Sparse projections onto the simplex

Anastasios Kyrillidis   ANASTASIOS.KYRILLIDIS@EPFL.CH
Stephen Becker   STEPHEN.BECKER@UPMC.FR
Volkan Cevher   VOLKAN.CEVHER@EPFL.CH
Christoph Koch   CHRISTOPH.KOCH@EPFL.CH

Abstract

Most learning methods with rank or sparsity constraints use convex relaxations, which lead to optimization with the nuclear norm or the ℓ_1-norm. However, several important learning applications cannot benefit from this approach as they feature these convex norms as constraints in addition to the non-convex rank and sparsity constraints. In this setting, we derive efficient sparse projections onto the simplex and its extension, and illustrate how to use them to solve high-dimensional learning problems in quantum tomography, sparse density estimation and portfolio selection with non-convex constraints.

1. Introduction

We study the following *sparse* Euclidean projections:

Problem 1. *(Simplex) Given* $\mathbf{w} \in \mathbb{R}^p$*, find a Euclidean projection of* $\mathbf{w}$ *onto the intersection of* k*-sparse vectors* $\Sigma_k = \{\boldsymbol{\beta} \in \mathbb{R}^p : |\{i : \beta_i \neq 0\}| \leq k\}$ *and the simplex* $\Delta_\lambda^+ = \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_i \geq 0,\ \sum_i \beta_i = \lambda\}$*:*

$$\mathcal{P}(\mathbf{w}) \in \underset{\boldsymbol{\beta} : \boldsymbol{\beta} \in \Sigma_k \cap \Delta_\lambda^+}{\operatorname{argmin}} \|\boldsymbol{\beta} - \mathbf{w}\|_2. \qquad (1)$$

Problem 2. *(Hyperplane) Replace* Δ_λ^+ *in* (1) *with the hyperplane constraint* $\Delta_\lambda = \{\boldsymbol{\beta} \in \mathbb{R}^p : \sum_i \beta_i = \lambda\}$*.*

We prove that it is possible to compute such projections in quasilinear time via simple greedy algorithms.

Our motivation with these projectors is to address important learning applications where the standard sparsity/low-rank heuristics based on the ℓ_1/nuclear-norm are either given as a constraint or conflicts with the problem constraints. For concreteness, we

highlight quantum tomography, density learning, and Markowitz portfolio design problems as running examples. We then illustrate provable non-convex solutions to minimize quadratic loss functions

$$f(\boldsymbol{\beta}) := \|\mathbf{y} - \boldsymbol{\mathcal{A}}(\boldsymbol{\beta})\|^2 \qquad (2)$$

subject to the constraints in Problem 1 and 2 with our projectors. In (2), we assume that $\mathbf{y} \in \mathbb{R}^m$ is given and the (known) operator $\boldsymbol{\mathcal{A}} : \mathbb{R}^p \to \mathbb{R}^m$ is linear.

For simplicity of analysis, our minimization approach is based on the projected gradient descent algorithm:

$$\boldsymbol{\beta}^{i+1} = \mathcal{P}(\boldsymbol{\beta}^i - \mu^i \nabla f(\boldsymbol{\beta}^i)), \qquad (3)$$

where $\boldsymbol{\beta}^i$ is the i-th iterate, $\nabla f(\cdot)$ is the gradient of the loss function, μ^i is a step-size, and $\mathcal{P}(\cdot)$ is based on Problem 1 or 2. When the linear map $\boldsymbol{\mathcal{A}}$ in (2) provides bi-Lipschitz embedding for the constraint sets, we can derive rigorous approximation guarantees for the algorithm (3); cf., (Garg & Khandekar, 2009).[1]

To the best of our knowledge, explicitly sparse Euclidean projections onto the simplex and hyperplane constraints have not been considered before. The closest work to ours is the paper (Kyrillidis & Cevher, 2012). In (Kyrillidis & Cevher, 2012), the authors propose an alternating projection approach in regression where the true vector is already sparse and within a convex norm-ball constraint. In contrast, we consider the problem of projecting an *arbitrary* given vector onto convex-based and sparse constraints *jointly*.

At the time of this submission, we become aware of (Pilanci et al., 2012), which considers cardinality regularized loss function minimization subject to simplex constraints. Their convexified approach relies on solving a lower-bound to the objective function and has $\mathcal{O}(p^4)$ complexity, which is not scalable. We also note that *regularizing* with the cardinality constraints is generally easier: e.g., our projectors become simpler.

---

[1] Surprisingly, a recent analysis of this algorithm along with similar assumptions indicates that rigorous guarantees can be obtained for minimization of general loss functions other than the quadratic (Bahmani et al., 2011).

Notation: Plain and boldface lowercase letters represent scalars and vectors, resp. The i-th entry of a vector $\mathbf{w}$ is w_i, and $[w_i]_+ = \max(w_i, 0)$, while $\boldsymbol{\beta}^i$ is the model estimate at the i-th iteration of an algorithm. Given a set $\mathcal{S} \subseteq \mathcal{N} = \{1, \ldots, p\}$, the complement $\mathcal{S}^c$ is defined with respect to $\mathcal{N}$, and the cardinality is $|\mathcal{S}|$. The support set of $\mathbf{w}$ is $\mathrm{supp}(\mathbf{w}) = \{i : w_i \neq 0\}$. Given a vector $\mathbf{w} \in \mathbb{R}^p$, $\mathbf{w}_{\mathcal{S}}$ is the projection (in $\mathbb{R}^p$) of $\mathbf{w}$ onto S, i.e. $(\mathbf{w}_{\mathcal{S}})_{\mathcal{S}^c} = 0$, whereas $\mathbf{w}_{|\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$ is $\mathbf{w}$ limited to $\mathcal{S}$ entries. The all-ones column vector is $\mathbb{1}$, with dimensions apparent from the context. We define Σ_k as the set of all k-sparse subsets of $\mathcal{N}$, and we sometimes write $\boldsymbol{\beta} \in \Sigma_k$ to mean $\mathrm{supp}(\boldsymbol{\beta}) \in \Sigma_k$. The trace of a matrix $\mathbf{X}$ is written $\mathrm{tr}(\mathbf{X})$.

2. Preliminaries

Basic definitions: Without loss of generality, assume $\mathbf{w}$ is sorted in descending order, so w_1 is the largest element. We denote $\mathcal{P}_{\lambda+}$ for the (convex) Euclidean projector onto the standard simplex Δ_λ^+, and $\mathcal{P}_\lambda$ for its extension to Δ_λ. The (non-convex) Euclidean projector onto the set Σ_k is $\mathcal{P}_{\Sigma_k}$, which retains the k-largest in magnitude elements. In contrast to $\mathcal{P}_\lambda$, the projection $\mathcal{P}_{\Sigma_k}$ need not be unique.

