparsity regularizers (LASSO and p-shrinkage) perform especially poorly in the group S and local neighborhood settings as the true support is not especially sparse overall, whereas SWAGGER is agnostic to the overall sparsity level. Furthermore, the E-LASSO \( |\ell_1^2| \) regularizer performs poorly when groups are overlapping, whereas SWAGGER handles this gracefully and without any additional effort required. We also see that the MSE within the correct support is particularly low with SWAGGER, as it implicitly de-biases the solutions (this can be readily seen from the \(-||x||_2^2 \) term in (4)). Of particular interest are the results where \( \lambda \) is tuned to result in the correct sparsity level. The SWAGGER results are particularly favorable in this setting, and this is the most typical use case; it is roughly equivalent to solving the constrained SWAGGER problem in (4), which only allows changes in the derivative to occur with some minimum separation, compared to standard total variation (TV) and Moreau-enhanced total variation (M-TV) compared to denoising.

B. Local Neighborhood Total Variation

The well established total variation (TV) penalty, \( \|Dx\|_1 \) where multiplying by \( D \) takes finite differences, aims to sparsify the changes along the solution vector \( x \). We propose here a local neighborhood total variation (LN-TV) penalty using SWAGGER. LN-TV is achieved by assigning \( B = D \) and using an \( S \) that promotes sparsity in local neighborhoods as defined in Section IV-C. LN-TV simply ensures that changes in the vector occur at least some minimum distance apart, while allowing changes of any amplitude. This is a prior which suggests the vector is piecewise-constant with a certain minimum segment length. If a typical minimum separation distance between significant changes in a vector is known, LN-TV can outperform TV and Moreau-enhanced TV (a convex-nonconvex variation) in a denoising setting, as seen in Fig. 2.

For more general problems, where a piecewise approximation is helpful but the segments do not necessarily have a minimum length well known \textit{a priori}, the LN-TV penalty can be combined with the traditional TV by solving a problem of the form:

\[
\hat{x} = \arg\min_x \frac{1}{2} \|Ax - y\|_2^2 + \lambda_1 \|Dx^T S D x\|_1 + \lambda_2 \|Dx\|_1 \text{TV} \tag{9}
\]

where the \( \lambda \)'s are parameters that control the strength of the regularization. If \( A = I \), this acts to denoise the input \( y \), but different choices of \( A \) can be used to solve deconvolution or deblurring problems, super-resolution, and so on. The addition of the LN-TV term can help avoid the ‘staircasing’ artefacts that often arise when using TV, and allow for weaker TV regularization, which reduces the bias in the result. We present an example of these favorable effects in Fig. 3. Here, the \( S \) matrix used promotes sparsity in \( D x \) within five indices either side of a nonzero component. This formulation is distinct from the one discussed in (4), as instead of a constraint we are using the SWAGGER term as a regularization with a strength prescribed by \( \lambda_1 \). This allows us to have some more nuance
by having a linear ramp from 1 to 0.5 in the band around the diagonal, to make close changes less likely than ones a greater distance away.

We can extend this formulation to a 2D setting also. Using the penalty formulation in (10), we promote horizontal and vertical differences that occur only with some minimal separation distance:

$$R(x) = \lambda (|D_h x^T S_h |_{D_h x} + |D_v x^T S_v |_{D_v x})$$

where $D_h$ and $D_v$ are matrices that take finite differences along the vertical columns and horizontal rows, respectively, and $S_h$ and $S_v$ impose local neighborhood sparsity in the vertical and horizontal directions. We choose $\Phi(x) = |x|^\kappa$ which introduces another parameter $\kappa$ as discussed in Section V.

We find that, in general, $\kappa$ should be kept somewhat smaller than 1 if the initial estimate is especially blurred or noisy, and larger than 1 if there exist hard edges in the initial estimate that should be maintained. For the results shown in Fig. 4 $\kappa = 0.75$.

