Pattern Information Processing: Constrained Least-Squares

Masashi Sugiyama
(Department of Computer Science)

Contact: W8E-505
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi/

Over-fitting

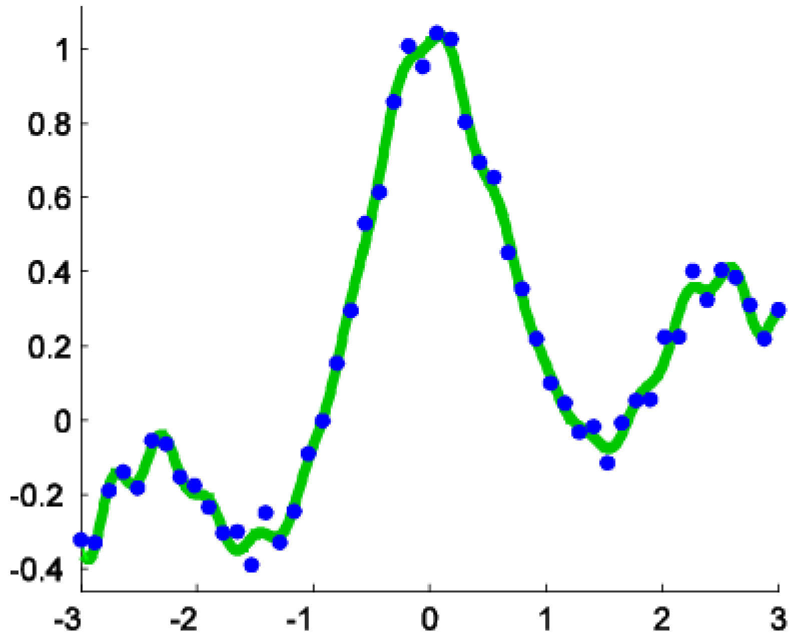
- LS was proved to be a good learning method:
  - Unbiased and BLUE in realizable cases.
  - Asymptotically unbiased and asymptotically efficient in unrealizable cases.
- However, a learned function can over-fit to noisy examples (e.g., when the noise level is high).

Over-fitting

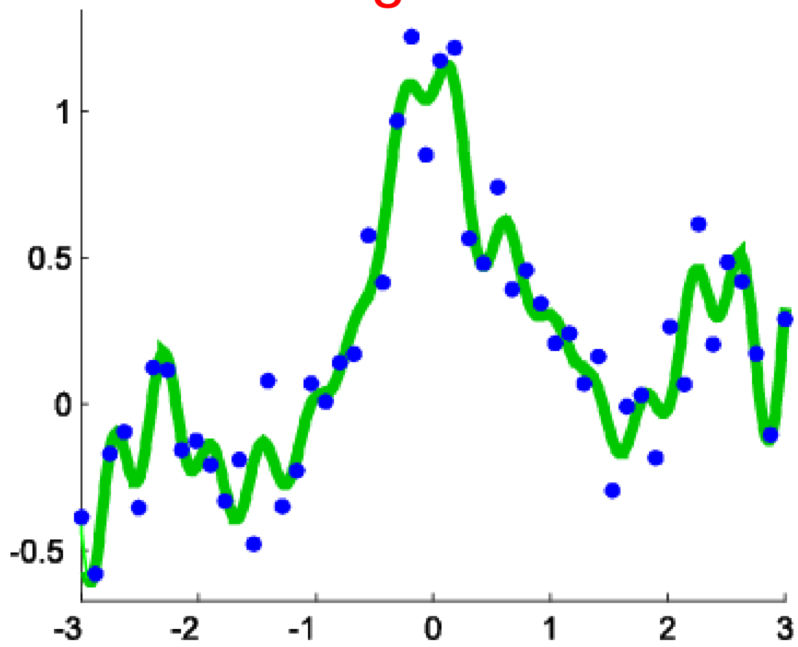
- Trigonometric polynomial model: $f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i \varphi_i(\boldsymbol{x})$

$$\varphi_i(x) = \{1, \sin x, \cos x, \ldots, \sin 15x, \cos 15x\}$$

Small noise



Large noise



- In order to prevent over-fitting, model (search space) should be restricted appropriately.