Definition 2.1 (Operator $\mathcal{P}_{L_k}$). *We define $\mathcal{P}_{L_k}(\mathbf{w})$ as the operator that keeps the k-largest entries of $\mathbf{w}$ (not in magnitude) and sets the rest to zero. This operation can be computed in $\mathcal{O}(p\min(k, \log(p)))$-time.*

Definition 2.2 (Euclidean projection $\mathcal{P}_{\lambda+}$). *The projector onto the simplex is given by*

$$(\mathcal{P}_{\lambda+}(w))_i = [w_i - \tau]_+, \text{ where } \tau := \frac{1}{\rho}\left(\sum_{i=1}^\rho w_i - \lambda\right)$$

for $\rho := \max\{j : w_j > \frac{1}{j}(\sum_{i=1}^j w_i - \lambda)\}$.

Definition 2.3 (Euclidean projection $\mathcal{P}_\lambda$). *The projector onto the extended simplex is given by*

$$(\mathcal{P}_\lambda(w))_i = w_i - \tau, \quad \text{where} \quad \tau = \frac{1}{p}\left(\sum_{i=1}^p w_i - \lambda\right).$$

Definition 2.4 (Restricted isometry property (RIP) (Candès et al., 2006)). *A linear operator $\boldsymbol{\mathcal{A}} : \mathcal{R}^p \to \mathcal{R}^m$ satisfies the k-RIP with constant $\delta_k \in (0,1)$ if*

$$1 - \delta_k \leq \|\boldsymbol{\mathcal{A}}(\boldsymbol{\beta})\|_2^2 / \|\boldsymbol{\beta}\|_2^2 \leq 1 + \delta_k, \ \forall \boldsymbol{\beta} \in \Sigma_k. \quad (4)$$

Guarantees of the gradient scheme (3): Let $\mathbf{y} = \boldsymbol{\mathcal{A}}(\boldsymbol{\beta}^\star) + \boldsymbol{\varepsilon} \in \mathbb{R}^m$, $(m \ll p)$, be a generative model where $\boldsymbol{\varepsilon}$ is an additive perturbation term and $\boldsymbol{\beta}^\star$ is the k-sparse true model generating $\mathbf{y}$. If the RIP assumption (4) is satisfied, then the projected gradient

descent algorithm in (3) features the following invariant on the objective (Garg & Khandekar, 2009):

$$f(\boldsymbol{\beta}^{i+1}) \leq \frac{2\delta_{2k}}{1 - \delta_{2k}} f(\boldsymbol{\beta}^i) + c_1 \|\boldsymbol{\varepsilon}\|_2, \quad (5)$$

for $c_1 > 0$ and stepsize $\mu^i = \frac{1}{1+\delta_{2k}}$. Hence, for $\delta_{2k} < 1/3$, the iterations of the algorithm are contractive and (3) obtains a good approximation on the loss function. In addition, (Foucart, 2010) shows that we can guarantee approximation on the true model via

$$\|\boldsymbol{\beta}^{i+1} - \boldsymbol{\beta}^\star\|_2 \leq 2\delta_{3k}\|\boldsymbol{\beta}^i - \boldsymbol{\beta}^\star\|_2 + c_2\|\boldsymbol{\varepsilon}\|_2, \quad (6)$$

for $c_1 > 0$ and $\mu^i = 1$. Similarly, when $\delta_{3k} < 1/2$, the iterations of the algorithm are contractive. Different step size μ^i strategies result in different guarantees; c.f., (Foucart, 2010; Garg & Khandekar, 2009; Kyrillidis & Cevher, 2011) for a more detailed discussion. Note that to satisfy a given RIP constant δ, random matrices with sub-Gaussian entries require $m = \mathcal{O}\left(k \log(p/k)/\delta^2\right)$. In low rank matrix cases, similar RIP conditions for (3) can be derived with approximation guarantees; cf., (Meka et al., 2010).

3. Underlying discrete problems

Let $\boldsymbol{\beta}^\star$ be a projection of $\mathbf{w}$ onto $\Sigma_k \cap \Delta_\lambda^+$ or $\Sigma_k \cap \Delta_\lambda$. We now make the following elementary observation:

Remark 1. *The Problem 1 and 2 statements can be equivalently transformed into the following nested minimization problem: $\{\mathcal{S}^\star, \boldsymbol{\beta}^\star\} =$*

$$\underset{\mathcal{S}:\mathcal{S}\in\Sigma_k}{\mathrm{argmin}} \left[\underset{\substack{\boldsymbol{\beta}:\boldsymbol{\beta}_\mathcal{S}\in\Delta_\lambda^+ \text{ or } \Delta_\lambda, \\ \boldsymbol{\beta}_{\mathcal{S}^c}=0}}{\mathrm{argmin}} \|(\boldsymbol{\beta} - \mathbf{w})_{|\mathcal{S}}\|_2^2 + \|\mathbf{w}_{|\mathcal{S}^c}\|_2^2 \right],$$

where $\mathrm{supp}(\boldsymbol{\beta}^\star) = \mathcal{S}^\star$ and $\boldsymbol{\beta}^\star \in \Delta_\lambda^+$ or Δ_λ.

Therefore, given $\mathcal{S}^\star = \mathrm{supp}(\boldsymbol{\beta}^\star)$, we can find $\boldsymbol{\beta}^\star$ by projecting $\mathbf{w}_{\mathcal{S}^\star}$ onto Δ_λ^+ or Δ_λ within the k-dimensional space. Thus, the difficulty is finding $\mathcal{S}^\star$. Hence, we split the problem into the task of finding the support and then finding the values on the support.

3.1. Problem 1

Given any support $\mathcal{S}$, the unique corresponding estimator is $\widehat{\boldsymbol{\beta}}_{|\mathcal{S}} = \mathcal{P}_{\lambda+}(\mathbf{w}_{|\mathcal{S}})$. We conclude that $\boldsymbol{\beta}^\star$ satisfies $\boldsymbol{\beta}^\star_{(\mathcal{S}^\star)^c} = 0$ and $\boldsymbol{\beta}^\star_{|\mathcal{S}^\star} = \mathcal{P}_{\lambda+}(\mathbf{w}_{|\mathcal{S}^\star})$, and

$$\mathcal{S}^\star \in \underset{\mathcal{S}:\mathcal{S}\in\Sigma_k}{\mathrm{argmin}} \|\mathcal{P}_{\lambda+}(\mathbf{w}_{|\mathcal{S}}) - \mathbf{w}_{|\mathcal{S}}\|_2^2 + \|\mathbf{w}_{|\mathcal{S}^c}\|_2^2$$

$$= \underset{\mathcal{S}:\mathcal{S}\in\Sigma_k}{\mathrm{argmax}} F_+(\mathcal{S}) \quad (7)$$

where $F_+(\mathcal{S}) := \sum_{i\in\mathcal{S}} \left(w_i^2 - ((\mathcal{P}_{\lambda+}(\mathbf{w}_{|\mathcal{S}}))_i - w_i)^2\right)$.