C. Modeling Mutual Surface Occlusions in Non-Line-of-Sight (NLOS) Imaging

The aim of NLOS imaging is to form reconstructions of scenes hidden from the direct line of sight of the observer. Typical methods to achieve this rely on the use of ultra-fast pulsed lasers and time-resolved single photon counting to infer the hidden geometry by measuring round-trip distances [22], [23], [24], [25]. Often, to recover a 3D scene, the hidden area will be discretized into patches or surfaces in some manner. Some of these patches may occlude others: if a surface is actually present in the scene, light coming from others behind it may have no paths to the measurement device. As the discretization is prescribed beforehand, the surfaces occluded by each other are known a priori, and they should be mutually exclusive. Therefore, we can use SWAGGER to ensure that groups of surfaces that occlude one another will be one-sparse in the recovered output. This provides two benefits. Firstly, the recovered output will be physically plausible as surfaces that cannot contribute any light in the measurements will not be present in the solution. Secondly, ensuring this physical constraint is met during the optimization process should lead to better estimations overall.

This idea extends to any situation in which an ideal discretization of some physical area, field, etc., involves some mutual exclusivity between elements of the discretization. Typically, to deal with this one may instead use a discretization that is less desirable but avoids or reduces this exclusivity requirement, use post-processing to make the recovery fit the expected structure, or simply ignore it. Using SWAGGER can instead ensure that reconstructions both fit the expected structure prescribed by the underlying physics of the problem, and also enjoy improved estimates due to the knowledge of the structure being used within the optimization procedure.

One NLOS imaging system, Edge-Resolved Transient Imaging [26], uses a laser to illuminate the floor at numerous positions in an arc around a vertical edge, such as a doorway, to form a 3D reconstruction of the room beyond. For each measurement position, a histogram of the arrival times of photons reaching a single photon avalanche diode (SPAD) detector is accumulated using time-correlated single photon counting. Each subsequent laser position illuminates more of the hidden scene, and by taking differences between measurement histograms one can recover a noisy measurement containing photon arrivals originating mostly from within a single wedge in the hidden area.

From this difference histogram measurement, one can attempt to fit surfaces to the scene within a wedge, to reconstruct...
Photon difference

(a) Experimental measurement input
(b) Ground truth LASSO + SWAGGER
(c) Sparsity matrix

0 1 2 3 4 ns 6 ns 8 ns 10 ns 12 ns 14 ns 16 ns 18 ns 20 ns 22 ns 24 ns 26 ns

Fig. 5. (a) Experimental measurement data containing predominantly photon returns from a single wedge in a hidden area. The data corresponds to the 18th wedge in the 'Staircase' scene, Figure 4 in [26]. (b) The ground truth scene and the reconstruction using SWAGGER. The displayed reflectivity estimate is enhanced by scaling based on the amount of the surface that is occluded. (c) The sparsity matrix used to model occlusion and mutual exclusivity of surfaces occupying the same space. (d) Reconstructions without using SWAGGER, with increasing regularization strength. No choice of λ gives accuracy comparable to the reconstruction using SWAGGER in (b).

the hidden area. It is assumed that surfaces always start at the ground, extend to a certain height, and are a certain distance from the corner. We can form a reconstruction by solving the following problem:

\[
x = \arg \min_{x \geq 0} \frac{1}{2} \|Ax - y\|^2_2 + \lambda_1 \|x\|_1
\]

s.t. \[|x^T S x| = 0\] (11)

where y is a difference histogram measurement and A is a matrix governing the light transport to and from surfaces at different distances from the edge, and with different heights. Fig. 5 shows reconstructions of a wedge using experimental data from [26]. One reconstruction uses only the \(\|x\|_1\) term and no constraints, and a second that includes the SWAGGER constraint. In [26], a Markov chain Monte Carlo (MCMC) method was used to form the reconstructions as it could implicitly include occlusions model and needed to only estimate parameters for a small number of surfaces, rather than use a full scene discretization. Here, we use SWAGGER and a full scene discretization to ensure that surfaces which should occlude those behind them, and vice versa, are accurately modeled. Secondly, it permits us to use a discretization with mutually exclusive elements—surfaces in the same position but with different heights—where only one should be present in the solution. As such, the result is a physically plausible reconstruction, which is quite accurate in support, height, and reflectivity.

