Implicit regularization

Pontus Giselsson

Outline

- **Implicit regularization**
- Early termination

Generalization and implicit regularization

- What is implicit regularization?
  - Assume infinitely many solutions on large manifolds exist
    - overparameterized neural networks
    - least squares with fewer examples than features
  - Algorithm selects solution with small(est) desired norm
- Implicit regularization may give model with better generalization
- SGD is believed to have good implicit regularization
- Adaptive scaling methods might have worse

Example – Gradient method and least squares

- We will consider least squares

$$\underset{x}{\text{minimize}}\ \tfrac{1}{2}\|Ax-b\|_2^2$$

  for which $\bar{x}$ exists such that $A\bar{x}=b$

- We will show that scaled gradient method

$$x_{k+1}=x_k-H^{-1}\nabla f(x_k)$$

  converges, if $x_0=0$, to minimum $\|\cdot\|_H$ solution

- This gives implicit regularization of gradient method
- Compare to explicit regularization that penalizes norm in problem

Least squares problem

- Consider least squares problem of the form

$$\underset{x}{\text{minimize}}\, \frac{1}{2}\|Ax - b\|_2^2$$

  where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $m < n$, and $\exists \bar{x}$ such that $A\bar{x} = b$
- Solution set $X = \{x : Ax = b\} = \{x : \bar{x} + v\}$ where:
  - $\bar{x} \in X$ is such that $A\bar{x} = b$
  - v any vector such that $Av = 0$, i.e. in nullspace of A, $\mathcal{N}(A)$

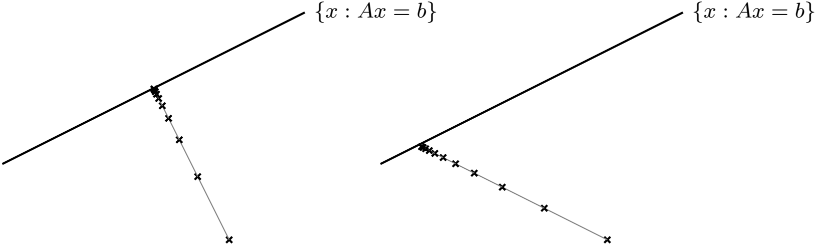
  why? cost satisfies for all x: $\frac{1}{2}\|Ax - b\|_2^2 \geq 0$ and

$$\frac{1}{2}\|A(\bar{x} + v) - b\|_2^2 = \|A\bar{x} + Av - b\|_2^2 = \|b - b\|_2^2 = 0$$

- If $v \in \mathcal{N}(A)$ so is tv ($A(tv) = tAv = 0$) for all $t \in \mathbb{R}$
- Since $m < n$, $\mathcal{N}(A)$ is (at least) $n - m$-dimensional subspace

Graphical interpretation

- What happens with scaled gradient method?
- Solution set X extends infinitely
  - sequence is perpendicular to X in scalar product $(Hx)^T y$
  - algorithm converges to projection point $\operatorname{argmin}_{x \in X}(\|x - x_0\|_H)$



Gradient method with $\|\cdot\|_2$ Gradient method with scaled $\|\cdot\|_H$

Convergence to minimum norm solution

- The scaled gradient method with $\gamma \in (0, \frac{2}{\beta})$

$$x_{k+1} = x_k - \gamma H^{-1} A^T (A x_k - b)$$

  converges to a point $x_k \to \bar{x}$ such that $A\bar{x} = b$
- Letting $\lambda_k = -\sum_{l=0}^{k} \gamma (A x_l - b) \in \mathbb{R}^m$ and unfolding iteration:

$$H x_{k+1} = H x_0 - \sum_{l=0}^{k} \gamma_l A^T (A x_l - b) = H x_0 + A^T \lambda_k$$

- The unique projection point $\hat{x} = \underset{x \in X}{\operatorname{argmin}} (\|x - x_0\|_H)$ if and only if

$$H\hat{x} - H x_0 + A^T \lambda = 0 \qquad \text{and} \qquad A\hat{x} = b$$

- There is only one such $\hat{x}$, so we must have $\lambda_k \to \lambda$ and $\bar{x} = \hat{x}$
- If $x_0 = 0$, the algorithm converges to $\underset{x \in X}{\operatorname{argmin}} (\|x\|_H)$

Comparison to Tikhonov regularization

- Tikhonov adds $\|\cdot\|_2^2$ norm penalty for better generalization
- Standard gradient method converges to minimium $\|\cdot\|_2$ norm
  - Similar to using Tikhonov regularization
- Scaled gradient converges to minimium $\|\cdot\|_H$ norm solution
- If H very skewed that can happen e.g. in
  - Newton, quasi-newton, Adagrad, Adam

  maybe not as good generalization
- Analysis in convex least squares setting
- Some evidence that same thing holds in nonconvex setting
  - That suggests using SGD instead of, e.g., Adam

Outline

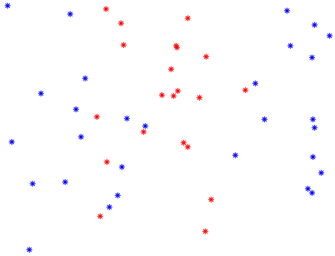
- Implicit regularization
- **Early termination**

Early termination

- Another implicit regularization is to terminate algorithm early
- Sometimes generalization deteriorates with higher accuracy
- Can happen if model too complex for data

Early termination – Example

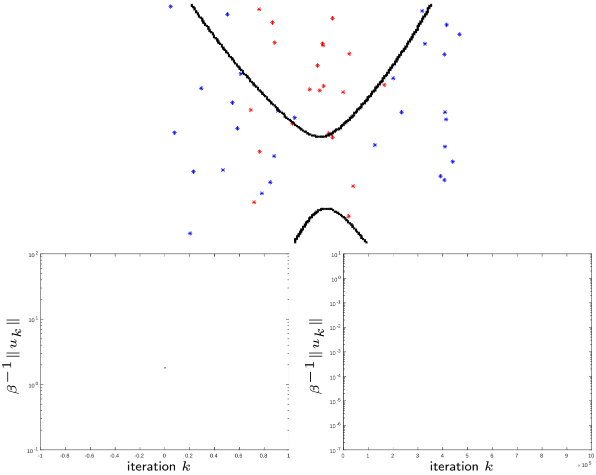
- Will consider SVM with small regularization on this problem data



- Will see:
  - best generalization after only a few iterations at medium accuracy
  - high accuracy takes many iterations but poor generalization

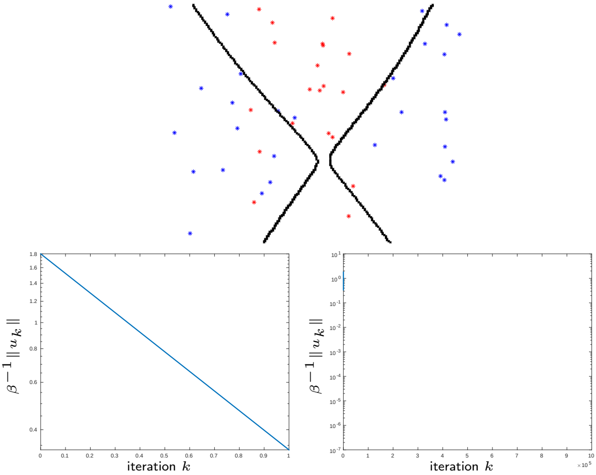
Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 1 Residual norm: $\beta^{-1}\|u_k\|_2 = 6.6e^{-1}$