Algorithm 1 GSSP

1: $\mathcal{S}^\star = \mathrm{supp}(\mathcal{P}_{L_k}(\mathbf{w}))$ {Select support}
2: $\boldsymbol{\beta}_{|\mathcal{S}^\star} = \mathcal{P}_{\lambda^+}(\mathbf{w}_{|\mathcal{S}^\star}), \boldsymbol{\beta}_{|(\mathcal{S}^\star)^c} = 0$ {Final projection}

Algorithm 2 GSHP

1: $\ell = 1$, $\mathcal{S} = j$, $j \in \arg\max_i [\lambda w_i]$ {Initialize}
2: Repeat: $\ell \leftarrow \ell + 1$, $\mathcal{S} \leftarrow \mathcal{S} \cup j$, where

    $j \in \arg\max_{i \in \mathcal{N} \setminus \mathcal{S}} \left| w_i - \frac{\sum_{j \in \mathcal{S}} w_j - \lambda}{\ell - 1} \right|$ {Grow}
3: Until $\ell = k$, set $\mathcal{S}^\star \leftarrow \mathcal{S}$ {Terminate}
4: $\boldsymbol{\beta}_{|\mathcal{S}^\star} = \mathcal{P}_\lambda(\mathbf{w}_{|\mathcal{S}^\star})$, $\boldsymbol{\beta}_{|(\mathcal{S}^\star)^c} = 0$ {Final projection}

This set function can be simplified to

$$F_+(\mathcal{S}) = \sum_{i \in \mathcal{S}} (w_i^2 - \tau^2), \tag{8}$$

where τ (which depends on S) is as in Lemma 1.

Lemma 1. *Let $\boldsymbol{\beta} = \mathcal{P}_{\lambda^+}(\mathbf{w})$ where $\beta_i = [w_i - \tau]_+$. Then, $w_i \geq \tau$ for all $i \in \mathcal{S} = \mathrm{supp}(\boldsymbol{\beta})$. Furthermore, $\tau = \frac{1}{|\mathcal{S}|} \left(\sum_{i \in \mathcal{S}} w_i - \lambda \right)$.*

Proof. Directly from the definition of τ in Definition 2.2. The intuition is quite simple: the "threshold" τ should be smaller than the smallest entry in the selected support, or we unnecessarily shrink the coefficients that are larger without introducing any new support to the solution. Same arguments apply to inflating the coefficients to meet the simplex budget. $\square$

3.2. Problem 2

Similar to above, we conclude that $\boldsymbol{\beta}^\star$ satisfies $\boldsymbol{\beta}^\star_{|\mathcal{S}^\star} = \mathcal{P}_\lambda(\mathbf{w}_{|\mathcal{S}^\star})$ and $\boldsymbol{\beta}^\star_{|(\mathcal{S}^\star)^c} = 0$, where

$$\mathcal{S}^\star \in \underset{\mathcal{S}: \mathcal{S} \in \Sigma_k}{\arg\min} \|\mathbf{z} - \mathbf{w}\|_2 = \underset{\mathcal{S}: \mathcal{S} \in \Sigma_k}{\arg\max} F(\mathcal{S}) \tag{9}$$

where $\mathbf{z} \in \mathbb{R}^p$ with $\mathbf{z}_{|\mathcal{S}} = \mathcal{P}_\lambda(\mathbf{w}_{|\mathcal{S}})$ and $\mathbf{z}_{|\mathcal{S}^c} = 0$ and $F(\mathcal{S}) := \left(\sum_{i \in \mathcal{S}} w_i^2 \right) - \frac{1}{|\mathcal{S}|} (\sum_{i \in \mathcal{S}} w_i - \lambda)^2$.

4. Sparse projections onto Δ_λ^+ and Δ_λ

Algorithm 1 below suggests an obvious greedy approach for the projection onto $\Sigma_k \cap \Delta_\lambda^+$. We select the set $\mathcal{S}^\star$ by naively projecting $\mathbf{w}$ as $\mathcal{P}_{L_k}(\mathbf{w})$. Remarkably, this gives the correct support set for Problem 1, as we prove in Section 5.1. We call this algorithm the greedy selector and simplex projector (GSSP). The overall complexity of GSSP is dominated by the sort operation in p-dimensions.

Unfortunately, the GSSP fails for Problem 2. As a result, we propose Algorithm 2 for the $\Sigma_k \cap \Delta_\lambda$ case

which is non-obvious. The algorithm first selects the index of the largest element that has the same sign as λ. It then grows the index set one at a time by finding the farthest element from the current mean, as adjusted by lambda. Surprisingly, the algorithm finds the correct support set, as we prove in Section 5.2. We call this algorithm the greedy selector and hyperplane projector (GSHP), whose overall complexity is similar to GSSP.

5. Main results

Remark 2. *When the symbol $\mathcal{S}$ is used as $\mathcal{S} = \mathrm{supp}(\bar{\boldsymbol{\beta}})$ for any $\bar{\boldsymbol{\beta}}$, then if $|\mathcal{S}| < k$, we enlarge $\mathcal{S}$ until it has k elements by taking the first $k - |\mathcal{S}|$ elements that are not already in $\mathcal{S}$, and setting $\bar{\boldsymbol{\beta}} = 0$ on these elements. The lexicographic approach is used to break ties when there are multiple solutions.*

5.1. Correctness of GSSP

Theorem 1. *Algorithm 1 exactly solves Problem 1.*

Proof. Intuitively, the k-largest coordinates should be in the solution. To see this, suppose that $\mathbf{u}$ is the projection of $\mathbf{w}$. Let w_i be one of the k-(most positive) coordinates of $\mathbf{w}$ and $u_i = 0$. Also, let $w_j < w_i$, $i \neq j$ such that $u_j > 0$. We can then construct a new vector $\mathbf{u}'$ where $u'_j = u_i = 0$ and $u'_i = u_j$. Therefore, $\mathbf{u}'$ satisfies the constraints, and it is closer to $\mathbf{w}$, i.e., $\|\mathbf{w} - \mathbf{u}\|_2^2 - \|\mathbf{w} - \mathbf{u}'\|_2^2 = 2u_j(w_i - w_j) > 0$. Hence, $\mathbf{u}$ cannot be the projection.