VIII. CONCLUSION

Using SWAGGER as a constraint allows one to recover solutions to problems that fit a known sparsity structure, including any number of one-sparse overlapping groups within the solution vector or a transform thereof. This is useful in modeling a wide variety of phenomena, such as occlusions in imaging problems. We illustrate the usefulness of the proposed penalty for some example use cases where, by simply constructing a sparsity structure matrix S, one can ensure estimates fit known hard structured sparsity constraints derived from domain knowledge of the problem. This results in solutions that are both physically plausible and also improved overall. Similarly, SWAGGER can be used as a regularization term with some strength, to promote solutions that fit a sparsity structure whilst allowing more nuance, which gives rise to local-neighborhood total variation. SWAGGER can be readily combined with other priors, regularization terms and constraints to improve estimations. A proximal (sub)gradient algorithm (and accelerated algorithm) have been proposed to quickly solve SWAGGER-constrained and SWAGGER-regularized problems.

APPENDIX

A. Proof of Proposition 1

Suppose both conditions on Proposition 1 hold. By condition 2), every entry of \(\Phi(Bx)\) is nonnegative. By condition 1), every entry of S is also nonnegative. Therefore, each entry in the matrix triple product \(\Phi(Bx)^T S \Phi(Bx)\) is nonnegative, ensuring that the result is nonnegative.

B. Proximity Operator Algorithm for \(\ell_1^2\)

Algorithm 2 reproduces the method from [17] to compute the proximal operator for the \(\ell_1^2\) norm (with \(O(N \log N)\) complexity).

Algorithm 2 Proximity operator of \(\ell_1^2\)

Input: \(z \in \mathbb{R}^N, \lambda \geq 0\)

Output: \(z = \arg \min_{q} \|z - q\|^2_2 + \frac{1}{2}\lambda \|q\|^2_2\)

1: sort entries of \(|z|\) into \(y\) s.t. \((y_1 \geq \cdots \geq y_N)\)
2: set \(\rho = \max\{j \in \{1, \ldots, N\}; y_j - \frac{\lambda}{1+\lambda} \sum_{r=1}^j y_r > 0\}\)
3: return \(z = \max(\{z - \tau, 0\}) \text{sign}(z)\), where \(\tau = \frac{\lambda}{1+\lambda} \sum_{r=1}^\rho y_r\)

C. Accelerated SWAGGER Algorithm

In [18], a monotonic accelerated proximal gradient method is outlined that is guaranteed to converge to a stationary point for nonconvex problems. Algorithm 3 applies this acceleration to Algorithm 1, where \(\alpha_x, \alpha_y\) and \(\alpha_\lambda\) are step sizes that can be fixed or computed dynamically with a back-tracking scheme. The convergence proof requires that \(f(x)\) is a proper function with Lipschitz continuous gradients (e.g., the least
squares loss), and the SWAGGER penalty is proper and lower semicontinuous, which is the case for $\Phi(x) = |x|$.

**Algorithm 3** Accelerated proximal subgradient method for SWAGGER

**Input:** $z_1 = x_1 = x_0$, $t_1 = 1$, $t_0 = 0$, $\overline{S} = I$.

**Output:** $x$

1. **while** not converged **do**
   2. $y^k = x^k + \frac{1}{t^k} (z^k - x^k) + \frac{1}{t^k} (x^k - x^{k-1})$
   3. $z^{k+1} = P_{\alpha y} (x^k - \alpha y \nabla_y (J + \lambda_k |y^k|^T S |y^k|))$
   4. $v^{k+1} = P_{\alpha x} (y^k - \alpha x \nabla_x (J + \lambda_k |x^k|^T S |x^k|))$
   5. $x^{k+1} = \begin{cases} v^{k+1} & F(v^{k+1}) \leq F(y^{k+1}) \\ z^{k+1} & \text{otherwise} \end{cases}$
   6. $t^{k+1} = \frac{1}{2} \left( \sqrt{1 + 1} + 1 \right)$
   7. $\lambda^{k+1} = \lambda^k + \alpha \lambda (|x^k|^T S |x^k|)$
   8. **end while**

9. **return** $x = x^{k+1}$

### D. Generating Test Vectors

For the experiments in Section VII-A we must generate a test vector $x$ for each trial which fits the sparsity structure described by a specific $S$ matrix. To do so we use Algorithm 3 where in Step 4, $u_i$ is drawn from the continuous uniform distribution on $[0.5, 1.5]$.

