# Analyzing Task Driven Learning Algorithms

## Mid Year Status

Mike Pekala

December 6, 2011

**Advisor:** Prof. Doron Levy (dlevy at math.umd.edu)
UMD Dept. of Mathematics & Center for Scientific Computation and Mathematical Modeling (CSCAMM)

# Today


1. Overview and Status
   - Overview
   - Schedule
2. Least Angle Regression
   - Geometric View
   - Rank 1 Updating
   - Validation
3. Dictionary Learning
   - Preliminary Results
4. Summary & Next Steps

# Overview

The underlying notion of *sparse coding* is that, in many domains, data vectors can be concisely represented as a sparse linear combination of basis elements or dictionary atoms. Recent results suggest that, for many tasks, performance improvements can be obtained by explicitly *learning* dictionaries directly from the data (vs. using predefined dictionaries, such as wavelets). Further results suggest that additional gains are possible by *jointly optimizing* the dictionary for both the data and the task (e.g. classification, denoising).

Consider the Task Driven Learning algorithm [Mairal et al., 2010]:

- Outer loop: Stochastic gradient descent to learn dictionary atoms, classification weight vector. We talked about this at the kickoff.
- Inner loop: Sparse approximation via penalized least squares. Authors use the Least Angle Regression (LARS) algorithm for this purpose. Vague at kickoff - we'll talk about this a bit more today.

# Schedule and Milestones

Schedule and milestones from the kickoff:

- Phase I: Algorithm development (Sept 23 - Jan 15)
  - Phase Ia: Implement LARS (Sept 23 $\sim$ Oct 24)
    - ✓ Milestone: LARS code available
  - Phase Ib: Validate LARS (Oct 24 $\sim$ Nov 14)
    - ✓ Milestone: results on diabetes data and hand-crafted problems
  - Phase Ic: Implement SGD framework (Nov 14 $\sim$ Dec 15)
    - 90% Milestone: Initial SGD code available
  - Phase Id: Validate SGD framework (Dec 15 $\sim$ Jan 15)
    - 25% Milestone: TDDL results on MNIST/USPS
- Phase II: Analysis on new data sets (Jan 15 - May 1)
  - Milestone: Preliminary results on selected dataset ($\sim$ Mar 1)
  - Milestone: Final report and presentation ($\sim$ May 1)

# Today


1. Overview and Status
   - Overview
   - Schedule
2. Least Angle Regression
   - Geometric View
   - Rank 1 Updating
   - Validation
3. Dictionary Learning
   - Preliminary Results
4. Summary & Next Steps

# Problem: Constrained Least Squares

Recall the Lasso: given $X = [x_1, \ldots, x_m] \in \mathbb{R}^{n \times m}, t \in \mathbb{R}_+$, solve:

$$\min_{\beta} \; ||y - X\beta||_2^2 \;\; s.t. \;\; ||\beta||_1 \leq t$$

which has an equivalent unconstrained formulation:

$$\min_{\beta} \; ||y - X\beta||_2^2 + \lambda ||\beta||_1$$

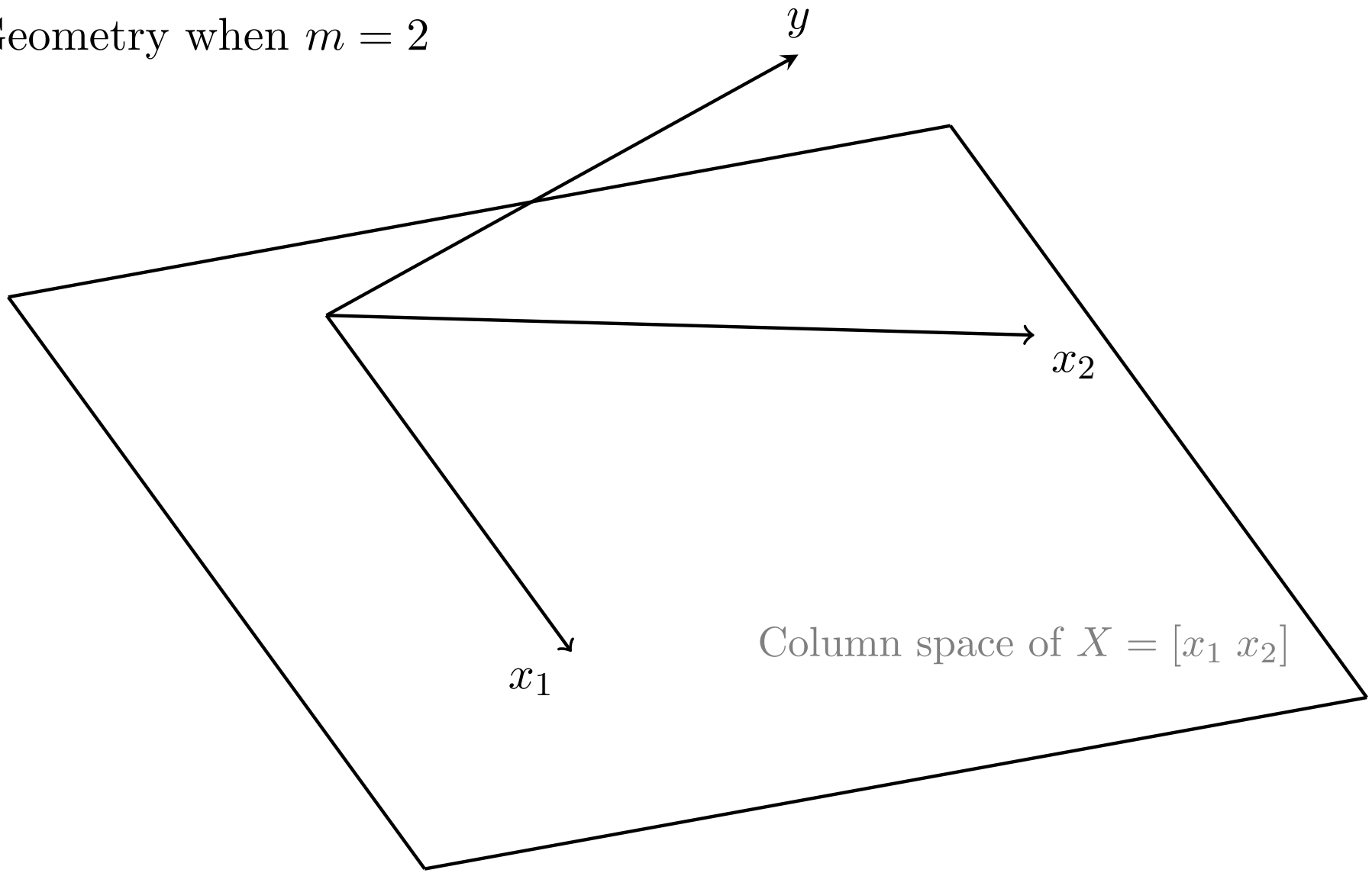
for some scalar $\lambda \geq 0$. The $L_1$ penalty improves upon OLS by introducing parsimony (feature selection) and regularization (improved generality).

