Combinatorial Selection and Least Absolute Shrinkage via the CLASH Algorithm

Anastasios Kyrillidis and Volkan Cevher
Laboratory for Information and Inference Systems, École Polytechnique Fédérale de Lausanne

Abstract—The least absolute shrinkage and selection operator (LASSO) for linear regression exploits the geometric interplay of the $\ell_2$-data error objective and the $\ell_1$-norm constraint to arbitrarily select sparse models. Guiding this uninformed selection process with sparsity models has been precisely the center of attention over the last decade in order to improve learning performance. To this end, we alter the selection process of LASSO to explicitly leverage combinatorial sparsity models (CSMs) via the combinatorial selection and least absolute shrinkage (CLASH) operator. We provide concrete guidelines how to leverage combinatorial constraints within CLASH, and characterize CLASH’s guarantees as a function of the set restricted isometry constants of the sensing matrix. Finally, our experimental results show that CLASH can outperform both LASSO and model-based compressive sensing in sparse estimation.

I. INTRODUCTION

The least absolute shrinkage and selection operator (LASSO) is the de facto standard algorithm for regression [1]. LASSO estimates sparse linear models by minimizing the empirical data error via:

$$\hat{x}_{\text{LASSO}} = \arg \min \left\{ \|y - \Phi x\|_2^2 : \|x\|_1 \leq \lambda \right\},$$  \(1\)

where $\| \cdot \|_r$ is the $\ell_r$-norm. In (1), $\Phi \in \mathbb{R}^{m \times n}$ is the sensing matrix, $y \in \mathbb{R}^m$ are the responses (or observations), $x \in \mathbb{R}^n$ is the loading vector and $\lambda \in \mathbb{R}^{++}$ governs the sparsity of the solution. Along with many efficient algorithms for its solution, the LASSO formulation is now backed with a rather mature theory for the generalization of its solutions as well as its variable selection consistency [2]–[5].

While the long name attributed to (1) is apropos,\(^1\) it does not capture the LASSO’s arbitrariness in subset selection via shrinkage to best explain the responses. In fact, this uninformed selection process not only prevents interpretability of results in many problems, but also fails to exploit key prior information that could radically improve learning performance. Based on this premise, approaches to guide the selection process of the LASSO are now aplenty.

Surprisingly, while the prior information in many regression problems generate fundamentally discrete constraints (e.g., on the sparsity patterns or the support of the LASSO solution), the majority of the existing approaches that enforce such constraints in selection are inherently continuous. For instance, a prevalent approach is to tailor a sparsity inducing norm to the constraints on the support set (c.f., [6]). That is, we create a structured convex norm by mixing basic norms with weights over pre-defined groups or using the Lovász extension of non-decreasing submodular set functions of the support. As many basic norms have well-understood behavior in sparse selection, reverse engineering such norms is quite intuitive.

While such structure inducing, convex norm-based approaches on the LASSO are impressive, our contention in this paper is that, in order to truly make an impact in structured sparsity problems, we must fully leverage explicitly combinatorial approaches to guide LASSO’s subset selection process. To achieve this, we show how Euclidean projections with structured sparsity constraints correspond to an integer linear program (ILP), which can be exactly or approximately solved subject to matroid (via the greedy algorithm), and certain linear inequality constraints (via convex relaxation or multi-knapsack solvers). A key actor in this process is a polynomial-time combinatorial algorithm that goes beyond simple selection heuristics towards provable solution quality as well as runtime/space bounds.

Furthermore, we introduce our combinatorial selection and least absolute shrinkage (CLASH) operator and theoretically characterize its estimation guarantees. CLASH enhances the model-based compressive sensing (model-CS) framework [7] by combining $\ell_1$-norm and combinatorial constraints on the regression vector. Therefore, CLASH uses a combination of shrinkage and hard thresholding operations to significantly outperform the model-CS approach, LASSO, or continuous structured sparsity approaches in learning performance of sparse linear models. Furthermore, CLASH establishes a regression framework where the underlying tractability of approximation in combinatorial selection is directly reflected in the algorithm’s estimation and convergence guarantees.

The organization of the paper is as follows. In Sections II and III, we set up the notation and the exact projections with structured sparsity constraints. We develop CLASH in Section IV and highlight the key components of its convergence proof in Section V. We present numerical results in Section VI. We provide our conclusions in Section VII.

