Quantum Algorithms for the Pathwise Lasso

João F. Doriguello*[1], Debbie Lim†[1,2], Chi Seng Pun‡[3], Patrick Rebentrost§[1,4], and Tushar Vaidya¶[3]

[1]Centre for Quantum Technologies, National University of Singapore, Singapore
[2]Center for Quantum Computer Science, Faculty of Computing, University of Latvia, Latvia
[3]School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
[4]Department of Computer Science, National University of Singapore, Singapore

December 22, 2023

Abstract

We present a novel quantum high-dimensional linear regression algorithm with an ℓ_1-penalty based on the classical LARS (Least Angle Regression) pathwise algorithm. Similarly to available classical numerical algorithms for Lasso, our quantum algorithm provides the full regularisation path as the penalty term varies, but quadratically faster per iteration under specific conditions. A quadratic speedup on the number of features/predictors d is possible by using the simple quantum minimum-finding subroutine from Dürr and Høyer (arXiv'96) in order to obtain the joining time at each iteration. We then improve upon this simple quantum algorithm and obtain a quadratic speedup both in the number of features d and the number of observations n by using the recent approximate quantum minimum-finding subroutine from Chen and de Wolf (ICALP'23). In order to do so, we construct, as one of our main contributions, a quantum unitary based on quantum amplitude estimation to approximately compute the joining times to be searched over by the approximate quantum minimum-finding subroutine. Since the joining times are no longer exactly computed, it is no longer clear that the resulting approximate quantum algorithm obtains a good solution. As our second main contribution, we prove, via an approximate version of the KKT conditions and a duality gap, that the LARS algorithm (and therefore our quantum algorithm) is robust to errors. This means that it still outputs a path that minimises the Lasso cost function up to a small error if the joining times are only approximately computed. Finally, in the model where the observations are generated by an underlying linear model with an unknown coefficient vector, we prove bounds on the difference between the unknown coefficient vector and the approximate Lasso solution, which generalises known results about convergence rates in classical statistical learning theory analysis.

*joaofd@nus.edu.sg †limhueychih@gmail.com ‡cspun@ntu.edu.sg §patrick@comp.nus.edu.sg ¶tushar.vaidya@ntu.edu.sg

Contents

1 Introduction

One of the most important research topics that has gotten renewed attention by statisticians and the machine learning community is high-dimensional data analysis [BEM13, BvdG11, GG08, JT09, KKMR22, WM22, ZY06]. For linear regression, this means that the number of data points is less than the number of explanatory variables (features). In machine learning parlance, this is usually referred to as overparameterisation. Solutions to such problems are, however, ill defined. Having more variables to choose from is a double-edged sword. While it gives us freedom of choice, computational cost considerations favour choosing sparse models. Some sort of regularisation in the linear model, usually in the form of a penalty term, is needed to favour sparse models.

Let us consider the regression problem defined as follows. Assume to be given a fixed design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and a vector of observations $\mathbf{y} \in \mathbb{R}^n$. The setting is such that $d \geqslant n$, i.e., more features than observations are given. The un-regularized ℓ_2-regression problem is defined as $\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$. Generally, different norms could be used in the penalty term. Essentially,

the three popular sparse regression models reduce to three prototypes recalled below with ℓ_0, ℓ_1, ℓ_2-norms and tuning parameter λ:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0 \quad \text{for all } \lambda > 0 \quad \text{(best subset selection)},$$

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \quad \text{for all } \lambda > 0 \quad \text{(Lasso regression)},$$

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2 \quad \text{for all } \lambda > 0 \quad \text{(Ridge regression)}.$$

The easiest is Ridge regression which has an analytical solution and the problem is convex, but tends to give non-sparse solutions. The problem with the ℓ_0-norm is that it is not convex. The best subset selection is in general NP-hard [DK08]. Lasso regression with ℓ_1-penalty is a convex relaxation of this problem and will be our central object of study throughout this paper. Lasso stands for Least Absolute Shrinkage and Selection operator and has been studied extensively in the literature because of its convexity and parsimonious solutions. Generally, the optimal tuning parameter also has to be determined from the data and as such we would like an endogenous model that determines the optimal $\boldsymbol{\beta}$ along with a specific λ. The key point is that we want a path of solutions as λ varies [EHJT04, FHT10a, ZLZ18].

For large language models that use deep learning, overparameterisation is actually desired. In high-dimensional linear regression models, in contrast, parsimonious models are desired. The conventional statistical paradigm is to keep the number of features fixed and study the asymptotics as the sample size increases. The other paradigm is predictor selection given a large choice that fits into high-dimensional regression. Selecting which of the many variables available is part of model choice and techniques that help us in choosing the right model have significant value. Having a larger number of variables in statistical models makes it simpler to pick the right model and achieve improved predictive performance. Overparametrisation is a *virtue* and increases the likelihood of including the essential features in the model. Yet this should be balanced with regularisation [EHJT04, FL10, LTTT14]. Introducing an ℓ_1-penalty confers a number of advantages [CDS01, HTW15, Tib96]:

1. The ℓ_1-penalty provides models that can be interpreted in a simple way and thus have advantages compared to black boxes (deep neural networks) in terms of transparency and intelligibility [MSK$^+$19, RCC$^+$22];

2. If the original model generating the data is sparse, then an ℓ_1-penalty provides the right framework to recover the original signal;

3. Lasso selects the true model consistently even with noisy data [ZY06], and this remains an active area of research [MOK23];

4. As ℓ_1-penalties are convex, sparsity leads to computational advantages. For example, take two million features and only 200 observations, then estimating two million parameters becomes extremely difficult.

Lasso-type methods are used across disciplines and in many different fields where sparsity is a desired prerequisite: genetics, compressed sensing, and portfolio optimisation [CDS01, CRT06, FNW07, PL22, UGH09, WCH$^+$09]. Advocating Lasso-type algorithms does not preclude the fact that Lasso methods can be married with neural networks [TDK23], though it is not the scope of our work here. Lasso techniques are part of the repertoire of tools available to theoreticians and practitioners in machine learning and data analysis.

1.1 Least Angle Regression algorithm

The pathwise Lasso regression problem is the problem when the regression parameter λ varies. The Least Angle Regression (LARS) algorithm, proposed and named by Efron *et al.* [EHJT04], outputs the Lasso solution $\hat{\boldsymbol{\beta}}(\lambda)$ for all $\lambda > 0$. To be more precise, consider the Lasso estimator described above,

$$\hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \quad \text{for all } \lambda > 0. \tag{1}$$

The ℓ_1-penalty term is $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^d |\beta_i|$, whereas the predictive loss is the standard Euclidean squared norm. In this work, we are interested in the Lasso solution $\hat{\boldsymbol{\beta}}(\lambda)$ as a function of the regularisation parameter $\lambda > 0$. For such we define the optimal regularisation path $\mathcal{P}$:

$$\mathcal{P} \triangleq \{\hat{\boldsymbol{\beta}}(\lambda) : \lambda > 0\}. \tag{2}$$

Efron *et al.* [EHJT04] showed that the optimal regularisation path $\mathcal{P}$ is *piecewise linear* and *continuous* with respect to λ. This means that there exist an $m \in \mathbb{N}$ and $\infty > \lambda_0 > \cdots > \lambda_{m-1} > \lambda_m = 0$ and $\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_m \in \mathbb{R}^d$ such that[1]

$$\hat{\boldsymbol{\beta}}(\lambda) = \hat{\boldsymbol{\beta}}(\lambda_t) + (\lambda_t - \lambda)\boldsymbol{\theta}_t \quad \text{for } \lambda_{t+1} < \lambda \leqslant \lambda_t \quad (t \in \{0, \ldots, m-1\}). \tag{3}$$

There is a maximal value $\lambda_{\max} = \lambda_0$ where $\hat{\boldsymbol{\beta}}(\lambda) = 0$ for all $\lambda \geqslant \lambda_0$ and $\hat{\beta}_j(\lambda) \neq 0$ for some $j \in [d]$ and $\lambda < \lambda_0$. Such value is $\lambda_0 = \|\mathbf{X}^\top \mathbf{y}\|_\infty$. The $m+1$ points $\lambda_0, \ldots, \lambda_m$ where $\partial\hat{\boldsymbol{\beta}}(\lambda)/\partial\lambda$ changes are called *kinks*, and the path $\{\hat{\boldsymbol{\beta}}(\lambda) : \lambda_{t+1} < \lambda \leqslant \lambda_t\}$ between two consecutive kinks λ_{t+1} and λ_t defines a linear segment. The fact that the regularisation path is piecewise linear and the locations of the kinks can be derived from the Karush-Kuhn-Tucker (KKT) optimality conditions (see Section 3 for more details).

Since the regularisation path is piecewise linear, one needs only to compute all kinks $(\lambda_t, \hat{\boldsymbol{\beta}}(\lambda_t))$ for $t \in \{0, \ldots, m-1\}$, from which the whole regularisation path follows by linear interpolation. That is exactly what the Least Angle Regression (LARS) algorithm proposed and named by Efron *et al.* [EHJT04] does (we note that a very similar idea appeared earlier in the works of Osborne *et al.* [OPT00a, OPT00b]). Starting at $\lambda = \infty$ (or λ_0) where the Lasso solution is $\hat{\boldsymbol{\beta}}(\lambda) = \mathbf{0} \in \mathbb{R}^d$, the LARS algorithm decreases the regularisation parameter λ and computes the regularisation path by finding the kinks along the way with the aid of the KKT conditions. Each kink corresponds to an iteration of the algorithm. More specifically, at each iteration t, the LARS algorithm computes the next kink λ_{t+1} by finding the closest point (to the previous kink) where the KKT conditions break and thus need to be updated. This is done by performing a search over the *active set* $\mathcal{A} \triangleq \{i \in [d] : \hat{\beta}_i \neq 0\}$ and another search over the *inactive set* $\mathcal{I} \triangleq [d]\backslash\mathcal{A}$. The search over $\mathcal{A}$ finds the next point $\lambda_{t+1}^{\text{cross}}$, called *crossing time*, where a variable i_{t+1}^{cross} must leave $\mathcal{A}$ and join $\mathcal{I}$. The search over $\mathcal{I}$ finds the next point $\lambda_{t+1}^{\text{join}}$, called *joining time*, where a variable i_{t+1}^{join} must leave $\mathcal{I}$ and join $\mathcal{A}$. The next kink is thus $\lambda_{t+1} = \max\{\lambda_{t+1}^{\text{cross}}, \lambda_{t+1}^{\text{join}}\}$. The overall complexity of the LARS algorithm per iteration is $O(nd + |\mathcal{A}|^2)$: the two searches over the active and inactive sets require $O(n|\mathcal{A}|)$ and $O(n|\mathcal{I}|)$ time, respectively, while computing the new direction $\partial\hat{\boldsymbol{\beta}}/\partial\lambda = \boldsymbol{\theta}_{t+1}$ of the regularisation path, which involves the computation of the pseudo-inverse of a submatrix of $\mathbf{X}$ specified by $\mathcal{A}$, requires $O(n|\mathcal{A}| + |\mathcal{A}|^2)$ time. When the Lasso solution is unique, it is known [Tib13] that $|\mathcal{A}| \leqslant \min\{n, d\} = n$ throughout the LARS algorithm.

[1] Define $[n] \triangleq \{1, \ldots, n\}$.

1.2 Our work

In the high-dimensional setting where $d \gg n$, the search over the inactive set $\mathcal{I}$, and thus the computation of the joining time, is by far the most costly step per iteration. In this work, we propose quantum algorithms for the pathwise Lasso regression problem based on the Least Angle Regression (LARS) algorithm. To the best of our knowledge, this work is the first to present a quantum version of the LARS algorithm. We propose mainly two quantum algorithms based on the LARS algorithm to speedup the computation of the joining time. Our first quantum algorithm, called simple quantum LARS algorithm, is a straightforward improvement that utilizes the well-known quantum minimum-finding subroutine from Dürr and Høyer [DH96] to perform the search over $\mathcal{I}$. We assume that the design matrix $\mathbf{X}$ is stored in a quantum-readable read-only memory (QROM) and can be accessed in time $O(\operatorname{poly}\log(nd))$. We also assume that data can be written into quantum-readable classical-writable classical memories (QRAM) of size $O(n)$, which can be later queried in time $O(\operatorname{poly}\log n)$. The complexity of computing the joining time is now $O(n\sqrt{|\mathcal{I}|})$, where $O(n)$ is the time required to compute, via classical circuits, the joining times to be maximised over $\mathcal{I}$. The final runtime of our simple quantum algorithm is $\widetilde{O}(n\sqrt{|\mathcal{I}|}+n|\mathcal{A}|+|\mathcal{A}|^2)$ per iteration, where we omit polylog factors in n and d.

We then improve upon our simple quantum algorithm by approximately computing the joining times to be maximised over $\mathcal{I}$, thus reducing the factor $O(n)$ to $O(\sqrt{n})$. This is done by constructing a quantum unitary based on quantum amplitude estimation [BHMT02] that approximately computes the joining times of all the variables in the inactive set $\mathcal{I}$. We then employ the recently developed approximate quantum minimum-finding subroutine from Chen and de Wolf [CdW23] to find an estimate ϵ-close to the true maximum joining time. The result is our approximate quantum LARS algorithm that achieves a quadratic speedup in both the number of features d and the number of observations n. We also propose a sampling-based classical LARS algorithm. Our improved quantum algorithm (and the sampling-based classical algorithm), however, introduces an error in computing the joining time $\lambda_{t+1}^{\text{join}}$. As one of our main contributions, we prove that the LARS algorithm is robust to errors. More specifically, we show that, by slightly adapting the LARS algorithm, it returns (and so our approximate quantum algorithm) an *approximate regularisation path* with error proportional to the error from computing $\lambda_{t+1}^{\text{join}}$. We first define an approximate regularisation path.

Definition 1. *Given $\epsilon : \mathbb{R}_{>0} \to \mathbb{R}_{\geqslant 0}$, a vector $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^d$ is an approximate Lasso solution with error $\epsilon(\lambda)$ if*

$$\frac{1}{2}\|\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}\|_2^2+\lambda\|\widetilde{\boldsymbol{\beta}}\|_1-\left(\min_{\boldsymbol{\beta}\in\mathbb{R}^d}\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1\right)\leqslant\epsilon(\lambda)\quad \textit{where } \lambda>0.$$

A set $\widetilde{\mathcal{P}} \triangleq \{\widetilde{\boldsymbol{\beta}}(\lambda) \in \mathbb{R}^d : \lambda > 0\}$ is an approximate regularisation path with error $\epsilon(\lambda)$ if, for all $\lambda > 0$, the point $\widetilde{\boldsymbol{\beta}}(\lambda)$ of $\widetilde{\mathcal{P}}$ is an approximate Lasso solution with error $\epsilon(\lambda)$.

Result 2 (Informal version of Theorem 25)**.** *Let $\epsilon \in [0, 2)$. Consider an approximate* LARS *algorithm that returns a path $\widetilde{\mathcal{P}} = \{\widetilde{\boldsymbol{\beta}}(\lambda) \in \mathbb{R}^d : \lambda > 0\}$ and wherein, at each iteration t, the joining time $\lambda_{t+1}^{\text{join}}$ is approximated by $\widetilde{\lambda}_{t+1}^{\text{join}}$ such that $\lambda_{t+1}^{\text{join}} \leqslant \widetilde{\lambda}_{t+1}^{\text{join}} \leqslant (1-\epsilon/2)^{-1}\lambda_{t+1}^{\text{join}}$. Then $\widetilde{\mathcal{P}}$ is an approximate regularisation path with error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}(\lambda)\|_1$.*

To prove the above result, we consider an approximate version of the KKT conditions and use a duality gap for the Lasso regression. As far as we are aware, this is the first direct result on the robustness of the LARS algorithm. Compared to the work of Mairal and Yu [MY12], who also introduced an approximate LARS algorithm and employed similar techniques to analyse its

correctness, the computation of the joining time in their case is exact, and errors only arise by utilising a first-order optimisation method to find an approximate solution when kinks happen to be too close.

Result 2 guarantees the correctness of all our algorithms. Their time complexities, on the other hand, are given by the theorem below and summarised in Table 1. In the following, let $\mathbf{X}^+$ be the Moore–Penrose inverse of $\mathbf{X}$. Also, given some matrix $\mathbf{A}$, let $\|\mathbf{A}\|_2$ be its the spectral norm, $\|\mathbf{A}\|_1 = \max_j \sum_i |A_{ij}|$, and $\|\mathbf{A}\|_{\max} = \max_{i,j} |A_{ij}|$.

Result 3 (Informal version of Theorems 22, 26, 27)**.** *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Assume that $\mathbf{X}$ is stored in a* QROM *and we have access to* QRAMs *and classical-samplable structures of memory size $O(n)$. Let $\delta, \epsilon \in (0,1)$, and $T \in \mathbb{N}$. Let $\alpha, \gamma \in (0,1]$ such that*

$$\max_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T} \|\mathbf{X}_{\mathcal{A}}^{+}\mathbf{X}_{\mathcal{A}^c}\|_1 \leqslant 1-\alpha \quad \text{and} \quad \min_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T} \frac{\|\mathbf{X}^\top(\mathbf{I}-\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^{+})\mathbf{y}\|_\infty}{\|\mathbf{X}\|_{\max}\|(\mathbf{I}-\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^{+})\mathbf{y}\|_1} \geqslant \gamma,$$

where $\mathcal{A}^c = [d]\backslash\mathcal{A}$ and $\mathbf{X}_{\mathcal{A}} \in \mathbb{R}^{n\times|\mathcal{A}|}$ is the matrix formed by the columns of $\mathbf{X}$ in $\mathcal{A}$.

- *There is a quantum* LARS *algorithm that returns an optimal regularisation path with T kinks with probability at least $1-\delta$ and in time*

$$\widetilde{O}\big(n\sqrt{|\mathcal{I}|} + n|\mathcal{A}| + |\mathcal{A}|^2\big)$$

  per iteration, where $\mathcal{A}$ and $\mathcal{I}$ are the active and inactive sets of the corresponding iteration.

- *There is a quantum* LARS *algorithm that returns an approximate regularisation path with additive error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}\|_1$ and T kinks with probability at least $1-\delta$ and in time*

$$\widetilde{O}\left(\frac{\gamma^{-1}+\sqrt{n}\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2}{\alpha\epsilon}\sqrt{|\mathcal{I}|} + n|\mathcal{A}| + |\mathcal{A}|^2\right)$$

  per iteration, where $\mathcal{A}$ and $\mathcal{I}$ are the active and inactive sets of the corresponding iteration.

- *There is a classical* LARS *algorithm that returns an approximate regularisation path with additive error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}\|_1$ and T kinks with probability at least $1-\delta$ and in time*

$$\widetilde{O}\left(\frac{\gamma^{-2}+n\|\mathbf{X}\|_{\max}^2\|\mathbf{X}^+\|_2^2}{\alpha^2\epsilon^2}|\mathcal{I}| + n|\mathcal{A}| + |\mathcal{A}|^2\right)$$

  per iteration, where $\mathcal{A}$ and $\mathcal{I}$ are the active and inactive sets of the corresponding iteration.

The notation $\widetilde{O}(\cdot)$ omits $\operatorname{poly}\log$ *terms in n, d, T, and δ.*

The complexity of our approximate quantum LARS algorithm depends on a few properties of the design matrix $\mathbf{X}$: the quantity $\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2$ and the parameters α and γ. We note that $\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2 \leqslant \|\mathbf{X}\|_2\|\mathbf{X}^+\|_2$ is upper bounded by the condition number of $\mathbf{X}$, which is simply a constant for well-behaved matrices. The existence of the parameter $\alpha \in (0,1]$ is often called *(mutual) incoherence* or *strong irrepresentable condition* in the literature [ZY06, HMZ08, Wai19] and measures the level of orthogonality between columns of $\mathbf{X}$. If the column space of $\mathbf{X}_{\mathcal{A}}$ is orthogonal to $\mathbf{X}_i$, then $\|\mathbf{X}_{\mathcal{A}}^{+}\mathbf{X}_i\|_1 = 0$. On the other hand, if $\mathbf{X}_i$ is contained in the column space of

Algorithm	Error	Time complexity per iteration
Classical LARS 1	None	$O(n\lvert\mathcal{I}\rvert + n\lvert\mathcal{A}\rvert + \lvert\mathcal{A}\rvert^2)$
Simple quantum LARS 2	None	$\widetilde{O}(n\sqrt{\lvert\mathcal{I}\rvert} + n\lvert\mathcal{A}\rvert + \lvert\mathcal{A}\rvert^2)$
Approximate classical LARS 4	$\lambda\epsilon\lVert\widetilde{\boldsymbol{\beta}}\rVert_1$	$\widetilde{O}\left(\frac{\gamma^{-2}+n\lVert\mathbf{X}\rVert_{\max}^2\lVert\mathbf{X}^+\rVert_2^2}{\alpha^2\epsilon^2}\lvert\mathcal{I}\rvert + n\lvert\mathcal{A}\rvert + \lvert\mathcal{A}\rvert^2\right)$
Approximate quantum LARS 3	$\lambda\epsilon\lVert\widetilde{\boldsymbol{\beta}}\rVert_1$	$\widetilde{O}\left(\frac{\gamma^{-1}+\sqrt{n}\lVert\mathbf{X}\rVert_{\max}\lVert\mathbf{X}^+\rVert_2}{\alpha\epsilon}\sqrt{\lvert\mathcal{I}\rvert} + n\lvert\mathcal{A}\rvert + \lvert\mathcal{A}\rvert^2\right)$

Table 1: Summary of results. Throughout this work, n is the number of observations, d is the number of features, $|\mathcal{A}|$ is the size of the active set, and $|\mathcal{I}|$ is the size of the inactive set. In addition, $\delta \in (0,1)$ is an upper bound on the failure probability, $\epsilon \in (0,1)$, $\mathbf{X}^+$ is the Moore–Penrose inverse of $\mathbf{X}$, and the parameters $\alpha, \gamma \in (0,1]$ are defined in Result 3.

$\mathbf{X}_\mathcal{A}$, then $\|\mathbf{X}_\mathcal{A}^+\mathbf{X}_i\|_1 = 1$. Mutual incoherence has been considered by several previous works [DH01, EB02, Fuc04, Tro06, ZY06, MB06, Wai09]. Refs. [ZY06, MB06] provide several examples of families of matrices that satisfy mutual incoherence. Finally, the parameter γ, introduced and named by us as *mutual overlap (between* $\mathbf{y}$ *and* $\mathbf{X}$*)*, measures the overlap between projections of the observation vector $\mathbf{y}$ and the columns of $\mathbf{X}$. We note that $\gamma \leqslant 1$ by Hölder's inequality. In Section 4.4, we bound these three quantities for the case when $\mathbf{X}$ is the standard Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ entries. We show that, with high probability, $\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2$ is a constant (which follows easily from the condition number of Gaussian random matrices [CD05]), the mutual incoherence is at least $1/2$, and the mutual overlap is $\Omega(1/\sqrt{n \log d})$ (which maintains the overall complexity of the approximate algorithms).

In the case when $\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2$ is a constant and α and γ are bounded away from 0, we retrieve the complexity $\widetilde{O}(\sqrt{n|\mathcal{I}|}/\epsilon + n|\mathcal{A}| + |\mathcal{A}|^2)$ per iteration. For iterations when $|\mathcal{A}| = O(n)$, we obtain the overall quadratic improvement $\widetilde{O}(\sqrt{nd})$ over the classical LARS algorithm. Our approximate quantum LARS algorithm, however, depends on the design matrix being well-behaved in order to bound the quantities mentioned above. Regarding our approximate classical LARS algorithm, its complexity is very similar to the usual LARS algorithm. It is possible to obtain some advantage, though, if α, γ are bounded away from 0 and $\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2 = o(1)$.

Compared to the previous work of Chen and de Wolf [CdW23] on quantum algorithms for Lasso, our work is different in that we study the pathwise Lasso regression problem as the tuning parameter λ varies, and not focus on a single point λ. Both works are thus incomparable.

The rest of the paper is organized as follows. In Section 2, we introduce important notations, the computational model, and quantum subroutines that will be used in the paper. Next, in Section 3, we describe the steps to obtain a Lasso path, conditions for the uniqueness of the Lasso solution, and analyse the per-iteration time complexity of the classical LARS algorithm. In Section 4, we describe our two quantum LARS algorithms and a classical equivalent algorithm based on sampling. We discuss bounds in the noisy regime in Section 5 and summarize our work with some possible future directions in Section 6.

1.3 Related work

Seeing the importance for variable selection in the high-dimensional data context, algorithms such as best subsets, forward selection, and backward elimination are well studied and widely used to produce sparse solutions [Mao02, Mao04, WFL00, VK05, BT19, RS14, TFZB08, ZJ96, KM14, WB06]. Apart from the aforementioned feature selection algorithms, Ridge regression, proposed by Hoerl and Kennard [HK70], is another popular and well-studied method [McD09, Dor14, HK70,

MS75, vW15, HKB75, SGV98, Kib03, Vin78].