There are multiple ways to solve this problem:

1. Directly, via convex optimization (can be expensive)
2. Iterative techniques
   - Forward selection ("matching pursuit"), forward stagewise, others.
   - Least Angle Regression (LARS) [Efron et al., 2004]

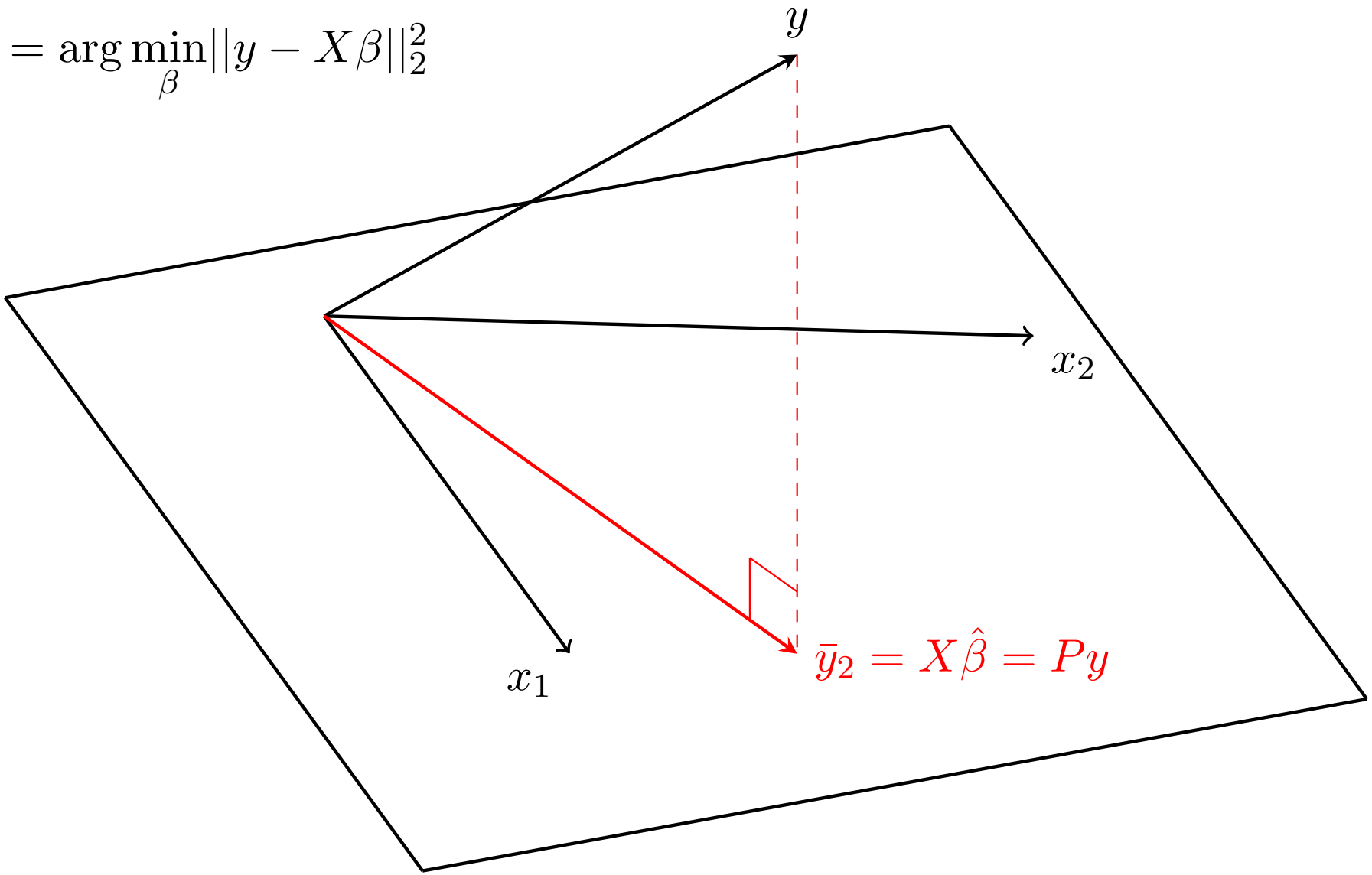
# Visualizing the algorithm

Geometry when $m = 2$



# Ordinary Least Squares (OLS)

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||_2^2$$

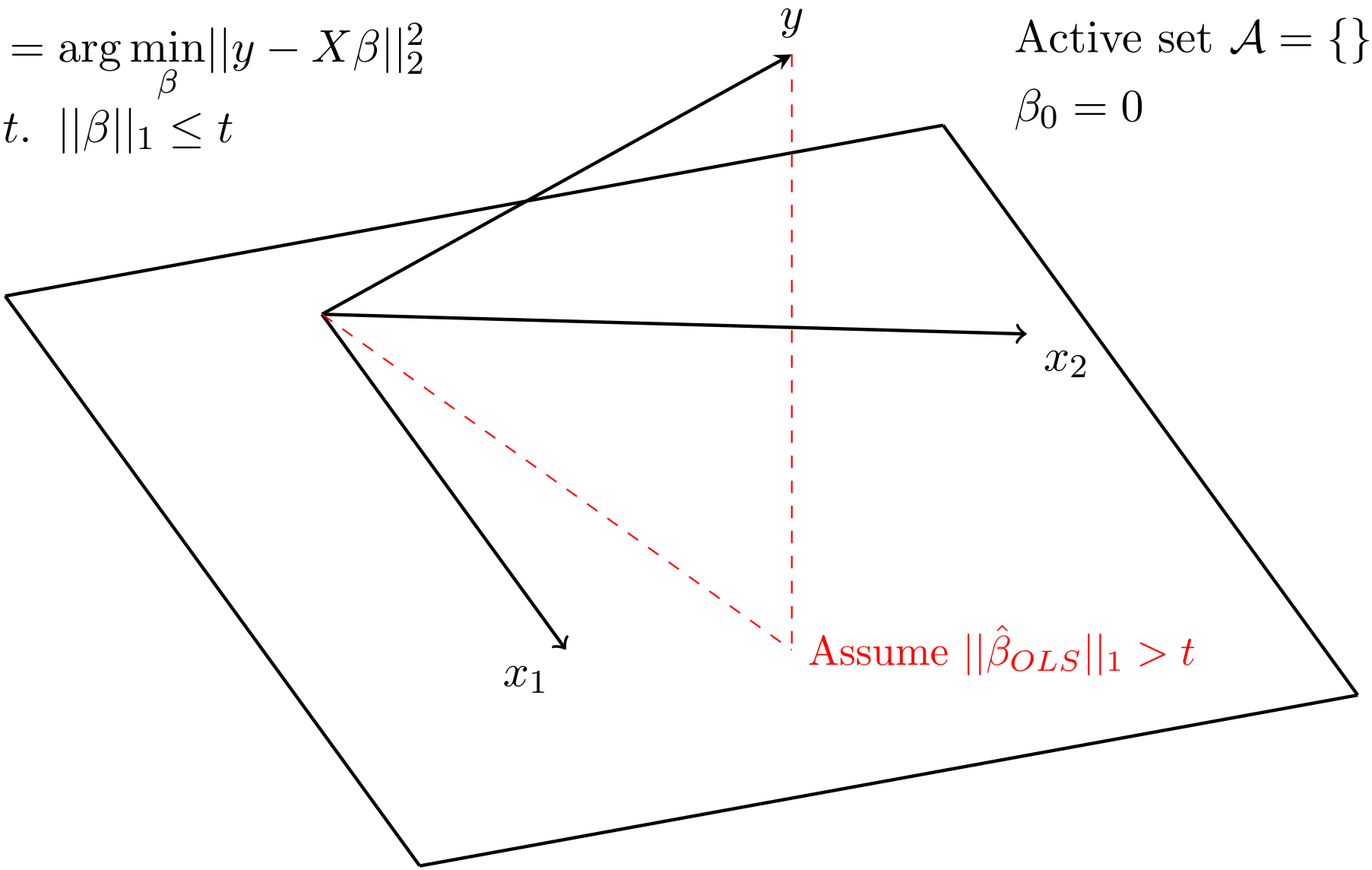


# Least Angle Regression (LARS)

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||_2^2$$
$$s.t. \;\; ||\beta||_1 \leq t$$