II. PRELIMINARIES

Notation: We use $[x]_j$ to denote the $j$-th element of $x$, and let $x_i$ represent the $i$-th iterate of CLASH. The index set of $n$
dimensions is denoted as $N = \{1, 2, \ldots, n\}$. Given $S \subseteq N$, we define the complement set $S^c = N \setminus S$. Moreover, given a set $S \subseteq N$ and a vector $x \in \mathbb{R}^n$, $(x)_S \in \mathbb{R}^{|S|}$ denotes a vector with the following properties: \((x)_S \mid_S = [x]_S\) and \((x)_{S^c} = 0\). The support set of $x$ is defined as $\supp(x) = \{i : |x_i| \neq 0\}$. We use $|S|$ to denote the cardinality of the set $S$. The empirical data error is denoted as $f(x) = \|y - \Phi x\|_2^2$ with gradient defined as $\nabla f(x) = -2\Phi^T(y - \Phi x), \Phi^T$ is the transpose operation. The notation $\nabla f(x)$ is shorthand for $(\nabla f(x))_S$. $\mathbb{I}$ represents the identity matrix.

**Combinatorial notions of sparsity:** We provide some definitions on combinatorial sparse models, and elaborate on a subset of interesting models with algorithmic implications.

**Definition 1** (Combinatorial sparsity models (CSMs)). We define a combinatorial sparsity model $C_k = \{S_m : \forall m, S_m \subseteq N, |S_m| \leq k\}$ with the sparsity parameter $k$ as a collection of distinct index subsets $S_m$.

Throughout the paper, we assume that any CSM $C_k$ is downward compatible, i.e., removing any subset of indices of any given element in $C_k$, it is still in $C_k$.

**Properties of the regression matrix:** Deriving approximation guarantees for CLASH behooves us to assume the restricted isometry property (RIP) (defined below) on the regression matrix $\Phi$ [8]. While the RIP and other similar conditions for deriving consistency properties of LASSO and its variants, such as the unique/exact representation property or the irrepresentable condition [5], [6], [9]–[11], are unverifiable a priori without exhaustive search, many random matrices satisfy them with high probability.

**Definition 2** (RIP [7], [8]). The regression matrix has the $k$-RIP with an isometry constant $\delta_k$ when
\[
(1 - \delta_k)|x|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_k)|x|_2^2, \quad \forall \supp(x) \in C_k,
\]
where $\delta_k = \max_{S \in C_k} \|\Phi_S^T\Phi_S - \mathbb{I}\|_{2\to2}$ and $\Phi_S$ is a submatrix of $\Phi$ as column-indexed by $S$.

Here, we also comment on the scaling of $(k, m, n)$ for the desired level of isometry. When the entries of $\Phi$ can be modeled as independent and identically distributed (iid) with respect to a sub-Gaussian distribution, we can show that $m = \mathcal{O}(\delta_k^{-2}(\log(2M) + k \log(12\delta_k^{-1})))$ with overwhelming probability [7]. Here, $M$ is the minimum number of subspaces covering $C_k$. While $m$ explicitly depends on $n$, for certain restricted CSs, such as the rooted connected tree of [7], this dependence can be quite weak, e.g., $m = \mathcal{O}(k)$.

### III. Exact and Approximate Projections onto CSs

The workhorse of the model-CS approach is the following non-convex projection problem onto CSs, as defined by $C_k$, which is a basic subset selection problem:
\[
P_{C_k}(x) = \arg \min_{w \in \mathbb{R}^n} \{\|w - x\|_2^2 : \supp(w) \in C_k\}, \quad (3)
\]
where $P_{C_k}(\cdot)$ is the projection operator. [7] shows that as long as $P_{C_k}(\cdot)$ is exactly computed in polynomial time for a CSM, their sparse recovery algorithms inherit strong approximation guarantees for that CSM. To better identify the CSs that live within the model-CS assumptions, we first state the following key observation—the proof can be found in [12].