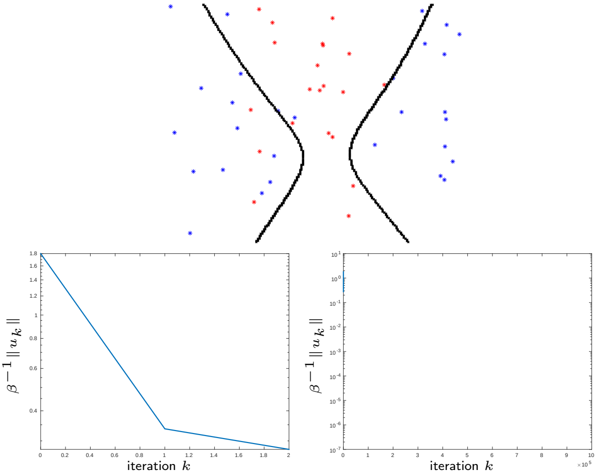
Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 2 Residual norm: $\beta^{-1}\|u_k\|_2 = 4.7e^{-1}$



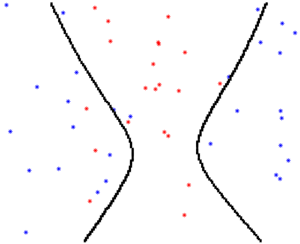
Early termination – Example

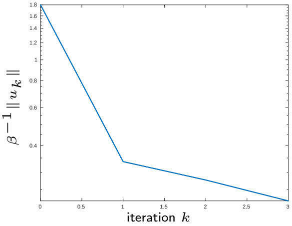
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 3 Residual norm: $\beta^{-1}\|u_k\|_2 = 3.5e^{-1}$

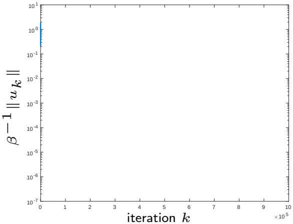


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 4 Residual norm: $\beta^{-1}\|u_k\|_2 = 2.8e^{-1}$

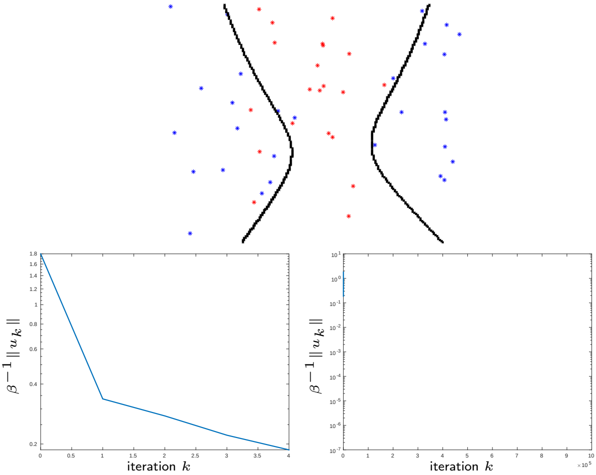






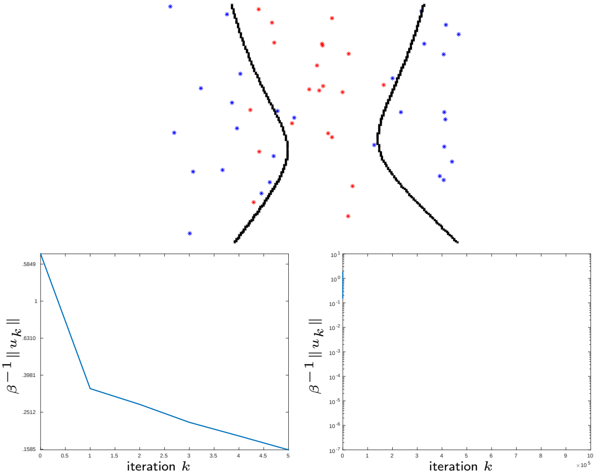
Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 5 Residual norm: $\beta^{-1}\|u_k\|_2 = 2.3e^{-1}$



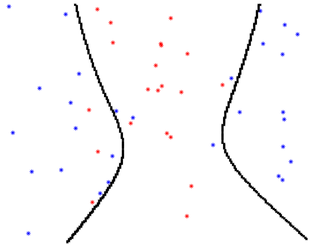
Early termination – Example

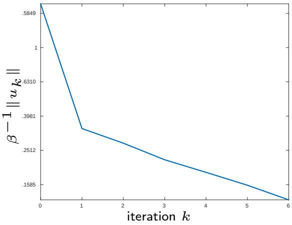
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 6 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.9e^{-1}$

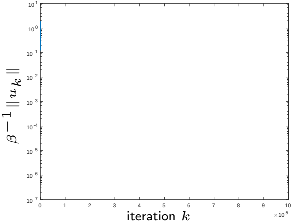


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 7 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.5e^{-1}$

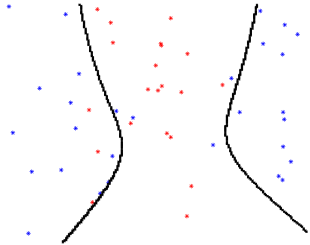


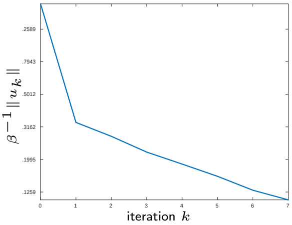


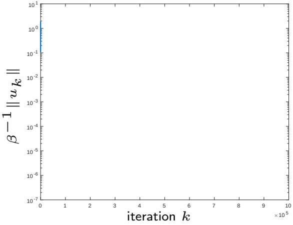


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 8 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.3e^{-1}$

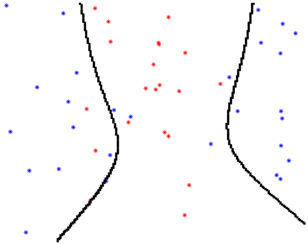


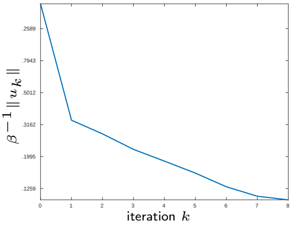


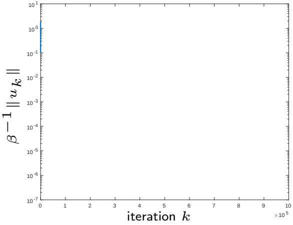


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 9 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.2e^{-1}$

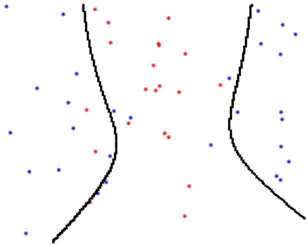


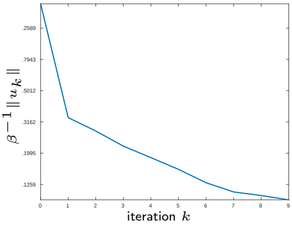


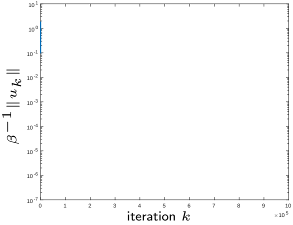


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 10 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.1e^{-1}$

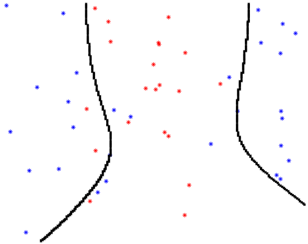


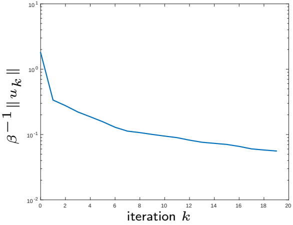


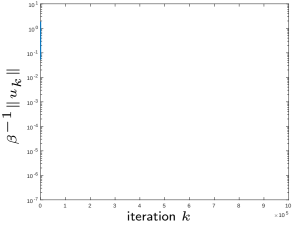


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 20 Residual norm: $\beta^{-1}\|u_k\|_2 = 5.8e^{-2}$