Active set $\mathcal{A} = \{\}$

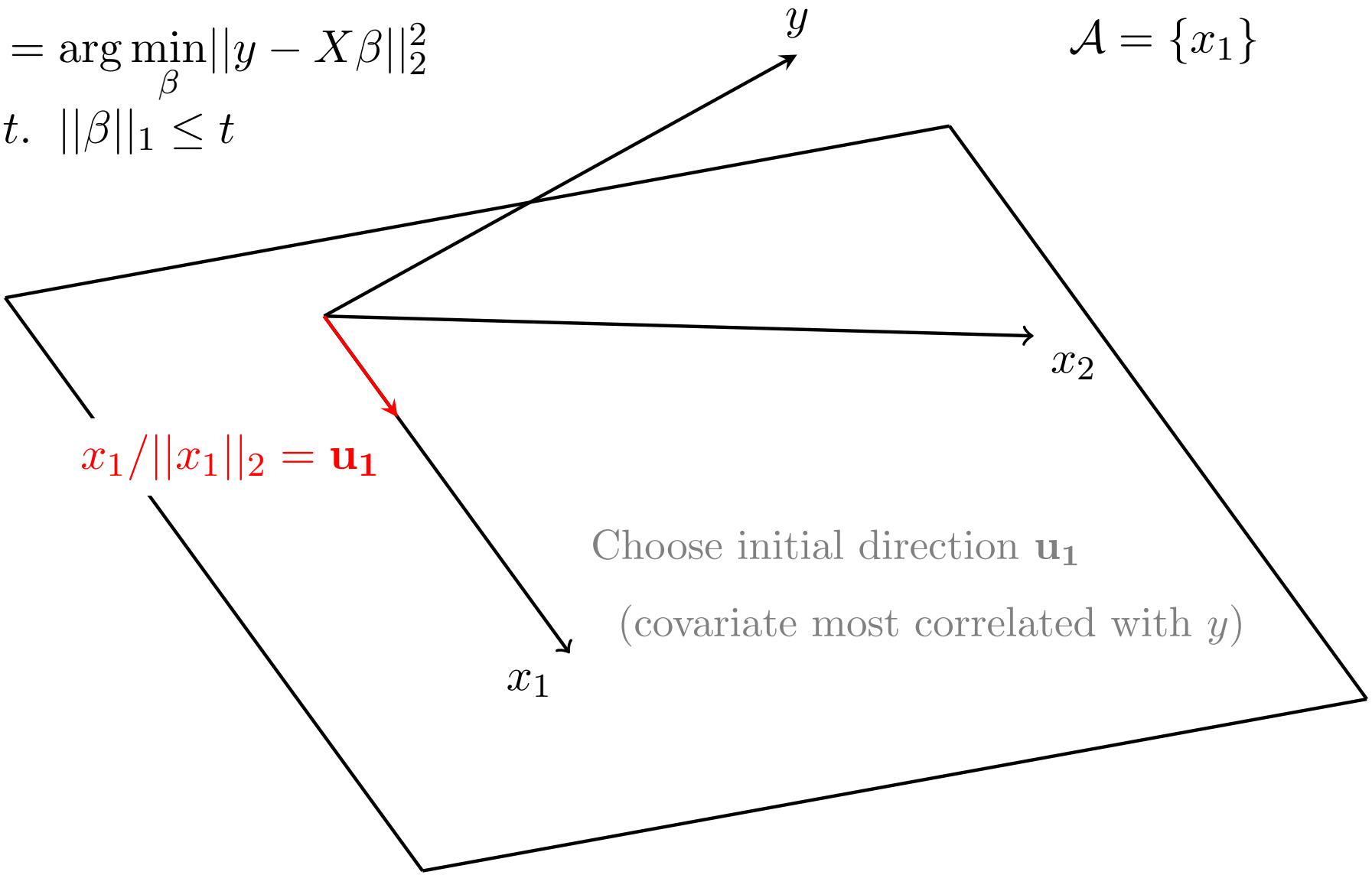
$\beta_0 = 0$



# Least Angle Regression (LARS)

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||_2^2$$
$$s.t. \;\; ||\beta||_1 \leq t$$