**Lemma 1** (Euclidean projections onto CSs). The support of the Euclidean projection onto $C_k$ in (3) can be obtained as a solution to the following discrete optimization problem:
\[
supp(P_{C_k}(x)) = \arg \max_{S \in C_k} \mathcal{F}(S; x), \quad (4)
\]
where $\mathcal{F}(S; x) = \|x\|_2^2 - \|(x)_{S^c} - x\|_2^2 = \sum_{i \in S} |x_i|^2$ is the modular, variance reduction set function. Moreover, let $\hat{S} \subseteq C_k$ be the minimizer of the discrete problem. Then, it holds that $P_{C_k}(x) = (x)_{\hat{S}}$, which corresponds to hard thresholding.

The following proposition refines this observation to further accentuate the algorithmic implications for CSs:

**Proposition 1** (CSM projections via ILP’s). The problem (4) is equivalent to the following integer linear program (ILP):
\[
supp \arg \min_{w} \{w^T z : |w_i| = -|\{z_i\}|^2, \supp(w) \in C_k\}
\]
where $[z]_i, (i = 1, \ldots, n)$, are support indicator variables.

The proof of Proposition 1 is straightforward and is omitted.

Regardless of whether we use a dynamic program, a greedy combinatorial algorithm, or an ILP solver, the formulations (4) or (5) make the underlying tractability of the combinatorial selection explicit. We highlight this notion via the polynomial-time modular $\epsilon$-approximation property (PMAP$_\epsilon$):

**Definition 3** (PMAP$_\epsilon$ [12]). A CSM has the PMAP$_\epsilon$ with constant $\epsilon$, if the modular subset selection problem (4) or the ILP (5) admit an $\epsilon$-approximation scheme with polynomial or pseudo-polynomial time complexity as a function of $n$, $\forall x \in \mathbb{R}^n$. Denoting the $\epsilon$-approximate solution of (4) or (5) as $S_\epsilon$, this means $\mathcal{F}(S_\epsilon; x) \geq (1 - \epsilon) \max_{S \in C_k} \mathcal{F}(S; x)$.

In this paper, we focus and elaborate on CSs with PMAP$_0$.

### A. Example CSs with PMAP$_0$

**Matroids:** By matroid, we mean that $C_k = (N, I)$ is a finite collection of subsets of $N$ that satisfies three conditions: (i) $\emptyset \in I$, (ii) if $S$ is in $I$, then any subset of $S$ is also in $I$, and (iii) for $S_1, S_2 \in I$ and $|S_1| > |S_2|$, there is an element $s \in S_1 \setminus S_2$ such that $S_2 \cup \{s\}$ is in $I$. As a simple example, the unstructured sparsity model (i.e., $x$ is $k$-sparse) forms a uniform matroid as it is defined as the union of all subsets of $N$ with cardinality $k$ or less. When $C_k$ forms a matroid, the greedy basis algorithm can efficiently compute (3) by solving (4) [13] where sorting and selecting the $k$ largest elements in absolute value is sufficient to obtain the exact projection.

Moreover, it turns out that this particular perspective provides a principled and tractable approach to encode an interesting class of matroid-structured sparsity models. The recipe is quite simple: we seek the intersection of a structure provider
matroid (e.g., partition, cographic/graphic, disjoint path, or matching matroid) with the sparsity provider uniform matroid. While the intersection of two matroids is not a matroid in general, we can prove that the intersection of the uniform matroid with any other matroid satisfies the conditions above.

Linear support constraints: Many interesting CSMs $C_k$ can be encoded using linear support constraints of the form:

$$C_k = \bigcup_{z \in 3} \text{supp}(z), 3 := \{ [z]_i \in \{0, 1\} : Az \leq b \},$$

where $[A, b]$ is an integral matrix, and the first row of $A$ is all 1’s and $[b]_1 = k$. As a basic example, the neuronal spike model of [14] is based on linear support constraints where each spike respects a minimum refractory distance to each other.

A key observation is that if each of the nonempty faces of $3$ contains an integral point (i.e., forming an integral polyhedra), then convex relaxation methods can exactly obtain the correct integer solutions in polynomial time. In general, checking the integrality of $3$ is NP-Hard. However, if $3$ is integral and non-empty for all integral $b$, then a necessary condition is that $A$ be a totally unimodular (TU) matrix [13]. A matrix is totally unimodular if the determinant of each square submatrix is equal to 0, 1, or -1. Example TU matrices include interval, perfect, and network matrices [13]. As expected, the constraint matrix $A$ of [14] is TU. Moreover, it is easy to verify that the sparse disjoint group model of [15] also defines a TU constraint, where groups have individual sparsity budgets.