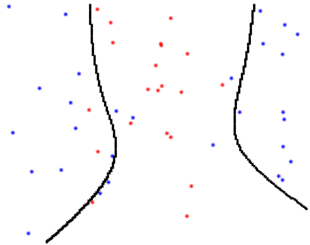


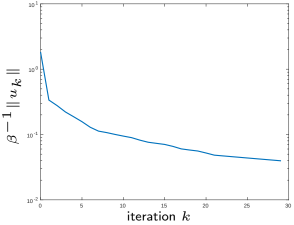


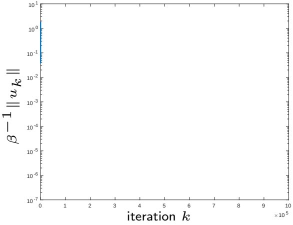


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 30 Residual norm: $\beta^{-1}\|u_k\|_2 = 4.1e^{-2}$

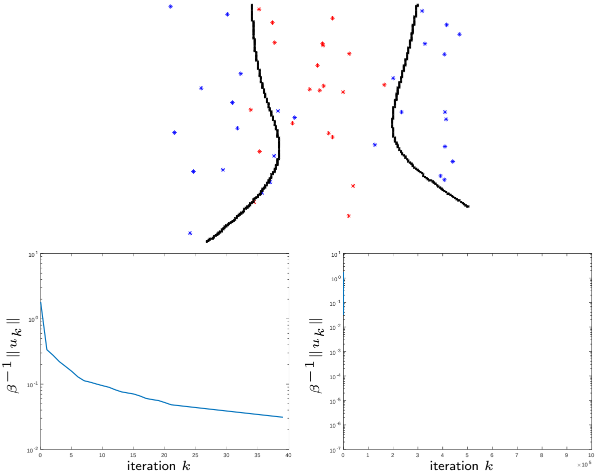






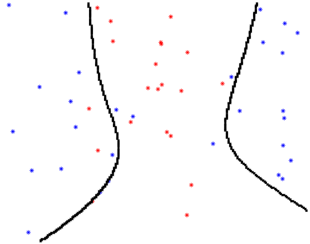
Early termination – Example

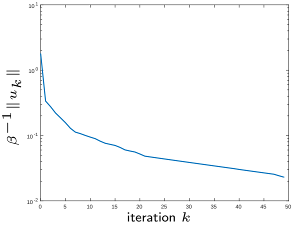
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 40 Residual norm: $\beta^{-1}\|u_k\|_2 = 3.2e^{-2}$

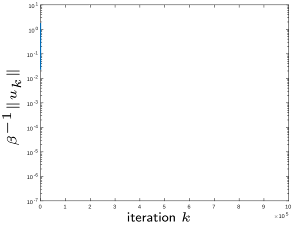


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 50 Residual norm: $\beta^{-1}\|u_k\|_2 = 2.4e^{-2}$

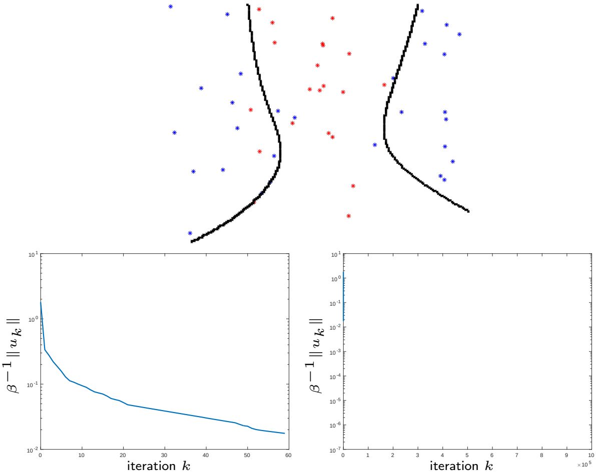






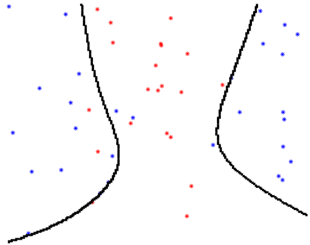
Early termination – Example

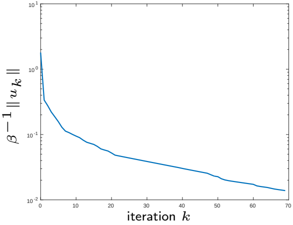
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 60 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.8e^{-2}$

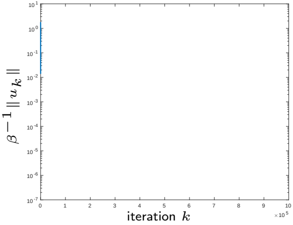


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 70 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.4e^{-2}$

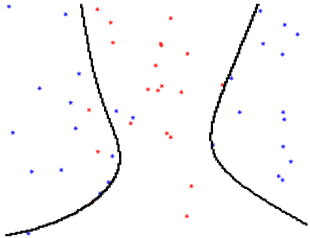


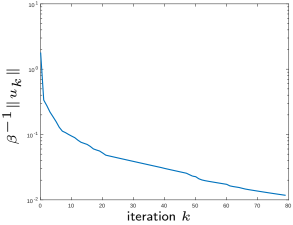


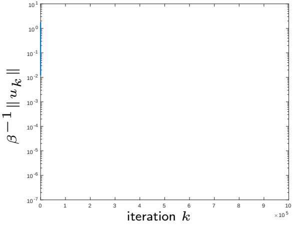


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 80 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.2e^{-2}$

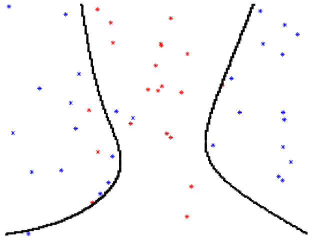


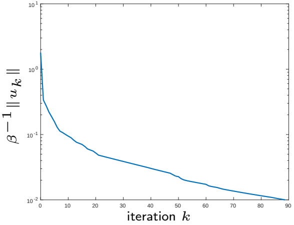


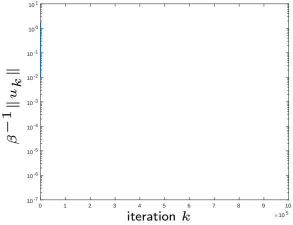


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 90 Residual norm: $\beta^{-1}\|u_k\|_2 = 1e^{-2}$

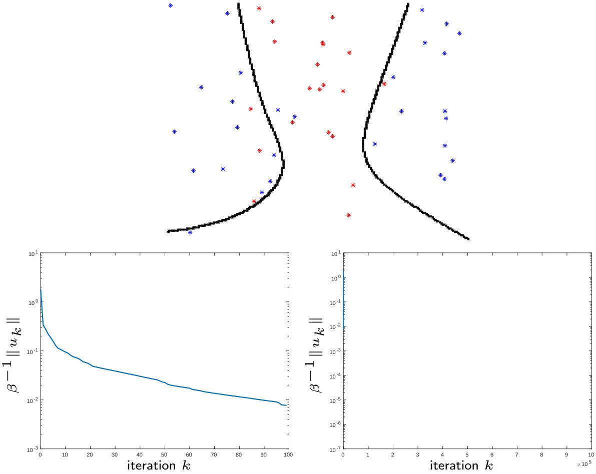






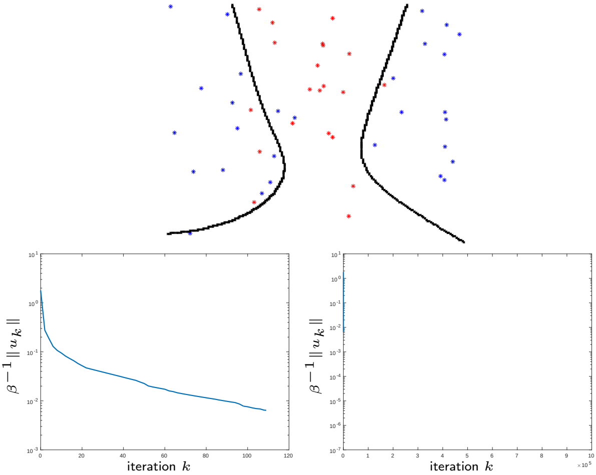
Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 100 Residual norm: $\beta^{-1}\|u_k\|_2 = 7.8e^{-3}$