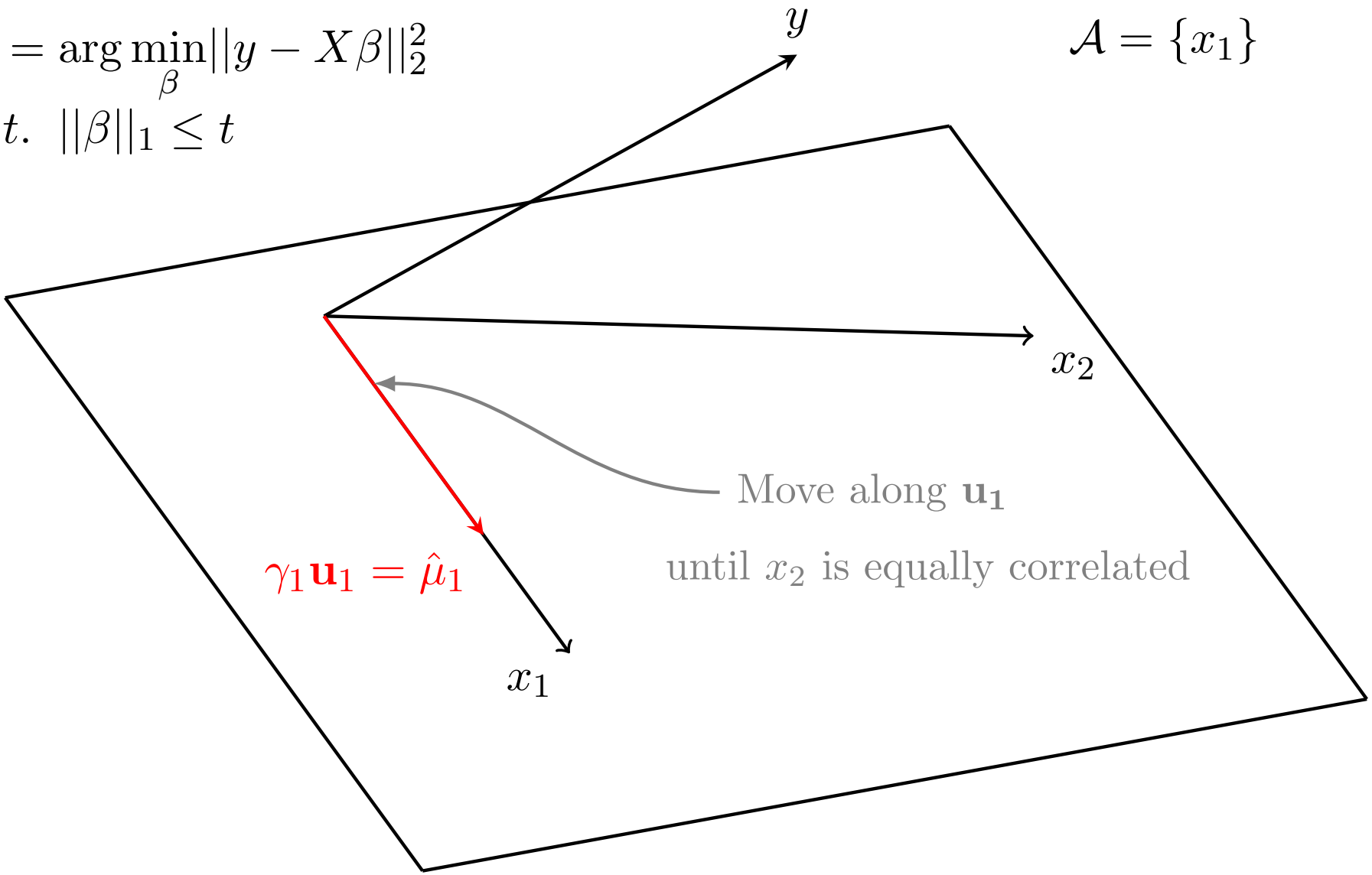
$\mathcal{A} = \{x_1\}$



# Least Angle Regression (LARS)

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||_2^2$$
$$s.t. \;\; ||\beta||_1 \leq t$$

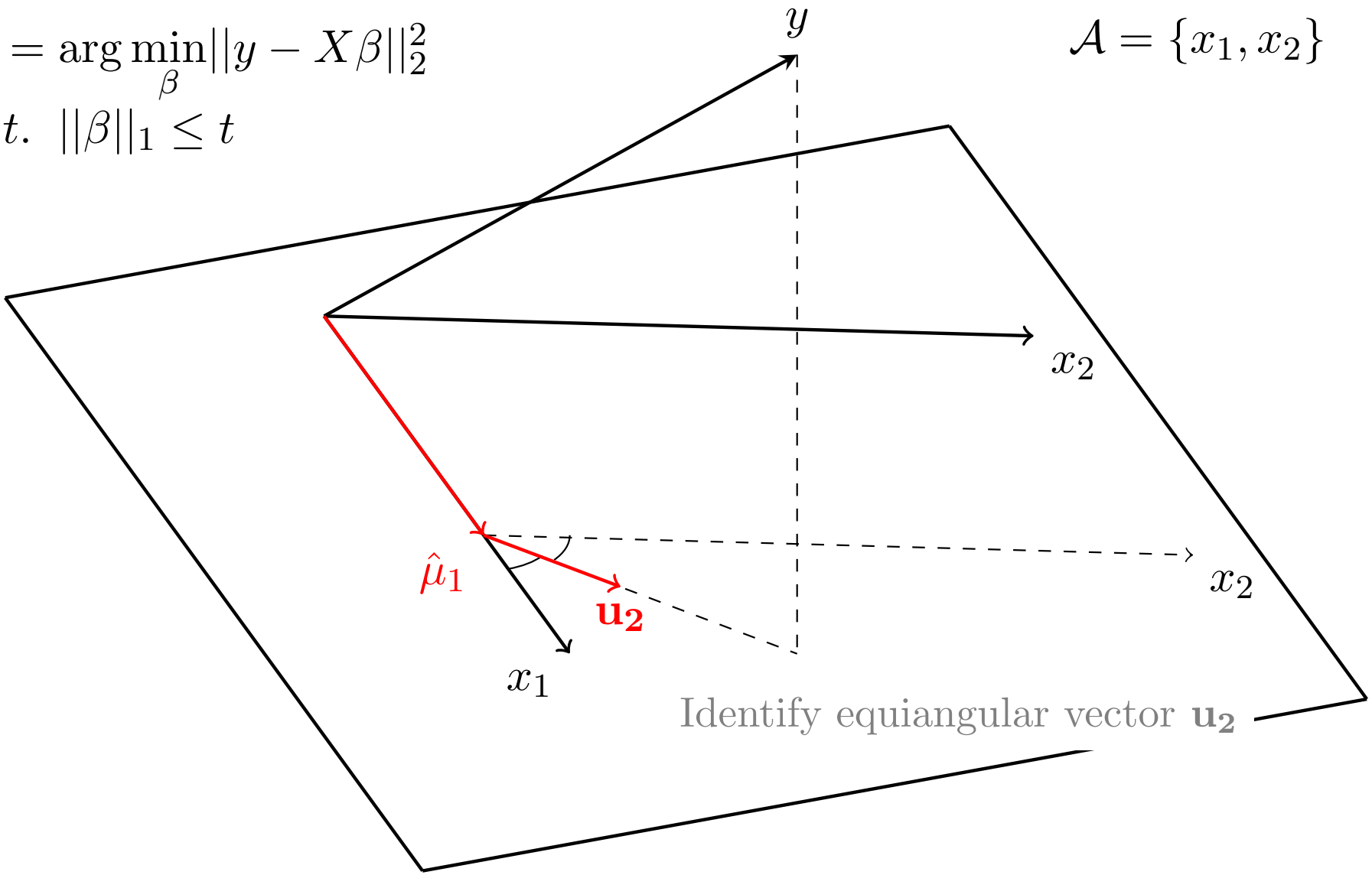
$\mathcal{A} = \{x_1\}$



# Least Angle Regression (LARS)

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||_2^2$$
$$s.t. \;\; ||\beta||_1 \leq t$$