To be complete in the proof, we also need to show that the cardinality k solutions are as good as any other solution with cardinality less than k. Suppose there exists a solution $\mathbf{u}$ with support $|\mathcal{S}| < k$. Now add *any* elements to $\mathcal{S}$ to form $\hat{\mathcal{S}}$ with size k. Then consider $\mathbf{w}$ restricted to $\hat{\mathcal{S}}$, and let $\hat{\mathbf{u}}$ be its projection onto the simplex. Because this is a projection, $\|\hat{\mathbf{u}}_{|\hat{\mathcal{S}}} - \mathbf{w}_{|\hat{\mathcal{S}}}\| \leq \|\mathbf{u}_{|\hat{\mathcal{S}}} - \mathbf{w}_{|\hat{\mathcal{S}}}\|$, hence $\|\hat{\mathbf{u}} - \mathbf{w}\| \leq \|\mathbf{u} - \mathbf{w}\|$. $\square$

5.2. Correctness of GSHP

Theorem 2. *Algorithm 2 exactly solves Problem 2.*

Proof. To motivate the support selection of GSHP, we now identify a key relation that holds for any $\mathbf{b} \in \mathbb{R}^k$:

$$\sum_{i=1}^k b_i^2 - \frac{\left(\sum_{i=1}^k b_i - \lambda \right)^2}{k} =$$

$$\lambda(2b_1 - \lambda) + \sum_{j=2}^k \frac{j-1}{j} \left(b_j - \frac{\sum_{i=1}^{j-1} b_i - \lambda}{j-1} \right)^2. \tag{10}$$

By its left-hand side, this relation is invariant under permutation of $\mathbf{b}$. Moreover, the summands in the sum

over k are certainly non-negative for $k \geq 2$, so without loss of generality the solution sparsity of the original problem is $\|\boldsymbol{\beta}^\star\|_0 = k$. For $k = 1$, F is maximized by picking an index i that maximizes λw_i, which is what the algorithm does.

For the sake of clarity (and space), we first describe the proof of the case $k \geq 2$ for $\lambda = 0$ and then explain how it generalizes for $\lambda \neq 0$. In the sequel, let us use the shortcut $\mathrm{avg}(S) = \frac{1}{|S|} \sum_{j \in S} w_j$.

Let $\mathcal{S}$ be an optimal solution index set and let I be the result computed by the algorithm. For a proof (of the case $k \geq 2, \lambda = 0$) by contradiction, assume that I and $\mathcal{S}$ differ. Let e be the first element of $I \backslash \mathcal{S}$ in the order of insertion into I by the algorithm. Let e' be the element of $\mathcal{S} \backslash I_0$ that lies closest to e. Without loss of generality, we may assume that $w_e \neq w_{e'}$, otherwise we could have chosen $\mathcal{S} \backslash \{e'\} \cup \{e\}$ rather than $\mathcal{S}$ as solution in the first place. Let $I_0 \subseteq I \cap \mathcal{S}$ be the indices added to I by the algorithm before e. Assume that I_0 is nonempty. We will later see how to ensure this.

Let $a := \mathrm{avg}(I_0)$ and $a' := \mathrm{avg}(\mathcal{S} \backslash \{e'\})$. There are three ways in which w_e, $w_{e'}$ and a' can be ordered relative to each other:

1. e' lies between e and a', thus $|w_{e'} - a'| < |w_e - a'|$ since $w_e \neq w_{e'}$.

2. a' lies between e and e'. But then, since there are no elements of $\mathcal{S}$ between e and e', $\mathcal{S} \backslash I_0$ moves the average a' beyond a away from e towards e', so $|w_{e'} - a'| < |w_{e'} - a|$ and $|w_e - a| < |w_e - a'|$. But we know that $|w_{e'} - a| < |w_e - a|$ since $e = \mathrm{argmax}_{i \in I_0} |w_i - a|$ by the choice of the greedy algorithm and $w_e \neq w_{e'}$. Thus $|w_{e'} - a'| < |w_e - a'|$.

3. $|w_e - a'| < |w_{e'} - a'|$, i.e., e lies between a' and e'. But this case is impossible: compared to a, a' averages over additional values that are closer to a than e, and e' is one of them. So a' must be on the same side as e' relative to e, not the opposite side.

So $|w_{e'} - a'| < |w_e - a'|$ is assured in all cases. Note in particular that if $|S| \geq 1$, $|w_i - \mathrm{avg}(S)| \; \theta \; |w_j - \mathrm{avg}(S)|$, then

$$F(S \cup \{i\}) = F(S) + \frac{k-1}{k}\left(w_i - \mathrm{avg}(S)\right)^2$$
$$\theta \; F(S) + \frac{k-1}{k}\left(w_j - \mathrm{avg}(S)\right)^2$$
$$= F(S \cup \{j\}), \qquad (11)$$

where θ is either '=' or '<'. By inequality (11), $F(\mathcal{S}) < F((\mathcal{S} \backslash \{e'\}) \cup \{e\})$. But this means that $\mathcal{S}$ is not a solution: contradiction.

We have assumed that I_0 is nonempty; this is ensured because any solution $\mathcal{S}$ must contain at least an index $i \in \mathrm{argmax}_j w_j$. Otherwise, we could replace a maximal index w.r.t. w in $\mathcal{S}$ by this i and get, by (11), a larger F value. This would be a contradiction with our assumption that $\mathcal{S}$ is a solution. Note that this maximal index is also picked (first) by the algorithm. This completes the proof for the case $\lambda = 0$. Let us now consider the general case where λ is unrestricted.

We reduce the general problem to the case that $\lambda = 0$. Let us write $F_{w,\lambda}$ to make the parameters w and λ explicit when talking of F. Let $w'_{i*} := w_{i*} - \lambda$ for one i^* for which λw_{i*} is maximal, and let $w'_i := w_i$ for all other i. We use the fact that, by the definition of F,

$$F_{w,\lambda}(\mathcal{S}) = 2\lambda w'_{i*} + \lambda^2 + F_{w',0}(\mathcal{S})$$

when $\mathcal{S}$ contains such an element $i^* \in \mathrm{argmax}_j(\lambda w_j)$. Clearly, i^* is an extremal element w.r.t. w and w_{i*} has maximum distance from $-\lambda$, so

$$i^* \in \mathrm{argmax}_j \left| w_j - \frac{\sum_{i \neq j} w_i - \lambda}{j - 1} \right|.$$

By (10), i^* must be in the optimal solution for $F_{w,\lambda}$. Also, $F_{w',0}(\mathcal{S})$ and $2\lambda w'_{i*} + \lambda^2 + F_{w',0}(\mathcal{S})$ are maximized by the same index sets $\mathcal{S}$ when $i^* \in \mathcal{S}$ is required. Thus,

$$\mathrm{argmax}_{\mathcal{S}} F_{w,\lambda}(\mathcal{S}) = \mathrm{argmax}_{\mathcal{S}: j \in \mathcal{S}} F_{w',0}(\mathcal{S}).$$

Now observe that our previous proof for the case $\lambda = 0$ also works if one adds a constraint that one or more indices be part of the solution: If the algorithm computes these elements as part of its result I, they are in $I_0 = I \cap \mathcal{S}$. But this is what the algorithm does on input (w, λ); it chooses i^* in its first step and then proceeds as if maximizing $F_{w',0}$. Thus we have established the algorithm's correctness. $\qquad \square$

6. Application: Quantum tomography

Problem: In quantum tomography (QT), we aim to learn a *density matrix* $\mathbf{X}^\star \in \mathbb{C}^{d \times d}$, which is Hermitian (i.e., $(\mathbf{X}^\star)^H = \mathbf{X}^\star$), positive semi-definite (i.e., $\mathbf{X}^\star \succeq 0$) and has $\mathrm{rank}(\mathbf{X}^\star) = r$ and $\mathrm{tr}(\mathbf{X}^\star) = 1$. The QT measurements are $\mathbf{y} = \mathcal{A}(\mathbf{X}^\star) + \boldsymbol{\eta}$, where $(\mathcal{A}(\mathbf{X}^\star))_i = \mathrm{tr}(\mathbf{E}_i \mathbf{X}^\star) + \eta_i$, and η_i is zero-mean Gaussian. The operators $\mathbf{E}_i$'s are the tensor product of the 2×2 Pauli matrices (Liu, 2011).