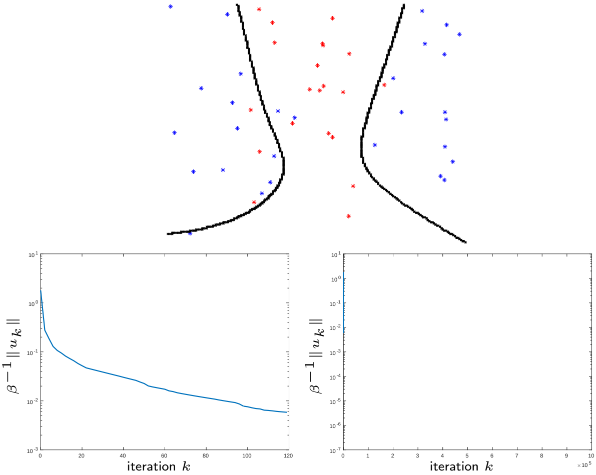
Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 110 Residual norm: $\beta^{-1}\|u_k\|_2 = 6.5e^{-3}$



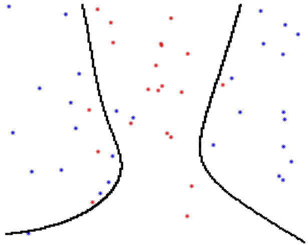
Early termination – Example

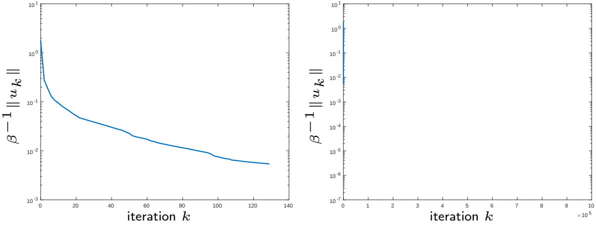
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 120 Residual norm: $\beta^{-1}\|u_k\|_2 = 5.9e^{-3}$



Early termination – Example

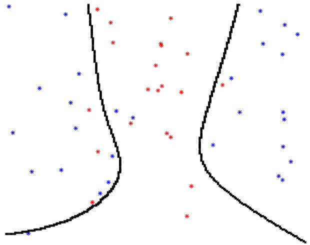
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 130 Residual norm: $\beta^{-1}\|u_k\|_2 = 5.5e^{-3}$

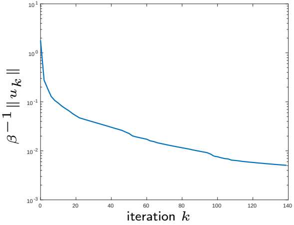


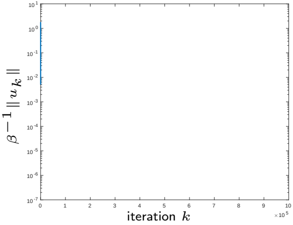


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 140 Residual norm: $\beta^{-1}\|u_k\|_2 = 5.1e^{-3}$

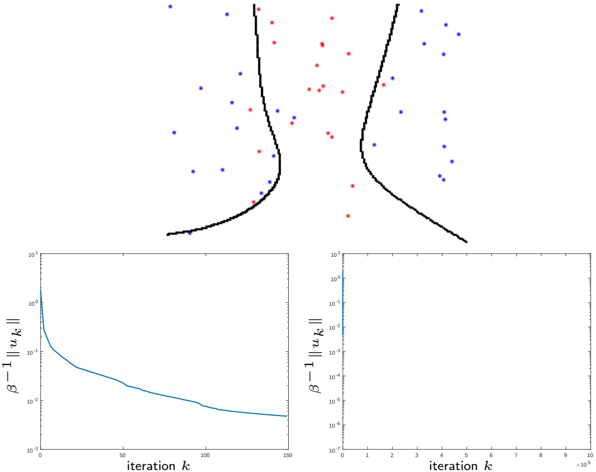






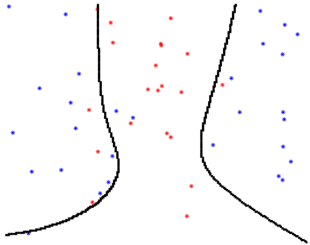
Early termination – Example

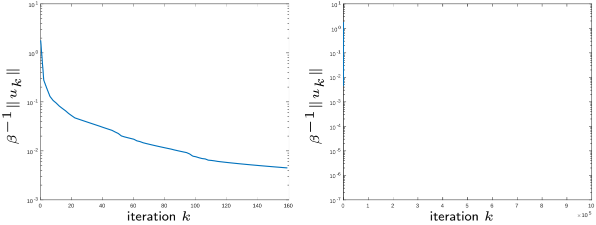
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 150 Residual norm: $\beta^{-1}\|u_k\|_2 = 4.8e^{-3}$



Early termination – Example

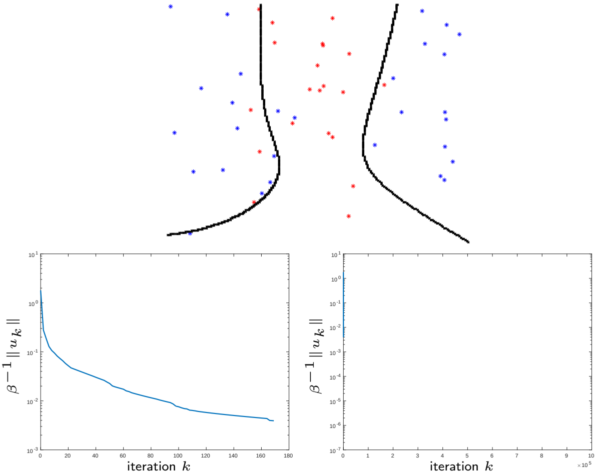
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 160 Residual norm: $\beta^{-1}\|u_k\|_2 = 4.5e^{-3}$





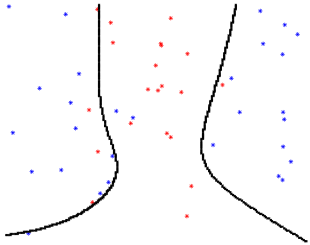
Early termination – Example

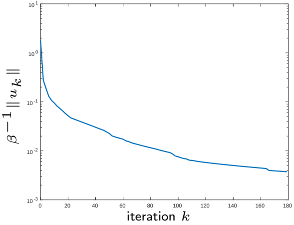
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 170 Residual norm: $\beta^{-1}\|u_k\|_2 = 3.9e^{-3}$

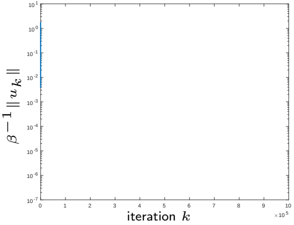


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 180 Residual norm: $\beta^{-1}\|u_k\|_2 = 3.8e^{-3}$