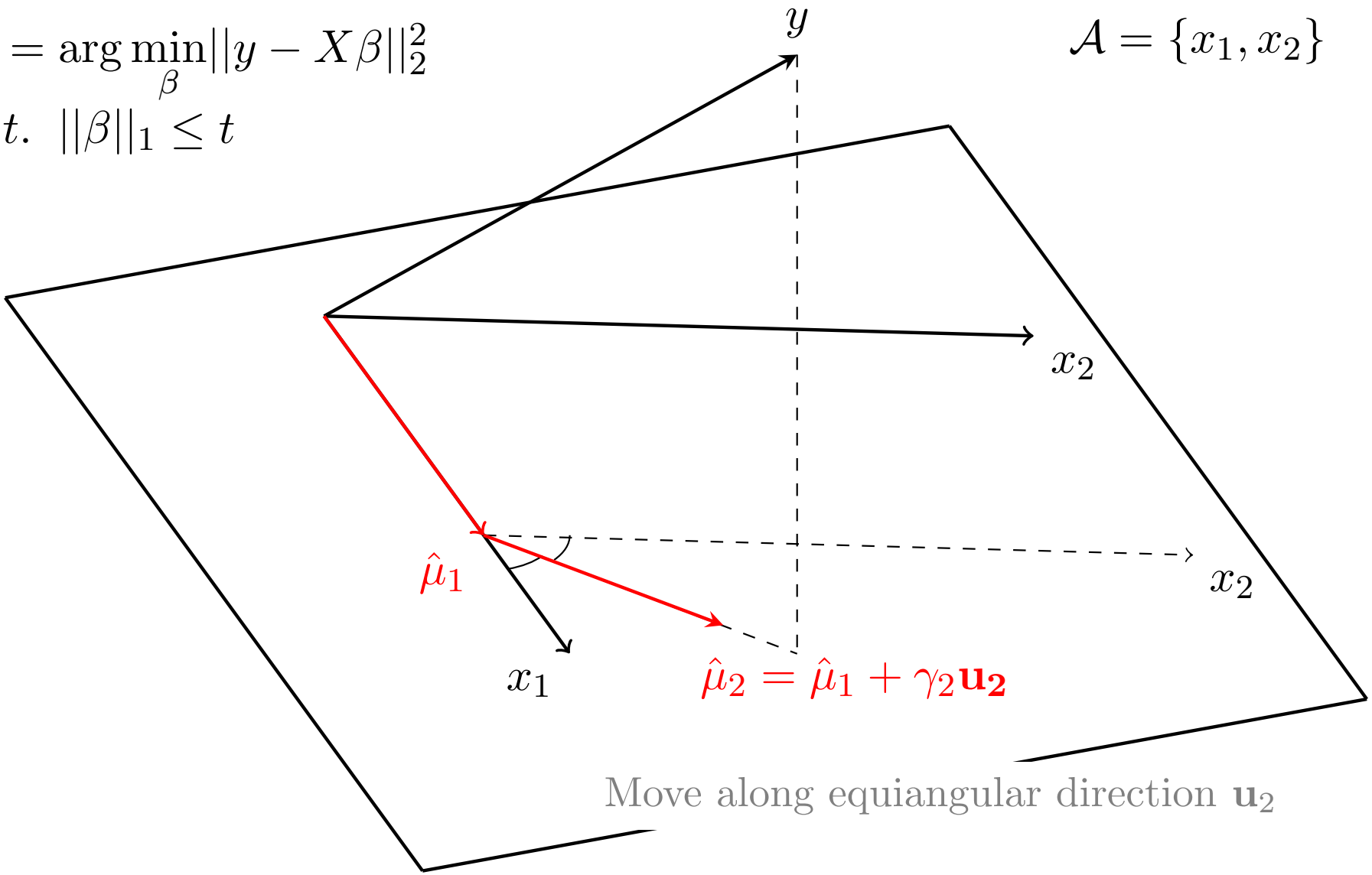
$$\mathcal{A} = \{x_1, x_2\}$$



# Least Angle Regression (LARS)

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||_2^2$$
$$s.t. \;\; ||\beta||_1 \leq t$$

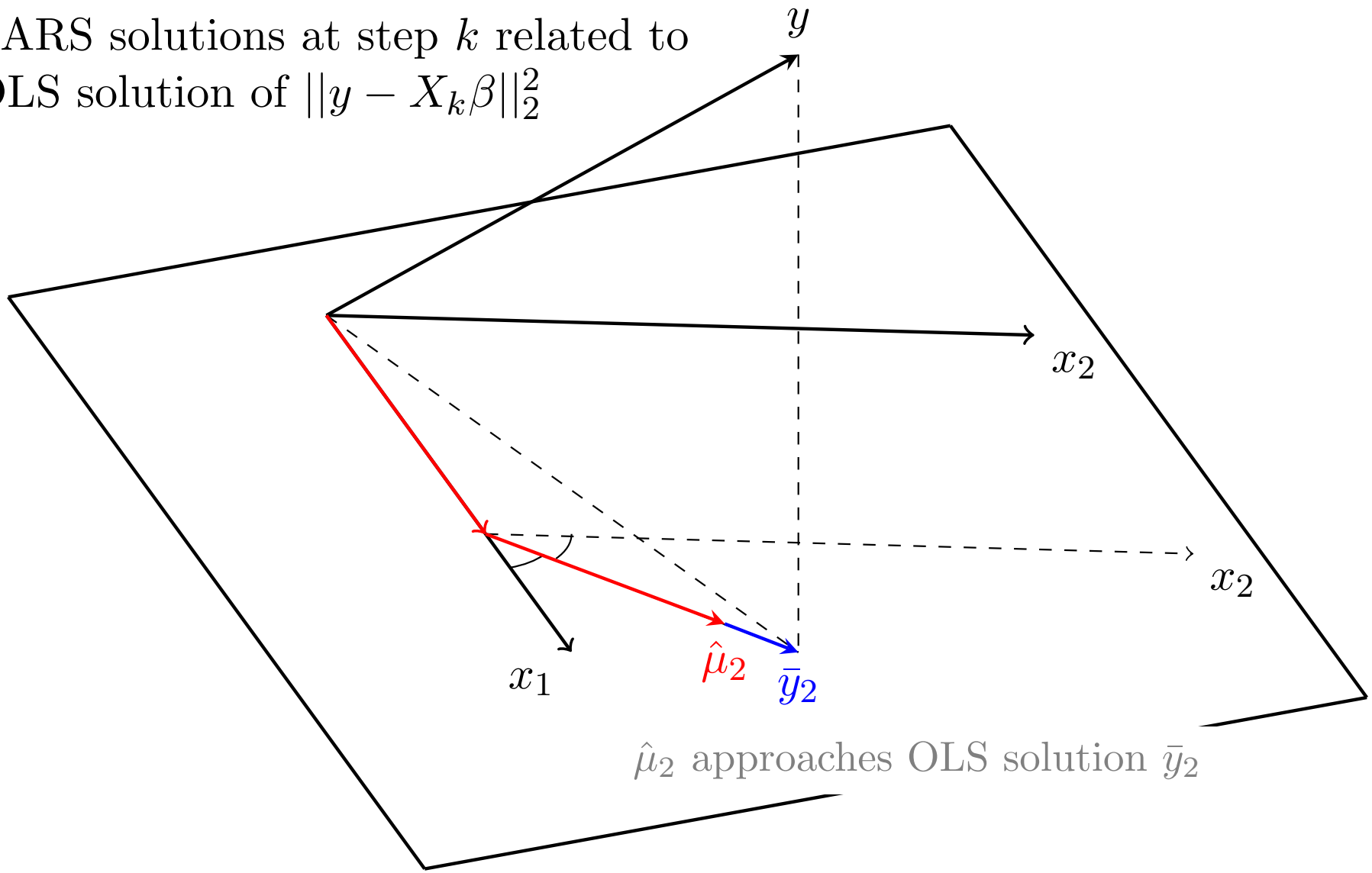
$$\mathcal{A} = \{x_1, x_2\}$$




Move along equiangular direction $\mathbf{u}_2$

# Relationship to OLS

LARS solutions at step $k$ related to OLS solution of $||y - X_k\beta||_2^2$



# Some Algorithm Properties

Full details in [Efron et al., 2004]

- **(2.22)** Successive LARS estimates $\hat{\mu}_k$ always approach but never reach the OLS estimate $\bar{y}_k$ (except maybe on the final iteration).

# Some Algorithm Properties

Full details in [Efron et al., 2004]

- **(2.22)** Successive LARS estimates $\hat{\mu}_k$ always approach but never reach the OLS estimate $\bar{y}_k$ (except maybe on the final iteration).
- **(Theorem 1)** With a small modification to the LARS step size calculation, and assuming covariates are added/removed one at a time from the active set, the complete LARS solution path yields *all* Lasso solutions.