In closer relation to our work, the homotopy method of Osborne *et al.* [OPT00b, OPT00a] and the Least Angle Regression (LARS) algorithm of Efron *et al.* [EHJT04] are feature selection algorithms which use a less greedy approach as compared to classical forward selection approaches. The LARS algorithm can be modified to give rise to algorithms for Lasso and for the efficient implementation of Forward Stagewise linear regression. Based on the work of Efron *et al.* [EHJT04], Rosset and Zhu [RZ07] gave a general characterization of loss-penalty pairs to allow for efficient generation of the full regularised coefficient paths. Mairal and Yu [MY12] later provided an upper bound on the number of linear segments of the Lasso regularisation path. They also showed that an ϵ-approximate path for Lasso with at most $O(1/\sqrt{\varepsilon})$ segments can be obtained by developing an approximate homotopy algorithm based on newly defined ϵ-approximate optimality conditions. Furthermore, Tibshirani [MY12] provided conditions for the uniqueness of the Lasso solution and extended the LARS algorithm to the setting where the Lasso solution is non-unique. Other works on algorithms for Lasso and its variants include Refs. [Rot04, Sto13, KKK08, GKZ18, AT16].

In the quantum setting, considerable amount of work has been done on linear regression using the (ordinary/unregularised) least squares approach. Kaneko *et al.* [KMTY21] proposed a quantum algorithm for linear regression which outputs classical estimates of the regression coefficients. This improves upon the earlier works [WBL12, Wan17] where the regression coefficients are encoded in the amplitudes of a quantum state. Chakraborty *et al.* [CGJ19] devised quantum algorithms for the weighted and generalised variants of the least squares problem by using block-encoding techniques and Hamiltonian simulation [LC19] to improve the quantum liner systems solver. As an application of linear regression to machine learning tasks, Schuld *et al.* [SSP16] designed a quantum pattern recognition algorithm based on the ordinary least squares approach, using ideas from [HHL09] and [LMR14]. Kerenidis and Prakash [KP20] provided a quantum stochastic gradient descent algorithm for the weighted least squares problem using a quantum linear systems solver in the QRAM data structure model [Pra14, KP17], which allows quantum state preparation to be performed efficiently.

While quantum linear regression has been relatively well studied in the unregularised setting, research on algorithms for regression using regularised least squares is not yet well established. In the sparse access model, Yu *et al.* [YGW19] presented a quantum algorithm for Ridge regression using a technique called parallel Hamiltonian simulation. This technique is then used to develop the quantum analogue of K-fold cross-validation [Reg70], which serves as an efficient performance estimator of Ridge regression. Other works on quantum Ridge regression algorithms in the sparse access model include [SX20, CYGL23]. Chen and de Wolf designed quantum algorithms for Lasso and Ridge regression from the perspective of empirical loss minimization. Both of their quantum algorithms output a classical vector whose loss is ϵ-close to the minimum achievable loss. The authors also proved quantum query lower bounds for Lasso and Ridge regression. Around the same time, Bellante and Zanero [BZ22] gave a polynomial speedup for the classical matching-pursuit algorithm, which is a heuristic algorithm for the best subset selection model, i.e., linear regression with an ℓ_0-regulariser. More recently, Chakraborty *et al.* [CMP23] gave the first quantum algorithms for least squares with general ℓ_2-norm regularisation, which includes regularised versions of quantum ordinary least squares, quantum weighted least squares, and quantum generalised least squares. Their algorithms use block-encoding techniques and the framework of quantum singular value transformation [GSLW19], and demonstrate substantial improvement as compared to previous results on quantum Ridge regression [SX20, CYGL23, YGW19].

2 Preliminaries

2.1 Notations

For a positive integer $n \in \mathbb{N}$, we use $[n]$ to denote the set $\{1, \ldots, n\}$. For a vector $\mathbf{u} \in \mathbb{R}^n$, we denote its i-th entry as u_i for $i \in [n]$. Let $\|\mathbf{u}\|_p \triangleq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ and by $\mathcal{D}_{\mathbf{u}}$ we denote the distribution over $[n]$ with probability density function $\mathcal{D}_{\mathbf{u}}(i) = |u_i|/\|\mathbf{u}\|_1$. Given a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, we denote its i-th column as $\mathbf{X}_i$ for $i \in [d]$. Given $\mathcal{A} = \{i_1, \ldots, i_k\} \subseteq [d]$, define $\mathcal{A}^c = [d] \backslash \mathcal{A}$ and let $\mathbf{X}_{\mathcal{A}} \in \mathbb{R}^{n \times |\mathcal{A}|}$ be the matrix $\mathbf{X}_{\mathcal{A}} = [\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_k}]$ formed by the columns of $\mathbf{X}$ in $\mathcal{A}$. The previous notation naturally extends to vectors, i.e., $\mathbf{u}_{\mathcal{A}} = [u_{i_1}, \ldots, u_{i_k}]$. Given $\mathcal{A} \subseteq [d]$ and $S \subseteq [n]$, $\mathbf{X}_{S\mathcal{A}}$ is the submatrix of $\mathbf{X}$ with rows in S and columns in $\mathcal{A}$. We denote by $\mathrm{col}(\mathbf{X})$, $\mathrm{row}(\mathbf{X})$, $\mathrm{null}(\mathbf{X})$, and $\mathrm{rank}(\mathbf{X})$ the column space, row space, null space, and the rank, respectively, of $\mathbf{X}$. We denote by $\mathbf{X}^+$ the Moore-Penrose inverse of $\mathbf{X}$. If $\mathbf{X}$ has linearly independent columns, then $\mathbf{X}^+ = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$. Note that $\mathbf{X}_{\mathcal{A}} \mathbf{X}_{\mathcal{A}}^+$ is the projection matrix onto $\mathrm{col}(\mathbf{X}_{\mathcal{A}})$. Let $\|\mathbf{X}\|_p \triangleq \sup_{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\|_p = 1} \|\mathbf{X}\mathbf{u}\|_p$ be the p-norm of $\mathbf{X}$, $p \in [1, \infty]$. As special cases, $\|\mathbf{X}\|_1 = \max_{j \in [d]} \sum_{i=1}^n |X_{ij}|$ and $\|\mathbf{X}\|_\infty = \max_{i \in [n]} \sum_{j=1}^d |X_{ij}|$. Let also $\|\mathbf{X}\|_{\max} \triangleq \max_{i \in [n], j \in [d]} |X_{ij}|$. We use $|\bar{0}\rangle$ to denote the state $|0\rangle \otimes \cdots \otimes |0\rangle$, where the number of qubits is clear from the context. We use $\widetilde{O}(\cdot)$ to hide polylogarithmic factors, i.e., $\widetilde{O}(f(n)) = O(f(n) \cdot \mathrm{poly}\log(f(n)))$. See Appendix A for a summary of symbols.

2.2 Concentration bounds

We revise a few notions of probability theory and concentration bounds, starting with the concept of sub-Gaussian variable.

Definition 4 (Sub-Gaussianity)**.** *A random variable Z is sub-Gaussian with parameter σ^2 if*

$$\mathbb{E}\left[e^{t(Z - \mathbb{E}[Z])}\right] \leqslant \exp\left(\frac{t^2 \sigma^2}{2}\right) \quad \text{for all } t \in \mathbb{R}.$$

We note that any sub-Gaussian random variable satisfies a Chernoff bound.

Fact 5. *If Z is a sub-Gaussian random variable with parameter σ^2, then*

$$\mathbb{P}[|Z - \mathbb{E}[Z]| > t] \leqslant 2 \exp\left(-\frac{t^2}{2\sigma^2}\right) \quad \text{for all } t \in \mathbb{R}.$$

Fact 6. *If each coordinate of $\mathbf{w} \in \mathbb{R}^n$ is an independent sub-Gaussian random variable with parameter σ^2, then, for any $\mathbf{u} \in \mathbb{R}^n$, $\mathbf{u}^\top \mathbf{w}$ is sub-Gaussian with parameter $\sigma^2 \|\mathbf{u}\|_2^2$.*

Proof. $\mathbb{E}\left[\exp\left(t\, \mathbf{u}^\top \mathbf{w}\right)\right] = \prod_{i=1}^n \mathbb{E}\left[\exp\left(t\, u_i w_i\right)\right] \leqslant \prod_{i=1}^n \exp\left(\frac{t^2 \sigma^2 u_i^2}{2}\right) = \exp\left(\frac{t^2 \sigma^2 \|\mathbf{u}\|_2^2}{2}\right)$, $\forall t \in \mathbb{R}$. □

More generally, one can use Hoeffding's inequality to bound the sum of bounded independent random variables.

Fact 7 (Hoeffding's bound)**.** *Let $Z_1, \ldots, Z_n$ be independent random variables such that, for $i \in [n]$, $a_i \leqslant Z_i \leqslant b_i$ almost surely. Let $Z \triangleq \sum_{i=1}^n Z_i$. Then*

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geqslant t] \leqslant 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \quad \text{for all } t > 0.$$

We will also make use of the following lemma about the relative approximation of a ratio of two real numbers that are given by relative approximations.

Lemma 8. *Let $\widetilde{a}, \widetilde{b} \in \mathbb{R}$ be estimates of $a, b \in \mathbb{R}\backslash\{0\}$, respectively, such that $|a - \widetilde{a}| \leqslant \epsilon_a$ and $|b - \widetilde{b}| \leqslant \epsilon_b$, where $\epsilon_a \geqslant 0$ and $\epsilon_b \in [0, |b|/2]$. Then*

$$\left|\frac{\widetilde{a}}{\widetilde{b}} - \frac{a}{b}\right| \leqslant 2\frac{|a|}{|b|}\left(\frac{\epsilon_a}{|a|} + \frac{\epsilon_b}{|b|}\right).$$

Proof. First note that

$$|b| - |\widetilde{b}| \leqslant \epsilon_b \implies \frac{1}{|\widetilde{b}|} \leqslant \frac{1}{|b| - \epsilon_b} \leqslant \frac{2}{|b|}.$$

Then

$$\left|\frac{\widetilde{a}}{\widetilde{b}} - \frac{a}{b}\right| = \left|\frac{\widetilde{a}b - ab + ab - a\widetilde{b}}{b\widetilde{b}}\right| \leqslant \left|\frac{\widetilde{a} - a}{\widetilde{b}}\right| + \left|\frac{a(b - \widetilde{b})}{b\widetilde{b}}\right| \leqslant \frac{2\epsilon_a}{|b|} + \frac{|a|}{|b|}\frac{2\epsilon_b}{|b|} = 2\frac{|a|}{|b|}\left(\frac{\epsilon_a}{|a|} + \frac{\epsilon_b}{|b|}\right). \qquad \square$$

2.3 Computational model

Classical computational model. Our classical computational model is a classical random-access machine. The input to the Lasso problem is a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and a vector $\mathbf{y} \in \mathbb{R}^n$, which are stored in a classical-readable read-only memory (ROM). For simplicity, reading any entry of $\mathbf{X}$ and $\mathbf{y}$ takes constant time. The classical computer can write bits to a classical-samplable-and-writable memory. More specifically, the classical computer can write a vector $\mathbf{u} \in \mathbb{R}^m$ (or parts of it) into a low-overhead samplable data structure in time $O(m)$, which allows it to sample an index $i \in [m]$ from the probability distribution $\mathcal{D}_{\mathbf{u}}$ in time $O(\operatorname{poly}\log m)$. We will often refer to the ability to sample from $\mathcal{D}_{\mathbf{u}}$ as having sampling access to the vector $\mathbf{u}$. In this work, we shall assume that all classical-samplable-and-writable memories have size $m = O(\operatorname{poly} n)$. We assume an arithmetic model in which we ignore issues arising from the fixed-point representation of real numbers. All basic arithmetic operations in this model take constant time.

Quantum computational model. Our quantum computational model is a classical random-access machine with access to a quantum computer. The input to the Lasso problem is a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and a vector $\mathbf{y} \in \mathbb{R}^n$, which are stored in a classical-readable read-only memory (ROM). For simplicity, reading any entry of $\mathbf{X}$ and $\mathbf{y}$ takes constant time. We assume that $\mathbf{X}$ is also stored in a quantum-readable read-only memory (QROM), whose single entries can be queried. This means that the quantum computer has access to an oracle $\mathcal{O}_{\mathbf{X}}$ that performs the mapping

$$\mathcal{O}_{\mathbf{X}} : |i, j\rangle|\bar{0}\rangle \mapsto |i, j\rangle|X_{ij}\rangle, \quad \forall i \in [n], j \in [d],$$

in time $O(\operatorname{poly}\log(nd))$. The classical computer can also write bits to a quantum-readable classical-writable classical memory (QRAM) [GLM08a, GLM08b, ABD+23], which can be accessed in superposition by the quantum computer. More specifically, the classical computer can write a vector $\mathbf{u} \in \mathbb{R}^m$ (or parts of it) into the memory of a QRAM in time $O(m)$, which allows the quantum computer to invoke an oracle $\mathcal{U}_{\mathbf{u}}$ that performs the mapping

$$\mathcal{U}_{\mathbf{u}} : |i\rangle|\bar{0}\rangle \mapsto |i\rangle|u_i\rangle, \quad \forall i \in [m],$$

in time $O(\operatorname{poly}\log m)$. We will often refer to the ability to invoke $\mathcal{U}_{\mathbf{u}}$ as having quantum access to the vector $\mathbf{u}$. In this work, we shall assume that all QRAMs have size $m = O(\operatorname{poly} n)$. We do not necessarily assume that $\mathbf{y}$ is stored in a QROM. Instead, it can be stored in a QRAM in time $O(n)$ by the classical computer.

The classical computer can send the description of a quantum circuit to the quantum computer, which is a sequence of quantum gates from a universal gate set plus queries to the bits stored in the QROM and/or QRAM; the quantum computer runs the circuit, performs a measurement in the computational basis, and returns the measurement outcome to the classical computer. We refer to the runtime of a classical/quantum computation as the number of basic gates performed plus the time complexity of all the calls to the QROM and QRAMs. We assume an arithmetic model in which we ignore issues arising from the fixed-point representation of real numbers. All basic arithmetic operations in this model take constant time.

Kerenidis and Prakash [Pra14, KP17] introduced a classical data structure (sub-sequentially called KP-tree) to store a vector $\mathbf{u} \in \mathbb{R}^m$ to enable the efficient preparation of the state

$$\sum_{i=1}^{m} \sqrt{\frac{|u_i|}{\|\mathbf{u}\|_1}} |i\rangle |\mathrm{sign}(u_i)\rangle.$$

Our quantum algorithms shall commonly build KP-trees of auxiliary vectors in order to prepare the above states.

Fact 9 (KP-tree [Pra14, KP17, CdW23])**.** *Let $\mathbf{u} \in \mathbb{R}^m$ be a vector with $w \in \mathbb{N}$ non-zero entries. There is a data structure called* KP*-tree of size $O(w \operatorname{poly}\log m)$ that stores each input (i, u_i) in time $O(\operatorname{poly}\log m)$. After the data structure has been constructed, there is a quantum algorithm that prepares the state $\sum_{i=1}^{m} \sqrt{\frac{|u_i|}{\|\mathbf{u}\|_1}} |i\rangle|\mathrm{sign}(u_i)\rangle$ up to negligible error in time $O(\operatorname{poly}\log m)$.*

2.4 Quantum subroutines

In this section, we review some useful quantum subroutines that shall be used in the rest of the paper, starting with the quantum minimum-finding algorithm from Dürr and Høyer [DH96].

Fact 10 (Quantum minimum finding [DH96])**.** *Given quantum access to a vector $\mathbf{u} \in \mathbb{R}^m$, there is a quantum algorithm that finds $\min_{i\in[m]} u_i$ with success probability $1-\delta$ in time $\widetilde{O}\left(\sqrt{m}\log\frac{1}{\delta}\right)$.*

The above subroutine was later generalised by Chen and de Wolf [CdW23] in the case when one has quantum access to the entries of $\mathbf{u}$ up to some additive error. We note that a similar result, but in a different setting, appeared in [QCR20].

Fact 11 ([CdW23, Theorem 2.4])**.** *Let $\delta_1, \delta_2 \in (0,1)$ such that $\delta_2 = O\big(\delta_1^2/(m\log(1/\delta_1))\big)$, $\epsilon > 0$, and $\mathbf{u} \in \mathbb{R}^m$. Suppose access to a unitary that maps $|k\rangle|\bar{0}\rangle \mapsto |k\rangle|R_k\rangle$ such that, for every $k \in [m]$, after measuring the state $|R_k\rangle$, with probability at least $1-\delta_2$ the first register r_k of the measurement outcome satisfies $|r_k - u_k| \leqslant \epsilon$. Then there is a quantum algorithm that finds an index k such that $u_k \leqslant \min_{j\in[m]} u_j + 2\epsilon$ with probability at least $1-\delta_1$ and in time $\widetilde{O}(\sqrt{m}\log(1/\delta_1))$.*

We shall also need the amplitude estimation subroutine from Brassard *et al.* [BHMT02].

Fact 12 ([BHMT02, Theorem 12])**.** *Given a natural number M and access to an $(n+1)$-qubit unitary U satisfying*

$$U|0^n\rangle|0\rangle = \sqrt{a}|\psi_1\rangle|1\rangle + \sqrt{1-a}|\psi_0\rangle|0\rangle,$$

where $|\psi_0\rangle$ and $|\psi_1\rangle$ are arbitrary n-qubit states and $a \in (0,1)$, there is a quantum algorithm that uses $O(M)$ applications of U and $U^\dagger$ and $\widetilde{O}(M)$ elementary gates, and outputs a state $|\Lambda\rangle$ such that, after measuring $|\Lambda\rangle$, with probability at least $9/10$, the first register λ of the outcome satisfies

$$|a-\lambda| \leqslant \frac{\sqrt{a(1-a)}}{M} + \frac{1}{M^2}.$$

Finally, the following result, which is based on [CdW23, Theorem 3.4], constructs the mapping $|j\rangle|\bar{0}\rangle \mapsto |j\rangle|R_j\rangle$ such that R_j holds an approximation to $\mathbf{A}_j^\top \mathbf{u}$ up to an additive error, where $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{u} \in \mathbb{R}^n$ can be accessed quantumly via KP-trees.

Lemma 13. *Given $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{u} \in \mathbb{R}^n$, assume they are stored in* KP*-trees and we have quantum access to their entries. Let $\delta \in (0,1)$ and $\epsilon > 0$. Then there is a quantum algorithm that implements the mapping $|j\rangle|\bar{0}\rangle \mapsto |j\rangle|R_j\rangle$ such that, for all $j \in [d]$, after measuring the state $|R_j\rangle$, with probability at least $1-\delta$, the outcome r_j of the first register satisfies*

$$\left|r_j - \mathbf{A}_j^\top \mathbf{u}\right| \leqslant \epsilon \min(\|\mathbf{A}_j\|_\infty \|\mathbf{u}\|_1, \|\mathbf{A}_j\|_1 \|\mathbf{u}\|_\infty)$$

in time $O(\epsilon^{-1} \log(1/\delta) \operatorname{poly}\log(nd))$.

Proof. Fix $j \in [d]$. Consider $\mathbf{A}_j^\top \mathbf{u} = \sum_{i=1}^n A_{ij} u_i$. Apply the operator from Fact 9 to create

$$\sum_{i=1}^n \sqrt{\frac{|u_i|}{\|\mathbf{u}\|_1}} |i\rangle|s_i\rangle|\bar{0}\rangle|0\rangle, \tag{4}$$

in time $O(\operatorname{poly}\log n)$, where $s_i \triangleq \operatorname{sign}(u_i)$. On the other hand, using query access to $\mathbf{A}$ create the mapping $|i\rangle|s_i\rangle|\bar{0}\rangle \mapsto |i\rangle|s_i\rangle|s_i A_{ij}\rangle$ costing time $O(\operatorname{poly}\log(nd))$. Using this operator on the first three registers in Eq. (4), we obtain

$$\sum_{i=1}^n \sqrt{\frac{|u_i|}{\|\mathbf{u}\|_1}} |i\rangle|s_i\rangle|s_i A_{ij}\rangle|0\rangle. \tag{5}$$

Now, define the positive-controlled rotation for each $a \in \mathbb{R}$ such that

$$U_{\mathrm{CR}^+} : |a\rangle|0\rangle \mapsto \begin{cases} |a\rangle(\sqrt{a}|1\rangle + \sqrt{1-a}|0\rangle) & \text{if } a \in (0,1], \\ |a\rangle|0\rangle & \text{otherwise.} \end{cases}$$

Applying U_{CR^+} on the last two registers in Eq. (5), we obtain

$$\begin{aligned}
&\sum_{\substack{i\in[n] \\ s_i A_{ij} > 0}} \sqrt{\frac{|u_i|}{\|\mathbf{u}\|_1}} |i\rangle|s_i\rangle|s_i A_{ij}\rangle \left(\sqrt{\frac{s_i A_{ij}}{\|\mathbf{A}_j\|_\infty}}|1\rangle + \sqrt{1 - \frac{s_i A_{ij}}{\|\mathbf{A}_j\|_\infty}}|0\rangle\right) + \sum_{\substack{i\in[n] \\ s_i A_{ij} \leqslant 0}} \sqrt{\frac{|u_i|}{\|\mathbf{u}\|_1}} |i\rangle|s_i\rangle|s_i A_{ij}\rangle|0\rangle \\
&= \sum_{\substack{i\in[n] \\ s_i A_{ij} > 0}} \sqrt{\frac{A_{ij} u_i}{\|\mathbf{A}_j\|_\infty \|\mathbf{u}\|_1}} |i\rangle|s_i\rangle|s_i A_{ij}\rangle|1\rangle \\
&\qquad + \left(\sum_{\substack{i\in[n] \\ s_i A_{ij} > 0}} \sqrt{\frac{|u_i|}{\|\mathbf{u}\|_1} - \frac{A_{ij} u_i}{\|\mathbf{A}_j\|_\infty \|\mathbf{u}\|_1}} |i\rangle|s_i\rangle|s_i A_{ij}\rangle + \sum_{\substack{i\in[n] \\ s_i A_{ij} \leqslant 0}} \sqrt{\frac{|u_i|}{\|\mathbf{u}\|_1}} |i\rangle|s_i\rangle|s_i A_{ij}\rangle \right)|0\rangle \\
&= \sqrt{a_+}|\phi_1\rangle|1\rangle + \sqrt{1-a_+}|\phi_0\rangle|0\rangle,
\end{aligned}$$

where

$$a_+ = \sum_{\substack{i\in[n] \\ s_i A_{ij} > 0}} \frac{A_{ij} u_i}{\|\mathbf{A}_j\|_\infty \|\mathbf{u}\|_1} \quad \text{and} \quad |\phi_1\rangle = \sum_{\substack{i\in[n] \\ s_i A_{ij} > 0}} \sqrt{\frac{A_{ij} u_i}{\sum_{k\in[n], s_k A_{kj} > 0} A_{kj} u_k}} |i\rangle|s_i\rangle|s_i A_{ij}\rangle.$$

Here, $|\phi_1\rangle$ and $|\phi_0\rangle$ are unit vectors. We now use Fact 12 to get an estimate $\widetilde{a}_+$ such that

$$|\widetilde{a}_+ - a_+| \leqslant \frac{\epsilon}{2}\sqrt{a_+} \implies \left| \|\mathbf{A}_j\|_\infty \|\mathbf{u}\|_1 \widetilde{a}_+ - \sum_{\substack{i\in[n] \\ s_i A_{ij}>0}} A_{ij}u_i \right| \leqslant \frac{\epsilon}{2}\sqrt{\|\mathbf{A}_j\|_\infty \|\mathbf{u}\|_1 \sum_{\substack{i\in[n] \\ s_i A_{ij}>0}} A_{ij}u_i}$$
$$\leqslant \frac{\epsilon}{2}\|\mathbf{A}_j\|_\infty \|\mathbf{u}\|_1$$

in time $O(\epsilon^{-1}\log(1/\delta)\operatorname{poly}\log(nd))$ and probability at least $1-\delta/2$. Notice that we know $\|\mathbf{u}\|_1$. Similarly, we estimate

$$a_- = -\sum_{\substack{i\in[n] \\ s_i A_{ij}<0}} \frac{A_{ij}u_i}{\|\mathbf{A}_j\|_\infty \|\mathbf{u}\|_1}$$

with additive error $\frac{\epsilon}{2}\|\mathbf{A}_j\|_\infty\|\mathbf{u}\|_1$. Thus $\mathbf{A}_j^\top \mathbf{u} = \sum_{i=1}^n A_{ij}u_i = \|\mathbf{A}_j\|_\infty\|\mathbf{u}\|_1(a_+ - a_-)$ is estimated with additive error $\epsilon\|\mathbf{A}_j\|_\infty\|\mathbf{u}\|_1$ and success probability at least $1-\delta$ in time $O(\epsilon^{-1}\log(1/\delta)\operatorname{poly}\log(nd))$.