**Algorithm 4** Generate a vector $x_S$ s.t. $|x|^T S |x| \approx 0$

**Input:** $x^0 \sim N(0, 1), S$, $\mu \ll 1$

**Output:** $x_S$

1. **while** $|x^0|^T S |x^0| \geq \mu$ **do**
   2. $x_{i+1} = \arg \min_x \|x^0 - x\| + |x|^T S |x|$
   3. **end while**

4. **return** $x_{S,j} = \begin{cases} 0, & x_{j+1} = 0 \\ \text{sgn}(x_{j+1}) u_j, & x_{j+1} \neq 0 \end{cases}$

### E. The Choice of $c$ for Convex Nonconvexity

The Hessian of the cost function

$$\hat{x} = \arg \min_x \frac{1}{2} \|Ax - y\|^2 + \lambda \|x|^T S |x|$$

is given by $A^T A + 2 \lambda S S^T X$ (where $X = \text{diag}(|\text{sgn}(x)|))$. We introduce a term $cI$ to the $S$ matrix:

$$A^T A + 2 \lambda X(S - c I) X = A^T A + 2 \lambda X S X - 2 \lambda c X I = A^T A + 2 \lambda X S X - 2 \lambda c I.$$

To maintain convexity, we wish to have

$$\lambda_{\min}(A^T A + 2 \lambda X X) - 2 \lambda c \geq 0,$$

or equivalently

$$2 \lambda c \leq \lambda_{\min}(A^T A + 2 \lambda X X).$$

Since $A^T A$ and $2 \lambda X X$ are both Hermitian,

$$\lambda_{\min}(A^T A + 2 \lambda X X) \geq \lambda_{\min}(A^T A) + \lambda_{\min}(2 \lambda X X)$$

by Weyl’s inequality [27]. Finally, we note

$$\min_x \lambda_{\min}(2 \lambda X X) = 2 \lambda \lambda_{\min}(S),$$

providing us with a bound for the choice of $c$:

$$c \leq \frac{1}{2 \lambda} \lambda_{\min}(A^T A) + \lambda_{\min}(S).$$

### ACKNOWLEDGMENT

We thank Y. Ma for comments on the manuscript, and S. W. Seidel, J. Rapp, J. Murray-Bruce and I. Selesnick for discussions.