# Some Algorithm Properties

Full details in [Efron et al., 2004]

- **(2.22)** Successive LARS estimates $\hat{\mu}_k$ always approach but never reach the OLS estimate $\bar{y}_k$ (except maybe on the final iteration).
- **(Theorem 1)** With a small modification to the LARS step size calculation, and assuming covariates are added/removed one at a time from the active set, the complete LARS solution path yields *all* Lasso solutions.
- **(Sec. 3.1)** With a change to the covariate selection rule, LARS can be modified to solve the *Positive Lasso* problem:

$$\min_{\beta} \; ||y - X\beta||_2^2 \qquad s.t. \; ||\beta||_1 \leq t$$

$$0 \leq \beta_j$$

# Some Algorithm Properties

Full details in [Efron et al., 2004]

- **(2.22)** Successive LARS estimates $\hat{\mu}_k$ always approach but never reach the OLS estimate $\bar{y}_k$ (except maybe on the final iteration).
- **(Theorem 1)** With a small modification to the LARS step size calculation, and assuming covariates are added/removed one at a time from the active set, the complete LARS solution path yields *all* Lasso solutions.
- **(Sec. 3.1)** With a change to the covariate selection rule, LARS can be modified to solve the *Positive Lasso* problem:

$$\min_{\beta} \; ||y - X\beta||_2^2 \qquad s.t. \; ||\beta||_1 \leq t$$

$$0 \leq \beta_j$$

- **(Sec. 7)** The cost of LARS is comprable to that of a least squares fit on $m$ variables. The LARS sequence incrementally generates a Cholesky factorization of $X^T X$ in a very specific order.

# LARS & Cholesky Decomposition

At iteration $k$, to determine the equiangular vector $\mathbf{u}_k$, one must invert the $k \times k$ matrix $G_k := X_k^T X_k$

Well, don't really invert. Generate Cholesky decomposition $G_k = R^T R$ and solve triangular linear systems.

(Recall: $G_k$ symm. pos. semi-definite and s.p.d. if $X_k$ is full rank):

$$\forall z \in \mathbb{R}^k, z \neq 0, \quad z^T G_k z = z^T X_k^T X_k z = (X_k z)^T (X_k z) = ||X_k z||_2^2 \geq 0$$

Could call `chol`$(G_k)$ each iteration, but there's a more efficient way...

# QR Rank 1 Updates

(Golub and Van Loan [1996, sec. 12.5.2])

Recall: $A = QR$, then $A^T A = (R^T Q^T)(QR) = R^T R$

Suppose we have $Q, R, \mathbf{z}$ and want the QR decomposition of:

$$\tilde{A} = [a_1, \ldots, a_k, \mathbf{z}, a_{k+1} \ldots, a_n]$$

Let $\mathbf{w} := Q^T \mathbf{z}$. Then,

$$Q^T \tilde{A} = \left[Q^T a_1, \ldots, Q^T a_k, \mathbf{w}, Q^T a_{k+1}, \ldots Q^T a_n\right]$$

e.g.

$$\begin{bmatrix} \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & 0 & \times \\ 0 & 0 & 0 & \times & 0 & 0 \\ 0 & 0 & 0 & \times & 0 & 0 \end{bmatrix}$$

# QR Rank 1 Updates

(Golub and Van Loan [1996, sec. 12.5.2])

Use Given's rotations to remove the "spike" introduced by $\mathbf{w}$:

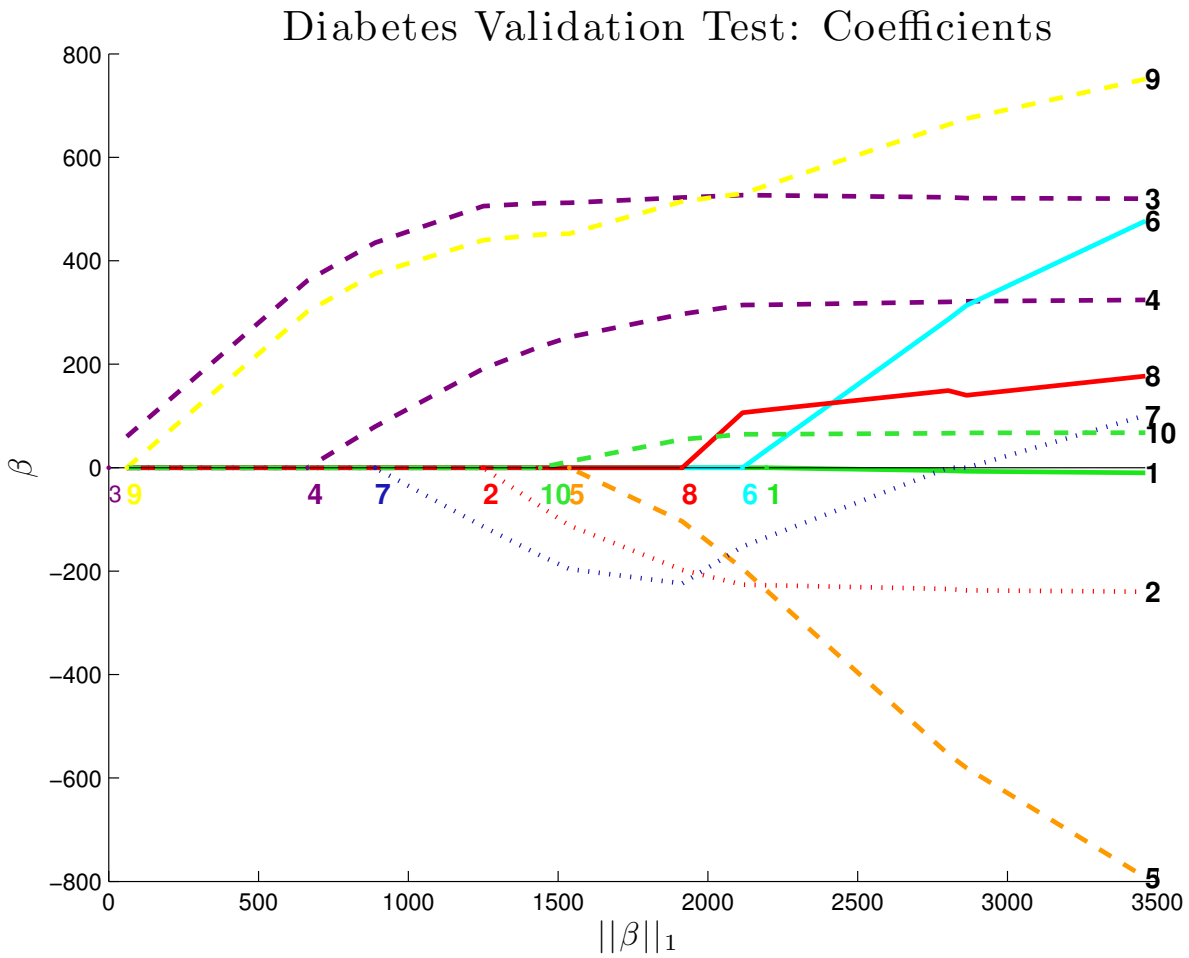
$$\begin{bmatrix} \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & 0 & \times \\ 0 & 0 & 0 & \times & 0 & 0 \\ 0 & 0 & 0 & \times & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- Operation requires $mn$ flops.
- Analogous approach to downdate $R$ when removing a column.
- Matlab functions `qrinsert()`, `qrdelete()` (Octave has `cholupdate()`...).
- Ran into trouble with non-uniqueness of QR decomp; used:

$$QR = QIR = QI_k^T I_k R$$

to swap sign of $R_{k,k}$ (where $I_k$ is ident. mat. w/ $I_{k,k} = -1$)

# Validation: Diabetes Data Set




$m = 10, n = 442$; Compares well with Figure 1 in [Efron et al., 2004]
Also validated by comparing orthogonal designs with theoretical result.