Today's Plan

- Two approaches to restricting models:
  - Subspace LS
  - Quadratically constrained LS
- Sparseness and model choice.

- We focus on linear/kernel models.

$$f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i \varphi_i(\boldsymbol{x})$$

Subspace LS

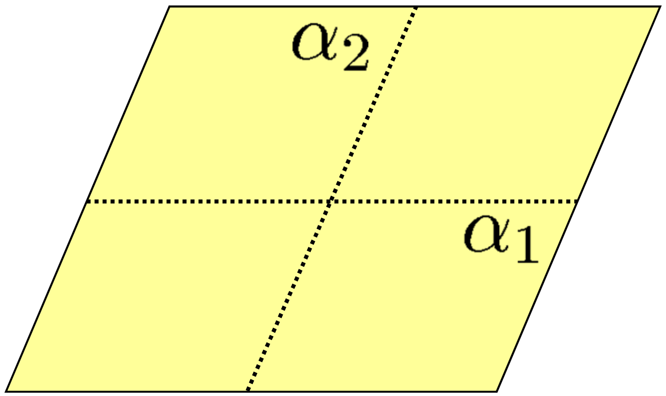
- Restrict the search space within a subspace

$$\hat{\boldsymbol{\alpha}}_{SLS} = \mathop{\mathrm{argmin}}_{\boldsymbol{\alpha}\in\mathbb{R}^b} J_{LS}(\boldsymbol{\alpha})$$
$$\text{subject to } \boldsymbol{P}\boldsymbol{\alpha} = \boldsymbol{\alpha}$$

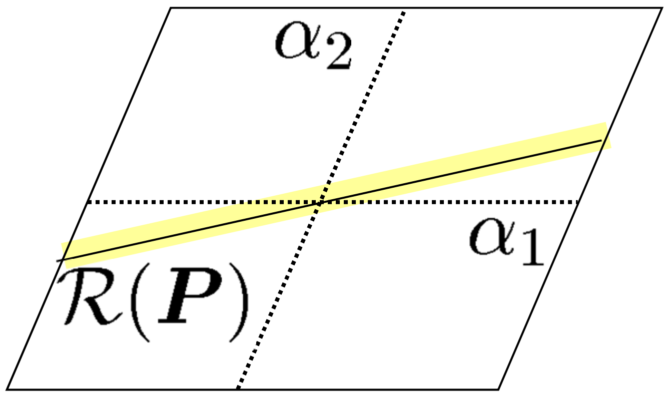
$\boldsymbol{P}$: orthogonal projection onto a subspace

$$J_{LS}(\boldsymbol{\alpha}) = \|\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{y}\|^2$$

$$\boldsymbol{P}^2 = \boldsymbol{P} \qquad \boldsymbol{P}^\top = \boldsymbol{P}$$



Ordinary LS



Subspace LS

How to Obtain A Solution

- Since

$$J_{LS}(\boldsymbol{\alpha}) = \|\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{y}\|^2$$

just replacing $\boldsymbol{X}$ with $\boldsymbol{X}\boldsymbol{P}$ gives a solution:

$$\begin{aligned}\boldsymbol{L}_{SLS} &= (\boldsymbol{P}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{P})^\dagger \boldsymbol{P}\boldsymbol{X}^\top \\ &= (\boldsymbol{X}\boldsymbol{P})^\dagger\end{aligned}$$