Recently, (Liu, 2011) showed that almost all such tensor constructions of $\mathcal{O}(rd \log^6 d)$ Pauli measurements satisfy the so-called rank-r restricted isometry property (RIP) for all $\mathbf{X} \in$

$\{\mathbf{X} \in \mathbb{C}^{d \times d} : \mathbf{X} \succeq 0, \text{rank}(\mathbf{X}) \leq r, \|\mathbf{X}\|_* \leq \sqrt{r}\|\mathbf{X}\|_F\}$:

$$(1 - \delta_r) \|\mathbf{X}\|_F^2 \leq \|\mathcal{A}(\mathbf{X})\|_F^2 \leq (1 + \delta_r) \|\mathbf{X}\|_F^2, \quad (12)$$

where $\|\cdot\|_*$ is the nuclear norm (i.e., the sum of singular values), which reduces to $\text{tr}(\mathbf{X})$ since $\mathbf{X} \succeq 0$. This key observation enables us to leverage the recent theoretical and algorithmic advances in low-rank matrix recovery from a few affine measurements.

The standard matrix-completion based approach to recover $\mathbf{X}^\star$ from $\mathbf{y}$ is the following convex relaxation:

$$\underset{\mathbf{X} \succeq 0}{\text{minimize}} \, \|\mathcal{A}(\mathbf{X}) - \mathbf{y}\|_F^2 + \lambda \|\mathbf{X}\|_*. \quad (13)$$

This convex approach is both tractable and amenable to theoretical analysis (Gross et al., 2010; Liu, 2011). As a result, we can provably reduce the number of samples m from $\mathcal{O}(d^2)$ to $\tilde{\mathcal{O}}(rd)$ (Liu, 2011).

Unfortunately, this convex approach fails to account for the physical constraint $\|\mathbf{X}\|_* = 1$. To overcome this difficulty, the relaxation parameter λ is tuned to obtain solutions with the desired rank followed by normalization to heuristically meet the trace constraint.

In this section, we demonstrate that one can do significantly better via the non-convex algorithm based on (3). A key ingredient then is the following projection:

$$\widehat{\mathbf{B}} \in \underset{\mathbf{B} \succeq 0}{\text{argmin}} \, \|\mathbf{B} - \mathbf{W}\|_F^2 \text{ s.t. } \text{rank}(\mathbf{B}) = r, \, \text{tr}(\mathbf{B}) = 1,$$
$$(14)$$

for a given *Hermitian* matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$. Since the RIP assumption holds here, we can obtain rigorous guarantees based on a similar analysis to (Foucart, 2010; Garg & Khandekar, 2009; Meka et al., 2010).

To obtain the solution, we compute the eigenvalue decomposition $\mathbf{W} = \mathbf{U}\Lambda_{\mathbf{W}}\mathbf{U}^H$ and then use the unitary invariance of the problem to solve $\mathbf{D}^\star \in \text{argmin}_{\mathbf{D}} \|\mathbf{D} - \Lambda_{\mathbf{W}}\|_F$ subject to $\|\mathbf{D}\|_* \leq 1$ and $\text{rank}(\mathbf{D}) \leq r$, and from $\mathbf{D}^\star$ form $\mathbf{U}\mathbf{D}^\star\mathbf{U}^H$ to obtain a solution. In fact, we can constrain $\mathbf{D}$ to be diagonal, and thus reduce the matrix problem to the vector version for $\mathbf{D} = \text{diag}(\mathbf{d})$, where the projector in Problem 1 applies. This reduction follows from the well-known result:

Proposition 6.1 ((Mirsky, 1960)). *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ and $q = \min\{m, n\}$. Let $\sigma_i(\mathbf{A})$ be the singular values of $\mathbf{A}$ in descending order (similarly for $\mathbf{B}$). Then,*

$$\sum_{i=1}^{q} \left(\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})\right)^2 \leq \|\mathbf{A} - \mathbf{B}\|_F^2.$$

Equation (14) has a solution if $r \geq 1$ since the constraint set is non-empty and compact (Weierstrass's theorem). As the vector reduction achieves the lower bound, it is an optimal projection.
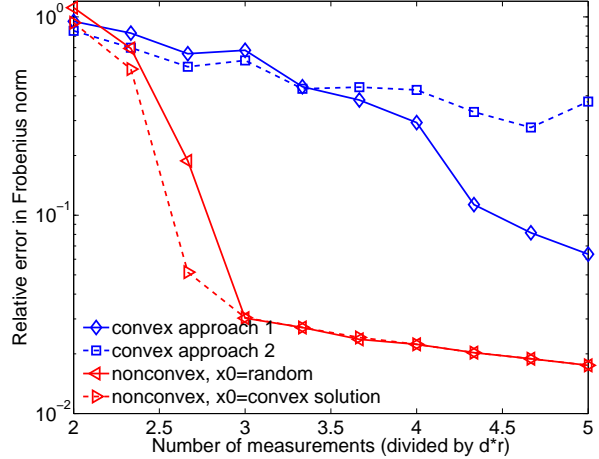

Figure 1. Quantum tomography with 8 qubits and 30 dB SNR: Each point is the median over 10 random realizations. Convex approach 1 refers to (13) and approach 2 is (15).
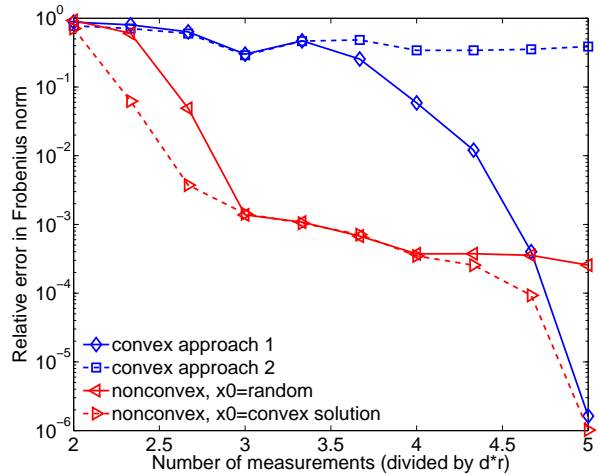

Figure 2. Same as Figure 1 but with 7 qubits, no noise.