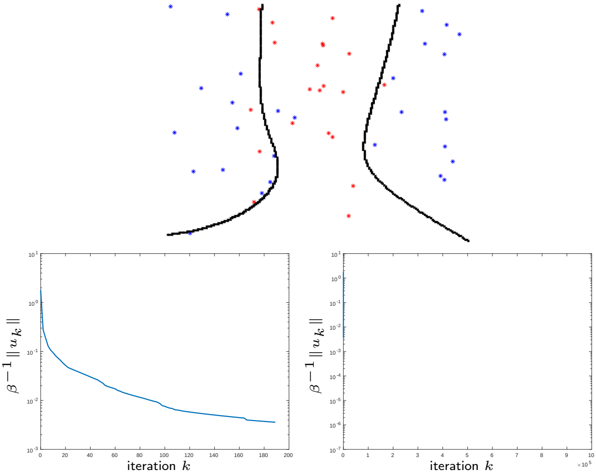






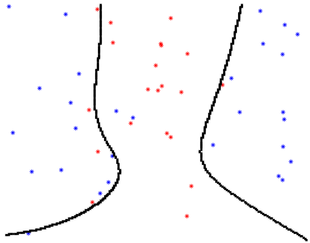
Early termination – Example

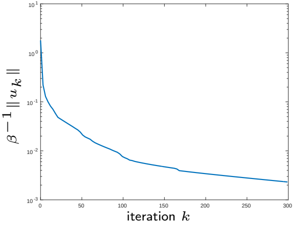
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 190 Residual norm: $\beta^{-1}\|u_k\|_2 = 3.6e^{-3}$

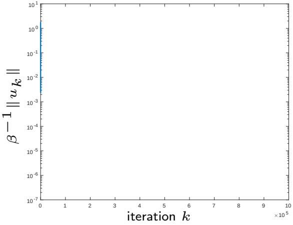


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 300 Residual norm: $\beta^{-1}\|u_k\|_2 = 2.3e^{-3}$

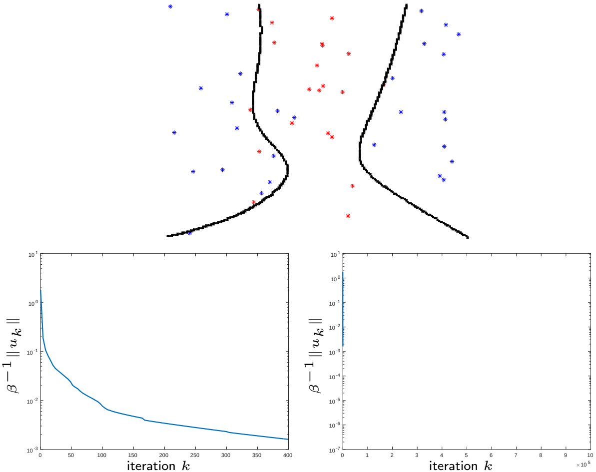






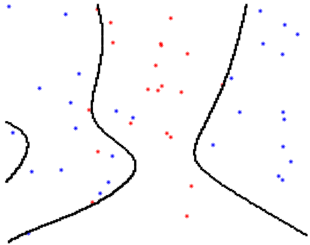
Early termination – Example

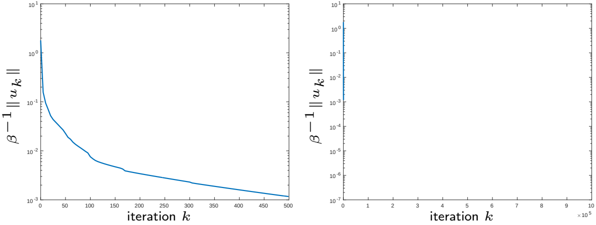
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 400 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.6e^{-3}$



Early termination – Example

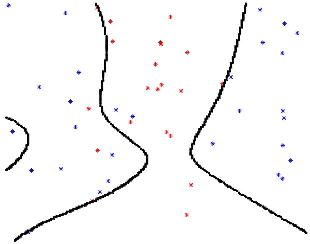
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 500 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.2e^{-3}$

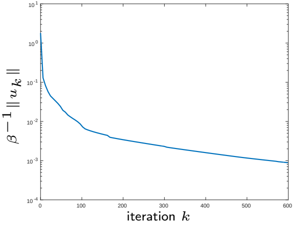


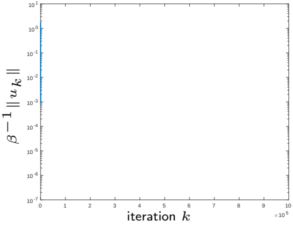


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 600 Residual norm: $\beta^{-1}\|u_k\|_2 = 8.9e^{-4}$

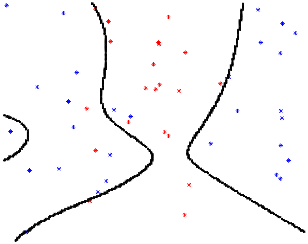


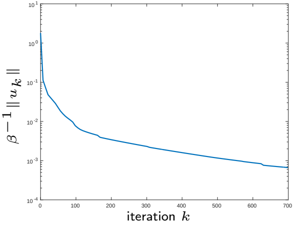


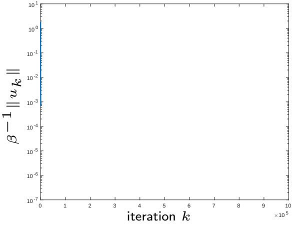


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 700 Residual norm: $\beta^{-1}\|u_k\|_2 = 6.7e^{-4}$

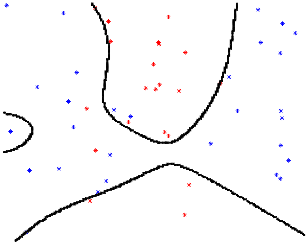


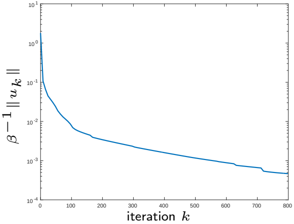


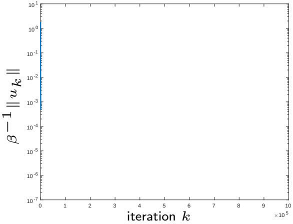


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 800 Residual norm: $\beta^{-1}\|u_k\|_2 = 4.7e^{-4}$

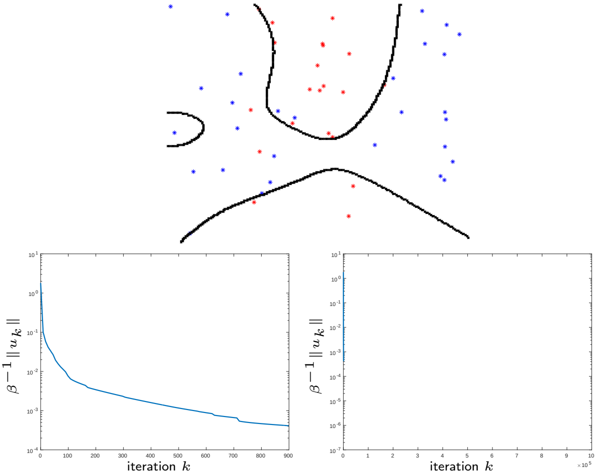






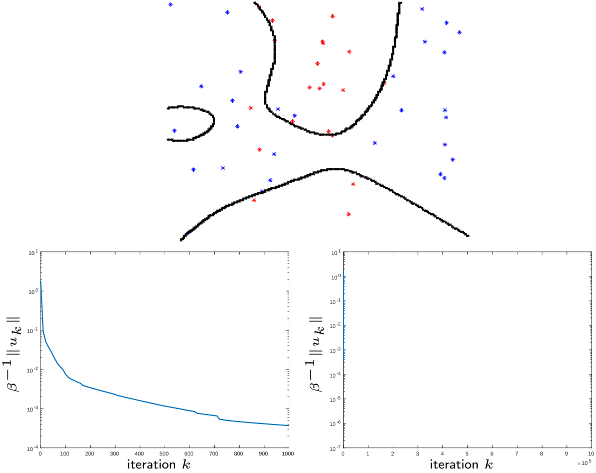
Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 900 Residual norm: $\beta^{-1}\|u_k\|_2 = 4.1e^{-4}$