$$\boldsymbol{X}_{i,j} = \varphi_j(\boldsymbol{x}_i)$$

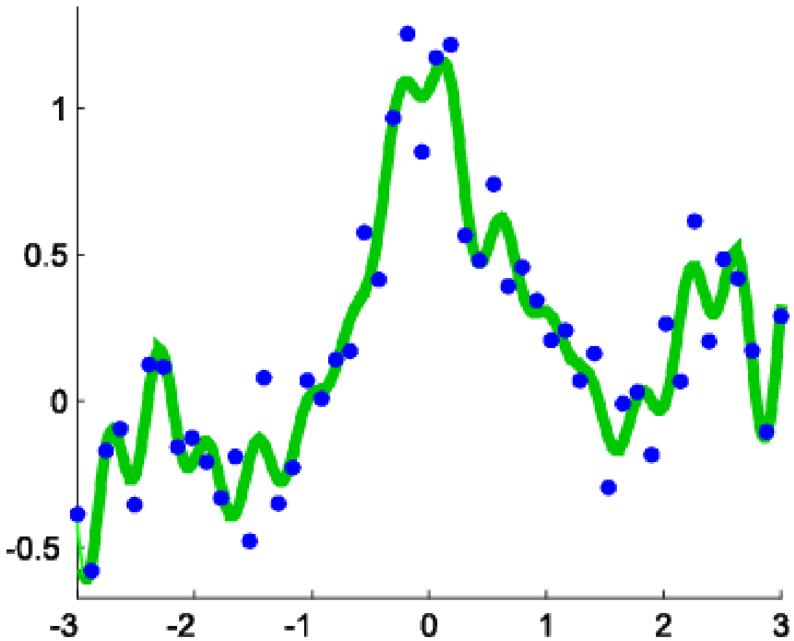
- † :Moore-Penrose generalized inverse

$$\boldsymbol{B} = \boldsymbol{A}^\dagger \iff \begin{cases} \boldsymbol{A}\boldsymbol{B}\boldsymbol{A} = \boldsymbol{A} \\ \boldsymbol{B}\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B} \\ (\boldsymbol{A}\boldsymbol{B})^\top = \boldsymbol{A}\boldsymbol{B} \\ (\boldsymbol{B}\boldsymbol{A})^\top = \boldsymbol{B}\boldsymbol{A} \end{cases}$$

$$\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}^\dagger = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/3 \end{pmatrix}$$

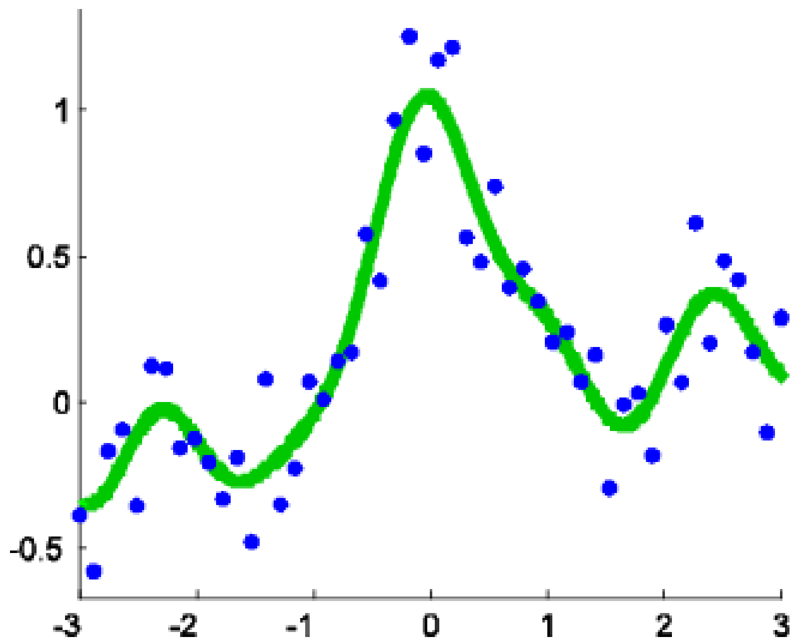
$$\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}^\dagger = \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \end{pmatrix}$$

Example of SLS



Full LS

$1, \ldots, \sin 15x, \cos 15x$



Subspace LS

$1, \ldots, \sin 5x, \cos 5x$

- Over-fit can be avoided by properly choosing a subspace.