Numerical experiments: We numerically demonstrate that the ability to project onto trace and rank constraints jointly can radically improve recovery even with the simple gradient descent algorithm as in (3). We follow the same approach in (Gross et al., 2010): we generate random Pauli measurements and add white noise. The experiments that follow use a 8 qubit system ($d = 2^8$) with noise SNR at 30 dB (so the absolute noise level changes depending on the number of measurements), and a 7 qubit noiseless system.

The measurements are generated using a random real-valued matrix $\mathbf{X}^\star$ with rank 2, although the algorithms also work with complex-valued matrices. A $d \times d$ rank r real-valued matrix has $dr - r(r-1)/2 \approx dr$ degrees of freedom, hence we need at least $2dr$ number of measurements to recover $\mathbf{X}^\star$ from noiseless measurements

(due to the null-space of the linear map). To test the various approaches, we vary the number of measurements between $2dr$ and $5dr$. We assume r is known, though other computational experience suggests that estimates of r return good answers as well.

The convex problem (13) depends on a parameter λ. We solve the problem for different λ in a bracketing search until we find the first λ that provides a solution with numerical rank r. Like (Flammia et al., 2012), we normalize the final estimate to ensure the trace is 1. Additionally, we test the following convex approach:

$$\underset{\mathbf{X} \succeq 0, \|\mathbf{X}\|_* \leq 1}{\text{minimize}} \|\mathcal{A}(\mathbf{X}) - \mathbf{y}\|_F^2. \qquad (15)$$

Compared to (13), no parameters are needed since we exploit prior knowledge of the trace, but there is no guarantee on the rank. Both convex approaches can be solved with proximal gradient descent; we use the TFOCS package (Becker et al., 2011) since it uses a sophisticated line search and Nesterov acceleration.

To illustrate the power of the combinatorial projections, we solve the following non-convex formulation:

$$\underset{\mathbf{X} \succeq 0, \|\mathbf{X}\|_* \leq 1, \text{rank}(\mathbf{X})=r}{\text{minimize}} \|\mathcal{A}(\mathbf{X}) - \mathbf{y}\|_F^2. \qquad (16)$$

Within the projected gradient algorithm (3), we use the GSSP algorithm as described above. The stepsize is $\mu^i = 3/\|\mathcal{A}\|^2$ where $\| \cdot \|$ is the operator norm; we can also apply Nesterov acceleration to speed convergence, but we use (3) for simplicity. Due to the non-convexity, the algorithm depends on the starting value $\mathbf{X}_0$. We try two strategies: (i) random initialization, and (ii) initializing with the solution from (15). Both initializations often lead to the same stationary point.

Figure 1 shows the relative error $\|\mathbf{X} - \mathbf{X}^\star\|_F/\|\mathbf{X}^\star\|_F$ of the different approaches. All approaches do poorly when there are only $2dr$ measurements since this is near the noiseless information-theoretic limit. For higher numbers of measurements, the non-convex approach substantially outperforms both convex approaches. For $2.4dr$ measurements, it helps to start $\mathbf{X}_0$ with the convex solution, but otherwise the two non-convex approaches are nearly identical.

Between the two convex solutions, (13) outperforms (15) since it tunes λ to achieve a rank r solution. Neither convex approach is competitive with the non-convex approaches since they do not take advantage of the prior knowledge on trace and rank.

Figure 2 shows more results on a 7 qubit problem without noise. Again, the non-convex approach gives better results, particularly when there are fewer measurements. As expected, both approaches approach perfect recovery as the number of measurements increases.

Approach	mean time	standard deviation
convex	0.294 s.	0.030 s.
non-convex	0.192 s.	0.019 s.

Table 1. Time per iteration of convex and non-convex approaches for quantum state tomography with 8 qubits.

Here we highlight another key benefit of the non-convex approach: since the number of eigenvectors needed in the partial eigenvalue decomposition is at most r, it is quite scalable. In general, the convex approach has intermediate iterates which require eigenvalue decompositions close to the full dimension, especially during the first few iterations. Table 1 shows average time per iteration for the convex and non-convex approach (overall time is more complicated, since the number of iterations depends on linesearch and other factors). Even using Matlab's dense eigenvalue solver `eig`, the iterations of the non-convex approach are faster; problems that used an iterative Krylov subspace solver would show an even larger discrepancy.[2]

7. Application: Measure learning

Problem: We study the kernel density learning setting: Let $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)} \in \mathbb{R}^p$ be an n-size corpus of p-dimensional samples, drawn from an unknown probability density function (pdf) $\mu(\mathbf{x})$. Here, we will form an estimator $\hat{\mu}(\mathbf{x}) := \sum_{i=1}^n \beta_i \kappa_\sigma(\mathbf{x}, \mathbf{x}^{(i)})$, where $\kappa_\sigma(\mathbf{x}, \mathbf{y})$ is a Gaussian kernel with parameter σ. Let us choose $\hat{\mu}(\mathbf{x})$ to minimize the integrated squared error criterion: $\text{ISE} = \mathbb{E}\|\hat{\mu}(\mathbf{x}) - \mu(\mathbf{x})\|_2^2$. As a result, we can introduce a density learning problem as estimating a weight vector $\boldsymbol{\beta}^\star \in \Delta_1^+$. The objective can then be written as follows (Bunea et al., 2010; Kim, 1995)

$$\boldsymbol{\beta}^\star \in \underset{\boldsymbol{\beta} \in \Delta_1^+}{\text{argmin}} \left\{ \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} - \mathbf{c}^T \boldsymbol{\beta} \right\}, \qquad (17)$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ with $\Sigma_{ij} = \kappa_{\sqrt{2}\sigma}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$, and

$$c_i = \frac{1}{n-1} \sum_{j \neq i} \kappa_\sigma(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}), \ \forall i, j. \qquad (18)$$

While the combination of the $-\mathbf{c}^T \boldsymbol{\beta}$ term and the non-negativity constraint induces some sparsity, it may not be enough. To avoid overfitting or obtain interpretable results, one might control the level of solution sparsity (Bunea et al., 2010). In this context, we extend (17) to include cardinality constraints, i.e. $\boldsymbol{\beta}^\star \in \Delta_1^+ \cap \Sigma_k$.