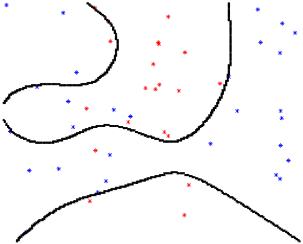
Early termination – Example

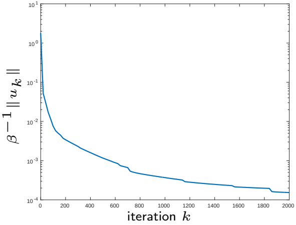
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 1000   Residual norm: $\beta^{-1}\|u_k\|_2 = 3.7e^{-4}$

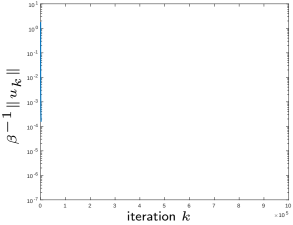


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 2000 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.5e^{-4}$

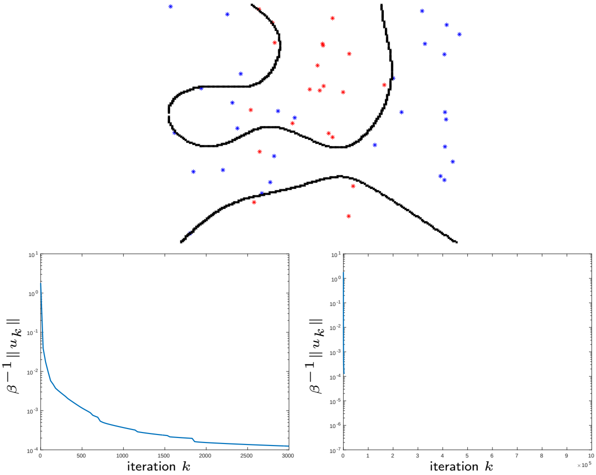






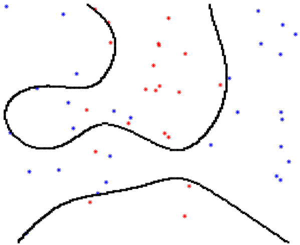
Early termination – Example

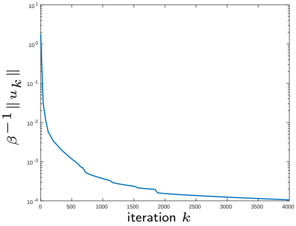
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 3000 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.2e^{-4}$

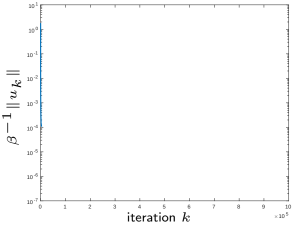


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 4000 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.1e^{-4}$

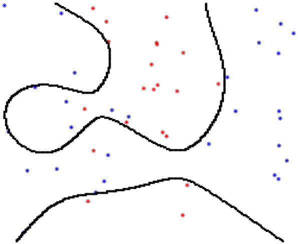


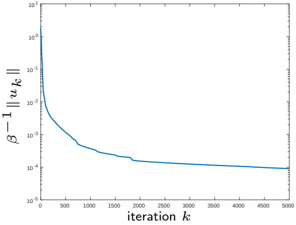


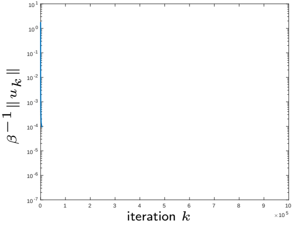


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 5000 Residual norm: $\beta^{-1}\|u_k\|_2 = 9e^{-5}$

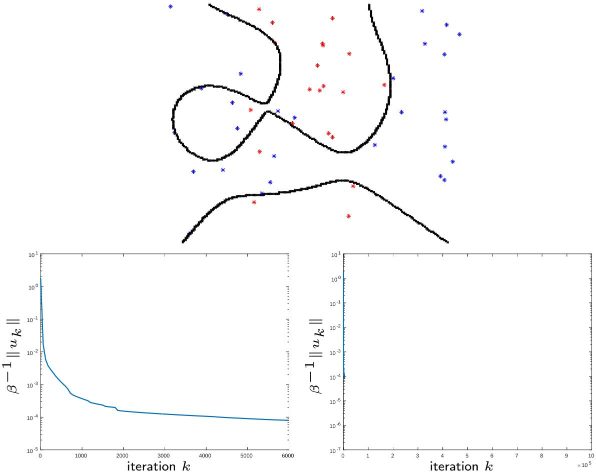






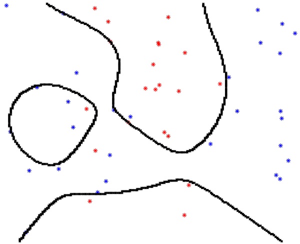
Early termination – Example

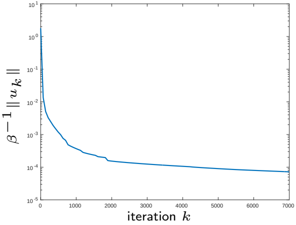
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 6000 Residual norm: $\beta^{-1}\|u_k\|_2 = 8e^{-5}$

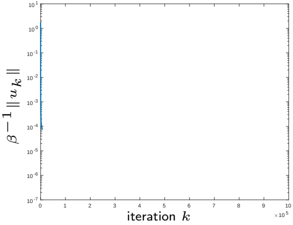


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 7000 Residual norm: $\beta^{-1}\|u_k\|_2 = 7.2e^{-5}$

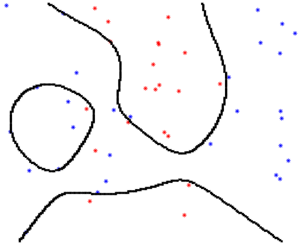


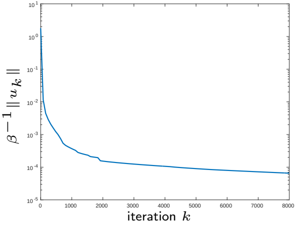


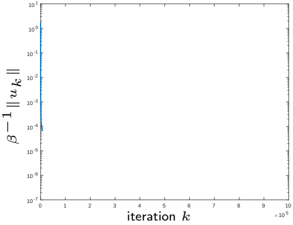


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 8000 Residual norm: $\beta^{-1}\|u_k\|_2 = 6.6e^{-5}$

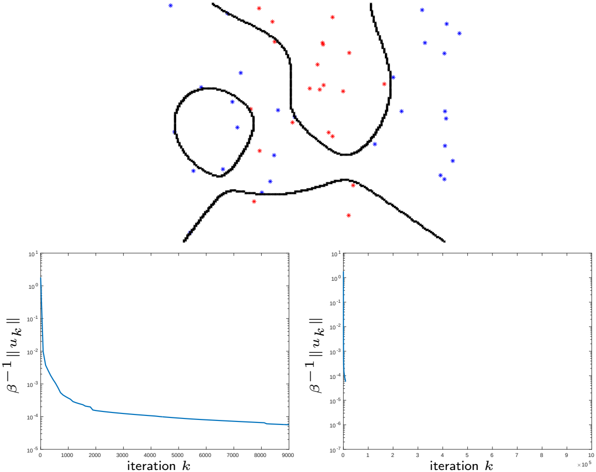






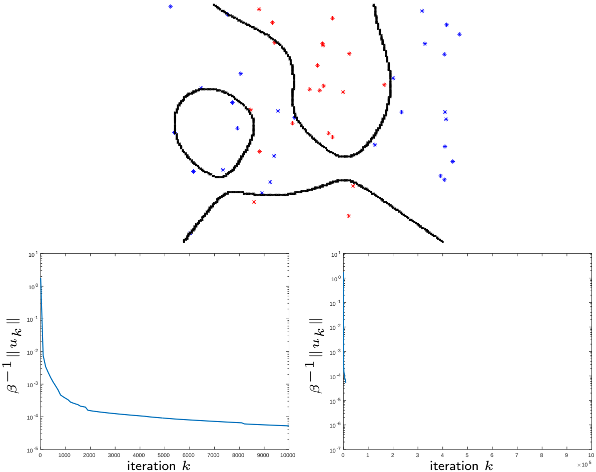
Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 9000 Residual norm: $\beta^{-1}\|u_k\|_2 = 5.6e^{-5}$