Principal Component Regression

- Choose the maximum-variance subspace:

$$\boldsymbol{P} = \sum_{k=1}^{m} \boldsymbol{\phi}_k \boldsymbol{\phi}_k^\top$$

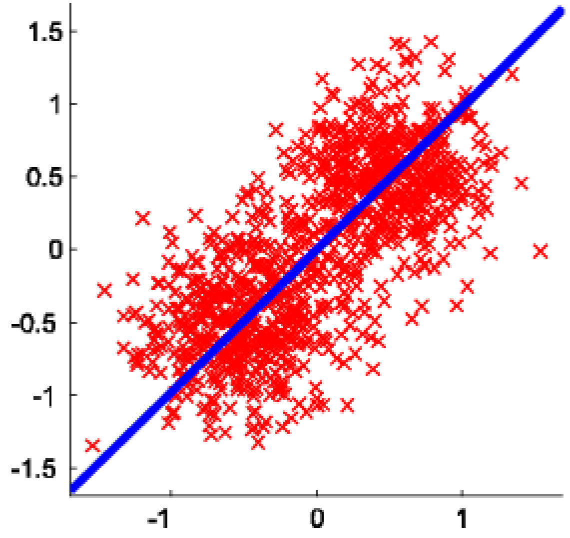
- Eigendecomposition of covariance matrix:

$$\boldsymbol{X}^\top \boldsymbol{H} \boldsymbol{H} \boldsymbol{X} \boldsymbol{\phi} = \lambda \boldsymbol{\phi} \qquad \boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n} \boldsymbol{1}_{n \times n}$$

- Eigenvalues: $\lambda_1 \geq \cdots \geq \lambda_b$
- Eigenvectors: $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_b$

$\boldsymbol{I}_n$: n-dimensional identity matrix

$\boldsymbol{1}_{n \times n}$: $n \times n$ matrix with all ones

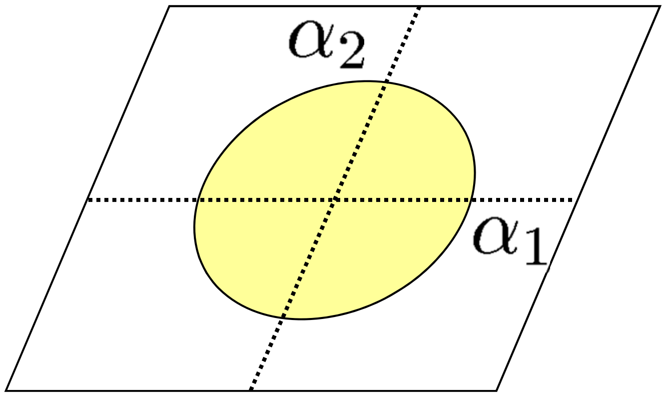


Quadratically Constrained LS

- Restrict the search space within a hyper-sphere.

$$\hat{\boldsymbol{\alpha}}_{QCLS} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^b}{\operatorname{argmin}} \, J_{LS}(\boldsymbol{\alpha})$$
$$\text{subject to } \|\boldsymbol{\alpha}\|^2 \leq C$$

$C \geq 0$



How to Obtain A Solution

- Lagrangian:

$$L(\boldsymbol{\alpha}, \lambda) = J_{LS}(\boldsymbol{\alpha}) + \lambda(\|\boldsymbol{\alpha}\|^2 - C)$$

  - λ : Lagrange multiplier

- Karush-Kuhn-Tucker (KKT) condition: for some $\lambda^* \geq 0$, the solution $\hat{\boldsymbol{\alpha}}_{QCLS}$ satisfies
  - $\dfrac{\partial L(\hat{\boldsymbol{\alpha}}_{QCLS}, \lambda^*)}{\partial \boldsymbol{\alpha}} = \mathbf{0}$
  - $\|\hat{\boldsymbol{\alpha}}_{QCLS}\|^2 - C \leq 0$
  - $\lambda^* \left(\|\hat{\boldsymbol{\alpha}}_{QCLS}\|^2 - C\right) = 0$

How to Obtain A Solution (cont.)