---

[2]Quantum state tomography does not easily benefit from iterative eigenvalue solvers, since the range of $\mathcal{A}^*$ is not sparse.
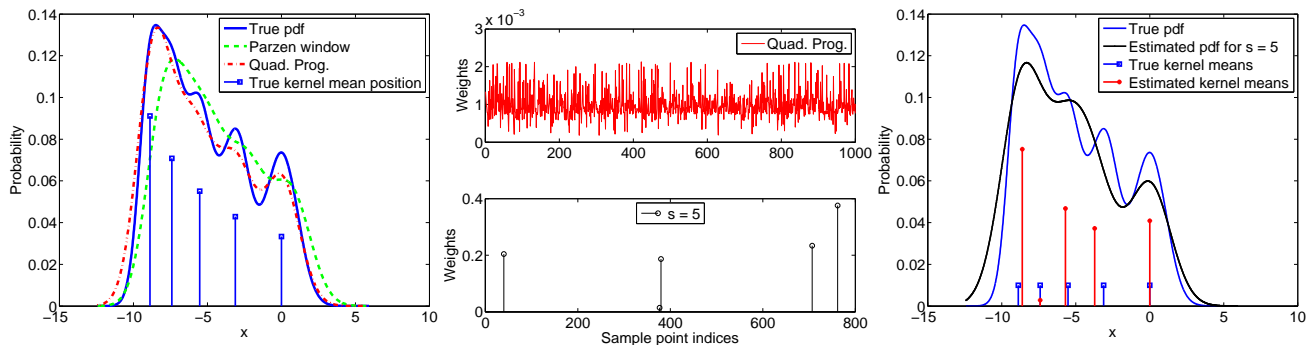
Figure 3. Density estimation results using the Parzen method (left), the quadratic program (17) (left and middle-top), and our approach (middle-bottom and right).
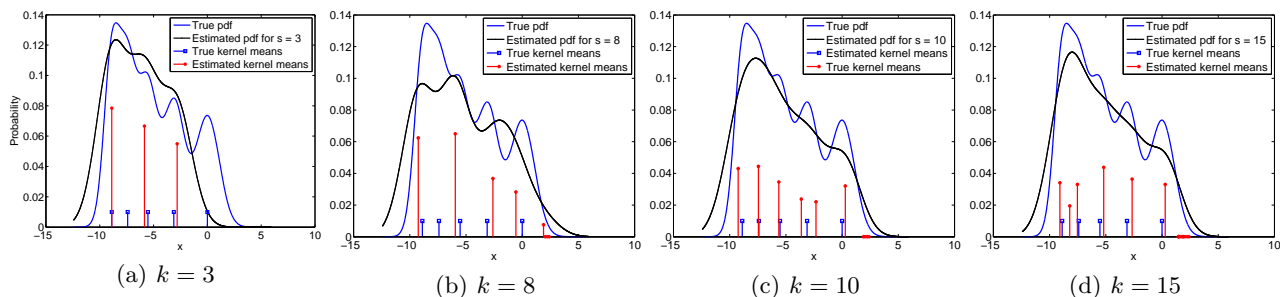

(a) $k = 3$ (b) $k = 8$ (c) $k = 10$ (d) $k = 15$

Figure 4. Estimation results for different k: Red spikes depict the estimated kernel means as well as the their relative contribution to the Gaussian mixture. As k increases, the additional nonzero coefficients in $\boldsymbol{\beta}^\star$ tend to have small weights.

Numerical experiments: We consider the following Gaussian mixture: $\mu(x) = \frac{1}{5} \sum_{i=1}^{5} \kappa_{\sigma_i}(\boldsymbol{\beta}_i, x)$, where $\sigma_i = (7/9)^i$ and $\boldsymbol{\beta}_i = 14(\sigma_i - 1)$. A sample of 1000 points is drawn from $\mu(x)$. We compare the density estimation performance of: (i) the Parzen method (Parzen, 1962), (ii) the quadratic programming formulation in (17), and (iii) our cardinality-constrained version of (17) using GSSP. While $\mu(x)$ is constructed by kernels with various widths, we assume a constant width during the kernel estimation. In practice, the width is not known *a priori* but can be found using cross-validation techniques; for simplicity, we assume kernels with width $\sigma = 1$.

Figure 3(left) depicts the true pdf and the estimated densities using the Parzen method and the quadratic programming approach. Moreover, the figure also includes a scaled plot of $1/\sigma_i$, indicating the height of the individual Gaussian mixtures. By default, the Parzen window method estimation interpolates 1000 Gaussian kernels with centers around the sampled points to compute the estimate $\hat{\mu}(x)$; unfortunately, neither the quadratic programming approach (as Figure 3 (middle-top) illustrates) nor the Parzen window estimator results are easily *interpretable* even though both approaches provide a good approximation of the true pdf.

Using our cardinality-constrained approach, we can significantly enhance the interpretability. This is because in the sparsity-constrained approach, we can control the number of estimated Gaussian components. Hence, if the model order is known *a priori*, the non-convex approach can be extremely useful.

To see this, we first show the coefficient profile of the sparsity based approach for $k = 5$ in Figure 3 (middle-bottom). Figure 3 (right) shows the estimated pdf for $k = 5$ along with the positions of weight coefficients obtained by our sparsity enforcing approach. Note that most of the weights obtained concentrate around the true means, fully exploiting our prior information about the ingredients of $\mu(x)$—this happens with rather high frequency in the experiments we conducted. Figure 4 illustrates further estimated pdf's using our approach for various k. Surprisingly, the resulting solutions are still approximately 5-sparse even if $k > 5$, as the over-estimated coefficients are extremely small, and hence the sparse estimator is reasonably robust to inaccurate estimates of k.

8. Application: Portfolio optimization

Problem: Given a sample covariance matrix $\boldsymbol{\Sigma}$ and expected mean $\boldsymbol{\mu}$, the return-adjusted Markowitz mean-variance (MV) framework selects a portfolio $\boldsymbol{\beta}^\star$ such that $\boldsymbol{\beta}^\star \in \text{argmin}_{\boldsymbol{\beta} \in \Delta_1^+} \left\{ \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} - \tau \boldsymbol{\mu}^T \boldsymbol{\beta} \right\}$, where

Δ_1^+ encodes the normalized capital constraint, and τ trades off risk and return (Brodie et al., 2009; DeMiguel et al., 2009). The solution $\boldsymbol{\beta}^\star \in \Delta_1^+$ is the distribution of investments over the p available assets.

While such solutions construct portfolios from scratch, a more realistic scenario is to incrementally adjust an existing portfolio as the market changes. Due to costs per transaction, we can naturally introduce cardinality constraints. In mathematical terms, let $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$ be the current portfolio selection. Given $\bar{\boldsymbol{\beta}}$, we seek to adjust the current selection $\boldsymbol{\beta} = \bar{\boldsymbol{\beta}} + \boldsymbol{\delta_\beta}$ such that $\|\boldsymbol{\delta_\beta}\|_0 \le k$. This leads to the following optimization problem:

$$\boldsymbol{\delta_\beta^*} \in \operatorname*{argmin}_{\boldsymbol{\delta_\beta} \in \Sigma_k \cap \Delta_\lambda} (\bar{\boldsymbol{\beta}} + \boldsymbol{\delta_\beta})^T \Sigma (\bar{\boldsymbol{\beta}} + \boldsymbol{\delta_\beta}) - \tau \mu^T (\bar{\boldsymbol{\beta}} + \boldsymbol{\delta_\beta}),$$

where λ is the level of update, and k controls the transactions costs. During an update, $\lambda = 0$ would keep the portfolio value constant while $\lambda > 0$ would increase it.