Swapping the roles of $\mathbf{A}_j$ and $\mathbf{u}$ and repeating the above steps leads to an estimate of $\mathbf{A}_j^\top\mathbf{u}$ with additive error $\epsilon\|\mathbf{A}_j\|_1\|\mathbf{u}\|_\infty$ in time $O(\epsilon^{-1}\log(1/\delta)\operatorname{poly}\log(nd))$. □

It is possible to prove a classical result analogous to the one above. More specifically, if we have sampling access to $\mathbf{A}\in\mathbb{R}^{n\times d}$ and $\mathbf{u}\in\mathbb{R}^n$, then there is a classical algorithm that allows to approximate $\mathbf{A}_j^\top\mathbf{u}$ up to an additive error.

Lemma 14. *Let $\mathbf{A}\in\mathbb{R}^{n\times d}$ and $\mathbf{u}\in\mathbb{R}^n$. Assume sampling access to $\mathbf{u}$ and to the columns of $\mathbf{A}$. Let $\delta\in(0,1)$ and $\epsilon>0$. There is a classical algorithm that, for any $j\in[d]$, outputs r_j such that*

$$|r_j - \mathbf{A}_j^\top\mathbf{u}| \leqslant \epsilon\min(\|\mathbf{A}_j\|_\infty\|\mathbf{u}\|_1, \|\mathbf{A}_j\|_1\|\mathbf{u}\|_\infty)$$

with probability at least $1-\delta$ and in time $O(\epsilon^{-2}\log(1/\delta)\operatorname{poly}\log(nd))$.

Proof. Fix $j\in[d]$. Let Z be the random variable defined by

$$\mathbb{P}\left[Z = \|\mathbf{A}_j\|_1 u_i \operatorname{sign}(A_{ij})\right] = \frac{|A_{ij}|}{\|\mathbf{A}_j\|_1} \quad \text{for } i\in[n].$$

Then

$$\mathbb{E}[Z] = \sum_{i=1}^n \frac{|A_{ij}|}{\|\mathbf{A}_j\|_1}\|\mathbf{A}_j\|_1 u_i \operatorname{sign}(A_{ij}) = \mathbf{A}_j^\top\mathbf{u}.$$

Sample $q = 2\epsilon^{-2}\ln(2/\delta)$ indices $\{i_1,\dots,i_q\}\subseteq[n]$ from the distribution $\mathcal{D}_{\mathbf{A}_j}$ by using sampling access to $\mathbf{A}$ and set $Z_k = \|\mathbf{A}_j\|_1 u_{i_k}\operatorname{sign}(A_{i_k j})$, $k\in[q]$. The algorithm outputs $\widehat{Z} = \frac{1}{q}\sum_{k=1}^q Z_k$ as an estimate for $\mathbf{A}_j^\top\mathbf{u}$. Since $Z_1,\dots,Z_q$ are i.i.d. copies of Z, a Hoeffding's bound (Fact 7) gives

$$\mathbb{P}[|\widehat{Z} - \mathbf{A}_j^\top\mathbf{u}| \geqslant \epsilon\|\mathbf{A}_j\|_1\|\mathbf{u}\|_\infty] \leqslant 2e^{-\epsilon^2 q/2} = \delta,$$

so the classical algorithm approximates $\mathbf{A}_j^\top\mathbf{u}$ with additive error $\epsilon\|\mathbf{A}_j\|_1\|\mathbf{u}\|_\infty$ and success probability at least $1-\delta$. The final runtime is the number of samples q times the complexity $O(\operatorname{poly}\log(nd))$ of sampling each index. Finally, it is possible to repeat the same procedure but swapping the roles of $\mathbf{A}_j$ and $\mathbf{u}$, so that Z is now defined as $\mathbb{P}\left[Z = \|\mathbf{u}\|_1 A_{ij}\operatorname{sign}(u_i)\right] = |u_i|/\|\mathbf{u}\|_1$, $i\in[n]$. □

3 Lasso path and the LARS algorithm

3.1 Lasso solution

In this section, we review several known properties of the solution to the Lasso regression problem defined as

$$\hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \quad \text{for some } \lambda > 0, \tag{6}$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the design matrix and $\mathbf{y} \in \mathbb{R}^n$ is a vector of observations. We will cover optimality conditions for the Lasso problem, the general form for its solution, and continuity and uniqueness conditions. Most of these results can be found in [RZ07, MY12, Tib13, SCL+18]. We start with obtaining the optimality conditions for the Lasso problem.

Fact 15 ([Tib13, Lemma 1])**.** *Every Lasso solution $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ gives the same fitted value $\mathbf{X}\hat{\boldsymbol{\beta}}$.*

Fact 16. *A vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ is a solution to the Lasso problem if and only if*

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \lambda\mathbf{s} \quad \textit{where } s_i \in \begin{cases} \{\operatorname{sign}(\hat{\beta}_i)\} & \textit{if } \hat{\beta}_i \neq 0, \\ [-1, 1] & \textit{if } \hat{\beta}_i = 0, \end{cases} \quad \textit{for } i \in [d]. \tag{7}$$

Let $\mathcal{A} \triangleq \{i \in [d] : |\mathbf{X}_i^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})| = \lambda\}$, $\mathcal{I} \triangleq [d]\backslash\mathcal{A}$, and $\boldsymbol{\eta} \triangleq \operatorname{sign}(\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})) \in \{-1, 0, 1\}^d$. Any Lasso solution $\hat{\boldsymbol{\beta}}$ is of the form

$$\hat{\boldsymbol{\beta}}_{\mathcal{I}} = \mathbf{0} \quad \textit{and} \quad \hat{\boldsymbol{\beta}}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^+(\mathbf{y} - \lambda(\mathbf{X}_{\mathcal{A}}^+)^\top\boldsymbol{\eta}_{\mathcal{A}}) + \mathbf{b}, \tag{8}$$

where

$$\mathbf{b} \in \operatorname{null}(\mathbf{X}_{\mathcal{A}}) \quad \textit{and} \quad \eta_i\big([\mathbf{X}_{\mathcal{A}}^+(\mathbf{y} - \lambda(\mathbf{X}_{\mathcal{A}}^+)^\top\boldsymbol{\eta}_{\mathcal{A}})]_i + b_i\big) \geqslant 0 \textit{ for } i \in \mathcal{A}. \tag{9}$$

Proof. Eq. (7) can be obtained by considering the Karush–Kuhn–Tucker conditions, or sub-gradient optimality conditions. More specifically, a vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ is a solution to the Lasso problem if and only if (see e.g. [BL06, Proposition 3.1.5])

$$\mathbf{0} \in \{\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) - \lambda\mathbf{s} : \mathbf{s} \in \partial\|\hat{\boldsymbol{\beta}}\|_1\},$$

where $\partial\|\hat{\boldsymbol{\beta}}\|_1$ denotes the sub-gradient of the ℓ_1-norm at $\hat{\boldsymbol{\beta}}$. Eq. (7) then follows from the fact that the sub-gradients $\mathbf{s} \in \mathbb{R}^d$ of $\|\hat{\boldsymbol{\beta}}\|_1$ are vectors such that $s_i = \operatorname{sign}(\hat{\beta}_i)$ if $\hat{\beta}_i \neq 0$ and $|s_i| \leqslant 1$ otherwise.

Moving on, first note that $\boldsymbol{\eta}$ is a sub-gradient and that the uniqueness of $\mathbf{X}\hat{\boldsymbol{\beta}}$ (Fact 15) implies the uniqueness of $\mathcal{A}$ and $\boldsymbol{\eta}$. By the definition of the sub-gradients $\mathbf{s}$ in Eq. (7), we know that $\hat{\boldsymbol{\beta}}_{\mathcal{I}} = \mathbf{0}$, and therefore the block $\mathcal{A}$ of Eq. (7) can be written as

$$\mathbf{X}_{\mathcal{A}}^\top(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}) = \lambda\boldsymbol{\eta}_{\mathcal{A}}.$$

This means that $\boldsymbol{\eta}_{\mathcal{A}} \in \operatorname{row}(\mathbf{X}_{\mathcal{A}})$, and so $\boldsymbol{\eta}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^\top(\mathbf{X}_{\mathcal{A}}^\top)^+\boldsymbol{\eta}_{\mathcal{A}}$. Using this fact and rearranging the above equation, we obtain

$$\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^\top(\mathbf{y} - \lambda(\mathbf{X}_{\mathcal{A}}^\top)^+\boldsymbol{\eta}_{\mathcal{A}}).$$

Therefore, any solution $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ to the above equation is of the form

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^+(\mathbf{y} - \lambda(\mathbf{X}_{\mathcal{A}}^\top)^+\boldsymbol{\eta}_{\mathcal{A}}) + \mathbf{b}, \qquad \mathbf{b} \in \operatorname{null}(\mathbf{X}_{\mathcal{A}}),$$

where we used the identity $\mathbf{A}^+(\mathbf{A}^\top)^+\mathbf{A}^\top = \mathbf{A}^+$. Any $\mathbf{b} \in \operatorname{null}(\mathbf{X}_{\mathcal{A}})$ produces a valid Lasso solution $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ as long as $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ has the correct signs given by $\boldsymbol{\eta}_{\mathcal{A}}$, i.e., $\operatorname{sign}(\hat{\beta}_i) = \eta_i$ for $i \in \mathcal{A}$. This restriction can be written as $\hat{\beta}_i\eta_i \geqslant 0$, which is Eq. (9). This completes the proof. ∎

We shall abuse the definition of the Karush-Kuhn-Tucker (KKT) conditions and refer to Eq. (7) as the KKT conditions themselves for the Lasso problem. The set

$$\mathcal{A} \triangleq \{i \in [d] : |\mathbf{X}_i^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})| = \lambda\}$$

and the sub-gradient vector

$$\boldsymbol{\eta} \triangleq \operatorname{sign}(\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})) \in \{-1, 0, 1\}^d$$

are normally referred to as *active* (or *equicorrelation*) *set* and *equicorrelation signs*, respectively. The set $\mathcal{I} \triangleq [d] \backslash \mathcal{A}$ is called *inactive set*.

Fact 17 ([Tib13, Lemma 9])**.** *For any $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times d}$, and $\lambda > 0$, the Lasso solution*

$$\hat{\boldsymbol{\beta}}_{\mathcal{I}} = \mathbf{0} \quad \textit{and} \quad \hat{\boldsymbol{\beta}}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^{+}(\mathbf{y} - \lambda(\mathbf{X}_{\mathcal{A}}^{+})^\top \boldsymbol{\eta}_{\mathcal{A}})$$

has the minimum ℓ_2-norm over all Lasso solutions.

Proof. By Fact 16, any Lasso solution has ℓ_2-norm

$$\|\hat{\boldsymbol{\beta}}\|_2^2 = \|\mathbf{X}_{\mathcal{A}}^{+}(\mathbf{y} - \lambda(\mathbf{X}_{\mathcal{A}}^{+})^\top \boldsymbol{\eta}_{\mathcal{A}})\|^2 + \|\mathbf{b}\|^2,$$

since $\mathbf{b} \in \operatorname{null}(\mathbf{X}_{\mathcal{A}})$. Hence the ℓ_2-norm is minimised when $\mathbf{b} = \mathbf{0}$. □

It follows from Fact 16 that the Lasso solution is unique if $\operatorname{null}(\mathbf{X}_{\mathcal{A}}) = \{\mathbf{0}\}$ and is given by

$$\hat{\boldsymbol{\beta}}_{\mathcal{I}} = \mathbf{0} \quad \text{and} \quad \hat{\boldsymbol{\beta}}_{\mathcal{A}} = (\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}})^{-1}(\mathbf{X}_{\mathcal{A}}^\top \mathbf{y} - \lambda \boldsymbol{\eta}_{\mathcal{A}}). \tag{10}$$

It is known [Tib13, Lemma 14] that there is a Lasso solution such that $|\mathcal{A}| \leqslant \min\{n, d\}$. If the Lasso solution is unique, then the active set has size at most $\min\{n, d\}$. The sufficient condition $\operatorname{null}(\mathbf{X}_{\mathcal{A}}) = \{\mathbf{0}\}$ for uniqueness has appeared several times in the literature [OPT00b, Fuc05, Wai09, CP09, Tib13]. Since the active set $\mathcal{A}$ depends on the Lasso solution at $\mathbf{y}, \mathbf{X}, \lambda$, the condition that $\mathbf{X}_{\mathcal{A}}$ has full rank is somewhat circular. Because of this, more natural conditions are assumed which imply $\operatorname{null}(\mathbf{X}_{\mathcal{A}}) = \{\mathbf{0}\}$, e.g. the general position property [Ros04, Don06, Dos12, Tib13] and the matrix $\mathbf{X}$ being drawn from a continuous probability distribution [Tib13]. We say that a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has columns in *general position* if no k-dimensional subspace in $\mathbb{R}^n$, for $k < \min\{n, d\}$, contains more than $k + 1$ elements from the set $\{\pm\mathbf{X}_1, \ldots, \pm\mathbf{X}_d\}$, excluding antipodal pairs. It is known [Ros04, Don06, Dos12, Tib13] that this is enough to guarantee the uniqueness of the Lasso solution.

Fact 18 ([Tib13, Lemma 3])**.** *If the columns of $\mathbf{X} \in \mathbb{R}^{n \times d}$ are in general position, then the Lasso solution is unique for any $\mathbf{y} \in \mathbb{R}^n$ and $\lambda > 0$ and is given by Eq.* (10)*.*

Another sufficient condition was given by Tibshirani [Tib13], who proved the following.

Fact 19 ([Tib13, Lemma 4])**.** *If the entries of $\mathbf{X} \in \mathbb{R}^{n \times d}$ are drawn from a continuous probability distribution on $\mathbb{R}^{nd}$, then the Lasso solution is unique for any $\mathbf{y} \in \mathbb{R}^n$ and $\lambda > 0$ and is given by Eq.* (10) *with probability* 1*.*

It is well known that, when $\operatorname{null}(\mathbf{X}_{\mathcal{A}}) = \{\mathbf{0}\}$, the KKT conditions from Fact 16 imply that the regularisation path $\mathcal{P} \triangleq \{\hat{\boldsymbol{\beta}}(\lambda) \in \mathbb{R}^d : \lambda > 0\}$ comprising all the solutions for $\lambda > 0$ is continuous piecewise linear, which was first proven by Efron, Hastie, Johnstone, and Tibshirani [EHJT04], confer Eq. (3). We note that, according to Eq. (8), $\boldsymbol{\theta}_t = (\mathbf{X}_{\mathcal{A}_t}^\top \mathbf{X}_{\mathcal{A}_t})^{+} \boldsymbol{\eta}_{\mathcal{A}_t}$ for some active set $\mathcal{A}_t$. This fact was later extended to the case when $\operatorname{rank}(\mathbf{X}_{\mathcal{A}}) < |\mathcal{A}|$ by Tibshirani [Tib13], who proved that the path of solutions with $\mathbf{b} = \mathbf{0}$ in Eq. (8) is continuous and piecewise linear.

Fact 20 ([Tib13, Appendix A]). *The regularisation path* $\mathcal{P} = \{\hat{\boldsymbol{\beta}}(\lambda) \in \mathbb{R}^d : \lambda > 0\}$ *formed by the Lasso solution*

$$\hat{\boldsymbol{\beta}}_{\mathcal{I}}(\lambda) = 0 \quad \textit{and} \quad \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda) = \mathbf{X}_{\mathcal{A}}^{+}(\mathbf{y} - \lambda(\mathbf{X}_{\mathcal{A}}^{+})^{\top}\boldsymbol{\eta}_{\mathcal{A}}),$$

with $\mathcal{A} = \{i \in [d] : |\mathbf{X}_i^{\top}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda))| = \lambda\}$ *and* $\mathcal{I} = [d]\backslash\mathcal{A}$*, is continuous piecewise linear.*

We call *kink* a point where the direction $\partial\hat{\boldsymbol{\beta}}(\lambda)/\partial\lambda$ changes. The path $\{\hat{\boldsymbol{\beta}}(\lambda) : \lambda_{t+1} < \lambda \leqslant \lambda_t\}$ between two consecutive kinks λ_{t+1} and λ_t thus defines a linear segment. An important consequence of the above result is that any continuous piecewise linear regularisation path has at most 3^d linear segments, since two different linear segments have different equicorrelation signs $\boldsymbol{\eta} \in \{-1, 0, 1\}^d$. This result was improved slightly to $(3^d + 1)/2$ iterations by Mairal and Yu [MY12], who also exhibited a worst-case Lasso problem with exactly $(3^d + 1)/2$ number of linear segments.

3.2 The LARS algorithm

The Lasso solution path $\{\hat{\boldsymbol{\beta}}(\lambda) : \lambda > 0\}$ can be computed by the Least Angle Regression (LARS) algorithm[2] which was proposed and named by Efron *et al.* [EHJT04], although similar ideas have already appeared in the work of Osborne *et al.* [OPT00b, OPT00a]. Since its introduction, the LARS algorithm have been improved and generalised [RZ07, MY12, Tib13, SCL+18]. In compressed sensing literature, this algorithm is better known as the Homotopy method [FR13].

Starting at $\lambda = \infty$, where the Lasso solution is trivially $\mathbf{0} \in \mathbb{R}^d$, the LARS algorithm iteratively computes all the kinks in the Lasso solution path by decreasing the parameter λ and checking for points where the path's linear trajectory changes. This is done by making sure that the KKT conditions from Fact 16 are always satisfied. When a kink is reached, a coordinate from the solution $\hat{\boldsymbol{\beta}}(\lambda)$ will go from non-zero (being in the active set $\mathcal{A}$) to zero (join the inactive set $\mathcal{I}$), or vice-versa, i.e., from zero (leave $\mathcal{I}$) to non-zero (into $\mathcal{A}$). We shall now describe the LARS algorithm in more details, following [Tib13]. Algorithm 1 summarises the LARS algorithm.

The LARS algorithm starts, without loss of generality, at the first kink $\lambda_0 = \|\mathbf{X}^{\top}\mathbf{y}\|_{\infty}$, since $\hat{\boldsymbol{\beta}}(\lambda) = \mathbf{0}$ for $\lambda \geqslant \lambda_0$ (plug $\hat{\boldsymbol{\beta}}(\lambda_0) = \mathbf{0}$ in Eq. (7)). The corresponding active and inactive sets are $\mathcal{A} = \operatorname{argmax}_{j\in[d]} |\mathbf{X}_j^{\top}\mathbf{y}|$ and $\mathcal{I} = [d]\backslash\mathcal{A}$, respectively, while the equicorrelation sign is $\boldsymbol{\eta}_{\mathcal{A}} = \operatorname{sign}(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{y})$. Since every iteration after the initial setup is identical, assume by induction that the LARS algorithm has computed the first $t+1$ kinks $(\lambda_0, \hat{\boldsymbol{\beta}}(\lambda_0)), \ldots, (\lambda_t, \hat{\boldsymbol{\beta}}(\lambda_t))$. At the beginning of the t-th iteration, the regularisation parameter is $\lambda = \lambda_t$, the Lasso solution is $\hat{\boldsymbol{\beta}}(\lambda_t)$, and the active and inactive sets are $\mathcal{A}$ and $\mathcal{I}$. From the previous solution $\hat{\boldsymbol{\beta}}(\lambda_t)$ we obtain the equicorrelation signs

$$\boldsymbol{\eta}_{\mathcal{A}} = \operatorname{sign}\big(\mathbf{X}_{\mathcal{A}}^{\top}(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t))\big).$$

According to Eq. (8), for $\lambda \leqslant \lambda_t$ sufficiently close to λ_t, the Lasso solution is[3]

$$\hat{\boldsymbol{\beta}}_{\mathcal{I}}(\lambda) = \mathbf{0} \quad \text{and} \quad \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda) = \mathbf{X}_{\mathcal{A}}^{+}(\mathbf{y} - \lambda(\mathbf{X}_{\mathcal{A}}^{+})^{\top}\boldsymbol{\eta}_{\mathcal{A}}) = \boldsymbol{\mu} - \lambda\boldsymbol{\theta}, \tag{11}$$

where $\boldsymbol{\mu} = \mathbf{X}_{\mathcal{A}}^{+}\mathbf{y}$ and $\boldsymbol{\theta} = (\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}})^{+}\boldsymbol{\eta}_{\mathcal{A}}$. We now decrease λ while keeping Eq. (11). As λ decreases, two important checks must be made. First, we check when (i.e., we compute the next value of λ at which) a variable i_{t+1}^{join} in $\mathcal{I}$ should join the active set $\mathcal{A}$ because it has attain the maximum correlation $|\mathbf{X}_{i_{t+1}^{\text{join}}}^{\top}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})| = \lambda$. We call this event the next joining time $\lambda_{t+1}^{\text{join}}$. Second, we check

[2] By LARS algorithm we mean the version of the algorithm that computes the Lasso path and not the version that performs a kind of forward variable selection. [3] From now on, we shall assume $\mathbf{b} = \mathbf{0}$ in order to guarantee the continuity of the Lasso path.

when a variable i_{t+1}^{cross} in the active set $\mathcal{A}$ should leave $\mathcal{A}$ and join $\mathcal{I}$ because $\hat{\beta}_{i_{t+1}^{\text{cross}}}$ crossed through zero. We call this event the next crossing time $\lambda_{t+1}^{\text{cross}}$.

For the first check, for each $i \in \mathcal{I}$, consider the equation

$$\mathbf{X}_i^\top(\mathbf{y} - \mathbf{X}_\mathcal{A}\hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda)) = \pm\lambda \implies \mathbf{X}_i^\top(\mathbf{y} - \mathbf{X}_\mathcal{A}(\boldsymbol{\mu} - \lambda\boldsymbol{\theta})) = \pm\lambda.$$

The solution to the above equation for λ, interpreted as the joining time of the i-th variable, is

$$\Lambda_i^{\text{join}}(\mathcal{A}) \triangleq \frac{\mathbf{X}_i^\top(\mathbf{y} - \mathbf{X}_\mathcal{A}\boldsymbol{\mu})}{\pm 1 - \mathbf{X}_i^\top\mathbf{X}_\mathcal{A}\boldsymbol{\theta}} = \frac{\mathbf{X}_i^\top(\mathbf{I} - \mathbf{X}_\mathcal{A}\mathbf{X}_\mathcal{A}^+)\mathbf{y}}{\pm 1 - \mathbf{X}_i^\top(\mathbf{X}_\mathcal{A}^\top)^+\boldsymbol{\eta}_\mathcal{A}}, \tag{12}$$

where we used that $\mathbf{X}_\mathcal{A}\mathbf{X}_\mathcal{A}^+(\mathbf{X}_\mathcal{A}^\top)^+ = (\mathbf{X}_\mathcal{A}^\top)^+$. Only one of the possibilities $+1$ or -1 will lead to a value in the necessary interval $[0, \lambda_t]$, and the following expressions always consider that solution. Therefore, the next joining time $\lambda_{t+1}^{\text{join}}$ and coordinate i_{t+1}^{join} are

$$\lambda_{t+1}^{\text{join}} = \max_{i\in\mathcal{I}}\{\Lambda_i^{\text{join}}(\mathcal{A})\}, \quad i_{t+1}^{\text{join}} = \operatorname*{argmax}_{i\in\mathcal{I}}\{\Lambda_i^{\text{join}}(\mathcal{A})\}.$$

For the second check, for each $i \in \mathcal{A}$, we solve the equation

$$\hat{\beta}_i(\lambda) = 0 \implies \mu_i - \lambda\theta_i = 0.$$

The crossing time of the i-th variable, which is the solution to the above equation restricted to values $\lambda \leqslant \lambda_t$, is thus defined as

$$\Lambda_i^{\text{cross}}(\mathcal{A}) \triangleq \frac{\mu_i}{\theta_i} \cdot \mathbf{1}\left[\frac{\mu_i}{\theta_i} \leqslant \lambda_t\right] = \frac{[\mathbf{X}_\mathcal{A}^+\mathbf{y}]_i}{[(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^+\boldsymbol{\eta}_\mathcal{A}]_i} \cdot \mathbf{1}\left[\frac{[\mathbf{X}_\mathcal{A}^+\mathbf{y}]_i}{[(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^+\boldsymbol{\eta}_\mathcal{A}]_i} \leqslant \lambda_t\right].$$

Therefore, the next crossing time $\lambda_{t+1}^{\text{join}}$ and coordinate i_{t+1}^{join} are

$$\lambda_{t+1}^{\text{cross}} = \max_{i\in\mathcal{A}}\{\Lambda_i^{\text{cross}}(\mathcal{A})\}, \quad i_{t+1}^{\text{cross}} = \operatorname*{argmax}_{i\in\mathcal{A}}\{\Lambda_i^{\text{cross}}(\mathcal{A})\}.$$

Finally, the next kink of the regularisation path is $(\lambda_{t+1}, \hat{\boldsymbol{\beta}}(\lambda_{t+1}))$ where $\lambda_{t+1} = \max\{\lambda_{t+1}^{\text{join}}, \lambda_{t+1}^{\text{cross}}\}$ and $\hat{\boldsymbol{\beta}}_\mathcal{I}(\lambda_{t+1}) = \mathbf{0}$ and $\hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda_{t+1}) = \boldsymbol{\mu} - \lambda_{t+1}\boldsymbol{\theta}$. If $\lambda_{t+1}^{\text{join}} > \lambda_{t+1}^{\text{cross}}$, then the coordinate i_{t+1}^{join} should join the active set $\mathcal{A}$. Otherwise, the coordinate i_{t+1}^{cross} should leave the active set $\mathcal{A}$. This ends the t-th iteration of the LARS algorithm.