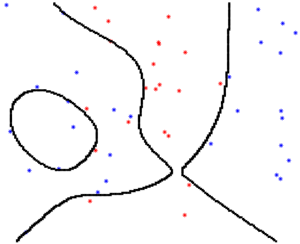
Early termination – Example

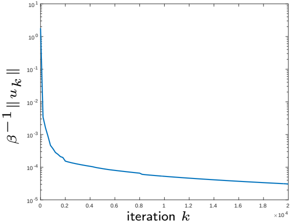
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 10000 Residual norm: $\beta^{-1}\|u_k\|_2 = 5.3e^{-5}$

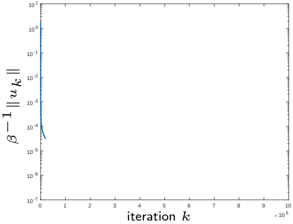


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 20000 Residual norm: $\beta^{-1}\|u_k\|_2 = 3.1e^{-5}$

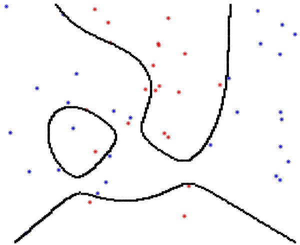


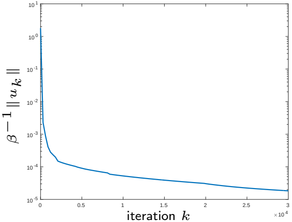


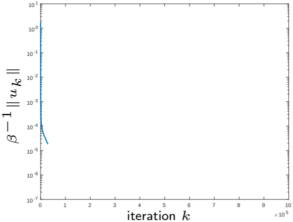


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 30000 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.8e^{-5}$

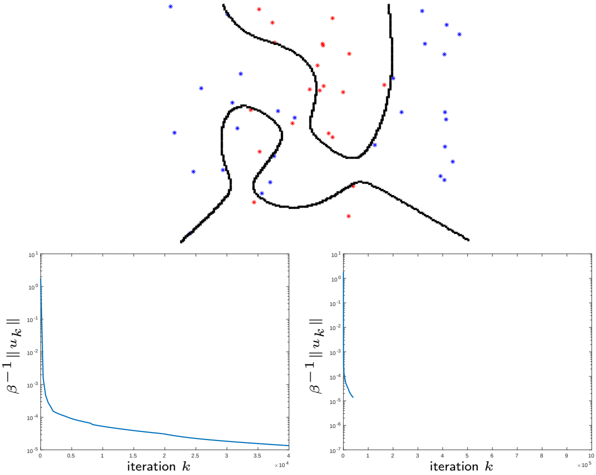






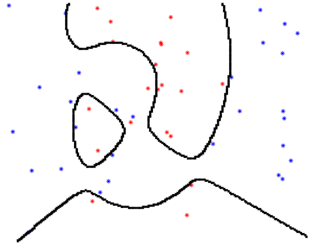
Early termination – Example

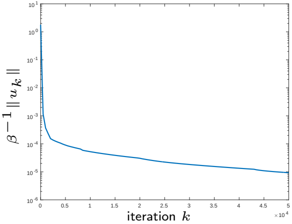
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 40000 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.3e^{-5}$

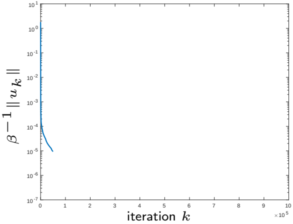


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 50000 Residual norm: $\beta^{-1}\|u_k\|_2 = 9.3e^{-6}$

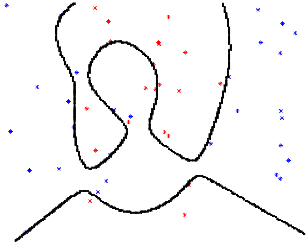


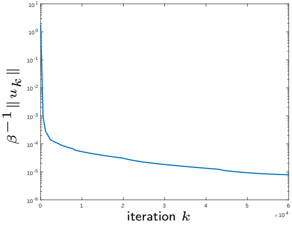


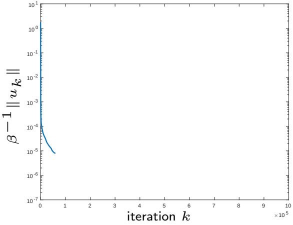


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 60000 Residual norm: $\beta^{-1}\|u_k\|_2 = 7.9e^{-6}$

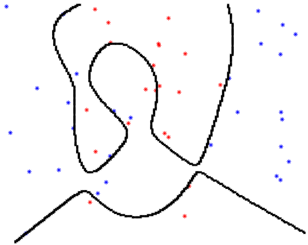


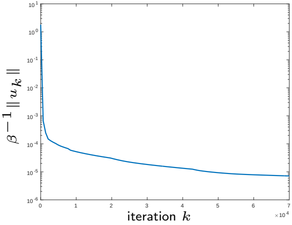


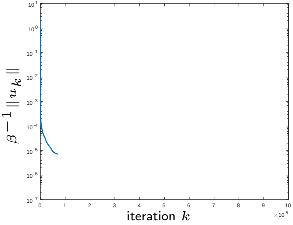


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 70000   Residual norm: $\beta^{-1}\|u_k\|_2 = 7.1e^{-6}$

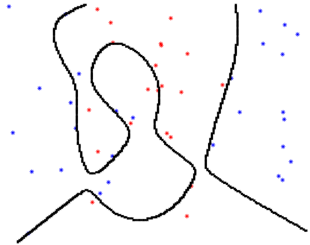


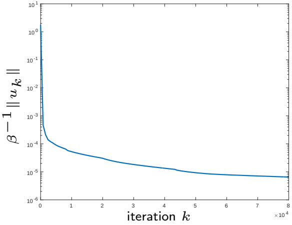


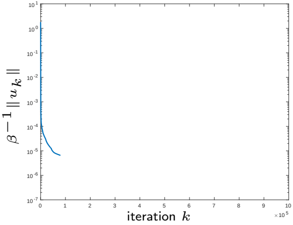


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 80000   Residual norm: $\beta^{-1}\|u_k\|_2 = 6.5e^{-6}$

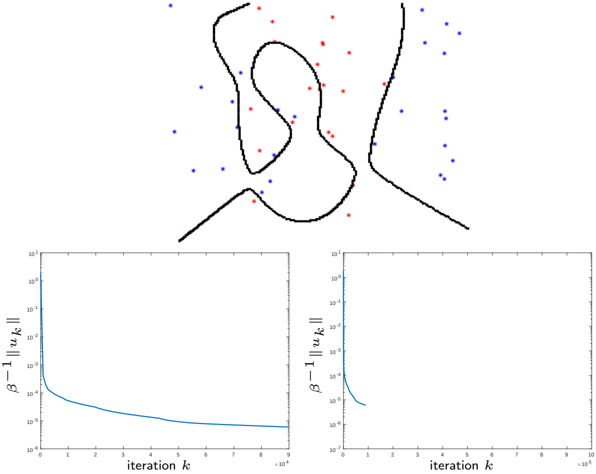






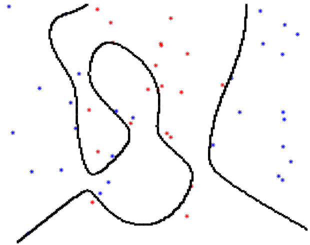
Early termination – Example

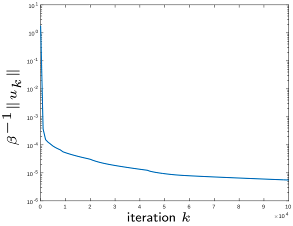
- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 90000 Residual norm: $\beta^{-1}\|u_k\|_2 = 6e^{-6}$