- $$\frac{\partial L(\hat{\boldsymbol{\alpha}}_{QCLS}, \lambda^*)}{\partial \boldsymbol{\alpha}} = \mathbf{0}$$

$$\Rightarrow \quad \hat{\boldsymbol{\alpha}}_{QCLS} = \boldsymbol{L}_{QCLS}\boldsymbol{y}$$

$$\boldsymbol{L}_{QCLS} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda^* \boldsymbol{I})^{-1} \boldsymbol{X}^\top$$

- We still need to determine λ^*, but this is not straightforward.
- In practice, we start from setting $\lambda\ (\geq 0)$ and solve

$$\hat{\boldsymbol{\alpha}}_{QCLS} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^b}{\mathrm{argmin}}\, J_{QCLS}(\boldsymbol{\alpha})$$

$$J_{QCLS}(\boldsymbol{\alpha}) = J_{LS}(\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|^2$$

Interpretation of QCLS

- QCLS tries to avoid overfitting by adding a penalty (regularizer) to the "goodness-of-fit" term.

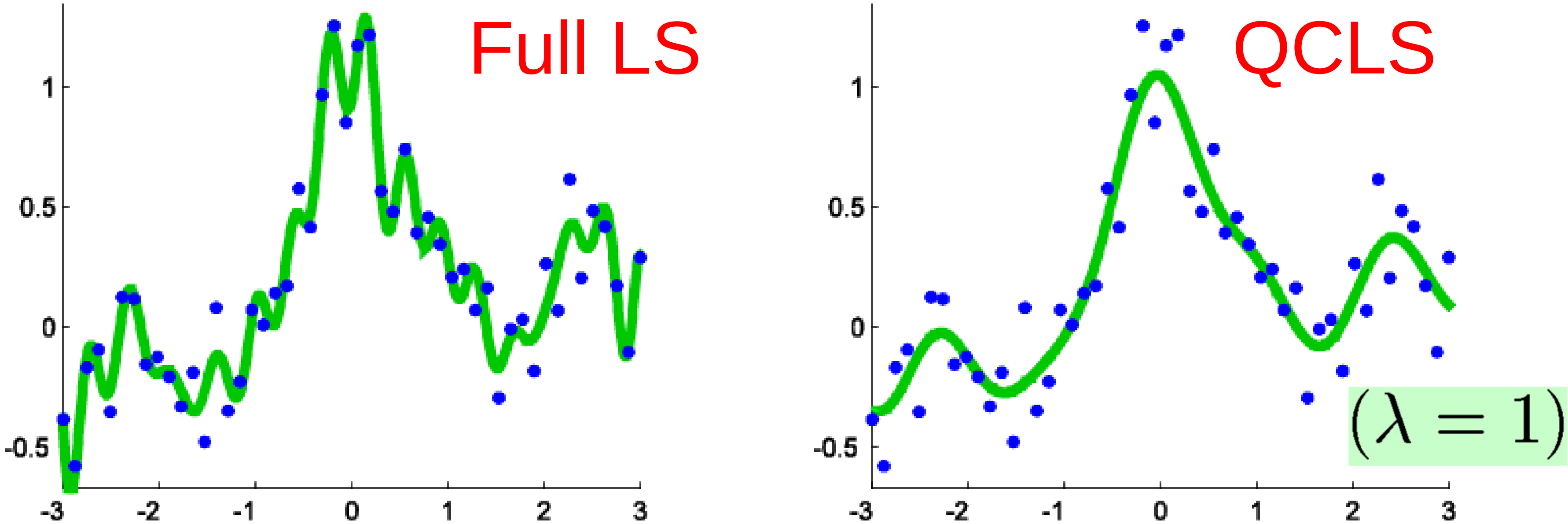
$$J_{QCLS}(\boldsymbol{\alpha}) = \underbrace{J_{LS}(\boldsymbol{\alpha})}_{\text{Goodness of fit}} + \underbrace{\lambda\|\boldsymbol{\alpha}\|^2}_{\text{Penalty (regularizer)}}$$