The time complexity of each iteration in the LARS algorithm is given by the following result.

Fact 21. *The time complexity per iteration of the* LARS *algorithm is $O(nd + |\mathcal{A}|^2)$, where $\mathcal{A}$ is the active set of the corresponding iteration.*

Proof. Let us start with the computation of $\mathbf{X}_\mathcal{A}^+$. Simply computing $\mathbf{X}_\mathcal{A}^+$ would require $O(n|\mathcal{A}|^2 + |\mathcal{A}|^3)$ time. However, the time complexity can be reduced to $O(n|\mathcal{A}| + |\mathcal{A}|^2)$ by using the Sherman–Morrison formula applied to the Moore-Penrose inverse. More specifically, suppose we have previously computed $\mathbf{X}_\mathcal{A}^+$ and the index i_{t+1}^{join} is then added to the active set $\mathcal{A}$ to obtain the new active set $\mathcal{A}' = \mathcal{A} \cup \{i_{t+1}^{\text{join}}\}$. Without lost of generality, write $\mathcal{A} = \{j_1, \ldots, j_k\}$ and $\mathcal{A}' = \{j_1, \ldots, j_k, i_{t+1}^{\text{join}}\}$, where $k = |\mathcal{A}|$. Let $\mathbf{A} \in \mathbb{R}^{n\times(|\mathcal{A}|+1)}$ be the matrix $\mathbf{X}_\mathcal{A}$ augmented with a zero column at the position corresponding to i_{t+1}^{join}, i.e., $\mathbf{A}_l = \mathbf{X}_{j_l}$ if $l \in [k]$, and $\mathbf{A}_l = \mathbf{0}$ if $l = k+1$. The augmented matrix is written using a rank-1 update as

$$\mathbf{X}_{\mathcal{A}'} = \mathbf{A} + \mathbf{X}_{i_{t+1}^{\text{join}}}\mathbf{e}_{k+1}^\top,$$

Algorithm 1: Classical LARS algorithm for the pathwise Lasso

Input: Vector $\mathbf{y} \in \mathbb{R}^n$ and matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$

Initialise $\mathcal{A} = \operatorname{argmax}_{j \in [d]} |\mathbf{X}_j^\top \mathbf{y}|$, $\mathcal{I} = [d] \backslash \mathcal{A}$, $\lambda_0 = \|\mathbf{X}^\top \mathbf{y}\|_\infty$, $\hat{\boldsymbol{\beta}}(\lambda_0) = \mathbf{0}$, $t = 0$

while $\mathcal{I} \neq \varnothing$ **do**

- $\boldsymbol{\eta}_\mathcal{A} \leftarrow \operatorname{sign}\left(\mathbf{X}_\mathcal{A}^\top (\mathbf{y} - \mathbf{X}_\mathcal{A} \hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda_t))\right)$
- $\boldsymbol{\mu} \leftarrow \mathbf{X}_\mathcal{A}^+ \mathbf{y}$ // $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{A}|}$
- $\boldsymbol{\theta} \leftarrow (\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^+ \boldsymbol{\eta}_\mathcal{A}$ // $\boldsymbol{\theta} \in \mathbb{R}^{|\mathcal{A}|}$
- $\Lambda_i^{\text{cross}} \leftarrow \frac{\mu_i}{\theta_i} \cdot \mathbf{1}\left[\frac{\mu_i}{\theta_i} \leqslant \lambda_t\right] \quad \forall i \in \mathcal{A}$
- $\Lambda_i^{\text{join}} \leftarrow \frac{\mathbf{X}_i^\top (\mathbf{y} - \mathbf{X}_\mathcal{A} \boldsymbol{\mu})}{\pm 1 - \mathbf{X}_i^\top \mathbf{X}_\mathcal{A} \boldsymbol{\theta}} \quad \forall i \in \mathcal{I}$
- $i_{t+1}^{\text{cross}} \leftarrow \operatorname{argmax}_{i \in \mathcal{A}} \{\Lambda_i^{\text{cross}}\}$ and $\lambda_{t+1}^{\text{cross}} \leftarrow \max_{i \in \mathcal{A}} \{\Lambda_i^{\text{cross}}\}$
- $i_{t+1}^{\text{join}} \leftarrow \operatorname{argmax}_{i \in \mathcal{I}} \{\Lambda_i^{\text{join}}\}$ and $\lambda_{t+1}^{\text{join}} \leftarrow \max_{i \in \mathcal{I}} \{\Lambda_i^{\text{join}}\}$
- $\lambda_{t+1} \leftarrow \max\{\lambda_{t+1}^{\text{join}}, \lambda_{t+1}^{\text{cross}}\}$
- $\hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda_{t+1}) \leftarrow \boldsymbol{\mu} - \lambda_{t+1} \boldsymbol{\theta}$
- **if** $\lambda_{t+1} = \lambda_{t+1}^{\text{cross}}$ **then**
  - Move i_{t+1}^{cross} from $\mathcal{A}$ to $\mathcal{I}$
- **else**
  - Move i_{t+1}^{join} from $\mathcal{I}$ to $\mathcal{A}$
- $t \leftarrow t + 1$

Output: Coefficients $[(\lambda_0, \hat{\boldsymbol{\beta}}(\lambda_0)), (\lambda_1, \hat{\boldsymbol{\beta}}(\lambda_1)), \dots]$

where $\mathbf{e}_l \in \{0,1\}^{|\mathcal{A}|+1}$ is the column vector with 1 in position l and 0 elsewhere. Therefore, by using the Sherman–Morrison formula applied to the Moore-Penrose inverse (see e.g. [MJ73]), the inverse $\mathbf{X}_{\mathcal{A}'}^+$ can be computed from the matrix $\mathbf{X}_\mathcal{A}^+$ and the vectors $\mathbf{X}_{i_{t+1}^{\text{join}}}$ and $\mathbf{e}_{k+1}$ using simple matrix and vector multiplication, which requires $O(n|\mathcal{A}| + |\mathcal{A}|^2)$ time. Regarding the other quantities,

- Computing $\boldsymbol{\eta}_\mathcal{A}$, $\boldsymbol{\mu}$, and $\boldsymbol{\theta}$ requires $O(n|\mathcal{A}| + |\mathcal{A}|^2)$ time (the updating argument can be used as well for $(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^+$);
- Finding i_{t+1}^{cross} and $\lambda_{t+1}^{\text{cross}}$ requires $O(|\mathcal{A}|)$ time;
- Finding i_{t+1}^{join} and $\lambda_{t+1}^{\text{join}}$ requires $O(n|\mathcal{I}|)$ time (the computations $\mathbf{y} - \mathbf{X}_\mathcal{A}\boldsymbol{\mu}$ and $\mathbf{X}_\mathcal{A}\boldsymbol{\theta}$ must be performed just once);
- Updating $\hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda_{t+1})$ requires $O(|\mathcal{A}|)$ time.

In total, a single iteration requires $O(n|\mathcal{I}| + n|\mathcal{A}| + |\mathcal{A}|^2) = O(nd + |\mathcal{A}|^2)$ time. □

Regarding the number of iterations taken by the LARS algorithm, heuristically, it is on average $O(n)$ [RZ07]. This can be understood as follows. If $n > d$, it would take $O(d)$ steps to add all the variables. If $n < d$, then we only add at most n variables in the case of a unique solution, since $|\mathcal{A}| \leqslant \min\{n, d\}$. Dropping variables is rare as λ is successfully decreased, usually $O(1)$ times. In the worst case, though, the number of iterations can be at most $(3^d + 1)/2$ as previously mentioned.

4 Quantum algorithms

In this section, we introduce our quantum algorithms based on the LARS algorithm. In Section 4.1, we propose our simple quantum LARS algorithm that exactly computes the pathwise Lasso, while in Section 4.2, we improve upon this simple algorithm and propose the approximate quantum LARS algorithm. In Section 4.3, we dequantise the approximate LARS algorithm using sampling techniques. Finally, in Section 4.4, we analyse the algorithms' complexity for the case when $\mathbf{X}$ is a standard Gaussian random matrix.

4.1 Simple quantum LARS algorithm

As a warm-up, we propose a simple quantum LARS algorithm for the Lasso path by quantising the search step of i_{t+1}^{join} and $\lambda_{t+1}^{\text{join}}$ using the quantum minimum-finding subroutine from Dürr and Høyer [DH96] (Fact 10). For this, we input the vectors $\mathbf{y} - \mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}$ and $\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}$ into KP-tree data structures accessed by QRAMs. Moreover, as described in Section 2.3, we assume the matrix $\mathbf{X}$ is stored in a QROM. The final result is an improvement in the runtime over the usual LARS algorithm to $\widetilde{O}(n\sqrt{|\mathcal{I}|} + n|\mathcal{A}| + |\mathcal{A}|^2)$ per iteration.

Algorithm 2: Simple quantum LARS algorithm for the pathwise Lasso

Input: Vector $\mathbf{y} \in \mathbb{R}^n$, matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\delta \in (0,1)$, and $T \in \mathbb{N}$
Initialise $\mathcal{A} = \operatorname{argmax}_{j\in[d]} |\mathbf{X}_j^\top \mathbf{y}|$, $\mathcal{I} = [d]\backslash\mathcal{A}$, $\lambda_0 = \|\mathbf{X}^\top\mathbf{y}\|_\infty$, $\hat{\boldsymbol{\beta}}(\lambda_0) = \mathbf{0}$, $t = 0$
while $\mathcal{I} \neq \varnothing$, $t \leqslant T$ **do**
  $\boldsymbol{\eta}_{\mathcal{A}} \leftarrow \operatorname{sign}\left(\mathbf{X}_{\mathcal{A}}^\top(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t))\right)$
  $\boldsymbol{\mu} \leftarrow \mathbf{X}_{\mathcal{A}}^{+}\mathbf{y}$ // $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{A}|}$
  $\boldsymbol{\theta} \leftarrow (\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{+}\boldsymbol{\eta}_{\mathcal{A}}$ // $\boldsymbol{\theta} \in \mathbb{R}^{|\mathcal{A}|}$
  Compute $\mathbf{y} - \mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}$ and $\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}$ classically and input them into QRAMs
  Define $\Lambda_i^{\text{join}} \triangleq \frac{\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu})}{\pm 1 - \mathbf{X}_i^\top\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}}$ and $\Lambda_i^{\text{cross}} \triangleq \frac{\mu_i}{\theta_i}\cdot\mathbf{1}\left[\frac{\mu_i}{\theta_i} \leqslant \lambda_t\right]$
  $i_{t+1}^{\text{cross}} \leftarrow \operatorname{argmax}_{i\in\mathcal{A}}\{\Lambda_i^{\text{cross}}\}$ and $\lambda_{t+1}^{\text{cross}} \leftarrow \max_{i\in\mathcal{A}}\{\Lambda_i^{\text{cross}}\}$
  $i_{t+1}^{\text{join}} \leftarrow \operatorname{argmax}_{i\in\mathcal{I}}\{\Lambda_i^{\text{join}}\}$ and $\lambda_{t+1}^{\text{join}} \leftarrow \max_{i\in\mathcal{I}}\{\Lambda_i^{\text{join}}\}$ with failure probability $\frac{\delta}{T}$ (Fact 10)
  $\lambda_{t+1} \leftarrow \max\{\lambda_{t+1}^{\text{join}}, \lambda_{t+1}^{\text{cross}}\}$
  $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_{t+1}) \leftarrow \boldsymbol{\mu} - \lambda_{t+1}\boldsymbol{\theta}$
  if $\lambda_{t+1} = \lambda_{t+1}^{\text{cross}}$ **then**
    Move i_{t+1}^{cross} from $\mathcal{A}$ to $\mathcal{I}$
  else
    Move i_{t+1}^{join} from $\mathcal{I}$ to $\mathcal{A}$
  $t \leftarrow t + 1$
Output: Coefficients $[(\lambda_0, \hat{\boldsymbol{\beta}}(\lambda_0)), (\lambda_1, \hat{\boldsymbol{\beta}}(\lambda_1)), \dots]$

Theorem 22. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Assume $\mathbf{X}$ is stored in a* QROM *and we have access to* QRAMs *of size $O(n)$. Let $\delta \in (0,1)$ and $T \in \mathbb{N}$. The simple quantum* LARS *algorithm 2 outputs, with probability at least $1-\delta$, at most T kinks of the optimal regularisation path in time*

$$O(n\sqrt{|\mathcal{I}|}\log(T/\delta)\operatorname{poly}\log(nd) + n|\mathcal{A}| + |\mathcal{A}|^2)$$

per iteration, where $\mathcal{A}$ and $\mathcal{I}$ are the active and inactive sets of the corresponding iteration.

Proof. The time complexity of Algorithm 2 is basically the same as Algorithm 1, the only difference being that uploading $\mathbf{y} - \mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}$ and $\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}$ into KP-trees takes $O(n \log n)$ time, and finding i_{t+1}^{join} and $\lambda_{t+1}^{\text{join}}$ now requires $O(n\sqrt{|\mathcal{I}|}\log(T/\delta)\operatorname{poly}\log(nd))$ time, since that accessing $\mathbf{X}_i$ and computing $\Lambda_i^{\text{join}}(\mathcal{A}) = \frac{\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu})}{\pm 1 - \mathbf{X}_i^\top \mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}}$ requires $O(n \operatorname{poly}\log(nd))$ and $O(n)$ time, respectively. □

4.2 Approximate quantum LARS algorithm

We now improve our previous simple quantum LARS algorithm. The main idea is to *approximately* compute the joining times $\Lambda_i^{\text{join}}(\mathcal{A})$, see Eq. (12), within the quantum minimum-finding algorithm instead of exactly computing them. This will lead to a quadratic improvement on the n dependence in the term $\widetilde{O}(n\sqrt{|\mathcal{I}|})$ to $\widetilde{O}(\sqrt{n|\mathcal{I}|})$. Such approximation, however, hinders our ability to exactly find the joining variables i_{t+1}^{join} and points $\lambda_{t+1}^{\text{join}}$. We can now only obtain a joining point $\widetilde{\lambda}_{t+1}^{\text{join}}$ which is ϵ-close to the true joining point $\lambda_{t+1}^{\text{join}}$. This imposes new complications on the correctness analysis of the LARS algorithm, since we can no longer guarantee that the KKT conditions are satisfied. In order to tackle this issue, we consider an approximate version of the KKT conditions. We note that a similar concept was already introduced in [MY12].

Definition 23. *Let $\epsilon \geqslant 0$. A vector $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^d$ satisfies the $\mathsf{KKT}_\lambda(\epsilon)$ condition if and only if, $\forall j \in [d]$,*

$$\begin{aligned} \lambda(1-\epsilon) \leqslant \mathbf{X}_j^\top(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})\operatorname{sign}(\widetilde{\beta}_j) \leqslant \lambda \quad & \text{if } \widetilde{\beta}_j \neq 0, \\ |\mathbf{X}_j^\top(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})| \leqslant \lambda \quad & \text{if } \widetilde{\beta}_j = 0. \end{aligned}$$

The reason for introducing the above approximate version of the KKT conditions is that it leads to approximate Lasso solutions, as proven in the next lemma.

Lemma 24. *If a vector $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^d$ satisfies the $\mathsf{KKT}_\lambda(\epsilon)$ condition for $\epsilon \geqslant 0$, then $\widetilde{\boldsymbol{\beta}}$ minimises the Lasso cost function up to an error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}\|_1$, i.e.,*

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widetilde{\boldsymbol{\beta}}\|_1 - \left(\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\right) \leqslant \lambda\epsilon\|\widetilde{\boldsymbol{\beta}}\|_1.$$

Proof. The primal problem $\min_{\boldsymbol{\beta}\in\mathbb{R}^d, \mathbf{z}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{y}-\mathbf{z}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \quad \text{s.t. } \mathbf{z} = \mathbf{X}\boldsymbol{\beta}$, with dummy variable $\mathbf{z}$, leads to the dual problem

$$\max_{\boldsymbol{\kappa}\in\mathbb{R}^n} -\frac{1}{2}\boldsymbol{\kappa}^\top\boldsymbol{\kappa} - \boldsymbol{\kappa}^\top\mathbf{y} \quad \text{s.t. } \|\mathbf{X}^\top\boldsymbol{\kappa}\|_\infty \leqslant \lambda.$$

It is known that, given a pair of feasible primal and dual variables $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\kappa}})$, the difference

$$\delta_\lambda(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\kappa}}) \triangleq \left(\frac{1}{2}\|\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widetilde{\boldsymbol{\beta}}\|_1\right) - \left(-\frac{1}{2}\widetilde{\boldsymbol{\kappa}}^\top\widetilde{\boldsymbol{\kappa}} - \widetilde{\boldsymbol{\kappa}}^\top\mathbf{y}\right)$$

is called a duality gap and provides an optimality guarantee (see e.g. [BL06, Section 4.3]):

$$0 \leqslant \left(\frac{1}{2}\|\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widetilde{\boldsymbol{\beta}}\|_1\right) - \left(\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\right) \leqslant \delta_\lambda(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\kappa}}).$$

The vector $\widetilde{\boldsymbol{\kappa}} = \mathbf{X}\widetilde{\boldsymbol{\beta}} - \mathbf{y}$ is feasible for the dual problem since $\widetilde{\boldsymbol{\beta}}$ satisfies the $\mathsf{KKT}_\lambda(\epsilon)$ condition ($\|\mathbf{X}^\top\widetilde{\boldsymbol{\kappa}}\|_\infty = \|\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})\|_\infty \leqslant \lambda$). We can thus compute the duality gap,

$$\begin{aligned} \delta_\lambda(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\kappa}}) &= \left(\frac{1}{2}\|\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widetilde{\boldsymbol{\beta}}\|_1\right) - \left(-\frac{1}{2}\widetilde{\boldsymbol{\kappa}}^\top\widetilde{\boldsymbol{\kappa}} - \widetilde{\boldsymbol{\kappa}}^\top\mathbf{y}\right) \\ &= \frac{1}{2}\widetilde{\boldsymbol{\kappa}}^\top\widetilde{\boldsymbol{\kappa}} + \lambda\|\widetilde{\boldsymbol{\beta}}\|_1 + \frac{1}{2}\widetilde{\boldsymbol{\kappa}}^\top\widetilde{\boldsymbol{\kappa}} + \widetilde{\boldsymbol{\kappa}}^\top\mathbf{y} \\ &= \lambda\|\widetilde{\boldsymbol{\beta}}\|_1 + (\widetilde{\boldsymbol{\kappa}} + \mathbf{y})^\top\widetilde{\boldsymbol{\kappa}}, \end{aligned}$$

but

$$(\widetilde{\boldsymbol{\kappa}}+\mathbf{y})^\top\widetilde{\boldsymbol{\kappa}}=\widetilde{\boldsymbol{\beta}}^\top\mathbf{X}^\top(\mathbf{X}\widetilde{\boldsymbol{\beta}}-\mathbf{y})=\sum_{i=1}^{d}|\widetilde{\beta}_i|\mathbf{X}_i^\top(\mathbf{X}\widetilde{\boldsymbol{\beta}}-\mathbf{y})\operatorname{sign}(\widetilde{\beta}_i)\leqslant-(\lambda-\lambda\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1,$$

using that $\widetilde{\boldsymbol{\beta}}$ satisfies the $\mathsf{KKT}_\lambda(\epsilon)$ condition. Therefore $\delta_\lambda(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\kappa}})\leqslant\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}\|_1$. □

Algorithm 3: Approximate quantum LARS algorithm for the pathwise Lasso

Input: Vector $\mathbf{y}\in\mathbb{R}^n$, matrix $\mathbf{X}\in\mathbb{R}^{n\times d}$, $\delta\in(0,1)$, $\epsilon\in(0,2)$, and $T\in\mathbb{N}$

Initialise $\mathcal{A}=\operatorname{argmax}_{j\in[d]}|\mathbf{X}_j^\top\mathbf{y}|$, $\mathcal{I}=[d]\backslash\mathcal{A}$, $\lambda_0=\|\mathbf{X}^\top\mathbf{y}\|_\infty$, $\widetilde{\boldsymbol{\beta}}(\lambda_0)=0$, $t=0$

while $\mathcal{I}\neq\varnothing$, $t\leqslant T$ **do**

- $\boldsymbol{\eta}_\mathcal{A}\leftarrow\frac{1}{\lambda_t}\mathbf{X}_\mathcal{A}^\top(\mathbf{y}-\mathbf{X}_\mathcal{A}\widetilde{\boldsymbol{\beta}}_\mathcal{A}(\lambda_t))$
- $\boldsymbol{\mu}\leftarrow\mathbf{X}_\mathcal{A}^+\mathbf{y}$ // $\boldsymbol{\mu}\in\mathbb{R}^{|\mathcal{A}|}$
- $\boldsymbol{\theta}\leftarrow(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^+\boldsymbol{\eta}_\mathcal{A}$ // $\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{A}|}$
- Classically compute $\mathbf{y}-\mathbf{X}_\mathcal{A}\boldsymbol{\mu}$ and $\mathbf{X}_\mathcal{A}\boldsymbol{\theta}$ and input them into QRAMs and KP-trees
- Define $\Lambda_i^{\text{join}}\triangleq\frac{\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_\mathcal{A}\boldsymbol{\mu})}{\pm1-\mathbf{X}_i^\top\mathbf{X}_\mathcal{A}\boldsymbol{\theta}}$ and $\Lambda_i^{\text{cross}}\triangleq\frac{\mu_i}{\theta_i}\cdot\mathbf{1}\left[\frac{\mu_i}{\theta_i}\leqslant\lambda_t\right]$
- $i_{t+1}^{\text{cross}}\leftarrow\operatorname{argmax}_{i\in\mathcal{A}}\{\Lambda_i^{\text{cross}}\}$ and $\lambda_{t+1}^{\text{cross}}\leftarrow\max_{i\in\mathcal{A}}\{\Lambda_i^{\text{cross}}\}$
- Obtain $\widetilde{i}_{t+1}^{\text{join}}\in\left\{j\in\mathcal{I}:\Lambda_j^{\text{join}}\geqslant(1-\epsilon/2)\max_{i\in\mathcal{I}}\{\Lambda_i^{\text{join}}\}\right\}$ with failure probability $\frac{\delta}{T}$ (Fact 11 and Lemma 13)
- $\widetilde{\lambda}_{t+1}^{\text{join}}\leftarrow(1-\epsilon/2)^{-1}\Lambda_{\widetilde{i}_{t+1}^{\text{join}}}^{\text{join}}$
- $\lambda_{t+1}\leftarrow\min\{\lambda_t,\max\{\lambda_{t+1}^{\text{cross}},\widetilde{\lambda}_{t+1}^{\text{join}}\}\}$
- $\widetilde{\boldsymbol{\beta}}_\mathcal{A}(\lambda_{t+1})\leftarrow\boldsymbol{\mu}-\lambda_{t+1}\boldsymbol{\theta}$
- **if** $\lambda_{t+1}=\lambda_{t+1}^{\text{cross}}$ **then**
  - Move i_{t+1}^{cross} from $\mathcal{A}$ to $\mathcal{I}$
- **else**
  - Move $\widetilde{i}_{t+1}^{\text{join}}$ from $\mathcal{I}$ to $\mathcal{A}$
- $t\leftarrow t+1$

Output: Coefficients $[(\lambda_0,\widetilde{\boldsymbol{\beta}}(\lambda_0)),(\lambda_1,\widetilde{\boldsymbol{\beta}}(\lambda_1)),\dots]$

We now present our approximate quantum LARS algorithm 3 for the pathwise Lasso. Its main idea is quite simple: we improve the search of the joining variable i_{t+1}^{join} and time $\lambda_{t+1}^{\text{join}}$ over $i\in\mathcal{I}$ by approximately computing the joining times $\Lambda_i^{\text{join}}(\mathcal{A})=\frac{\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_\mathcal{A}\boldsymbol{\mu})}{\pm1-\mathbf{X}_i^\top\mathbf{X}_\mathcal{A}\boldsymbol{\theta}}$ and using the approximate quantum minimum-finding subroutine from Chen and de Wolf [CdW23] (Fact 11). More specifically, the joining variable $\widetilde{i}_{t+1}^{\text{join}}$ is obtained to be any variable $j\in\mathcal{I}$ that maximises $\Lambda_j^{\text{join}}(\mathcal{A})$ up to some small relative error ϵ, i.e., at any iteration, Algorithm 3 randomly samples from the set

$$\left\{j\in\mathcal{I}:\Lambda_j^{\text{join}}(\mathcal{A})\geqslant(1-\epsilon/2)\max_{i\in\mathcal{I}}\{\Lambda_i^{\text{join}}(\mathcal{A})\}\right\}. \tag{13}$$

Since the corresponding joining time $\Lambda_{\widetilde{i}_{t+1}^{\text{join}}}^{\text{join}}(\mathcal{A})$ can be smaller than the true joining time $\lambda_{t+1}^{\text{join}}=\max_{i\in\mathcal{I}}\{\Lambda_i^{\text{join}}(\mathcal{A})\}$, we rescale it by the factor $(1-\epsilon/2)^{-1}$ and set $\widetilde{\lambda}_{t+1}^{\text{join}}=(1-\epsilon/2)^{-1}\Lambda_{\widetilde{i}_{t+1}^{\text{join}}}^{\text{join}}(\mathcal{A})$. This ensures that $\widetilde{\lambda}_{t+1}^{\text{join}}\geqslant\lambda_{t+1}^{\text{join}}$ and consequently that the current iteration stops before $\lambda_{t+1}^{\text{join}}$, which guarantees that the approximate KKT conditions are satisfied as shown in Theorem 25 below.