B. How about PMAP$_\epsilon$?

For completeness and due to lack of space, we only mention PMAP$_\epsilon$, which extends the breath of the model-CS approach. For a detailed treatment of PMAP$_\epsilon$ and CLASH, c.f. [12], which describes multi-knapsack CSMS as a concrete example. Moreover, for many of the PMAP$_0$ examples above, we can employ $\epsilon$-approximate—randomized—techniques to reduce computational cost.

IV. THE CLASH ALGORITHM

The new CLASH algorithm obtains approximate solutions to the LASSO problem in (1) with the added twist that the solution must live within the CSM, as defined by $C_k$:

$$\hat{x}_{\text{CLASH}} = \arg \min \{ f(x) : \| x \|_1 \leq \lambda, \text{supp}(x) \in C_k \}. \quad (6)$$

When available, using the CSM constraint $C_k$ in addition to the $\ell_1$-norm constraint enhances learning in two important ways. First, the combinatorial constraints restricts the LASSO solution to exhibit true model-based supports, increasing the interpretability of the solution without relaxing $C_k$ into a convex norm. Second, it empirically requires much fewer number of samples to obtain the true solution than both the LASSO and the model-CS approaches.

We provide a pseudo-code of an example implementation of CLASH in Algorithm 1. One can think of alternative ways of implementing CLASH, such as single gradient updates in Step 2, or removing Step 4 altogether. While such changes may lead to different—possibly better—approximation guarantees for the solution of (6), we observe degradation in the empirical performance of the algorithm as compared to this implementation, whose guarantees are as follows:

**Theorem 1** (Iteration invariant). Let $x^* \in \mathbb{R}^n$ be the true vector that satisfies the conditions of (6) and let $y = \Phi x^* + \varepsilon$ be the set of observations with additive error $\varepsilon \in \mathbb{R}^m$. Then, the $i$-th iterate $x_i$ of CLASH satisfies the following recursion:

$$\|x_{i+1} - x^*\|_2 \leq \rho \|x_i - x^*\|_2 + c_1(\delta_{2k}, \delta_{3k})\|\varepsilon\|_2$$

where $\rho \triangleq \frac{\delta_{2k} + \delta_{3k}}{\sqrt{1 - \delta_{3k}^2}} \sqrt{\frac{1 + \delta_{3k}^2}{1 - \delta_{3k}^2}}$ and $c_1(\delta_{2k}, \delta_{3k})$ is a constant defined in [16]. The iterations contract when $\delta_{3k} < 0.3658$.

**Remark 1.** [Model mismatch and selection] Let us assume a generative model $y = \Phi x^* + \varepsilon$. Let $x^*$ be the best approximation of $x$ in $C_k$ within $\ell_1$-ball of radius $\lambda$. Then, we can show that the iteration invariant of Theorem 1 still holds with $SNR = \frac{\|x^*\|_2}{\|\varepsilon\|_2}$, where $\|\varepsilon\|_2 \leq \|\varepsilon\|_2 + \|\Phi(x^* - x^*)\|_2$, where the latter quantity (the impact of mismatch) can be analyzed using the restricted amplification property of $\Phi$ [7]. For instance, when $C_k$ is the uniform sparsity model, then $\|\Phi(x^* - x^*)\|_2 \leq \sqrt{1 + \delta_{3k}} \left( \|\beta - x^*\|_2 + \sqrt{\frac{2(1 + \delta_{3k})}{\sqrt{k}}} \|\varepsilon\|_2 \right)$, which should presumably be small if the model is selected correctly.

In the absence of prior information, we automate the parameter selection by using the Donoho-Tanner phase transition [19] to choose the maximum $k$ allowed for a given $(m, n)$-pair, and then by using cross validation to pick $\lambda$ [20].

V. PROOF OF THEOREM 1

We sketch the proof of Theorem 1 a l`a [17] and [21] assuming the general case of PMAP$_\epsilon$. The details of the proof can be found in an extended version of the paper [16].