- For this reason, QCLS is also called quadratically regularized LS.
- λ is called the regularization parameter.

Example of QCLS

- Gaussian kernel model:

$$f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2}\right)$$



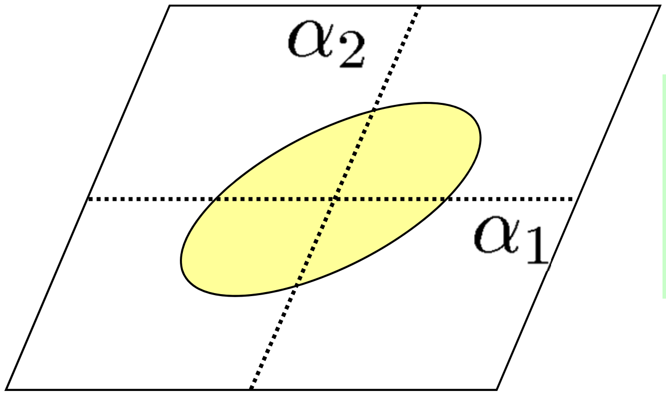
- Over-fit can be avoided by properly choosing the regularization parameter.

Generalization

- Restrict the search space within a hyper-ellipsoid.

$$\hat{\boldsymbol{\alpha}}_{QCLS} = \mathop{\mathrm{argmin}}_{\boldsymbol{\alpha} \in \mathbb{R}^b} J_{LS}(\boldsymbol{\alpha})$$
$$\text{subject to } \langle \boldsymbol{R}\boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle \leq C$$

$C \geq 0$



$\boldsymbol{R}$:Positive semi-definite matrix ("regularization matrix")

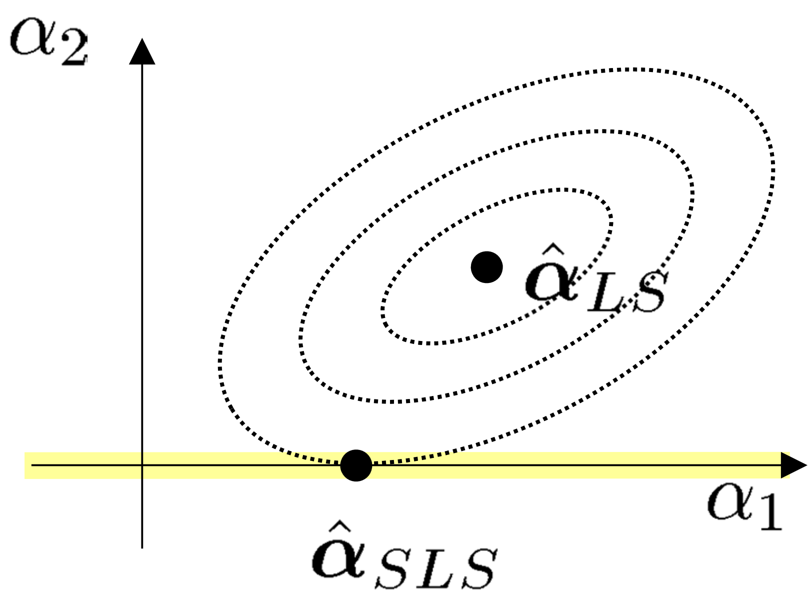
$\forall \boldsymbol{\alpha}, \langle \boldsymbol{R}\boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle \geq 0$

- Solution: (proof is homework!)

$$\boldsymbol{L}_{QCLS} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{R})^{-1} \boldsymbol{X}^\top$$

Sparseness of Solution

- In SLS, if the subspace is spanned by a subset of basis functions $\{\varphi_i(\boldsymbol{x})\}_{i=1}^{b}$, some of the parameters $\{\hat{\alpha}_i\}_{i=1}^{b}$ are exactly zero.