# Today


1. Overview and Status
   - Overview
   - Schedule
2. Least Angle Regression
   - Geometric View
   - Rank 1 Updating
   - Validation
3. Dictionary Learning
   - Preliminary Results
4. Summary & Next Steps

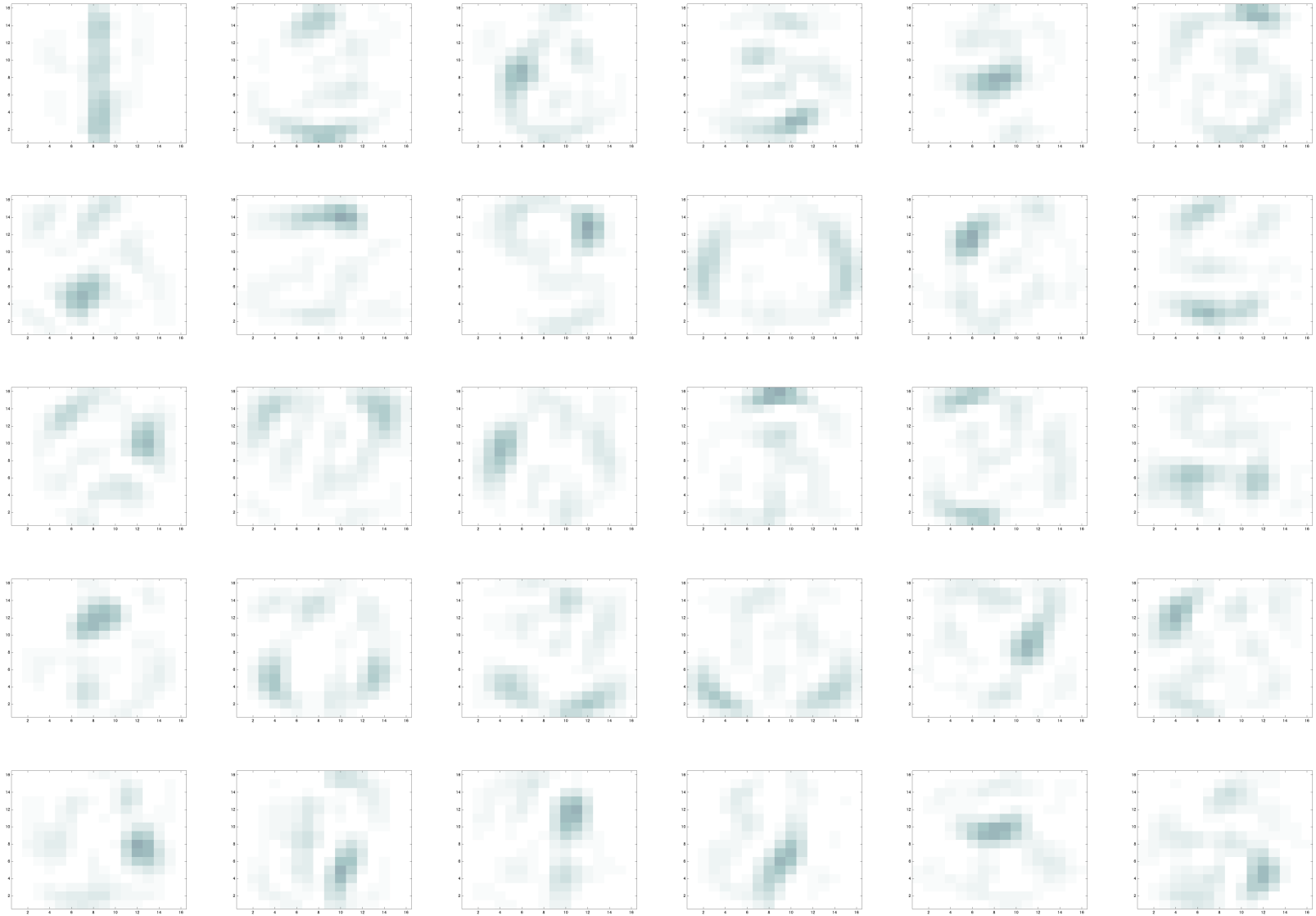
# Dictionary Learning: Progress

Warning:

- The following is preliminary - work in progress.
- No intelligent parameter selection, only looking at dictionary learning at the moment.