We notice that, since the joining variable is randomly sampled from the set in Eq. (13), it might be possible that $\widetilde{\lambda}^{\text{join}}_{t+2} > \widetilde{\lambda}^{\text{join}}_{t+1}$, i.e., the joining time at some later iteration is greater than the joining time at some earlier iteration. While this is counter to the original LARS algorithm, where the regularisation parameter λ always decreases, there is in principle no problem in allowing λ to increase within a small interval as long as the approximate KKT conditions are satisfied. Nonetheless, in order to avoid redundant segments in the Lasso path, we set the next iteration's regularisation parameter λ_{t+1} as the minimum between the current λ_t and $\max\{\lambda^{\text{cross}}_{t+1}, \widetilde{\lambda}^{\text{join}}_{t+1}\}$. The solution path can thus stay "stationary" for a few iterations. Ideally, one could just move all the variables from the set in Eq. (13) into $\mathcal{A}$ at once and the result would be the same. This is reminiscent to the approximate homotopy algorithm of Mairal and Yu [MY12].

We show next that the imprecision in finding $\lambda^{\text{join}}_{t+1}$ leads to a path that satisfies the approximate KKT conditions from Definition 23 and thus approximates the Lasso function up to a small error.

Theorem 25. *Let $\epsilon \in [0, 2)$. Consider an approximate* LARS *algorithm (e.g. Algorithm 3) that returns a solution path $\widetilde{\mathcal{P}} = \{\widetilde{\boldsymbol{\beta}}(\lambda) \in \mathbb{R}^d : \lambda > 0\}$ and wherein, at each iteration t, the joining variable $\widetilde{i}^{\text{join}}_{t+1}$ is taken from the set,*

$$\left\{ j \in \mathcal{I} : \Lambda^{\text{join}}_j(\mathcal{A}) \geqslant (1 - \epsilon/2) \max_{i \in \mathcal{I}} \{\Lambda^{\text{join}}_i(\mathcal{A})\} \right\},$$

where $\Lambda^{\text{join}}_j(\mathcal{A}) \triangleq \frac{\mathbf{X}_j^\top(\mathbf{y} - \mathbf{X}_\mathcal{A}\boldsymbol{\mu})}{\pm 1 - \mathbf{X}_j^\top \mathbf{X}_\mathcal{A}\boldsymbol{\theta}}$, and the corresponding joining time $\widetilde{\lambda}^{\text{join}}_{t+1}$ is

$$\widetilde{\lambda}^{\text{join}}_{t+1} = (1 - \epsilon/2)^{-1} \Lambda^{\text{join}}_{\widetilde{i}^{\text{join}}_{t+1}}(\mathcal{A}),$$

where $\mathcal{A}$ and $\mathcal{I}$ are the active and inactive sets at the corresponding iteration t and

$$\boldsymbol{\eta}_\mathcal{A} = \frac{1}{\lambda_t} \mathbf{X}_\mathcal{A}^\top (\mathbf{y} - \mathbf{X}_\mathcal{A} \widetilde{\boldsymbol{\beta}}_\mathcal{A}(\lambda_t)), \qquad \boldsymbol{\mu} = \mathbf{X}_\mathcal{A}^+ \mathbf{y}, \qquad \boldsymbol{\theta} = (\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^+ \boldsymbol{\eta}_\mathcal{A}.$$

Then $\widetilde{\mathcal{P}}$ is a continuous and piecewise linear approximate regularisation path with error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}(\lambda)\|_1$.

Proof. We shall prove that the Lasso solution $\widetilde{\boldsymbol{\beta}}(\lambda)$ satisfies the $\mathsf{KKT}_\lambda(\epsilon)$ condition for all $\lambda > 0$. According to Lemma 24, $\widetilde{\mathcal{P}}$ is then an approximate regularisation path with error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}(\lambda)\|_1$.

The proof is by induction on the iteration loop t. The case $t = 0$ is trivial. Assume then that the computed path through iteration $t - 1$ satisfies $\mathsf{KKT}_\lambda(\epsilon)$ for all $\lambda \geqslant \lambda_t$. Consider the t-th iteration with active set $\mathcal{A} = \{i \in [d] : \widetilde{\beta}_i \neq 0\}$ and equicorrelation signs $\boldsymbol{\eta}_\mathcal{A} = \frac{1}{\lambda_t} \mathbf{X}_\mathcal{A}^\top (\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}(\lambda_t))$. The induction hypothesis implies that the current Lasso solution $\widetilde{\boldsymbol{\beta}}(\lambda_t)$ satisfies $\mathsf{KKT}_{\lambda_t}(\epsilon)$ at λ_t, i.e.,

$$\begin{aligned} \lambda_t(1 - \epsilon) \leqslant \mathbf{X}_j^\top (\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}(\lambda_t)) \operatorname{sign}(\widetilde{\beta}_j(\lambda_t)) \leqslant \lambda_t \quad &\text{if } \widetilde{\beta}_j \neq 0, \\ |\mathbf{X}_j^\top (\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}(\lambda_t))| \leqslant \lambda_t \quad &\text{if } \widetilde{\beta}_j = 0. \end{aligned}$$

Recall that for $\lambda \leqslant \lambda_t$ the Lasso solution is

$$\widetilde{\boldsymbol{\beta}}_\mathcal{I}(\lambda) = \mathbf{0} \quad \text{and} \quad \widetilde{\boldsymbol{\beta}}_\mathcal{A}(\lambda) = \mathbf{X}_\mathcal{A}^+ (\mathbf{y} - \lambda (\mathbf{X}_\mathcal{A}^+)^\top \boldsymbol{\eta}_\mathcal{A}).$$

Thus

$$\begin{aligned} \mathbf{X}_\mathcal{A}^\top (\mathbf{y} - \mathbf{X}_\mathcal{A} \widetilde{\boldsymbol{\beta}}_\mathcal{A}(\lambda)) &= \mathbf{X}_\mathcal{A}^\top \mathbf{y} - \mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A} \mathbf{X}_\mathcal{A}^+ \mathbf{y} + \lambda \mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A} (\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^+ \boldsymbol{\eta}_\mathcal{A} \\ &= \lambda \mathbf{X}_\mathcal{A}^\top (\mathbf{X}_\mathcal{A}^\top)^+ \boldsymbol{\eta}_\mathcal{A} \\ &= \lambda \boldsymbol{\eta}_\mathcal{A}, \end{aligned}$$

where we used the identity $\mathbf{A}^\top \mathbf{A}\mathbf{A}^+ = \mathbf{A}^\top$ and the last equality holds as $\boldsymbol{\eta}_{\mathcal{A}} \in \mathrm{row}(\mathbf{X}_{\mathcal{A}})$. Since

$$\eta_j \,\mathrm{sign}(\widetilde{\beta}_j(\lambda)) = \eta_j \,\mathrm{sign}(\widetilde{\beta}_j(\lambda_t)) = \frac{1}{\lambda_t}\mathbf{X}_j^\top(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}(\lambda_t))\,\mathrm{sign}(\widetilde{\beta}_j(\lambda_t))$$

for λ sufficiently close to λ_t and since $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t)$ satisfies $\mathsf{KKT}_{\lambda_t}(\epsilon)$, we conclude that

$$\lambda(1-\epsilon) \leqslant \mathbf{X}_j^\top(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}(\lambda))\,\mathrm{sign}(\widetilde{\beta}_j(\lambda)) \leqslant \lambda$$

for $j \in \mathcal{A}$ as required. Therefore, as λ decreases, one of the following conditions must break: either $\|\mathbf{X}_{\mathcal{I}}^\top(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}(\lambda))\|_\infty \leqslant \lambda$ or $\eta_j \neq 0$ for all $j \in \mathcal{A}$. The first breaks at the next joining time $\lambda_{t+1}^{\mathrm{join}} = \max_{i\in\mathcal{I}}\{\Lambda_i^{\mathrm{join}}(\mathcal{A})\}$, and second breaks at the next crossing time $\lambda_{t+1}^{\mathrm{cross}} = \max_{i\in\mathcal{A}}\{\Lambda_i^{\mathrm{cross}}(\mathcal{A})\}$. Since we only decrease λ to $\lambda_{t+1} = \min\{\lambda_t, \max\{\lambda_{t+1}^{\mathrm{cross}}, \widetilde{\lambda}_{t+1}^{\mathrm{join}}\}\}$, we have verified that $\widetilde{\boldsymbol{\beta}}(\lambda)$ satisfies $\mathsf{KKT}_\lambda(\epsilon)$ for $\lambda \geqslant \lambda_{t+1}$.

We now prove that adding or deleting variables from the active set preserves the condition $\mathsf{KKT}_{\lambda_{t+1}}(\epsilon)$ at λ_{t+1}. More specifically, let $\mathcal{A}^*$ and $\boldsymbol{\eta}_{\mathcal{A}^*}^*$ denote the active set and equicorrelation signs at the beginning of the $(t+1)$-th iteration.

Case 1 (Deletion): Let us start with the case when a variable leaves the active set at $\lambda_{t+1} = \lambda_{t+1}^{\mathrm{cross}}$, i.e., $\mathcal{A}^* = \mathcal{A}\backslash\{i_{t+1}^{\mathrm{cross}}\}$ is formed by removing an element from $\mathcal{A}$. The Lasso solution before deletion with equicorrelation signs $\boldsymbol{\eta}_{\mathcal{A}}$ is

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_{t+1}) = \begin{bmatrix} \widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1}) \\ \widetilde{\beta}_{i_{t+1}^{\mathrm{cross}}}(\lambda_{t+1}) \end{bmatrix} = \begin{bmatrix} \big[\mathbf{X}_{\mathcal{A}}^+(\mathbf{y} - \lambda_{t+1}(\mathbf{X}_{\mathcal{A}}^+)^\top\boldsymbol{\eta}_{\mathcal{A}})\big]_{\mathcal{A}^*} \\ 0 \end{bmatrix}$$

since the variable i_{t+1}^{cross} crosses through zero at $\lambda_{t+1} = \lambda_{t+1}^{\mathrm{cross}}$. On the other hand, the Lasso solution after deletion is

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}^*(\lambda_{t+1}) = \begin{bmatrix} \widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}^*(\lambda_{t+1}) \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{\mathcal{A}^*}^+(\mathbf{y} - \lambda_{t+1}(\mathbf{X}_{\mathcal{A}^*}^+)^\top\boldsymbol{\eta}_{\mathcal{A}^*}^*) \\ 0 \end{bmatrix},$$

where $\boldsymbol{\eta}_{\mathcal{A}^*}^* = \frac{1}{\lambda_{t+1}}\mathbf{X}_{\mathcal{A}^*}^\top(\mathbf{y} - \mathbf{X}_{\mathcal{A}^*}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1}))$. Therefore

$$\begin{aligned}\mathbf{X}_{\mathcal{A}^*}^+(\mathbf{y} - \lambda_{t+1}(\mathbf{X}_{\mathcal{A}^*}^+)^\top\boldsymbol{\eta}_{\mathcal{A}^*}^*) &= \mathbf{X}_{\mathcal{A}^*}^+\mathbf{y} - \mathbf{X}_{\mathcal{A}^*}^+(\mathbf{X}_{\mathcal{A}^*}^+)^\top\mathbf{X}_{\mathcal{A}^*}^\top\mathbf{y} + \mathbf{X}_{\mathcal{A}^*}^+(\mathbf{X}_{\mathcal{A}^*}^+)^\top\mathbf{X}_{\mathcal{A}^*}^\top\mathbf{X}_{\mathcal{A}^*}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1}) \\ &= \mathbf{X}_{\mathcal{A}^*}^+\mathbf{X}_{\mathcal{A}^*}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1}),\end{aligned}$$

where we used the identity $\mathbf{A}^+(\mathbf{A}^+)^\top\mathbf{A}^\top = \mathbf{A}^+$. Therefore, the solution $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1})$ to the above equation must be

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1}) = \mathbf{X}_{\mathcal{A}^*}^+(\mathbf{y} - \lambda_{t+1}(\mathbf{X}_{\mathcal{A}^*}^+)^\top\boldsymbol{\eta}_{\mathcal{A}^*}^*) + \mathbf{b} = \widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}^*(\lambda_{t+1}) + \mathbf{b},$$

where $\mathbf{b} \in \mathrm{null}(\mathbf{X}_{\mathcal{A}^*})$. Since $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1})$ must have minimum ℓ_2-norm, $\mathbf{b} = \mathbf{0}$ and so

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}^*(\lambda_{t+1}) = \begin{bmatrix} \widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}^*(\lambda_{t+1}) \\ 0 \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1}) \\ 0 \end{bmatrix},$$

which implies that $\widetilde{\mathcal{P}}$ is continuous at λ_{t+1}. Finally,

$$\mathbf{X}_j^\top(\mathbf{y} - \mathbf{X}_{\mathcal{A}^*}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}^*(\lambda_{t+1})) = \mathbf{X}_j^\top(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_{t+1}))$$

for all $j \in [d]$, and so $\widetilde{\boldsymbol{\beta}}^*(\lambda_{t+1})$ satisfies the $\mathsf{KKT}_{\lambda_{t+1}}(\epsilon)$ condition.

Case 2 (Insertion): Now we look at the case when a variable joins the active set at $\lambda_{t+1} = \min\{\lambda_t, \widetilde{\lambda}^{\text{join}}_{t+1}\}$, i.e., $\mathcal{A}^* = \mathcal{A} \cup \{\widetilde{i}^{\text{join}}_{t+1}\}$ is formed by adding an element to $\mathcal{A}$. The Lasso solution before insertion is

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1}) = \begin{bmatrix} \widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_{t+1}) \\ 0 \end{bmatrix},$$

while the Lasso solution after insertion with equicorrelation signs

$$\boldsymbol{\eta}^*_{\mathcal{A}^*} = \frac{1}{\lambda_{t+1}} \mathbf{X}^\top_{\mathcal{A}^*}(\mathbf{y} - \mathbf{X}_{\mathcal{A}^*}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1})) = \frac{1}{\lambda_{t+1}} \mathbf{X}^\top_{\mathcal{A}^*}(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_{t+1}))$$

is

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}^*_{\mathcal{A}^*}(\lambda_{t+1}) &= \mathbf{X}^+_{\mathcal{A}^*}(\mathbf{y} - \lambda_{t+1}(\mathbf{X}^+_{\mathcal{A}^*})^\top \boldsymbol{\eta}^*_{\mathcal{A}^*}) \\
&= \mathbf{X}^+_{\mathcal{A}^*}\mathbf{y} - \mathbf{X}^+_{\mathcal{A}^*}(\mathbf{X}^+_{\mathcal{A}^*})^\top \mathbf{X}^\top_{\mathcal{A}^*}\mathbf{y} + \mathbf{X}^+_{\mathcal{A}^*}(\mathbf{X}^+_{\mathcal{A}^*})^\top \mathbf{X}^\top_{\mathcal{A}^*}\mathbf{X}_{\mathcal{A}^*}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1}) \\
&= \mathbf{X}^+_{\mathcal{A}^*}\mathbf{X}_{\mathcal{A}^*}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1}),
\end{aligned}$$

where we used the identity $\mathbf{A}^+(\mathbf{A}^+)^\top\mathbf{A}^\top = \mathbf{A}^+$. Therefore, the solution $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1})$ to the above equation must be

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1}) = \mathbf{X}^+_{\mathcal{A}^*}(\mathbf{y} - \lambda_{t+1}(\mathbf{X}^+_{\mathcal{A}^*})^\top \boldsymbol{\eta}^*_{\mathcal{A}^*}) + \mathbf{b} = \widetilde{\boldsymbol{\beta}}^*_{\mathcal{A}^*}(\lambda_{t+1}) + \mathbf{b},$$

where $\mathbf{b} \in \text{null}(\mathbf{X}_{\mathcal{A}^*})$. Since $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1})$ must have minimum ℓ_2-norm, $\mathbf{b} = \mathbf{0}$ and so

$$\widetilde{\boldsymbol{\beta}}^*_{\mathcal{A}^*}(\lambda_{t+1}) = \widetilde{\boldsymbol{\beta}}_{\mathcal{A}^*}(\lambda_{t+1}) = \begin{bmatrix} \widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_{t+1}) \\ 0 \end{bmatrix},$$

which implies that $\widetilde{\mathcal{P}}$ is continuous at λ_{t+1}. Also,

$$\mathbf{X}_j^\top(\mathbf{y} - \mathbf{X}_{\mathcal{A}^*}\widetilde{\boldsymbol{\beta}}^*_{\mathcal{A}^*}(\lambda_{t+1})) = \mathbf{X}_j^\top(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_{t+1}))$$

for all $j \in [d]$. It only remains to prove that the variable $\widetilde{i}^{\text{join}}_{t+1}$ satisfies the $\mathsf{KKT}_{\lambda_{t+1}}(\epsilon)$ condition once it is added to $\mathcal{A}$. Indeed, first define $\Delta^{\text{join}}_{t+1} \triangleq \min\{\lambda_t, \widetilde{\lambda}^{\text{join}}_{t+1}\} - \lambda^{\text{join}}_{t+1}$ and notice that $\Delta^{\text{join}}_{t+1} \leqslant \widetilde{\lambda}^{\text{join}}_{t+1}\epsilon/2$ since $\lambda_t \geqslant \lambda^{\text{join}}_{t+1}$ and $(1-\epsilon/2)\widetilde{\lambda}^{\text{join}}_{t+1} \leqslant \lambda^{\text{join}}_{t+1}$. Then, at the point $\lambda_{t+1} = \min\{\lambda_t, \widetilde{\lambda}^{\text{join}}_{t+1}\}$,

$$\begin{aligned}
\big|\mathbf{X}^\top_{\widetilde{i}^{\text{join}}_{t+1}}(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\widetilde{\lambda}^{\text{join}}_{t+1}))\big| &= \big|\mathbf{X}^\top_{\widetilde{i}^{\text{join}}_{t+1}}(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda^{\text{join}}_{t+1})) + \Delta^{\text{join}}_{t+1}\mathbf{X}^\top_{\widetilde{i}^{\text{join}}_{t+1}}(\mathbf{X}^+_{\mathcal{A}})^\top \boldsymbol{\eta}_{\mathcal{A}}\big| \\
&\geqslant \lambda^{\text{join}}_{t+1} - \frac{\epsilon\widetilde{\lambda}^{\text{join}}_{t+1}}{2}\big|\mathbf{X}^\top_{\widetilde{i}^{\text{join}}_{t+1}}(\mathbf{X}^+_{\mathcal{A}})^\top \boldsymbol{\eta}_{\mathcal{A}}\big| \\
&= \lambda^{\text{join}}_{t+1} - \frac{\epsilon\widetilde{\lambda}^{\text{join}}_{t+1}}{2\lambda_t}\big|\mathbf{X}^\top_{\widetilde{i}^{\text{join}}_{t+1}}(\mathbf{X}^+_{\mathcal{A}})^\top \mathbf{X}^\top_{\mathcal{A}}(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t))\big| \\
&\geqslant \lambda^{\text{join}}_{t+1} - \frac{\epsilon\widetilde{\lambda}^{\text{join}}_{t+1}}{2\lambda_t}\|\mathbf{X}^+_{\mathcal{A}}\mathbf{X}_{\widetilde{i}^{\text{join}}_{t+1}}\|_1\|\mathbf{X}^\top_{\mathcal{A}}(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t))\|_\infty \\
&\geqslant \lambda^{\text{join}}_{t+1} - \frac{\epsilon\widetilde{\lambda}^{\text{join}}_{t+1}}{2}\|\mathbf{X}^+_{\mathcal{A}}\mathbf{X}_{\widetilde{i}^{\text{join}}_{t+1}}\|_1 \\
&\geqslant \Big(1 - \frac{\epsilon}{2}\Big)\widetilde{\lambda}^{\text{join}}_{t+1} - \frac{\epsilon\widetilde{\lambda}^{\text{join}}_{t+1}}{2}\|\mathbf{X}^+_{\mathcal{A}}\mathbf{X}_{\widetilde{i}^{\text{join}}_{t+1}}\|_1 \\
&\geqslant (1-\epsilon)\widetilde{\lambda}^{\text{join}}_{t+1},
\end{aligned}$$

where we used that $\big|\mathbf{X}_{\widetilde{i}_{t+1}^{\text{join}}}^{\top}(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_{t+1}^{\text{join}}))\big| = \lambda_{t+1}^{\text{join}}$ by the definition of the joining time $\lambda_{t+1}^{\text{join}}$, that $\widetilde{\boldsymbol{\beta}}(\lambda_t)$ satisfies the $\mathsf{KKT}_{\lambda_t}(\epsilon)$ condition at λ_t, i.e., $\|\mathbf{X}_{\mathcal{A}}^{\top}(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t))\|_{\infty} \leqslant \lambda_t$, and finally that $\max_{j\in\mathcal{A}^c}\|\mathbf{X}_{\mathcal{A}}^{+}\mathbf{X}_j\|_1 \leqslant 1$. □

After asserting the correctness of Algorithm 3, we now analyse its complexity. Specifically, we show how to sample the joining variable $\widetilde{i}_{t+1}^{\text{join}}$ from the set in Eq. (13) by combining the approximate quantum minimum-finding subroutine from Chen and de Wolf [CdW23] (Fact 11) and the unitary map that approximates (parts of) the joining times $\Lambda_i^{\text{join}}(\mathcal{A}) = \frac{\mathbf{X}_i^{\top}(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu})}{\pm 1-\mathbf{X}_i^{\top}\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}}$ (Lemma 13). Our final complexity depends on the mutual incoherence between different column subspaces and the overlap between $\mathbf{y}$ and different column subspaces.

Theorem 26. *Let $\mathbf{X}\in\mathbb{R}^{n\times d}$ and $\mathbf{y}\in\mathbb{R}^n$. Assume $\mathbf{X}$ is stored in a* QROM *and we have access to* QRAMs *of size $O(n)$. Let $\delta\in(0,1)$, $\epsilon\in(0,2)$, and $T\in\mathbb{N}$. Let $\alpha,\gamma\in(0,1]$ be such that*

$$\max_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T}\|\mathbf{X}_{\mathcal{A}}^{+}\mathbf{X}_{\mathcal{A}^c}\|_1 \leqslant 1-\alpha \quad \text{and} \quad \min_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T}\frac{\|\mathbf{X}^{\top}(\mathbf{I}-\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^{+})\mathbf{y}\|_{\infty}}{\|\mathbf{X}\|_{\max}\|(\mathbf{I}-\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^{+})\mathbf{y}\|_1} \geqslant \gamma.$$

The approximate quantum LARS *algorithm 3 outputs, with probability at least $1-\delta$, a continuous piecewise linear approximate regularisation path $\widetilde{\mathcal{P}} = \{\widetilde{\boldsymbol{\beta}}(\lambda)\in\mathbb{R}^d : \lambda>0\}$ with error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}(\lambda)\|_1$ and at most T kinks in time*

$$O\left(\frac{\gamma^{-1}+\sqrt{n}\|\mathbf{X}\|_{\max}\|\mathbf{X}^{+}\|_2}{\alpha\epsilon}\sqrt{|\mathcal{I}|}\log^2(T/\delta)\,\mathrm{poly}\log(nd)+n|\mathcal{A}|+|\mathcal{A}|^2\right)$$

per iteration, where $\mathcal{A}$ and $\mathcal{I}$ are the active and inactive sets of the corresponding iteration.