Numerical experiments: To clearly highlight the impact of the non-convex projector, we create a synthetic portfolio update problem, where we know the solution. As in (Brodie et al., 2009), we cast this problem as a regression problem and synthetically generate $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star$ where $p = 1000$ such that $\boldsymbol{\beta}^\star \in \Delta_\lambda$ (λ is chosen randomly), and $\|\boldsymbol{\beta}^\star\|_0 = k$ for $k = 100$.

Since in general we do not expect RIP assumptions to hold in portfolio optimization, our goal here is to refine the sparse solution of a state-of-the-art convex solver via (3) in order to accommodate the strict sparsity and budget constraints. Hence, we first consider the basis pursuit criterion and solve it using SPGL1 (van den Berg & Friedlander, 2008):

$$\text{minimize } \|\boldsymbol{\beta}\|_1 \text{ s.t. } \begin{bmatrix} \mathbf{X} \\ \mathbb{1}^T/\sqrt{p} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \mathbf{y} \\ \lambda/\sqrt{p} \end{bmatrix}. \quad (19)$$

The normalization by $1/\sqrt{p}$ in the last equality gives the constraint matrix a better condition number, since otherwise it is too ill-conditioned for a first-order solver.

Almost none of the solutions to (19) return a k-sparse solution. Hence, we initialize (3) with the SPGL1 solution to meet the constraints. The update step in (3) uses the GSHP algorithm.

Figure 5 shows the resulting relative errors $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2/\|\boldsymbol{\beta}^\star\|_2$. We see that not only does (3) return a k-sparse solution, but that this solution is also closer to $\boldsymbol{\beta}^\star$, particularly when the sample size is small. As the sample size increases, the knowledge that $\boldsymbol{\beta}^\star$ is k-sparse makes up a smaller percentage of what we know about the signal, so the gap between (19) and (3) diminishes.

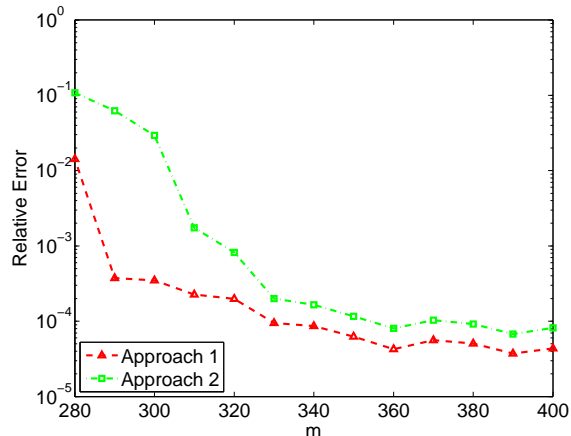*Figure 5.* Relative error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2/\|\boldsymbol{\beta}^\star\|_2$ comparison as a function of m: Approach 1 is the non-convex approach (3), and approach 2 is (19). Each point corresponds to the median value of 30 Monte-Carlo realizations.

9. Conclusions

While non-convexity in learning algorithms is undesirable according to conventional wisdom, avoiding it might be difficult in many problems. In this setting, we show how to efficiently obtain exact sparse projections onto the positive simplex and its hyperplane extension. We empirically demonstrate that our projectors provide substantial accuracy benefits in quantum tomography from fewer measurements and enable us to exploit prior non-convex knowledge in density learning. Moreover, we also illustrate that we can refine the solution of well-established state-of-the-art convex sparse recovery algorithms to enforce non-convex constraints in sparse portfolio updates. The quantum tomography example in particular illustrates that the non-convex solutions can be extremely useful; here, the non-convexity appears milder, since a fixed-rank matrix still has extra degrees of freedom from the choice of its eigenvectors.

10. Acknowledgements

References

Bahmani, S., Boufounos, P., and Raj, B. Greedy sparsity-constrained optimization. In *Signals, Systems and Computers (ASILOMAR), 2011 Confer-*

ence Record of the Forty Fifth Asilomar Conference on, pp. 1148–1152. IEEE, 2011.

Becker, Stephen, Candès, Emmanuel, and Grant, Michael. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, pp. 1–54, 2011. ISSN 1867-2949. URL http://dx.doi.org/10.1007/s12532-011-0029-5. 10.1007/s12532-011-0029-5.

Brodie, J., Daubechies, I., De Mol, C., Giannone, D., and Loris, I. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

Bunea, F., Tsybakov, A.B., Wegkamp, M.H., and Barbu, A. SPADES and mixture models. *The Annals of Statistics*, 38(4):2525–2558, 2010.

Candès, E., Romberg, J., and Tao, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Information Theory*, 52(2):489 – 509, February 2006.

DeMiguel, V., Garlappi, L., Nogales, F.J., and Uppal, R. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812, 2009.

Flammia, S.T., Gross, D., Liu, Y.K., and Eisert, J. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012. URL http://stacks.iop.org/1367-2630/14/i=9/a=095022.

Foucart, S. Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants. *Proceedings of the 13th International Conference on Approximation Theory*, 2010.

Garg, R. and Khandekar, R. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*. ACM, 2009.

Gross, D., Liu, Y.-K., Flammia, S. T., Becker, S., and Eisert, J. Quantum state tomography via compressed sensing. *Phys. Rev. Lett.*, 105(15):150401, Oct 2010. doi: 10.1103/PhysRevLett.105.150401.

Kim, D. *Least squares mixture decomposition estimation*. PhD thesis, 1995.

Kyrillidis, A. and Cevher, V. Recipes on hard thresholding methods. Dec. 2011.

Kyrillidis, A. and Cevher, V. Combinatorial selection and least absolute shrinkage via the Clash algorithm. In *IEEE International Symposium on Information Theory*, July 2012.

Liu, Y.K. Universal low-rank matrix recovery from Pauli measurements. In *NIPS*, pp. 1638–1646, 2011.

Meka, Raghu, Jain, Prateek, and Dhillon, Inderjit S. Guaranteed rank minimization via singular value projection. In *NIPS Workshop on Discrete Optimization in Machine Learning*, 2010.

Mirsky, L. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59, 1960.

Nemhauser, G.L. and Wolsey, L.A. *Integer and combinatorial optimization*, volume 18. Wiley New York, 1988.

Parzen, E. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

Pilanci, M., El Ghaoui, L., and Chandrasekaran, V. Recovery of sparse probability measures via convex programming. In *Advances in Neural Information Processing Systems 25*, pp. 2429–2437, 2012.

van den Berg, E. and Friedlander, M. P. Probing the Pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comput.*, 31(2):890–912, 2008.