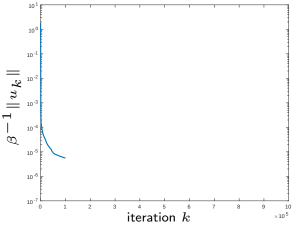


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 100000   Residual norm: $\beta^{-1}\|u_k\|_2 = 5.5e^{-6}$

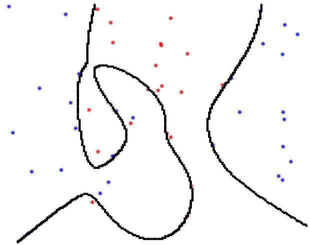


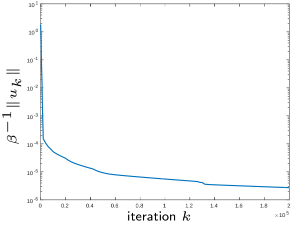


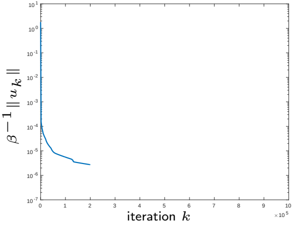


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 200000   Residual norm: $\beta^{-1}\|u_k\|_2 = 2.7e^{-6}$

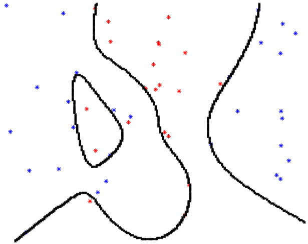


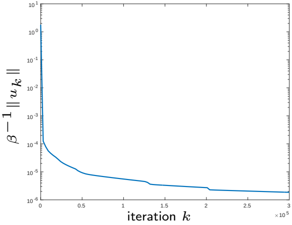


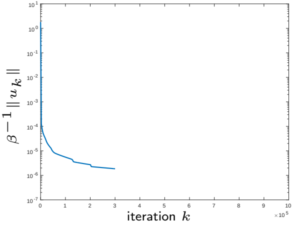


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 300000 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.9e^{-6}$

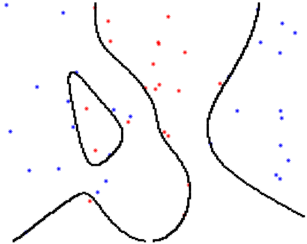


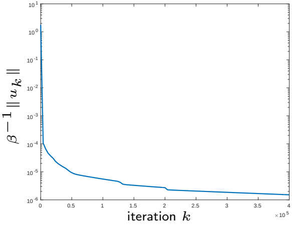


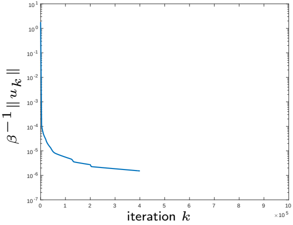


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 400000   Residual norm: $\beta^{-1}\|u_k\|_2 = 1.5e^{-6}$

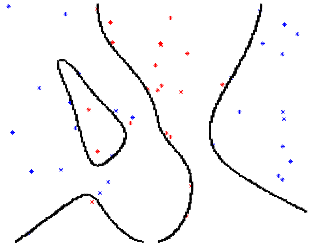


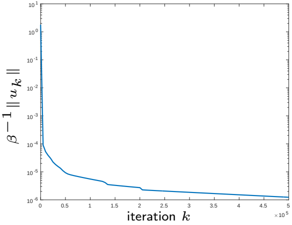


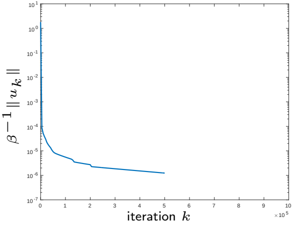


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 500000 Residual norm: $\beta^{-1}\|u_k\|_2 = 1.2e^{-6}$

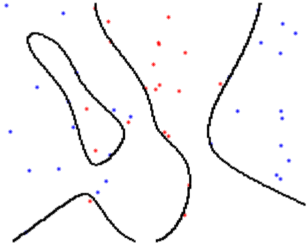


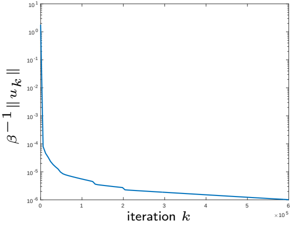


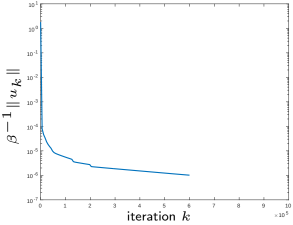


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 600000 Residual norm: $\beta^{-1}\|u_k\|_2 = 1e^{-6}$

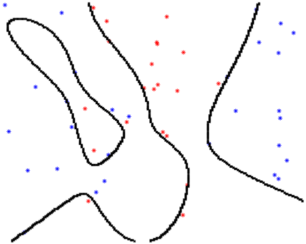


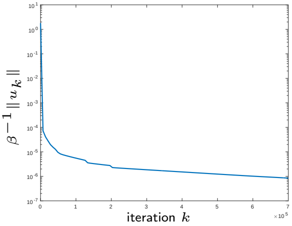


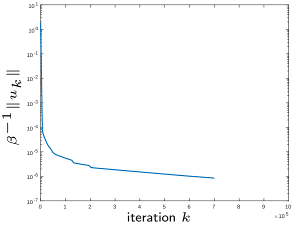


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 700000 Residual norm: $\beta^{-1}\|u_k\|_2 = 8.4e^{-7}$

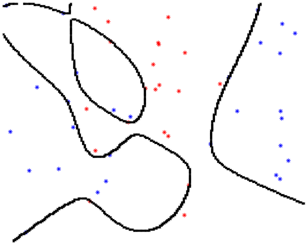


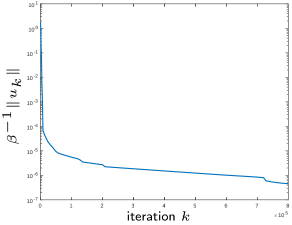


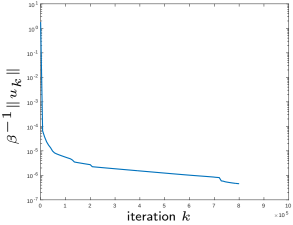


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 800000   Residual norm: $\beta^{-1}\|u_k\|_2 = 4.6e^{-7}$

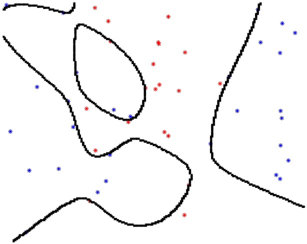


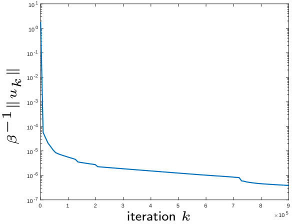


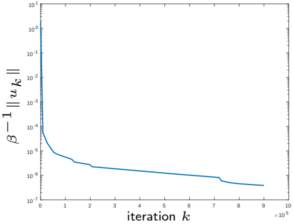


Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 900000 Residual norm: $\beta^{-1}\|u_k\|_2 = 3.9e^{-7}$







Early termination – Example

- SVM polynomial features of degree 6, $\lambda = 0.00001$
- Iteration number: 1000000 Residual norm: $\beta^{-1}\|u_k\|_2 = 3.4e^{-7}$