### REFERENCES

[1] R. Tibshirani, “Regression shrinkage and selection via the lasso,” J. Royal Statistical Society, Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.
[2] S. Bakin, “Adaptive regression and model selection in data mining problems,” Ph.D. dissertation, Australian National University, 1999.
[3] Y. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” J. Royal Statistical Society Series B, vol. 68, no. 1, pp. 49–67, Feb. 2006.
[4] J. Yoon and S. J. Hwang, “Combined group and exclusive sparsity for deep neural networks,” in Proc. 34th Int. Conf. Machine Learning, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70, 06–11 Aug 2017, pp. 3958–3966.
[5] F. Song, X. Tan, and S. Chen, “Exploiting relationship between attributes for improved face verification,” Computer Vision and Image Understanding, vol. 122, pp. 143–154, 2014.
[6] A. Wang, J. Cai, J. Lu, and T. J. Cham, “Modality and component aware feature fusion for RGB-D scene classification,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Jun. 2016.
[7] D. Paz-Linares, M. Vega-Hernández, P. A. Rojas-López, P. A. Valdés-Hernández, E. Martínez-Montes, and P. A. Valdés-Sosa, “Spatial temporal EEG source imaging with the hierarchical Bayesian elastic net and elitist lasso models,” Frontiers in Neuroscience, vol. 11, p. 635, 2017.
[8] N. Rao, C. Cox, R. Nowak, and T. T. Rogers, “Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis,” in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013, pp. 2202–2210.
[9] M. Kowalski, “Sparse regression using mixed norms,” Applied and Computational Harmonic Analysis, vol. 27, no. 3, pp. 303–324, 2009.
[10] M. Kowalski and B. Torrésani, “Sparsity and persistence: Mixed norms provide simple signal models with dependent coefficients,” Signal Image and Video Processing, vol. 3, pp. 251–264, Sep. 2009.
[11] Y. Zhou, R. Jin, and S. C. Hoi, “Exclusive lasso for multi-task feature selection,” in Proc. Thirteenth Int. Conf. Artificial Intelligence and Statistics, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9, 13–15 May 2010, pp. 988–995.
[12] Ilker Bayram and S. Bulek, “A penalty function promoting sparsity within and across groups,” IEEE Trans. Signal Process., vol. 65, no. 16, pp. 4238–4251, Aug 2017.
[13] Ilker Bayram and S. Bulek, “A penalty function promoting sparsity within and across groups,” IEEE Trans. Signal Processing, vol. 65, no. 2, pp. 288–292, Feb 2018.
[14] S. Walker and P. Van Mieghem, “On lower bounds for the largest eigenvalue of a symmetric matrix,” Linear Algebra and its Applications, vol. 429, no. 2, pp. 519 – 526, 2008.
[15] Z. B. Charles, M. Farber, C. R. Johnson, and L. Kennedy-Shaffer, “Nonpositive eigenvalues of hollow, symmetric, nonnegative matrices,” SIAM J. Matrix Analysis and Applications, vol. 34, no. 3, pp. 1384–1400, 2013.
[16] B. Eisele, “Sparse regularization via convex analysis,” IEEE Trans. Signal Process., vol. 65, no. 17, pp. 4481–4494, 2017.
[17] A. F. T. Martins, N. Smith, E. Xing, P. Aguiar, and M. Figueiredo, “Online learning of structured predictors with multiple kernels,” in Proc. Fourteenth Int. Conf. Artificial Intelligence and Statistics, G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15, 11–13 Apr 2011, pp. 507–515.
[18] H. Li and Z. Lin, “Accelerated proximal gradient methods for nonconvex programming,” in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 379–387.
[19] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Found. Trends Mach. Learn.*, vol. 3, no. 1, p. 1–122, Jan. 2011. [Online]. Available: [https://doi.org/10.1561/2200000016](https://doi.org/10.1561/2200000016)

[20] J. Woodworth and R. Chartrand, “Compressed sensing recovery via nonconvex shrinkage penalties,” *Inverse Problems*, vol. 32, no. 7, p. 075004, May 2016.

[21] I. Selesnick, “Total variation denoising via the Moreau envelope,” *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 216–220, 2017.

[22] A. Kirmani, T. Hutchison, J. Davis, and R. Raskar, “Looking around the corner using transient imaging,” in *Proc. IEEE Int. Conf. Computer Vision*, 2009, pp. 159–166.

[23] A. Velten, T. Willwacher, O. Gupta, A. Veeraraghavan, M. G. Bawendi, and R. Raskar, “Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging,” *Nat. Commun.*, vol. 3, 2012.

[24] F. Heide, L. Xiao, W. Heidrich, and M. B. Hullin, “Diffuse mirrors: 3D reconstruction from diffuse indirect illumination using inexpensive time-of-flight sensors,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 3222–3229.

[25] M. O’Toole, D. B. Lindell, and G. Wetzstein, “Confocal non-line-of-sight imaging based on the light-cone transform,” *Nature*, vol. 555, pp. 338–341, Mar. 2018.

[26] J. Rapp, C. Saunders, J. Tachella, J. Murray-Bruce, Y. Altmann, J.-Y. Tourneret, S. McLaughlin, R. M. A. Dawson, F. N. C. Wong, and V. K. Goyal, “Seeing around corners with edge-resolved transient imaging,” arXiv:2002.07118v1 [eess.IV], Feb. 2020.

[27] R. A. Horn and C. R. Johnson, *Matrix Analysis: Second Edition*. Cambridge University Press, 2012.