Sparseness and Model Choice

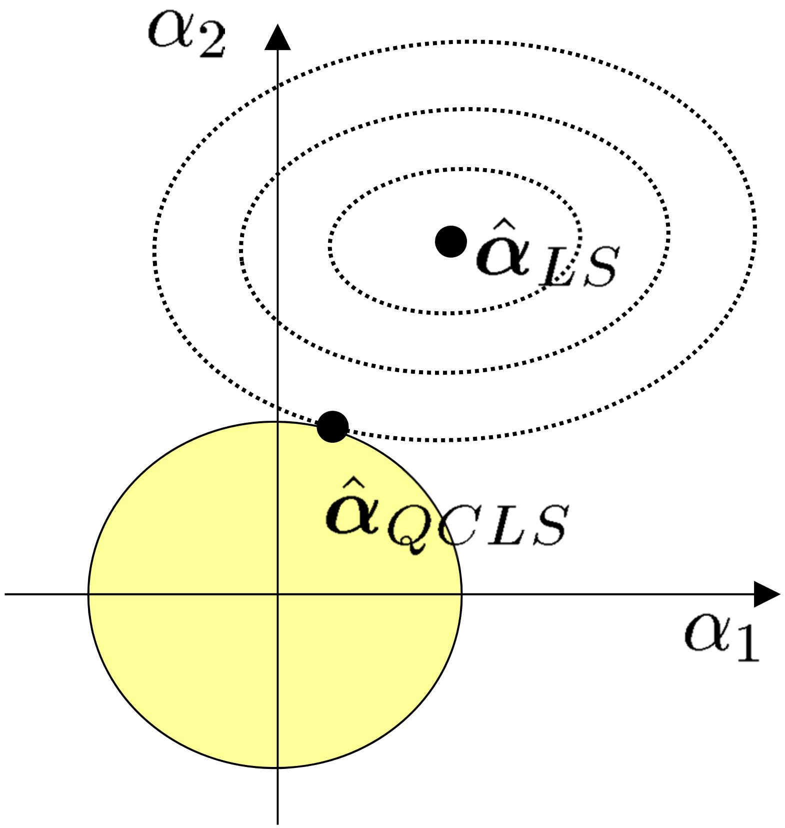
- Having sparsity is computationally attractive:
  - Calculating output values is easier.

$$f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i \varphi_i(\boldsymbol{x})$$

  - Computing the solution can potentially be easier.
- However, the number of possible subspaces is combinatorial, 2^b.
- It is computationally infeasible to find the best subset if b is large.

Sparseness and Model Choice (cont.)

- In QCLS, model choice is continuous: λ
- However, solution is not generally sparse.



Homework

1. Prove that the solution of

$$\hat{\boldsymbol{\alpha}}_{QCLS} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^b}{\mathrm{argmin}}\, J_{LS}(\boldsymbol{\alpha})$$

$$\text{subject to } \langle \boldsymbol{R}\boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle \leq C$$

is given by

$$\hat{\boldsymbol{\alpha}}_{QCLS} = \boldsymbol{L}_{QCLS}\boldsymbol{y}$$

$$\boldsymbol{L}_{QCLS} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{R})^{-1} \boldsymbol{X}^\top$$

Homework (cont.)

2. For your own toy 1-dimensional data, perform simulations using
   - Gaussian kernel models
   - Quadratically-constrained least-squares learning

   and analyze the results, e.g., changing
   - Target functions
   - Number of samples
   - Noise level
   - Width of Gaussian kernel
   - Regularization parameter/matrix