Proof. The correctness of the approximate quantum LARS algorithm 3 follows from Theorem 25, since it is a LARS algorithm wherein the joining variable $\widetilde{i}_{t+1}^{\text{join}}$ is taken from the set

$$\left\{j\in\mathcal{I} : \Lambda_j^{\text{join}}(\mathcal{A}) \geqslant (1-\epsilon/2)\max_{i\in\mathcal{I}}\{\Lambda_i^{\text{join}}(\mathcal{A})\}\right\} \tag{14}$$

and the corresponding joining point $\widetilde{\lambda}_{t+1}^{\text{join}}$ is

$$\widetilde{\lambda}_{t+1}^{\text{join}} = (1-\epsilon/2)^{-1}\Lambda_{\widetilde{i}_{t+1}^{\text{join}}}^{\text{join}}(\mathcal{A}).$$

Regarding the time complexity of the algorithm, the most expensive steps are computing $\mathbf{X}_{\mathcal{A}}^{+}$ in time $O(n|\mathcal{A}|+|\mathcal{A}|^2)$, inputting $\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}$ and $\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}$ into KP-trees in time $O(n\log n)$, and sampling $\widetilde{i}_{t+1}^{\text{join}}$. We now show how to obtain the joining variable $\widetilde{i}_{t+1}^{\text{join}}$, i.e., how it can be sampled from the set in Eq. (14).

The first step is to create a unitary operator that maps $|i\rangle|\bar{0}\rangle \mapsto |i\rangle|R_i\rangle$ such that, for all $i\in\mathcal{I}$, after measuring the state $|R_i\rangle$, with probability at least $1-\delta_0$, where $\delta_0 = O\big(\delta^2/(T^2|\mathcal{I}|\log(T/\delta))\big)$, the outcome r_i of the first register satisfies

$$|r_i - \Lambda_i^{\text{join}}(\mathcal{A})| \leqslant \epsilon_0. \tag{15}$$

For such, we use Lemma 13 to build maps $|i\rangle|\bar{0}\rangle \mapsto |i\rangle|\Psi_i\rangle$ and $|i\rangle|\bar{0}\rangle \mapsto |i\rangle|\Phi_i\rangle$ whose respective outcomes ψ_i and ϕ_i of the first registers satisfy

$$|\psi_i - \mathbf{X}_i^{\top}(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu})| \leqslant \epsilon_1\|\mathbf{X}_i\|_{\infty}\|\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}\|_1 \leqslant \epsilon_1\|\mathbf{X}\|_{\max}\|\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}\|_1,$$
$$|\phi_i - \mathbf{X}_i^{\top}\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}| \leqslant \epsilon_2\|\mathbf{X}_i\|_{\infty}\|\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}\|_1 \leqslant \epsilon_2\|\mathbf{X}\|_{\max}\|\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}\|_1,$$

in time $O(\epsilon_1^{-1}\log(1/\delta_0)\operatorname{poly}\log(nd))$ and $O(\epsilon_2^{-1}\log(1/\delta_0)\operatorname{poly}\log(nd))$, respectively. Then, by using Lemma 8, we can construct the desired map $|i\rangle|\bar{0}\rangle \mapsto |i\rangle|R_i\rangle$ as in Eq. (15) with

$$\epsilon_0 = \max_{i\in\mathcal{I}} \left\{ \Lambda_i^{\text{join}}(\mathcal{A}) \left(\frac{\epsilon_1 \|\mathbf{X}\|_{\max}\|\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}\|_1}{|\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu})|} + \frac{\epsilon_2\|\mathbf{X}\|_{\max}\|\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}\|_1}{|\pm 1-\mathbf{X}_i^\top\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}|} \right) \right\}. \tag{16}$$

Let us upper bound ϵ_0. First note that $\max_{i\in\mathcal{I}}\{\Lambda_i^{\text{join}}(\mathcal{A})\} = \lambda_{t+1}^{\text{join}}$ by the definition of $\lambda_{t+1}^{\text{join}}$. Regarding the other quantities,

$$\begin{aligned}
\|\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}\|_1 = \|(\mathbf{X}_{\mathcal{A}}^+)^\top\boldsymbol{\eta}_{\mathcal{A}}\|_1 &= \frac{1}{\lambda_t}\|(\mathbf{X}_{\mathcal{A}}^+)^\top\mathbf{X}_{\mathcal{A}}^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t))\|_1 \\
&\leqslant \frac{1}{\lambda_t}\|\mathbf{X}_{\mathcal{A}}^+\|_\infty\|\mathbf{X}_{\mathcal{A}}^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t))\|_\infty \\
&\leqslant \sqrt{n}\|\mathbf{X}_{\mathcal{A}}^+\|_2 \\
&\leqslant \sqrt{n}\|\mathbf{X}^+\|_2,
\end{aligned}$$

using that $\|\mathbf{X}_{\mathcal{A}}^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t))\|_\infty \leqslant \lambda_t$ since $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t)$ satisfies the $\mathsf{KKT}_{\lambda_t}(\epsilon)$ condition. Also,

$$\begin{aligned}
|\pm 1-\mathbf{X}_i^\top\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}| &\geqslant 1-|\mathbf{X}_i^\top\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}| \\
&= 1-\frac{1}{\lambda_t}|\mathbf{X}_i^\top(\mathbf{X}_{\mathcal{A}}^+)^\top\mathbf{X}_{\mathcal{A}}^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t))| \\
&\geqslant 1-\frac{1}{\lambda_t}\|\mathbf{X}_{\mathcal{A}}^+\mathbf{X}_i\|_1\|\mathbf{X}_{\mathcal{A}}^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t))\|_\infty \\
&\geqslant \alpha,
\end{aligned}$$

where we used Hölder's inequality and again that $\|\mathbf{X}_{\mathcal{A}}^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_t))\|_\infty \leqslant \lambda_t$. The above inequalities are sufficient to bound the second term in Eq. (16). Regarding the first term,

$$\begin{aligned}
\max_{i\in\mathcal{I}}\left\{\left|\frac{\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu})}{\pm 1-\mathbf{X}_i^\top\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}}\right| \frac{\epsilon_1\|\mathbf{X}\|_{\max}\|\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}\|_1}{|\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu})|}\right\} &= \epsilon_1 \max_{i\in\mathcal{I}}\left\{\frac{\|\mathbf{X}\|_{\max}\|\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}\|_1}{|\pm 1-\mathbf{X}_i^\top\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}|}\right\} \\
&\leqslant \epsilon_1\alpha^{-1}\|\mathbf{X}\|_{\max}\|\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}\|_1 \\
&= \epsilon_1\alpha^{-1}\|\mathbf{X}\|_{\max}\|\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}\|_1 \frac{\lambda_{t+1}^{\text{join}}}{\max_{i\in\mathcal{I}}\left|\frac{\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu})}{\pm 1-\mathbf{X}_i^\top\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}}\right|} \\
&\leqslant 2\epsilon_1\alpha^{-1}\lambda_{t+1}^{\text{join}}\frac{\|\mathbf{X}\|_{\max}\|\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}\|_1}{\max_{i\in\mathcal{I}}|\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu})|} \\
&\leqslant 2\epsilon_1\alpha^{-1}\gamma^{-1}\lambda_{t+1}^{\text{join}},
\end{aligned}$$

using that $\max_{i\in\mathcal{I}}|\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu})| = \|\mathbf{X}_{\mathcal{I}}^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu})\|_\infty = \|\mathbf{X}^\top(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu})\|_\infty \geqslant \gamma\|\mathbf{X}\|_{\max}\|\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}\|_1$ and that $\alpha \leqslant |\pm 1-\mathbf{X}_i^\top\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}| \leqslant 2$. All the above leads to

$$\epsilon_0 \leqslant \lambda_{t+1}^{\text{join}}\big(2\epsilon_1\alpha^{-1}\gamma^{-1} + \epsilon_2\alpha^{-1}\sqrt{n}\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2\big).$$

By taking (notice that $\epsilon \in (0,2)$ guarantees that ϵ_2 is within the required range from Lemma 8)

$$\epsilon_1 = \frac{\alpha\gamma\epsilon}{8} \quad \text{and} \quad \epsilon_2 = \frac{\alpha\epsilon}{4\sqrt{n}\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2},$$

then $\epsilon_0 \leqslant \epsilon\lambda_{t+1}^{\text{join}}/4$ in time $O\Big(\frac{\gamma^{-1}+\sqrt{n}\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2}{\alpha\epsilon}\log(1/\delta_0)\operatorname{poly}\log(nd)\Big)$. Finally, we apply the approximate quantum minimum-finding algorithm (Fact 11) to obtain an index $\widetilde{i}_{t+1}^{\text{join}} \in \mathcal{I}$ such that

$$\Lambda^{\text{join}}_{\widetilde{i}^{\text{join}}_{t+1}}(\mathcal{A}) \geqslant \max_{i\in\mathcal{I}}\{\Lambda_i^{\text{join}}(\mathcal{A})\} - 2\epsilon_0 \geqslant (1-\epsilon/2)\max_{i\in\mathcal{I}}\{\Lambda_i^{\text{join}}(\mathcal{A})\},$$

i.e., $\widetilde{i}_{t+1}^{\text{join}}$ belongs to the set from Eq. (14) with probability at least $1-\delta/T$ as promised. The total complexity is $\widetilde{O}(\sqrt{|\mathcal{I}|}\log(T/\delta))$ times the complexity of constructing the mapping $|i\rangle|\bar{0}\rangle \mapsto |i\rangle|R_i\rangle$. The final success probability follows from a union bound. □

If $\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2$ is constant and the parameters $\alpha, \gamma \in (0,1]$ are bounded away from 0, then the complexity of Algorithm 3 is $\widetilde{O}(\sqrt{n|\mathcal{I}|}/\epsilon + n|\mathcal{A}| + |\mathcal{A}|^2)$ per iteration. If $|\mathcal{A}| = O(n)$, then we obtain a fully quadratic advantage $\widetilde{O}(\sqrt{nd})$ over the classical LARS algorithm.

4.3 Approximate classical LARS algorithm

By using virtually the same techniques and results behind the approximate quantum LARS algorithm, it is possible to devise an analogous approximate classical LARS algorithm. The idea is again to approximately compute the joining times $\Lambda_i^{\text{join}}(\mathcal{A}) = \frac{\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_\mathcal{A}\boldsymbol{\mu})}{\pm 1 - \mathbf{X}_i^\top\mathbf{X}_\mathcal{A}\boldsymbol{\theta}}$, but by using the classical sampling procedure from Lemma 14 instead. For well-behaved design matrices, the complexity per iteration is $\widetilde{O}(n|\mathcal{I}|/\epsilon^2 + n|\mathcal{A}| + |\mathcal{A}|^2)$. The approximate classical LARS algorithm for the pathwise Lasso is shown in Algorithm 4.

Algorithm 4: Approximate classical LARS algorithm for the pathwise Lasso

Input: Vector $\mathbf{y} \in \mathbb{R}^n$, matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\delta \in (0,1)$, $\epsilon \in (0,2)$, and $T \in \mathbb{N}$
Initialise $\mathcal{A} = \operatorname{argmax}_{j\in[d]}|\mathbf{X}_j^\top\mathbf{y}|$, $\mathcal{I} = [d]\backslash\mathcal{A}$, $\lambda_0 = \|\mathbf{X}^\top\mathbf{y}\|_\infty$, $\widetilde{\boldsymbol{\beta}}(\lambda_0) = 0$, $t = 0$
while $\mathcal{I} \neq \varnothing$, $t \leqslant T$ **do**
  $\boldsymbol{\eta}_\mathcal{A} \leftarrow \frac{1}{\lambda_t}\mathbf{X}_\mathcal{A}^\top(\mathbf{y} - \mathbf{X}_\mathcal{A}\widetilde{\boldsymbol{\beta}}_\mathcal{A}(\lambda_t))$
  $\boldsymbol{\mu} \leftarrow \mathbf{X}_\mathcal{A}^+\mathbf{y}$ // $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{A}|}$
  $\boldsymbol{\theta} \leftarrow (\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^+\boldsymbol{\eta}_\mathcal{A}$ // $\boldsymbol{\theta} \in \mathbb{R}^{|\mathcal{A}|}$
  Input $\mathbf{y} - \mathbf{X}_\mathcal{A}\boldsymbol{\mu}$ and $\mathbf{X}_\mathcal{A}\boldsymbol{\theta}$ into classical-samplable memories
  Define $\Lambda_i^{\text{join}} \triangleq \frac{\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_\mathcal{A}\boldsymbol{\mu})}{\pm 1 - \mathbf{X}_i^\top\mathbf{X}_\mathcal{A}\boldsymbol{\theta}}$ and $\Lambda_i^{\text{cross}} \triangleq \frac{\mu_i}{\theta_i}\cdot\mathbf{1}\left[\frac{\mu_i}{\theta_i} \leqslant \lambda_t\right]$
  $i_{t+1}^{\text{cross}} \leftarrow \operatorname{argmax}_{i\in\mathcal{A}}\{\Lambda_i^{\text{cross}}\}$ and $\lambda_{t+1}^{\text{cross}} \leftarrow \max_{i\in\mathcal{A}}\{\Lambda_i^{\text{cross}}\}$
  Compute $\{r_i\}_{i\in\mathcal{I}}$ such that $|r_i - \Lambda_i^{\text{join}}| \leqslant \frac{\epsilon\lambda_{t+1}^{\text{join}}}{2}$ with failure probability $\frac{\delta}{T|\mathcal{I}|}$ (Lemma 14)
  $\widetilde{i}_{t+1}^{\text{join}} \leftarrow \operatorname{argmax}_{i\in\mathcal{I}}\{r_i\}$ and $\widetilde{\lambda}_{t+1}^{\text{join}} \leftarrow (1-\epsilon/2)^{-1}\max_{i\in\mathcal{I}}\{r_i\}$
  $\lambda_{t+1} \leftarrow \min\{\lambda_t, \max\{\lambda_{t+1}^{\text{cross}}, \widetilde{\lambda}_{t+1}^{\text{join}}\}\}$
  $\widetilde{\boldsymbol{\beta}}_\mathcal{A}(\lambda_{t+1}) \leftarrow \boldsymbol{\mu} - \lambda_{t+1}\boldsymbol{\theta}$
  if $\lambda_{t+1} = \lambda_{t+1}^{\text{cross}}$ **then**
    Move i_{t+1}^{cross} from $\mathcal{A}$ to $\mathcal{I}$
  else
    Move $\widetilde{i}_{t+1}^{\text{join}}$ from $\mathcal{I}$ to $\mathcal{A}$
  $t \leftarrow t+1$
Output: Coefficients $[(\lambda_0, \widetilde{\boldsymbol{\beta}}(\lambda_0)), (\lambda_1, \widetilde{\boldsymbol{\beta}}(\lambda_1)), \dots]$

Theorem 27. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Assume access to classical-samplable structures of size $O(n)$. Let $\delta \in (0,1)$, $\epsilon \in (0,2)$, and $T \in \mathbb{N}$. Let $\alpha, \gamma \in (0,1]$ be such that*

$$\max_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T} \|\mathbf{X}_{\mathcal{A}}^{+}\mathbf{X}_{\mathcal{A}^c}\|_1 \leqslant 1-\alpha \quad \textit{and} \quad \min_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T} \frac{\|\mathbf{X}^\top(\mathbf{I}-\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^{+})\mathbf{y}\|_\infty}{\|\mathbf{X}\|_{\max}\|(\mathbf{I}-\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^{+})\mathbf{y}\|_1} \geqslant \gamma.$$

The approximate classical LARS *algorithm 4 outputs, with probability at least $1-\delta$, a continuous piecewise linear approximate regularisation path $\widetilde{\mathcal{P}} = \{\widetilde{\boldsymbol{\beta}}(\lambda) \in \mathbb{R}^d : \lambda > 0\}$ with error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}(\lambda)\|_1$ and at most T kinks in time*

$$O\left(\frac{\gamma^{-2}+n\|\mathbf{X}\|_{\max}^2\|\mathbf{X}^+\|_2^2}{\alpha^2\epsilon^2}|\mathcal{I}|\log(Td/\delta)\operatorname{poly}\log n + n|\mathcal{A}| + |\mathcal{A}|^2\right)$$

per iteration, where $\mathcal{A}$ and $\mathcal{I}$ are the active and inactive sets of the corresponding iteration.

Proof. We start by noticing that the joining time computation step outputs $\widetilde{i}_{t+1}^{\text{join}} = \operatorname{argmax}_{i\in\mathcal{I}}\{r_i\}$ and $\widetilde{\lambda}_{t+1}^{\text{join}} = (1-\epsilon/2)^{-1}\max_{i\in\mathcal{I}}\{r_i\}$, where $|r_i - \Lambda_i^{\text{join}}(\mathcal{A})| \leqslant \epsilon\lambda_{t+1}^{\text{join}}/2$. Since $\max_{i\in\mathcal{I}}\{r_i\} \geqslant r_{i_{t+1}^{\text{join}}} \geqslant \Lambda_{i_{t+1}^{\text{join}}}^{\text{join}}(\mathcal{A}) - \epsilon\lambda_{t+1}^{\text{join}}/2 = (1-\epsilon/2)\lambda_{t+1}^{\text{join}}$, this means that the joining variable $\widetilde{i}_{t+1}^{\text{join}}$ is taken from the set

$$\left\{j \in \mathcal{I} : \Lambda_j^{\text{join}}(\mathcal{A}) \geqslant (1-\epsilon/2)\max_{i\in\mathcal{I}}\{\Lambda_i^{\text{join}}(\mathcal{A})\}\right\}.$$

Theorem 25 thus guarantees the correctness of the algorithm. Regarding the time complexity, the most expensive steps are computing $\mathbf{X}_{\mathcal{A}}^{+}$ in time $O(n|\mathcal{A}|+|\mathcal{A}|^2)$, inputting $\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}$ and $\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}$ into classical-samplable memories in time $O(n\log n)$, and computing the $|\mathcal{I}|$ values $\{r_i\}_{i\in\mathcal{I}}$. By following the exact same steps as in the proof of Theorem 26, approximating any quantity $\Lambda_i^{\text{join}}(\mathcal{A})$ within precision $\epsilon\lambda_{t+1}^{\text{join}}/2$ and failure probability at most $\delta/(T|\mathcal{I}|)$ requires time

$$O\left(\frac{\gamma^{-2}+n\|\mathbf{X}\|_{\max}^2\|\mathbf{X}^+\|_2^2}{\alpha^2\epsilon^2}\log(Td/\delta)\operatorname{poly}\log n\right)$$

(note that the term $O(\operatorname{poly}\log n)$ comes from sampling the vectors $\mathbf{y}-\mathbf{X}_{\mathcal{A}}\boldsymbol{\mu}$ and $\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}$). The final success probability follows from a union bound. □

4.4 Standard Gaussian random design matrix

The complexity of the approximate LARS algorithms depends on the quantities

$$\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2, \qquad \max_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T} \|\mathbf{X}_{\mathcal{A}}^{+}\mathbf{X}_{\mathcal{A}^c}\|_1, \quad \text{and} \quad \min_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T} \frac{\|\mathbf{X}^\top(\mathbf{I}-\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^{+})\mathbf{y}\|_\infty}{\|\mathbf{X}\|_{\max}\|(\mathbf{I}-\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^{+})\mathbf{y}\|_1}.$$

In this section, we shall bound these quantities for the specific case when the design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the standard Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ entries. According to Fact 19, in this situation the Lasso solution is unique, thus $\operatorname{null}(\mathbf{X}_{\mathcal{A}}) = \{\mathbf{0}\}$ and $\mathbf{X}_{\mathcal{A}}^{+} = (\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^\top$. As we shall prove below, with high probability, $\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2$ is a constant and the mutual incoherence is at least $1/2$. The mutual overlap between $\mathbf{y}$ and $\mathbf{X}$, on the other hand, is $\Omega(1/\sqrt{n\log d})$ with high probability, which preserves the overall $\widetilde{O}(\sqrt{n|\mathcal{I}|})$ complexity of the approximate quantum LARS algorithm.

Let us start our analysis with the quantity $\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2$. The next result states that it is basically a constant with high probability in the high-dimensional case.

Lemma 28. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ with $d \geqslant 2n$ be a standard Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Then, for any $\delta \in (0,1)$, $\mathbb{P}[\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2 \leqslant e^{3/\delta}] \geqslant 1-\delta$.*

Proof. In order to bound the quantity $\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2$, first note that it is upper bounded by the condition number $\kappa(\mathbf{X}) \triangleq \|\mathbf{X}\|_2\|\mathbf{X}^+\|_2$ of $\mathbf{X}$. The condition number of standard Gaussian random matrices was studied by Chen and Dongarra [CD05], who proved that $\mathbb{E}[\ln \kappa(\mathbf{X})] \leqslant 2.258 + \ln \frac{d}{|d-n|+1}$ (see also [ES05]). Thus, in the case when $d \geqslant 2n$, $\mathbb{E}[\ln \kappa(\mathbf{X})] \leqslant 2.258 + \ln 2 \leqslant 3$. Hence, by Markov's inequality, $\mathbb{P}[\kappa(\mathbf{X}) \leqslant e^{3/\delta}] \geqslant 1-\delta$ for any $\delta \in (0,1)$. □

We now turn our attention to

$$\max_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T} \|\mathbf{X}_{\mathcal{A}}^+\mathbf{X}_{\mathcal{A}^c}\|_1 = \max_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T} \|(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}^c}\|_1 = \max_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T} \|\mathbf{X}_{\mathcal{A}^c}^\top\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\|_\infty.$$

We prove in the next lemma that, if $T = O(\sqrt{n/\log d})$, then the mutual incoherence is a constant bounded away from 0 with high probability. Our result applies to the more general case when the rows of $\mathbf{X}$ are sampled i.i.d. according to a $\mathcal{N}(0,\mathbf{\Sigma})$ distribution, where $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$. A similar result can be found in [Wai19, Exercise 7.19].