**Lemma 2** (Active set expansion - Step 1). The support set $S_i$, where $|S_i| \leq 2k$, identifies a subspace in $C_{2k}$ such that:

$$\|(x_i - x^*)_{S_i}\|_2 \leq (\delta_{3k} + \delta_{2k} + \sqrt{1 + \delta_{3k}})\|x_i - x^*\|_2 + (\sqrt{2(1 + \delta_{3k})} + \sqrt{\|\varepsilon\|_2})\|\varepsilon\|_2$$

(7)
Lemma 2 states that, at each iteration, Step 1 of CLASH identifies a $2k$ support set such that the unrecovered energy of $x^*$ is bounded. For $\epsilon = 0$, CLASH exactly identifies the support where the projected gradient onto $C_k$ can make most impact on the loading vector in the support complement of its current solution, which are subsequently merged together.

**Lemma 3** (Greedy descent with least absolute shrinkage - Step 2). Let $S_i$ be a $2k$-sparse support set. Then, the least squares solution $v_i$ in step 2 of Algorithm 1 satisfies

$$\|v_i - x^*\|_2 \leq \frac{1}{\sqrt{1 - \delta_{2k}^2}} \| (x_i - x^*) S_i \|_2 + \sqrt{1 + \delta_{2k}} \| \epsilon \|_2.$$ 

We borrow the proof of Lemma 3 from [21]. This step improves the objective function $f(x)$ as much as possible on the active set in order to arbitrate the active set. The solution simultaneously satisfies the $\ell_1$-norm constraint.

Step 3 projects the solution onto $C_k$, whose action is characterized by the following lemma. Here, we show the $\epsilon$-approximate projection explicitly:

**Lemma 4** (Combinatorial selection - Step 3). Let $v_i$ be a $2k$-sparse proxy vector with indices in support set $S_i$, $C_k$ be a CSM and $\gamma_i$ the projection of $v_i$ under $C_k$. Then:

$$\|\gamma_i - v_i\|_2^2 \leq (1 - \epsilon) \| (v_i - x^*) S_i \|_2^2 + \epsilon \|v_i\|_2^2.$$ 

Step 4 requires the following Corollary to Lemma 3:

**Corollary 1** (De-bias - Step 4). Let $\Gamma_i$ be the support set of a proxy vector $\gamma_i$ where $|\Gamma_i| \leq k$. Then, the least squares solution $x_{i+1}$ in Step 4 satisfies

$$\|x_{i+1} - x^*\|_2 \leq \frac{1}{\sqrt{1 - \delta_{2k}^2}} \|\gamma_i - x^*\|_2 + \sqrt{1 + \delta_{2k}} \| \epsilon \|_2.$$ 

Step 4 de-biases the current result on the putative solution support. Its characterization connects Lemmas 3 and 4:

**Lemma 5.** Let $v_i$ be the least squares solution of the greedy descent step (step 5) and $\gamma_i$ a proxy vector to $v_i$ after applying Combinatorial selection step. Then, $\|\gamma_i - x^*\|_2$ can be expressed in terms of the distance from $v_i$ to $x^*$ as follows:

$$\|\gamma_i - x^*\|_2 \leq \sqrt{1 + \frac{1}{\sqrt{1 - \delta_{2k}^2}} \|v_i - x^*\|_2 + \sqrt{1 + \delta_{2k}} \| \epsilon \|_2}.$$ 

where $D_1, D_2, D_3$ are constants depending on $\epsilon, \delta_{2k}, \delta_{3k}$.

Finally, the proof of Theorem 1 follows by concatenating Corollary 1 with Lemmas 2, 3, and 5, and setting $\epsilon = 0$.

VI. EXPERIMENTS

In the following experiments, we compare algorithms from the following list: (i) the LASSO algorithm [1], (ii) the Basis Pursuit DeNoising (BPDN) [22], (iii) the sparse-CLASH algorithm, where $C_k$ is the index set of $k$-sparse signals, (iv) the model-CLASH algorithm\(^2\), which explicitly carries $C_k$, and (v) Subspace Pursuit (SP) algorithm [18], as integrated with the model-CS approach. We emphasize here that when $\lambda \to \infty$ in (6), CLASH must converge to the model-based SP solution.