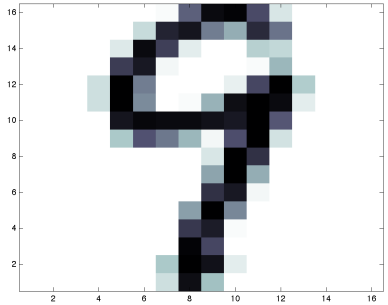
# Atoms: LARS+LASSO

time=6230.27 (sec), m=30, nIters=1000, $\lambda$=0.01

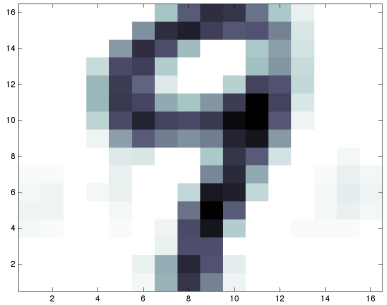


# Experiment: LARS+LASSO, USPS 5023

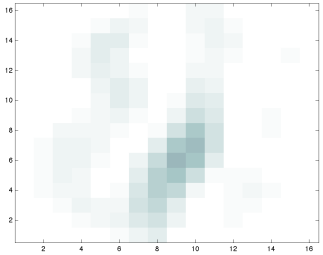
Num. atoms > 1e-4: 30

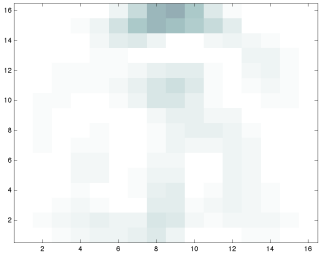


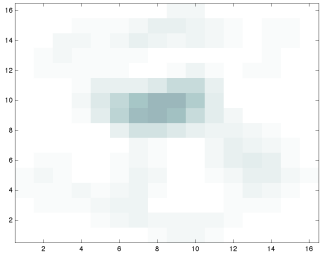
true digit

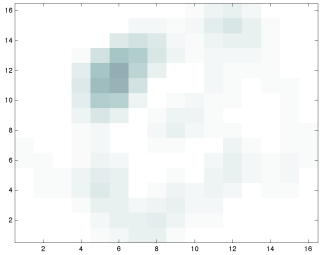


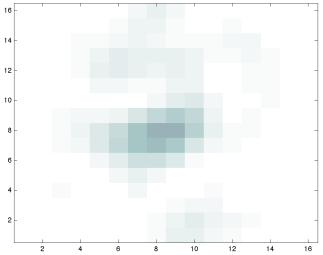
reconstruction







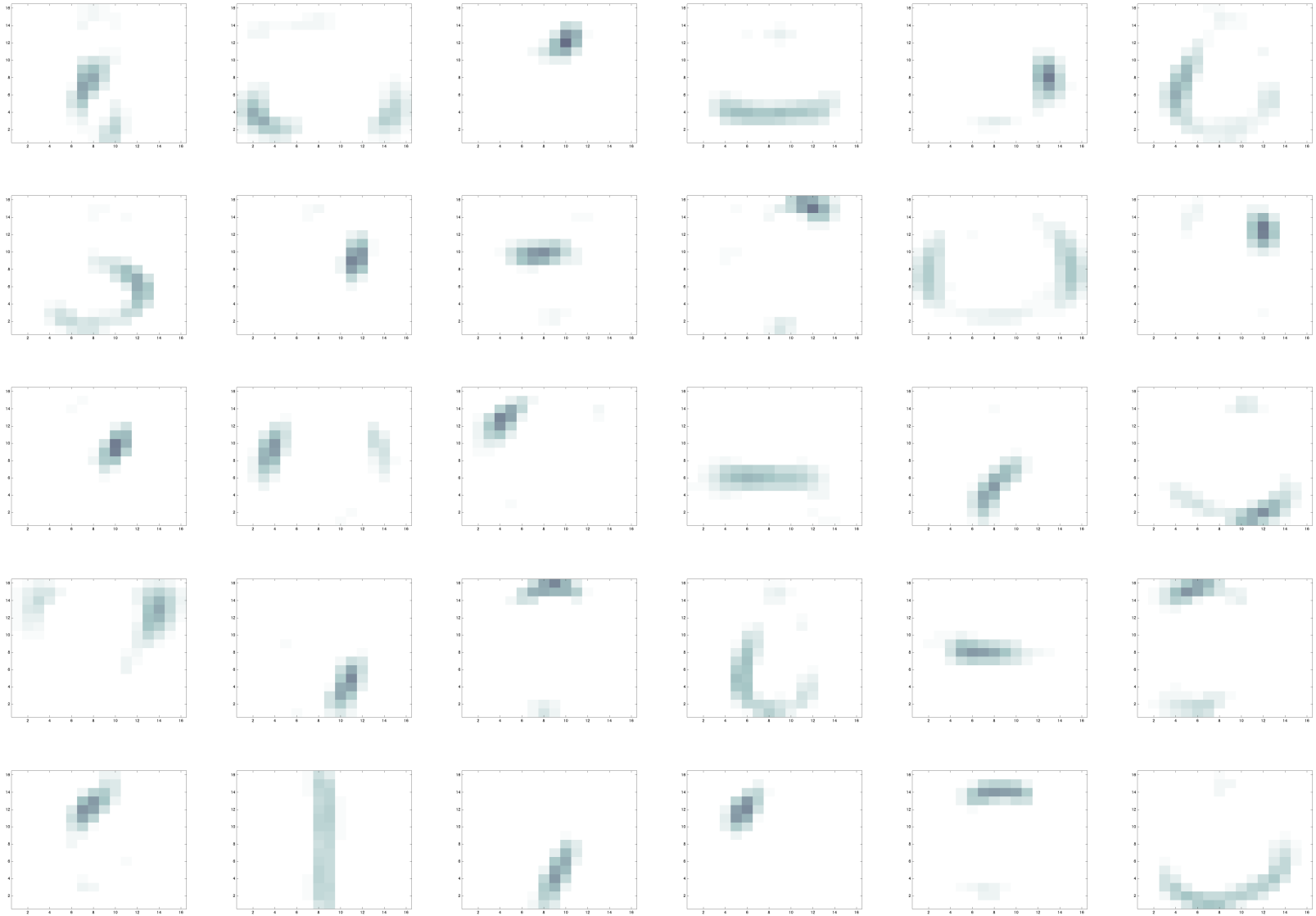




| | | | | |
|---|---|---|---|---|
| $a_{28} = 2.632$ | $a_{16} = 1.662$ | $a_{29} = 1.571$ | $a_{11} = 1.438$ | $a_5 = -1.276$ |
| (0.12 %) | (0.07 %) | (0.07 %) | (0.06 %) | (0.06 %) |

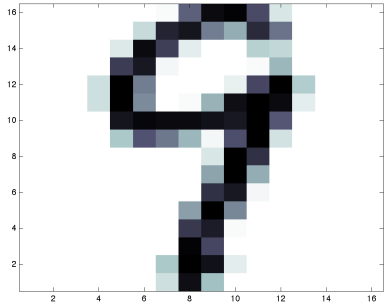
# Atoms: LARS+LASSO-NN

time=2435.80 (sec), m=30, nIters=1000, $\lambda$=0.01

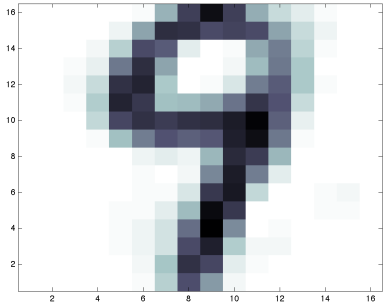


# Experiment: LARS+LASSO-NN, USPS 5023

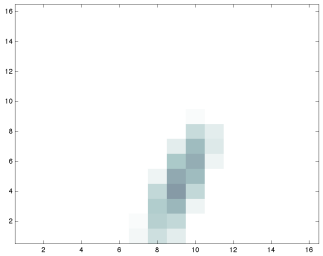
Num. atoms > 1e-4: 13

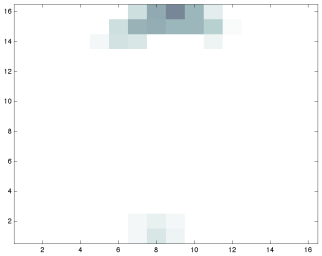


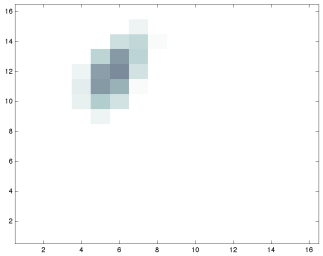
true digit

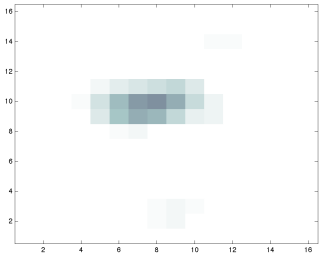


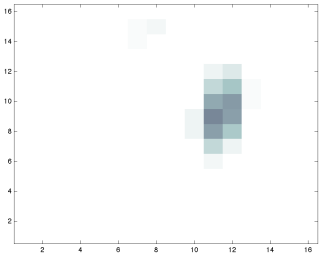
reconstruction











| | | | | |
|---|---|---|---|---|
| $a_{27} = 2.520$ | $a_{21} = 1.994$ | $a_{28} = 1.882$ | $a_9 = 1.822$ | $a_8 = 1.270$ |
| (0.20 %) | (0.16 %) | (0.15 %) | (0.14 %) | (0.10 %) |

# Today


1. Overview and Status
   - Overview
   - Schedule
2. Least Angle Regression
   - Geometric View
   - Rank 1 Updating
   - Validation
3. Dictionary Learning
   - Preliminary Results
4. Summary & Next Steps

# Summary

Progress

- Milestones met; currently on schedule.
- Also completed a few extra tasks not in the original plan (non-negative LARS, incremental Cholesky).

Near Term

- Finish Task Driven Learning Framework and validation
- On to hyperspectral data!

Optional Steps

- Parallel SGD (e.g. [Zinkevich et al., 2010])

# Bibliography I

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

Gene Golub and Charles Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

Julien Mairal, Francis Bach, and Jean Ponce. Task-Driven Dictionary Learning. Rapport de recherche RR-7400, INRIA, 2010.

Martin Zinkevich, Markus Weimer, Alex Smola, and Lihong Li. Parallelized stochastic gradient descent. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2595–2603, 2010.