Lemma 29. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a random matrix with rows sampled i.i.d. according to a $\mathcal{N}(0,\mathbf{\Sigma})$ distribution, where $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$. Let $\delta \in (0,1)$ and $\gamma_{\min}(\mathbf{\Sigma}) \in (0,1]$ be the minimum eigenvalue of $\mathbf{\Sigma}$. Suppose that the diagonal entries of $\mathbf{\Sigma}$ are at most 1 and that there is $\bar{\alpha} \in (0,1]$ such that*

$$\max_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T} \max_{j\in\mathcal{A}^c} \|\mathbf{\Sigma}_{j\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_1 \leqslant 1-\bar{\alpha},$$

for some $T \in \mathbb{N}$. If $T = O\big(\sqrt{n\bar{\alpha}^2\gamma_{\min}(\mathbf{\Sigma})/\log(d/\delta)}\big)$, then

$$\mathbb{P}\left[\max_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T} \|\mathbf{X}_{\mathcal{A}^c}^\top\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\|_\infty \leqslant 1-\frac{\bar{\alpha}}{2}\right] \geqslant 1-\delta.$$

Proof. For fixed $\mathcal{A}$ and any $\mathbf{u} \in \mathbb{R}^{|\mathcal{A}|}$ such that $\|\mathbf{u}\|_\infty = 1$, first note that

$$\|\mathbf{X}_{\mathcal{A}^c}^\top\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\|_\infty \geqslant \|\mathbf{X}_{\mathcal{A}^c}^\top\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{u}\|_\infty = \max_{j\in\mathcal{A}^c} |\mathbf{X}_j^\top\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{u}|.$$

Consider the quantity $|\mathbf{X}_j^\top\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{u}|$ for fixed $j \in \mathcal{A}^c$. The zero-mean Gaussian vector $\mathbf{X}_j$ can be decomposed into a linear prediction based on $\mathbf{X}_{\mathcal{A}}$ plus a prediction error as (see [Wai09, Section V.A] and [Wai19, Exercise 11.3])

$$\mathbf{X}_j^\top = \mathbf{\Sigma}_{j\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^\top + \mathbf{W}_j^\top,$$

where $\mathbf{W}_j \in \mathbb{R}^n$ is a vector with i.i.d. $\mathcal{N}(0, [\mathbf{\Sigma}_{\mathcal{A}^c|\mathcal{A}}]_{jj})$ entries. Here the (positive semi-definite) matrix $\mathbf{\Sigma}_{\mathcal{A}^c|\mathcal{A}} = \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}^c} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}^c}$ is the conditional co-variance matrix of $(\mathbf{X}_{\mathcal{A}^c}|\mathbf{X}_{\mathcal{A}})$. Thus

$$\begin{aligned}
|\mathbf{X}_j^\top\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{u}| &\leqslant |\mathbf{\Sigma}_{j\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{u}| + |\mathbf{W}_j^\top\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{u}| \\
&\leqslant \|\mathbf{\Sigma}_{j\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_1\|\mathbf{u}\|_\infty + |\mathbf{W}_j^\top\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{u}| \\
&\leqslant 1-\bar{\alpha} + |\mathbf{W}_j^\top\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{u}|.
\end{aligned}$$

Since $\mathbf{W}_j$ is independent of $\mathbf{X}_\mathcal{A}$, $\mathbf{W}_j^\top \mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\mathbf{u}$ is a sub-Gaussian random variable with parameter (Fact 6)

$$[\mathbf{\Sigma}_{\mathcal{A}^c|\mathcal{A}}]_{jj}\|\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\mathbf{u}\|_2^2 \leqslant \|\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\mathbf{u}\|_2^2 \\ = \mathbf{u}^\top(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\mathbf{u} \leqslant \|\mathbf{u}\|_2^2\|(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\|_2 \leqslant |\mathcal{A}|\|(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\|_2,$$

where we used that $[\mathbf{\Sigma}_{\mathcal{A}^c|\mathcal{A}}]_{jj} \leqslant 1$ since $\mathbf{\Sigma}_{\mathcal{A}^c|\mathcal{A}}$ is positive semi-definite and the diagonal entries of $\mathbf{\Sigma}$ are at most 1. We now relate the quantity $\|(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\|_2$ with $\gamma_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}) \geqslant \gamma_{\min}(\mathbf{\Sigma})$. According to [Wai09, Lemma 9],

$$\mathbb{P}\left[\|(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\|_2 \geqslant \frac{(1+t+\sqrt{|\mathcal{A}|/n})^2}{n\gamma_{\min}(\mathbf{\Sigma})}\right] \leqslant 2e^{-nt^2/2} \quad \text{for all } t > 0.$$

Define the event $\mathcal{E}(\mathbf{X}_\mathcal{A}) \triangleq \{\|(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\|_2 \geqslant 9/(n\gamma_{\min}(\mathbf{\Sigma}))\}$. Since $|\mathcal{A}| \leqslant n$, the event $\mathcal{E}(\mathbf{X}_\mathcal{A})$ happens with probability at most $2e^{-n/2}$. Therefore, with probability at least $1-2e^{-n/2}$, the quantity $\mathbf{W}_j^\top \mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\mathbf{u}$ is a sub-Gaussian random variable with parameter at most $9|\mathcal{A}|/(n\gamma_{\min}(\mathbf{\Sigma}))$. A Chernoff bound yields

$$\begin{aligned}\mathbb{P}\left[\max_{j\in\mathcal{A}^c}|\mathbf{W}_j^\top \mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\mathbf{u}| \geqslant \frac{\bar\alpha}{2}\right] &\leqslant \mathbb{P}\left[\max_{j\in\mathcal{A}^c}|\mathbf{W}_j^\top \mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\mathbf{u}| \geqslant \frac{\bar\alpha}{2} \mid \mathcal{E}^c(\mathbf{X}_\mathcal{A})\right] + \mathbb{P}[\mathcal{E}(\mathbf{X}_\mathcal{A})] \\ &\leqslant 2(d-|\mathcal{A}|)\exp\left(-\frac{\bar\alpha^2 n\gamma_{\min}(\mathbf{\Sigma})}{72|\mathcal{A}|}\right) + 2\exp(-n/2) \\ &\leqslant 4(d-|\mathcal{A}|)\exp\left(-\frac{\bar\alpha^2 n\gamma_{\min}(\mathbf{\Sigma})}{72|\mathcal{A}|}\right).\end{aligned}$$

Therefore, with high probability, $|\mathbf{X}_j^\top \mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\mathbf{u}| \leqslant 1-\bar\alpha/2$ for any $\mathbf{u} \in \mathbb{R}^{|\mathcal{A}|}$ such that $\|\mathbf{u}\|_\infty = 1$. Summing over all $\mathcal{A} \subseteq [d]$ such that $|\mathcal{A}| \leqslant T$, we conclude that

$$\begin{aligned}\mathbb{P}\left[\max_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T}\|\mathbf{X}_{\mathcal{A}^c}^\top \mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\|_\infty \geqslant 1-\frac{\bar\alpha}{2}\right] &\leqslant 4d\exp\left(-\frac{\bar\alpha^2 n\gamma_{\min}(\mathbf{\Sigma})}{72T}\right)\sum_{i=0}^{T}\binom{d}{i} \\ &\leqslant 4d\exp\left(-\frac{\bar\alpha^2 n\gamma_{\min}(\mathbf{\Sigma})}{72T} + d\mathsf{H}(T/d)\right),\end{aligned}$$

where $\mathsf{H}(p) \triangleq -p\ln p-(1-p)\ln(1-p)$ is the entropy function. Using the bound $\mathsf{H}(p) \leqslant p(1-\ln p)$, the above probability can be bounded as

$$\mathbb{P}\left[\max_{\mathcal{A}\subseteq[d]:|\mathcal{A}|\leqslant T}\|\mathbf{X}_{\mathcal{A}^c}^\top \mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A})^{-1}\|_\infty \geqslant 1-\frac{\bar\alpha}{2}\right] \leqslant 4d\exp\left(-\frac{\bar\alpha^2 n\gamma_{\min}(\mathbf{\Sigma})}{72T} + T(1+\ln(d/T))\right).$$

Hence, if

$$n \geqslant \frac{72T}{\bar\alpha^2\gamma_{\min}(\mathbf{\Sigma})}\left(T\left(1+\ln\frac{d}{T}\right)+\ln\frac{4d}{\delta}\right) = O\left(\frac{T^2\log(d/\delta)}{\bar\alpha^2\gamma_{\min}(\mathbf{\Sigma})}\right),$$

then the above failure probability is at most δ. The bound on n follows from our choice of T. □

The above lemma applies to standard Gaussian matrices with i.i.d. $\mathcal{N}(0,1)$ entries, in which case $\mathbf{\Sigma} = \mathbf{I}$. Hence $\bar\alpha = 1$ and $\gamma_{\min}(\mathbf{I}) = 1$, and thus the mutual incoherence of a standard Gaussian matrix is greater than $1/2$ with high probability.

Finally, we analyse the last quantity, the mutual overlap between $\mathbf{y}$ and $\mathbf{X}$. The next lemma states that the mutual overlap between $\mathbf{y}$ and $\mathbf{X}$ is $\Omega(1/\sqrt{n\log d})$ with high probability, even if we allow the size of the set $\mathcal{A}$ to be as large as $\lfloor d/2 \rfloor$.

Lemma 30. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a standard Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ entries and let $\mathbf{y} \in \mathbb{R}^n$. Let $\delta \in (0,1)$. Then*

$$\mathbb{P}\left[\min_{\mathcal{A}\subseteq[d]:|\mathcal{A}|<\lfloor d/2\rfloor} \frac{\|\mathbf{X}^\top(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}\|_\infty}{\|\mathbf{X}\|_{\max}\|(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}\|_1} \geqslant \frac{\delta^{2/d}}{8}\sqrt{\frac{\pi}{n\ln(4nd/\delta)}}\right] \geqslant 1-\delta.$$

Proof. Let us start by bounding $\|\mathbf{X}\|_{\max}$. By a simple Chernoff bound plus a union bound, $\mathbb{P}[\|\mathbf{X}\|_{\max} \geqslant \sqrt{2\ln(4nd/\delta)}] \leqslant \delta/2$. We thus move on to the remaining part of the expression. First, for fixed $\mathcal{A}$,

$$\|\mathbf{X}^\top(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}\|_\infty = \max_{j\in\mathcal{A}^c}|\mathbf{X}_j^\top(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}|.$$

Fix $j \in \mathcal{A}^c$. Since $\mathbf{X}_j$ is independent of $\mathbf{X}_\mathcal{A}$, the quantity $\mathbf{X}_j^\top(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}$ is a Gaussian random variable with variance $\|(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}\|_2^2$. Let $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\mathrm{d}t$, $z \geqslant 0$, be the error function. Then, for $t \in [0,1]$,

$$\mathbb{P}\left[\frac{|\mathbf{X}_j^\top(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}|}{\|(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}\|_1} \leqslant t\sqrt{\frac{\pi}{2n}}\right] = \operatorname{erf}\left(\frac{t\sqrt{\pi}\|(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}\|_1}{2\sqrt{n}\|(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}\|_2}\right) \leqslant t,$$

using that $\|(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}\|_1 \leqslant \sqrt{n}\|(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}\|_2$ and that $\operatorname{erf}(z) \leqslant 2z/\sqrt{\pi}$. By the independence of all $\mathbf{X}_j$, $j \in \mathcal{A}^c$, we thus have

$$\mathbb{P}\left[\max_{j\in\mathcal{A}^c}\frac{|\mathbf{X}_j^\top(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}|}{\|(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}\|_1} \leqslant t\sqrt{\frac{\pi}{2n}}\right] \leqslant t^{d-|\mathcal{A}|}.$$

Summing over all $\mathcal{A} \subseteq [d]$ such that $|\mathcal{A}| < \lfloor d/2\rfloor$, we conclude that

$$\mathbb{P}\left[\min_{\mathcal{A}\subseteq[d]:|\mathcal{A}|<\lfloor d/2\rfloor}\frac{\|\mathbf{X}^\top(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}\|_\infty}{\|(\mathbf{I}-\mathbf{X}_\mathcal{A}(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A})^{-1}\mathbf{X}_\mathcal{A}^\top)\mathbf{y}\|_1} \leqslant t\sqrt{\frac{\pi}{2n}}\right] \leqslant \sum_{|\mathcal{A}|=0}^{\lfloor d/2\rfloor-1}\binom{d}{|\mathcal{A}|}t^{d-|\mathcal{A}|} \leqslant 2^{d-1}t^{d/2}.$$

Thus, if $t = \delta^{2/d}/4$, then the above failure probability is at most $\delta/2$. By putting together both the above bound and the bound on $\|\mathbf{X}\|_{\max}$, we arrive at the desired result. □

5 Bounds in noisy regime

In this section, we assume that the observation vector $\mathbf{y} \in \mathbb{R}^n$ and the design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ are connected via the standard linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{w},$$

where the true solution $\boldsymbol{\beta}^* \in \mathbb{R}^d$ (also referred to as regression vector) is sparse and $\mathbf{w} \in \mathbb{R}^n$ is a noise vector. Our aim is to determine how close the Lasso solution produced by the LARS algorithm, especially its approximate version from Algorithm 3, is to the true solution $\boldsymbol{\beta}^*$. More specifically, we shall focus on bounding the *mean-squared prediction error* $\|\mathbf{X}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}})\|_2^2/n$. Throughout this section we shall assume the following.

Assumption 31. *Given the design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, then $\max_{i\in[d]}\|\mathbf{X}_i\|_2 \leqslant \sqrt{Cn}$ for some $C > 0$.*

The results from this section are standard in the Lasso literature [Wai19, BvdG11] for the case when the regularisation path from LARS algorithm exactly minimises the Lasso cost function. We generalise them for case when $\mathcal{P}$ is an approximate regularisation path with error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}\|_1$.

5.1 Slow rates

It is possible to obtain bounds on the mean-squared prediction error without barely any assumptions on the design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and the noise vector $\mathbf{w} \in \mathbb{R}^n$, as the next result shows.

Theorem 32. *Let $\lambda > 0$ and $\epsilon \in [0, 1)$. Any approximate solution $\widetilde{\boldsymbol{\beta}}$ with error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}\|_1$ of the Lasso with regularisation parameter $\lambda \geqslant \|\mathbf{X}^\top \mathbf{w}\|_\infty/(1-\epsilon)$ satisfies*

$$\|\mathbf{X}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}})\|_2^2 \leqslant 2(2-\epsilon)\lambda\|\boldsymbol{\beta}^*\|_1.$$

Proof. Notice first that, since $\widetilde{\boldsymbol{\beta}}$ is a minimiser up to additive error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}\|_1$, then

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\|_2^2 + \lambda(1-\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1 \leqslant \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_1.$$

Use the equality $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{w}$ in the above inequality to obtain

$$\frac{1}{2}\|\mathbf{X}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}}) + \mathbf{w}\|_2^2 + \lambda(1-\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1 \leqslant \frac{1}{2}\|\mathbf{w}\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_1,$$

from which we get the following inequalities

$$\begin{aligned}\frac{1}{2}\|\mathbf{X}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}})\|_2^2 + \lambda(1-\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1 + \frac{1}{2}\|\mathbf{w}\|_2^2 + \mathbf{w}^\top\mathbf{X}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}}) \leqslant \frac{1}{2}\|\mathbf{w}\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_1 \implies \\ \frac{1}{2}\|\mathbf{X}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}})\|_2^2 + \lambda(1-\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1 + \mathbf{w}^\top\mathbf{X}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}}) \leqslant \lambda\|\boldsymbol{\beta}^*\|_1\end{aligned}$$

and finally

$$\begin{aligned}\frac{1}{2}\|\mathbf{X}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}})\|_2^2 &\leqslant \mathbf{w}^\top\mathbf{X}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \lambda\|\boldsymbol{\beta}^*\|_1 - \lambda(1-\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1 \\ &\leqslant \|\mathbf{X}^\top\mathbf{w}\|_\infty\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \lambda\|\boldsymbol{\beta}^*\|_1 - \lambda(1-\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1 \\ &\leqslant \|\mathbf{X}^\top\mathbf{w}\|_\infty(\|\widetilde{\boldsymbol{\beta}}\| + \|\boldsymbol{\beta}^*\|_1) + \lambda\|\boldsymbol{\beta}^*\|_1 - \lambda(1-\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1 \\ &= \|\widetilde{\boldsymbol{\beta}}\|\left(\|\mathbf{X}^\top\mathbf{w}\|_\infty - \lambda(1-\epsilon)\right) + \|\boldsymbol{\beta}^*\|(\|\mathbf{X}^\top\mathbf{w}\|_\infty + \lambda) \\ &\leqslant \|\boldsymbol{\beta}^*\|(\|\mathbf{X}^\top\mathbf{w}\|_\infty + \lambda) \\ &\leqslant (2-\epsilon)\lambda\|\boldsymbol{\beta}^*\|,\end{aligned}$$

where the second inequality follows from Hölder's inequality and the third follows from the triangle inequality. □

We give some examples of the above result using common linear regression models.

Linear Gaussian model. In the classical linear Gaussian model, the design matrix is deterministic, meaning that $\mathbf{X} \in \mathbb{R}^{n \times d}$ is fixed, while the noise vector $\mathbf{w} \in \mathbb{R}^n$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. We then bound the probability that $\lambda \geqslant \|\mathbf{X}^\top\mathbf{w}\|_\infty/(1-\epsilon)$. Using that $\mathbf{X}_i^\top\mathbf{w}$ is a sub-Gaussian random variable with parameter $\sigma\|\mathbf{X}_i\|_2 \leqslant \sigma\sqrt{Cn}$ (Fact 6), we bound the complement of this probability:

$$\mathbb{P}\left[\|\mathbf{X}^\top\mathbf{w}\|_\infty > t\right] = \mathbb{P}\left[\max_{i\in[d]}|\mathbf{X}_i^\top\mathbf{w}| > t\right] \leqslant \sum_{i=1}^d \mathbb{P}[|\mathbf{X}_i^\top\mathbf{w}| > t] \leqslant 2d\exp\left(-\frac{t^2}{2C\sigma^2 n}\right),$$

where the first inequality follows from a union bound. Setting $t = \lambda(1-\epsilon) = \sqrt{2C\sigma^2 n \log(2d/\delta)}$, the above probability is at most δ. Consequently,

$$\mathbb{P}\left[\|\mathbf{X}^\top \mathbf{w}\|_\infty \leqslant \lambda(1-\epsilon)\right] \geqslant 1-\delta.$$

Substituting this value of λ in Theorem 32, with probability at least $1-\delta$,

$$\frac{\|\mathbf{X}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}})\|_2^2}{n} \leqslant \frac{2(2-\epsilon)\lambda\|\boldsymbol{\beta}^*\|}{n} = \frac{2-\epsilon}{1-\epsilon}\sqrt{\frac{8C\sigma^2\log(2d/\delta)}{n}}\|\boldsymbol{\beta}^*\|.$$

We see that the mean square error vanishes by a square root of n factor as n becomes large. Without additional assumptions, this is known as a slow rate.

Compressed sensing. In compressed sensing, it is common for the design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ to be chosen by the user, and a standard choice is the standard Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Assume further that the noise vector $\mathbf{w} \in \mathbb{R}^n$ is deterministic with $\|\mathbf{w}\|_\infty \leqslant \sigma$. Thus $\mathbf{X}_i^\top \mathbf{w}$ is a sub-Gaussian random variable with parameter $\|\mathbf{w}\|_2 \leqslant \sqrt{n}\|\mathbf{w}\|_\infty \leqslant \sqrt{n}\sigma$ (Fact 6). By following the exact same steps as the previous example (this time with $C = 1$), we conclude that, with probability at least $1-\delta$,

$$\frac{\|\mathbf{X}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}})\|_2^2}{n} \leqslant \frac{2-\epsilon}{1-\epsilon}\sqrt{\frac{8\sigma^2\log(2d/\delta)}{n}}\|\boldsymbol{\beta}^*\|.$$

So the mean square error also vanishes by a square root of n factor as n becomes large. We shall see next that, by imposing additional restrictions on the design matrix, faster rates can be obtained.

5.2 Fast Rates

To obtain faster rates, we need additional assumptions on the design matrix $\mathbf{X}$. The condition below originates actually in the compressed sensing literature. We assume the true solution $\boldsymbol{\beta}^*$ has a hard finite support $\mathrm{supp}(\boldsymbol{\beta}^*)$.

Definition 33 (Restricted Eigenvalue condition)**.** *For $S \subseteq [d]$ and $\zeta \in \mathbb{R}$, let*

$$\mathbb{C}_\zeta(S) \triangleq \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}_{S^c}\|_1 \leqslant \zeta\|\boldsymbol{\Delta}_S\|_1\}.$$

A matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ satisfies the Restricted Eigenvalue (RE) *condition over S with parameters (κ, ζ) if*

$$\frac{1}{n}\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 \geqslant \kappa\|\boldsymbol{\Delta}\|_2^2 \quad \forall \boldsymbol{\Delta} \in \mathbb{C}_\zeta(S).$$

An interpretation of the RE condition is that it bounds away the minimum eigenvalues of the matrix $\mathbf{X}^\top\mathbf{X}$ in specific directions. Ideally, we would like to bound the curvature of the cost function $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ in all directions, which would guarantee strong bounds on the mean square error. However, in the high-dimensional setting $d \gg n$, $\mathbf{X}^\top\mathbf{X}$ is a $d \times d$ matrix with rank at most n, so it is impossible to guarantee a positive curvature in all directions. Therefore, we must relax such stringent condition on the curvature and require that it holds only on a subset $\mathbb{C}_\zeta(S)$ of vectors. The RE condition yields the following stronger result.

Theorem 34. *Let $\lambda, \kappa > 0$ and $\epsilon \in [0, 1/4]$. Let $S \triangleq \mathrm{supp}(\boldsymbol{\beta}^*)$. Suppose $\mathbf{X} \in \mathbb{R}^{n\times d}$ satisfies the* RE *condition over S with parameters $(\kappa, 5)$. Any approximate solution $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^d$ with error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}\|_1$ of the Lasso with regularisation parameter $\lambda \geqslant 4\|\mathbf{X}^\top\mathbf{w}\|_\infty$ satisfies*

$$\|\mathbf{X}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + 3\lambda\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leqslant 81\lambda^2\frac{|S|}{n\kappa} + 18\lambda\epsilon\|\boldsymbol{\beta}^*_S\|_1.$$

The proof has many variations, and we give the most direct one below. Note that the above implies

$$\|\mathbf{X}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \leqslant 81\lambda^2 \frac{|S|}{n\kappa} + 18\lambda\epsilon\|\boldsymbol{\beta}^*_S\|_1, \qquad \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leqslant 27\lambda\frac{|S|}{n\kappa} + 6\epsilon\|\boldsymbol{\beta}^*_S\|_1.$$

Proof. We use the basic optimality of $\widetilde{\boldsymbol{\beta}}$,

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\|_2^2 + \lambda(1-\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1 \leqslant \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_1.$$

Define $\boldsymbol{\Delta} \triangleq \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Expanding out the terms like in Theorem 32,

$$\begin{aligned}
0 \leqslant 2\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 &\leqslant 4\mathbf{w}^\top\mathbf{X}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + 4\lambda\|\boldsymbol{\beta}^*\|_1 - 4\lambda(1-\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1 \\
&\leqslant 4\|\mathbf{X}^\top\mathbf{w}\|_\infty\|\boldsymbol{\Delta}\|_1 + 4\lambda\|\boldsymbol{\beta}^*\|_1 - 4\lambda(1-\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1 \\
&\leqslant \lambda\|\boldsymbol{\Delta}\|_1 + 4\lambda\|\boldsymbol{\beta}^*\|_1 - 4\lambda(1-\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1 \\
&\leqslant \lambda(\|\widetilde{\boldsymbol{\beta}}\|_1 + \|\boldsymbol{\beta}^*\|_1) + 4\lambda\|\boldsymbol{\beta}^*\|_1 - 4\lambda(1-\epsilon)\|\widetilde{\boldsymbol{\beta}}\|_1,
\end{aligned}$$

from which we infer that $\|\widetilde{\boldsymbol{\beta}}\|_1 \leqslant \frac{5}{3-4\epsilon}\|\boldsymbol{\beta}^*\|_1 \leqslant \frac{5}{2}\|\boldsymbol{\beta}^*\|_1$. Let us now return to the inequality $2\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 \leqslant \lambda\|\boldsymbol{\Delta}\|_1 + 4\lambda(\|\boldsymbol{\beta}^*\|_1 - \|\widetilde{\boldsymbol{\beta}}\|_1) + 4\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}\|_1$ and further decompose the right-hand side into the respective partitioning sets S and S^c by using that $\boldsymbol{\beta}^*_S + \boldsymbol{\Delta}_S = \widetilde{\boldsymbol{\beta}}_S$ and $\boldsymbol{\Delta}_{S^c} = \widetilde{\boldsymbol{\beta}}_{S^c}$, since $\boldsymbol{\beta}^*_{S^c} = 0$. We also use the inequality $\|\widetilde{\boldsymbol{\beta}}\|_1 \leqslant \frac{5}{2}\|\boldsymbol{\beta}^*\|_1 = \frac{5}{2}\|\boldsymbol{\beta}^*_S\|_1$. Therefore

$$\begin{aligned}
2\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 &\leqslant \lambda\|\boldsymbol{\Delta}\|_1 + 4\lambda(\|\boldsymbol{\beta}^*\|_1 - \|\widetilde{\boldsymbol{\beta}}\|_1) + 10\lambda\epsilon\|\boldsymbol{\beta}^*_S\|_1 \\
&\leqslant \lambda(\|\boldsymbol{\Delta}_S\|_1 + \|\boldsymbol{\Delta}_{S^c}\|_1) + 4\lambda(\|\boldsymbol{\beta}^*\|_1 - \|\widetilde{\boldsymbol{\beta}}\|_1) + 10\lambda\epsilon\|\boldsymbol{\beta}^*_S\|_1 \\
&= \lambda(\|\boldsymbol{\Delta}_S\|_1 + \|\boldsymbol{\Delta}_{S^c}\|_1) + 4\lambda(\|\boldsymbol{\beta}^*_S\|_1 - \|\boldsymbol{\beta}^*_S + \boldsymbol{\Delta}_S\|_1 - \|\boldsymbol{\Delta}_{S^c}\|_1) + 10\lambda\epsilon\|\boldsymbol{\beta}^*_S\|_1 \\
&\overset{(*)}{\leqslant} \lambda(\|\boldsymbol{\Delta}_S\|_1 + \|\boldsymbol{\Delta}_{S^c}\|_1) + 4\lambda(\|\boldsymbol{\Delta}_S\|_1 - \|\boldsymbol{\Delta}_{S^c}\|_1) + 10\lambda\epsilon\|\boldsymbol{\beta}^*_S\|_1 \\
&= \lambda(5\|\boldsymbol{\Delta}_S\|_1 - 3\|\boldsymbol{\Delta}_{S^c}\|_1) + 10\lambda\epsilon\|\boldsymbol{\beta}^*_S\|_1,
\end{aligned}$$

where the inequality $(*)$ follows from

$$\|\boldsymbol{\beta}^*_S\|_1 - \|\boldsymbol{\beta}^*_S + \boldsymbol{\Delta}_S\|_1 \leqslant \|\boldsymbol{\beta}^*_S\|_1 - (\|\boldsymbol{\beta}^*_S\|_1 - \|\boldsymbol{\Delta}_S\|_1) = \|\boldsymbol{\Delta}_S\|_1.$$

We now proceed by considering two cases: (Case 1) $\|\boldsymbol{\Delta}_S\|_1 \geqslant \epsilon\|\boldsymbol{\beta}^*_S\|_1$ or (Case 2) $\|\boldsymbol{\Delta}_S\|_1 < \epsilon\|\boldsymbol{\beta}^*_S\|_1$. Therefore it must hold that either (Case 1)

$$2\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 + 3\lambda\|\boldsymbol{\Delta}_{S^c}\|_1 \leqslant 15\lambda\|\boldsymbol{\Delta}_S\|_1$$

or (Case 2)

$$2\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 + 3\lambda\|\boldsymbol{\Delta}_{S^c}\|_1 \leqslant 15\lambda\epsilon\|\boldsymbol{\beta}^*_S\|_1.$$

In the first case, we find that $\boldsymbol{\Delta} \in \mathbb{C}_5(S)$, which means that we can now apply the RE condition. More specifically, from $2\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 + 3\lambda\|\boldsymbol{\Delta}_{S^c}\|_1 \leqslant 15\lambda\|\boldsymbol{\Delta}_S\|_1$ we get

$$2\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 + 3\lambda\|\boldsymbol{\Delta}\|_1 \leqslant 18\lambda\|\boldsymbol{\Delta}_S\|_1 \leqslant 18\lambda\sqrt{|S|}\|\boldsymbol{\Delta}\|_2 \leqslant 18\lambda\sqrt{\frac{|S|}{n\kappa}}\|\mathbf{X}\boldsymbol{\Delta}\|_2 \leqslant 81\lambda^2\frac{|S|}{n\kappa} + \|\mathbf{X}\boldsymbol{\Delta}\|_2^2,$$

where the last inequality follows from $2ab \leqslant a^2 + b^2$ for $a, b \in \mathbb{R}$. Thus $\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 + 3\lambda\|\boldsymbol{\Delta}\|_1 \leqslant 81\lambda^2\frac{|S|}{n\kappa}$.