The LASSO algorithm finds a solution to the problem defined in (1), where we use a Nesterov accelerated projected gradient algorithm. The BPDN algorithm in turn solves the following optimization problem:

$$\hat{x}_{BP DN} = \arg \min \{\|x\|_1 : \|Dx - y\|_2 \leq \sigma\},$$

where $\sigma$ represents prior knowledge on the energy of the additive noise term. To solve (9), we use the spectral projected gradient method SPGL1 algorithm [23].

In the experiments below, the nonzero coefficients of $x^*$ are generated iid according to the standard normal distribution with $\|x^*\|_2 = 1$. The BPDN algorithm is given the true $\sigma$ values. While CLASH is given the true value of $k$ for the experiments below, additional experiments (not shown) shows that our phase transition heuristics is quite good and the mismatch is graceful as indicated in Remark 1. All the algorithms use a high precision stopping tolerance $\eta = 10^{-6}$. 

**Experiment 1: Improving simple sparse recovery.** In this experiment, we generate random realizations of the model $y = \Phi x^* + \varepsilon$ for $n = 800$. Here, $\Phi$ is a dense random matrix whose entries are iid Gaussian with zero mean and variance $1/m$. We consider two distinct generative model settings: (i) with additive Gaussian white noise with $\|\varepsilon\|_2 = 0.05$, $m = 240$ and $k = 89$, and (ii) the noiseless model ($\|\varepsilon\|_2 = 0$), $m = 250$ and sparsity parameter $k = 93$. For this experiment, we perform 500 Monte Carlo model realizations.

We sweep $\lambda$ and illustrate the recovery performance of CLASH (6). Figures 1(a)-(b) illustrate that the combination of hard thresholding with norm constraints can improve the signal recovery performance significantly over convex-only and hard thresholding-only methods—both in noisy and noiseless problem settings. For $\|\varepsilon\|_2 = 0$, CLASH perfectly recovers the signal when $\lambda$ is close to the true value. When $\lambda \ll \|x^*\|_1$, the performance degrades due to the large norm mismatch.

**Experiment 2: Improving structured sparse recovery** We consider two signal CSMS: in the first model, we assume $k$-sparse signals that admit clustered sparsity with coefficients in $C$-contiguous blocks on an undirected, acyclic chain graph [24]. Without loss of generality, we use $C = 5$ (Figure 1(c)). The second model corresponds to a TU system where we partition the $k$-sparse signals into uniform blocks and force sparsity constraints on individual blocks; in this case, we solve the set optimization problem optimally via linear programming relaxation (Figure 1(d)). Here, the noise energy level satisfies $\|\varepsilon\|_2 = 0.05$, and $n = 500$, $m = 125$, and $k = 50$. In both cases, we conduct 100 Monte Carlo iterations and perform sparse estimation for a range of $\lambda$ values.

In Figure 1(c), we observe that clustered sparsity structure provides a distinct advantage in reconstruction compared to LASSO formulation and the sparse-CLASH algorithm. Furthermore, note that when $\lambda$ is large, norm constraints have no effect and the model-CLASH provides essentially the same results as the model-CS approach [7]. On the other hand, the sparse-
CLASH improves significantly beyond the LASSO solution thanks to the $\ell_1$-norm constraint.

In Figure 1(d) however, the situation is radically changed: while the TU constraints enhance the reconstruction of model-based CS approach over simple sparse recovery, the improvement becomes quite large as the $\ell_1$-norm constraint kicks in. We also observe the improvement in sparse-CLASH but it is not as accentuated as the model-CLASH.

VII. CONCLUSIONS

CLASH establishes a regression framework where efficient algorithms from combinatorial and convex optimization can interface for interpretable sparse solutions. Our experiments demonstrate that, while the model-based selection can greatly improve sparse recovery over the approaches based on uniform sparsity alone, the $\ell_1$-norm shrinkage operations have an undeniable, positive impact on the learning performance. Understanding the tradeoffs between the complexity of approximation and the recovery guarantees of CLASH is a promising theoretical and practical direction. Finally, joint combinatorial and norm constrained scenarios appear in many problems of interest such as portfolio optimization.

REFERENCES


Fig. 1. Median values of signal error $\|\hat{x} - x^*\|_2$. Top row: simple sparsity model under noisy $\|e\|_2 = 0.05$ (left column) and noiseless $\|e\|_2 = 0$ (right column) settings. Bottom row: the $(k, C)$-clustered sparsity model (left column) and the TU model (right column).