In the second case, from $2\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 + 3\lambda\|\boldsymbol{\Delta}_{S^c}\|_1 \leqslant 15\lambda\epsilon\|\boldsymbol{\beta}^*_S\|_1$ we get

$$2\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 + 3\lambda\|\boldsymbol{\Delta}\|_1 \leqslant 18\lambda\epsilon\|\boldsymbol{\beta}^*_S\|_1 \implies \|\mathbf{X}\boldsymbol{\Delta}\|_2^2 + 3\lambda\|\boldsymbol{\Delta}\|_1 \leqslant 18\lambda\epsilon\|\boldsymbol{\beta}^*_S\|_1.$$

Putting both Cases 1 and 2 together leads to the desired result. ∎

Linear Gaussian model and compressed sensing. Once again, we can apply the above result to common linear regression models. In the case when $\mathbf{X} \in \mathbb{R}^{n\times d}$ is fixed and $\mathbf{w} \in \mathbb{R}^n$ has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries, the same reasoning from the previous section yields

$$\mathbb{P}\left[\lambda \geqslant 4\|\mathbf{X}^\top \mathbf{w}\|_\infty\right] \geqslant 1-\delta$$

by setting $\lambda = \sqrt{32C\sigma^2 n\log(2d/\delta)}$. This value of λ leads to, with probability at least $1-\delta$,

$$\frac{\|\mathbf{X}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2^2}{n} \leqslant \frac{2592C\sigma^2|S|\log(2d/\delta)}{\kappa n} + 18\epsilon\sqrt{\frac{32C\sigma^2\log(2d/\delta)}{\kappa n}}\|\boldsymbol{\beta}^*_S\|_1.$$

The same bound (with $C=1$) holds for the model where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the standard Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ entries and $\mathbf{w} \in \mathbb{R}^n$ is fixed with $\|\mathbf{w}\|_\infty \leqslant \sigma$. Thus, under the RE condition, if $\epsilon = O(\sqrt{\log(d)/n})$, then the mean square error vanishes by an n factor as n becomes large, which can be substantially smaller than the $\sqrt{\log(d)/n}$ bound from the previous section. This is known as a fast rate.

6 Discussion and future work

We studied and quantised the LARS pathwise algorithm (or homotopy method) proposed by Efron *et al.* [EHJT04] and Osborne *et al.* [OPT00b, OPT00a] which produces the set of Lasso solutions for varying the penalty term λ. By assuming quantum access to the Lasso input, we proposed two quantum algorithms. The first one (Algorithm 1) simply replaces the classical search within the joining time calculation with the quantum minimum-finding subroutine from Dürr and Høyer [DH96]. Similar to the classical LARS algorithm, it outputs the exact Lasso path, but has an improved runtime by a quadratic factor in the number of features d. Our second quantum algorithm (Algorithm 3), on the other hand, outputs an approximate Lasso path $\widetilde{\mathcal{P}}$ by computing the joining times up to some error. This is done by approximating the joining times $\Lambda_i^{\text{join}}(\mathcal{A}) = \frac{\mathbf{X}_i^\top(\mathbf{y}-\mathbf{X}_\mathcal{A}\boldsymbol{\mu})}{\pm 1-\mathbf{X}_i^\top\mathbf{X}_\mathcal{A}\boldsymbol{\theta}}$ using quantum amplitude estimation [BHMT02] and finding their maximum via the approximate quantum minimum-finding subroutine from Chen and de Wolf [CdW23]. Consequently, the runtime of our approximate quantum LARS algorithm is quadratically improved in both the number of features d and observations n. We stabilised the correctness of Algorithm 3 by employing an approximate version of the KKT conditions and a duality gap, and showed that $\widetilde{\mathcal{P}}$ is a minimiser of the Lasso cost function up to additive error $\lambda\epsilon\|\widetilde{\boldsymbol{\beta}}(\lambda)\|_1$. Finally, we dequantised Algorithm 3 and proposed an approximate classical LARS algorithm (Algorithm 4) based on sampling that retains the quadratic improvement in the number of observations n.

The time complexity of our approximate LARS algorithms, both classical and quantum, directly depends on the design matrix $\mathbf{X}$ via the quantity $\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2$, the mutual incoherence, and the mutual overlap between the design matrix and the vector of observations $\mathbf{y}$. If $\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2$ is a constant and mutual incoherence and overlap are both bounded away from 0, then the approximate quantum LARS algorithm's complexity is $\widetilde{O}(\sqrt{n|\mathcal{I}|}/\epsilon + n|\mathcal{A}| + |\mathcal{A}|^2)$ per iteration. For small sizes of the active set, $|\mathcal{A}| = O(1)$, we obtain the overall quadratic improvement $\widetilde{O}(\sqrt{nd})$ over the classical LARS algorithm. However, for large sizes as $|\mathcal{A}| = \Omega(n)$, it is more advantageous to employ the simple quantum LARS algorithm since its complexity is just $\widetilde{O}(n\sqrt{d})$. The simple quantum LARS algorithm, moreover, does not depend on $\|\mathbf{X}\|_{\max}\|\mathbf{X}^+\|_2$ and the mutual incoherence and overlap, and therefore should be superior for a wider class of design matrices. We proved, nonetheless, that Algorithm 3 exhibits the complexity $\widetilde{O}(\sqrt{nd})$ per iteration for the case when $\mathbf{X}$ is a standard Gaussian random matrix with high probability.

We point out some future directions of research following our work. In Algorithms 3 and 4, the crossing times are computed exactly and the joining times are estimated up to some error. A natural extension would be to allow errors also in computing the crossing times. This would require new techniques in order to retain the path solution continuity. Another direction is to reduce the number of iterations of the LARS algorithm. In the classical setting, Mairal and Yu [MY12] proposed an approximate LARS algorithm with a maximum number of iterations by employing first-order optimisation methods when two kinks are too close to each other. Finally, it would be interesting to design a fully quantum LARS algorithm by using efficient quantum subroutines for matrix multiplication and matrix inversion, e.g. based on block-encoding techniques [LC19, GSLW19, CGJ19].

Our algorithms are designed to work only in the fault tolerant regime. Another future direction is to study the LARS algorithm in the NISQ (Noisy Intermediate-Scale Quantum) setting [Pre18], where algorithms operate on a relatively small number of qubits and have shallow circuit depths. In addition, while we have studied the plain-vanilla LARS algorithm, other Lasso related algorithms still remain unexplored in the quantum setting. For example, fused Lasso [TSR$^+$05, XKW$^+$16, Hoe10], which penalises the ℓ_1-norm of both the coefficients and their successive differences, is a generalisation of Lasso with applications to support vector classifier [Gun98, RML14, SV99]. On the other hand, grouped Lasso is a generalised model for linear regression with ℓ_1 and ℓ_2-penalties [SFHT13, FHT10b]. This model is used in settings where the design matrix can be decomposed into groups of submatrices. Simon *et al.* [SFHT13] proposed a sparse grouped Lasso algorithm which finds applications in sparse graphical modelling. We believe that similar techniques employed in this work could be applied to these alternative Lasso settings.

7 Acknowledgements

JFD specially thanks Hanzhong Liu, Julien Mairal, and Bin Yu for very useful clarifications regarding Ref. [MY12] and Lasso in general. JFD also thanks Yihui Quek for pointing out Ref. [QCR20] and Rahul Jain, Josep Lumbreras, and Marco Tomamichel for interesting discussions. DL acknowledges funding from the Latvian Quantum Initiative under EU Recovery and Resilience Facility under project no. 2.3.1.1.i.0/1/22/I/CFLA/001 in the final part of the project. CSP gratefully acknowledges Ministry of Education (MOE), AcRF Tier 2 grant (Reference No: MOE-T2EP20220-0013) for the funding of this research. This research is supported by the National Research Foundation, Singapore and A*STAR under its CQT Bridging Grant and its Quantum Engineering Programme under grant NRF2021-QEP2-02-P05.

References

[ABD$^+$23] Jonathan Allcock, Jinge Bao, João F. Doriguello, Alessandro Luongo, and Miklos Santha. Constant-depth circuits for Uniformly Controlled Gates and Boolean functions with application to quantum memory circuits. *arXiv preprint arXiv:2308.08539*, 2023. 10

[AT16] Taylor B. Arnold and Ryan J. Tibshirani. Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, 25(1):1–27, 2016. 8

[BEM13] Mohsen Bayati, Murat A Erdogdu, and Andrea Montanari. Estimating lasso risk and noise level. *Advances in Neural Information Processing Systems*, 26, 2013. 2

[BHMT02] Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. Quantum amplitude amplification and estimation. *Contemporary Mathematics*, 305:53–74, 2002. 5, 11, 35

[BL06] Jonathan Borwein and Adrian Lewis. *Convex Analysis*. Springer, 2006. 14, 20

[BT19] Giorgos Borboudakis and Ioannis Tsamardinos. Forward-backward selection with early dropping. *The Journal of Machine Learning Research*, 20(1):276–314, 2019. 7

[BvdG11] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011. 2, 31

[BZ22] Armando Bellante and Stefano Zanero. Quantum matching pursuit: A quantum algorithm for sparse representations. *Physical Review A*, 105(2):022414, 2022. 8

[CD05] Zizhong Chen and Jack J. Dongarra. Condition numbers of Gaussian random matrices. *SIAM Journal on Matrix Analysis and Applications*, 27(3):603–620, 2005. 7, 29

[CDS01] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001. 3

[CdW23] Yanlin Chen and Ronald de Wolf. Quantum algorithms and lower bounds for linear regression with norm constraints. In *50th International Colloquium on Automata, Languages, and Programming (ICALP 2023)*, volume 261 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 38:1–38:21, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. 5, 7, 11, 12, 21, 25, 35

[CGJ19] Shantanav Chakraborty, András Gilyén, and Stacey Jeffery. The power of block-encoded matrix powers: Improved regression techniques via faster hamiltonian simulation. *Proceedings of the 46th International Colloquium on Automata, Languages, and Programming (ICALP)*, 132, 2019. 8, 36

[CMP23] Shantanav Chakraborty, Aditya Morolia, and Anurudh Peduri. Quantum regularized least squares. *Quantum*, 7:988, 2023. 8

[CP09] Emmanuel J. Candès and Yaniv Plan. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009. 15

[CRT06] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006. 3

[CYGL23] Menghan Chen, Chaohua Yu, Gongde Guo, and Song Lin. Faster quantum ridge regression algorithm for prediction. *International Journal of Machine Learning and Cybernetics*, 14(1):117–124, 2023. 8

[DH96] Christoph Dürr and Peter Høyer. A quantum algorithm for finding the minimum. *arXiv preprint quant-ph/9607014*, 1996. 5, 11, 19, 35

[DH01] David L. Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *IEEE transactions on information theory*, 47(7):2845–2862, 2001. 7

[DK08] Abhimanyu Das and David Kempe. Algorithms for subset selection in linear regression. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 45–54, 2008. 3

[Don06] David L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(6):797–829, 2006. 15

[Dor14] Ashok Vithoba Dorugade. New ridge parameters for ridge regression. *Journal of the Association of Arab Universities for Basic and Applied Sciences*, 15:94–99, 2014. 8

[Dos12] Charles Dossal. A necessary and sufficient condition for exact sparse recovery by ℓ_1 minimization. *Comptes Rendus Mathematique*, 350(1-2):117–120, 2012. 15

[EB02] Michael Elad and Alfred M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 48(9):2558–2567, 2002. 7

[EHJT04] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004. 3, 4, 8, 15, 16, 35

[ES05] Alan Edelman and Brian D. Sutton. Tails of condition number distributions. *SIAM journal on matrix analysis and applications*, 27(2):547–560, 2005. 29

[FHT10a] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010. 3

[FHT10b] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010. 36

[FL10] Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010. 3

[FNW07] Mário A.T. Figueiredo, Robert D. Nowak, and Stephen J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of selected topics in signal processing*, 1(4):586–597, 2007. 3

[FR13] Simon Foucart and Holger Rauhut. *An invitation to compressive sensing*. Springer, 2013. 16

[Fuc04] J.-J. Fuchs. Recovery of exact sparse representations in the presence of noise. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages ii–533. IEEE, 2004. 7

[Fuc05] J.J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608, 2005. 15

[GG08] Pierre Garrigues and Laurent Ghaoui. An homotopy algorithm for the Lasso with online observations. *Advances in neural information processing systems*, 21, 2008. 2

[GKZ18] Brian R. Gaines, Juhyun Kim, and Hua Zhou. Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics*, 27(4):861–871, 2018. 8

[GLM08a] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Architectures for a quantum random access memory. *Physical Review A*, 78(5):052310, 2008. 10

[GLM08b] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Quantum random access memory. *Physical review letters*, 100(16):160501, 2008. 10

[GSLW19] András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 193–204, 2019. 8, 36

[Gun98] Steve R. Gunn. Support vector machines for classification and regression. Technical Report 1, University of Southampton, 1998. 36

[HHL09] Aram W. Harrow, Avinatan Hassidim, and Seth Lloyd. Quantum algorithm for linear systems of equations. *Physical review letters*, 103(15):150502, 2009. 8

[HK70] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. 7, 8

[HKB75] Arthur E. Hoerl, Robert W. Kannard, and Kent F. Baldwin. Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, 4(2):105–123, 1975. 8

[HMZ08] Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618, 2008. 6

[Hoe10] Holger Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010. 36

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, May 2015. 3

[JT09] Iain M. Johnstone and D. Michael Titterington. Statistical challenges of high-dimensional data, 2009. 2

[Kib03] B.M. Golam Kibria. Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computation*, 32(2):419–435, 2003. 8

[KKK08] Jinseog Kim, Yuwon Kim, and Yongdai Kim. A gradient-based optimization algorithm for LASSO. *Journal of Computational and Graphical Statistics*, 17(4):994–1009, 2008. 8

[KKMR22] Jonathan Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. Lower bounds on randomly preconditioned Lasso via robust sparse designs. *Advances in Neural Information Processing Systems*, 35:24419–24431, 2022. 2

[KM14] Vipin Kumar and Sonajharia Minz. Feature selection: a literature review. *SmartCR*, 4(3):211–229, 2014. 7

[KMTY21] Kazuya Kaneko, Koichi Miyamoto, Naoyuki Takeda, and Kazuyoshi Yoshino. Linear regression by quantum amplitude estimation and its extension to convex optimization. *Physical Review A*, 104(2):022430, 2021. 8

[KP17] Iordanis Kerenidis and Anupam Prakash. Quantum recommendation systems. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 49:1–49:21, Dagstuhl, Germany, 2017. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. 8, 11

[KP20] Iordanis Kerenidis and Anupam Prakash. Quantum gradient descent for linear systems and least squares. *Physical Review A*, 101(2):022316, 2020. 8

[LC19] Guang Hao Low and Isaac L. Chuang. Hamiltonian simulation by qubitization. *Quantum*, 3:163, 2019. 8, 36

[LMR14] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. Quantum principal component analysis. *Nature Physics*, 10(9):631–633, 2014. 8

[LTTT14] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413, 2014. 3

[Mao02] K.Z. Mao. Fast orthogonal forward selection algorithm for feature subset selection. *IEEE Transactions on Neural Networks*, 13(5):1218–1224, 2002. 7

[Mao04] K.Z. Mao. Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):629–634, 2004. 7

[MB06] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436 – 1462, 2006. 7

[McD09] Gary C. McDonald. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100, 2009. 8

[MJ73] Carl D. Meyer Jr. Generalized inversion of modified matrices. *SIAM Journal on Applied Mathematics*, 24(3):315–323, 1973. 18

[MOK23] Xiangming Meng, Tomoyuki Obuchi, and Yoshiyuki Kabashima. On model selection consistency of lasso for high-dimensional Ising models. In *International Conference on Artificial Intelligence and Statistics*, pages 6783–6805. PMLR, 2023. 3

[MS75] Donald W. Marquardt and Ronald D. Snee. Ridge regression in practice. *The American Statistician*, 29(1):3–20, 1975. 8

[MSK+19] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019. 3

[MY12] Julien Mairal and Bin Yu. Complexity analysis of the lasso regularization path. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1835–1842, 2012. 5, 8, 14, 16, 20, 22, 36

[OPT00a] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000. 4, 8, 16, 35

[OPT00b] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the LASSO and its dual. *Journal of Computational and Graphical statistics*, 9(2):319–337, 2000. 4, 8, 15, 16, 35

[PL22] Yiming Peng and Vadim Linetsky. Portfolio selection: A statistical learning approach. In *Proceedings of the Third ACM International Conference on AI in Finance*, pages 257–263, 2022. 3

[Pra14] Anupam Prakash. *Quantum algorithms for linear algebra and machine learning*. University of California, Berkeley, 2014. 8, 11

[Pre18] John Preskill. Quantum computing in the NISQ era and beyond. *Quantum*, 2:79, 2018. 36

[QCR20] Yihui Quek, Clement Canonne, and Patrick Rebentrost. Robust quantum minimum finding with an application to hypothesis selection. *arXiv preprint arXiv:2003.11777*, 2020. 11, 36

[RCC+22] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022. 3

[Reg70] Ridge Regression. Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. 8

[RML14] Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. Quantum support vector machine for big data classification. *Physical review letters*, 113(13):130503, 2014. 36

[Ros04] Saharon Rosset. Following curved regularized optimization solution paths. *Advances in Neural Information Processing Systems*, 17, 2004. 15

[Rot04] Volker Roth. The generalized lasso. *IEEE transactions on neural networks*, 15(1):16–28, 2004. 8

[RS14] Matthias Reif and Faisal Shafait. Efficient feature size reduction via predictive forward selection. *Pattern Recognition*, 47(4):1664–1673, 2014. 7

[RZ07] Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030, 2007. 8, 14, 16, 18

[SCL+18] Karl Sjöstrand, Line Harder Clemmensen, Rasmus Larsen, Gudmundur Einarsson, and Bjarne Ersbøll. SpaSM: A MATLAB toolbox for sparse statistical modeling. *Journal of Statistical Software*, 84:1–37, 2018. 14, 16

[SFHT13] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013. 36

[SGV98] Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998. 8

[SSP16] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. Prediction by linear regression on a quantum computer. *Physical Review A*, 94(2):022342, 2016. 8

[Sto13] Mihailo Stojnic. A framework to characterize performance of LASSO algorithms. *arXiv preprint arXiv:1303.7291*, 2013. 8

[SV99] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9:293–300, 1999. 36

[SX20] Changpeng Shao and Hua Xiang. Quantum regularized least squares solver with parameter estimate. *Quantum Information Processing*, 19:1–20, 2020. 8

[TDK23] Ryan Thompson, Amir Dezfouli, and Robert Kohn. The contextual lasso: Sparse linear models via deep neural networks. *arXiv preprint arXiv:2302.00878*, 2023. 3

[TFZB08] Feng Tan, Xuezheng Fu, Yanqing Zhang, and Anu G. Bourgeois. A genetic algorithm-based method for feature subset selection. *Soft Computing*, 12:111–120, 2008. 7

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996. 3

[Tib13] Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013. 4, 14, 15, 16

[Tro06] Joel A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, 52(3):1030–1051, 2006. 7

[TSR+05] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(1):91–108, 2005. 36

[UGH09] M. Graziano Usai, Mike E. Goddard, and Ben J. Hayes. LASSO with cross-validation for genomic selection. *Genetics research*, 91(6):427–436, 2009. 3

[Vin78] Hrishikesh D. Vinod. A survey of ridge regression and related techniques for improvements over ordinary least squares. *The Review of Economics and Statistics*, pages 121–131, 1978. 8

[VK05] Dimitrios Ververidis and Constantine Kotropoulos. Sequential forward feature selection with low computational cost. In *2005 13th European Signal Processing Conference*, pages 1–4. IEEE, 2005. 7

[vW15] Wessel N. van Wieringen. Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*, 2015. 8

[Wai09] Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009. 7, 15, 29, 30

[Wai19] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019. 6, 29, 31

[Wan17] Guoming Wang. Quantum algorithm for linear regression. *Physical review A*, 96(1):012335, 2017. 8

[WB06] Hua-Liang Wei and Stephen A. Billings. Feature subset selection and ranking for data dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):162–166, 2006. 7

[WBL12] Nathan Wiebe, Daniel Braun, and Seth Lloyd. Quantum algorithm for data fitting. *Physical review letters*, 109(5):050505, 2012. 8

[WCH+09] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009. 3

[WFL00] David C. Whitley, Martyn G. Ford, and David J. Livingstone. Unsupervised forward selection: a method for eliminating redundant variables. *Journal of chemical information and computer sciences*, 40(5):1160–1168, 2000. 7

[WM22] John Wright and Yi Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022. 2

[XKW+16] Bo Xin, Yoshinobu Kawahara, Yizhou Wang, Lingjing Hu, and Wen Gao. Efficient generalized fused lasso and its applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(4):1–22, 2016. 36

[YGW19] Chao-Hua Yu, Fei Gao, and Qiao-Yan Wen. An improved quantum algorithm for ridge regression. *IEEE Transactions on Knowledge and Data Engineering*, 33(3):858–866, 2019. 8

[ZJ96] Douglas Zongker and Anil Jain. Algorithms for feature selection: An evaluation. In *Proceedings of 13th international conference on pattern recognition*, volume 2, pages 18–22. IEEE, 1996. 7

[ZLZ18] Tuo Zhao, Han Liu, and Tong Zhang. Pathwise coordinate optimization for sparse learning: Algorithm and theory. *The Annals of Statistics*, 46(1):180 – 218, 2018. 3

[ZY06] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006. 2, 3, 6, 7

A Summary of symbols

The symbols and their corresponding concept are summarised in the table below.

Symbol	Explanation
n	Number of observations or sample points
d	Number of features
$\mathbf{y}$	Vector of observations
$\mathbf{X}$	Design matrix
$\boldsymbol{\beta}$	Optimisation variables
$\hat{\boldsymbol{\beta}}$	Lasso optimal solution
$\widetilde{\boldsymbol{\beta}}$	Lasso approximate solution
λ	Regularisation or penalty parameter
$\mathcal{P}$	Optimal regularisation Path
$\widetilde{\mathcal{P}}$	Approximate regularisation Path
$\mathcal{A}$	Active set
$\mathcal{I}$	Inactive set
$\boldsymbol{\eta}$	Equicorrelation signs
α	Mutual incoherence
γ	Mutual overlap

Table 2: Summary of symbols and their corresponding concept used in the paper.