

Foundations of Machine Learning

Regression

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu

Regression Problem

- **Training data:** sample drawn i.i.d. from set X according to some distribution D ,

$$S = ((x_1, y_1), \dots, (x_m, y_m)) \in X \times Y,$$

with $Y \subseteq \mathbb{R}$ is a measurable subset.

- **Loss function:** $L: Y \times Y \rightarrow \mathbb{R}_+$ a measure of closeness, typically $L(y, y') = (y' - y)^2$ or $L(y, y') = |y' - y|^p$ for some $p \geq 1$.

- **Problem:** find hypothesis $h: X \rightarrow \mathbb{R}$ in H with small generalization error with respect to target f

$$R_D(h) = \mathbb{E}_{x \sim D} [L(h(x), f(x))].$$

Notes

■ Empirical error:

$$\hat{R}_D(h) = \frac{1}{m} \sum_{i=1}^m L(h(x_i), y_i).$$

■ In much of what follows:

- $Y = \mathbb{R}$ or $Y = [-M, M]$ for some $M > 0$.
- $L(y, y') = (y' - y)^2 \longrightarrow$ **mean squared error.**

This Lecture

- Generalization bounds
- Linear regression
- Kernel ridge regression
- Support vector regression
- Lasso

Generalization Bound - Finite H

- **Theorem:** let H be a finite hypothesis set, and assume that L is bounded by M . Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\forall h \in H, R(h) \leq \hat{R}(h) + M \sqrt{\frac{\log |H| + \log \frac{2}{\delta}}{2m}}.$$

- **Proof:** By the union bound,

$$\Pr \left[\sup_{h \in H} |R(h) - \hat{R}(h)| > \epsilon \right] \leq \sum_{h \in H} \Pr \left[|R(h) - \hat{R}(h)| > \epsilon \right].$$

By Hoeffding's bound, for a fixed h ,

$$\Pr \left[|R(h) - \hat{R}(h)| > \epsilon \right] \leq 2e^{-\frac{2m\epsilon^2}{M^2}}.$$

Rademacher Complexity of L_p Loss

- **Theorem:** Let $p \geq 1$, $H_p = \{x \mapsto |h(x) - f(x)|^p : h \in H\}$. Assume that $\sup_{x \in X, h \in H} |h(x) - f(x)| \leq M$. Then, for any sample S of size m ,

$$\hat{\mathfrak{R}}_S(H_p) \leq pM^{p-1}\hat{\mathfrak{R}}_S(H).$$

Proof

■ **Proof:** Let $H' = \{x \mapsto h(x) - f(x) : h \in H\}$. Then, observe that $H_p = \{\phi \circ h : h \in H'\}$ with $\phi : x \mapsto |x|^p$.

- ϕ is pM^{p-1} - Lipschitz over $[-M, M]$, thus

$$\widehat{\mathfrak{R}}_S(H_p) \leq pM^{p-1} \widehat{\mathfrak{R}}_S(H').$$

- Next, observe that:

$$\begin{aligned} \widehat{\mathfrak{R}}_S(H') &= \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i h(x_i) + \sigma_i f(x_i) \right] \\ &= \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i h(x_i) \right] + \mathbb{E}_{\sigma} \left[\sum_{i=1}^m \sigma_i f(x_i) \right] = \widehat{\mathfrak{R}}_S(H). \end{aligned}$$

Rad. Complexity Regression Bound

■ **Theorem:** Let $p \geq 1$ and assume that $\|h - f\|_\infty \leq M$ for all $h \in H$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in H$,

$$\mathbb{E} \left[|h(x) - f(x)|^p \right] \leq \frac{1}{m} \sum_{i=1}^m |h(x_i) - f(x_i)|^p + 2pM^{p-1} \mathfrak{R}_m(H) + M^p \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

$$\mathbb{E} \left[|h(x) - f(x)|^p \right] \leq \frac{1}{m} \sum_{i=1}^m |h(x_i) - f(x_i)|^p + 2pM^{p-1} \widehat{\mathfrak{R}}_S(H) + 3M^p \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

■ **Proof:** Follows directly bound on Rademacher complexity and general Rademacher bound.

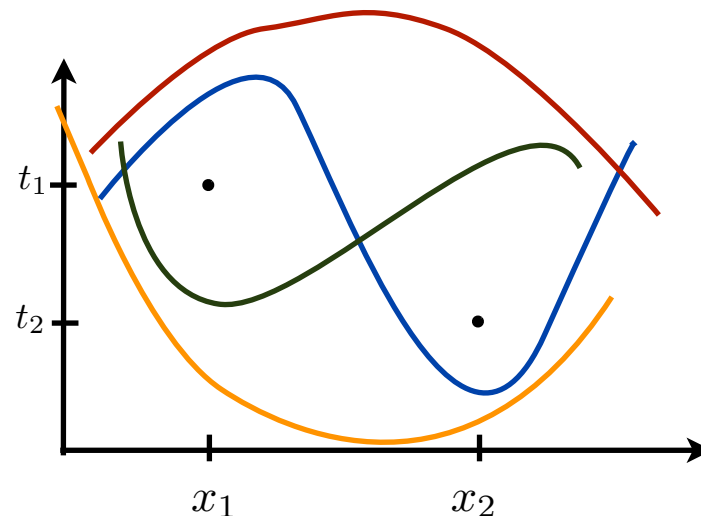
Notes

- As discussed for binary classification:
 - estimating the Rademacher complexity can be computationally hard for some H s.
 - can we come up instead with a combinatorial measure that is easier to compute?

Shattering

- **Definition:** Let G be a family of functions mapping from X to \mathbb{R} . $A = \{x_1, \dots, x_m\}$ is **shattered** by G if there exist $t_1, \dots, t_m \in \mathbb{R}$ such that

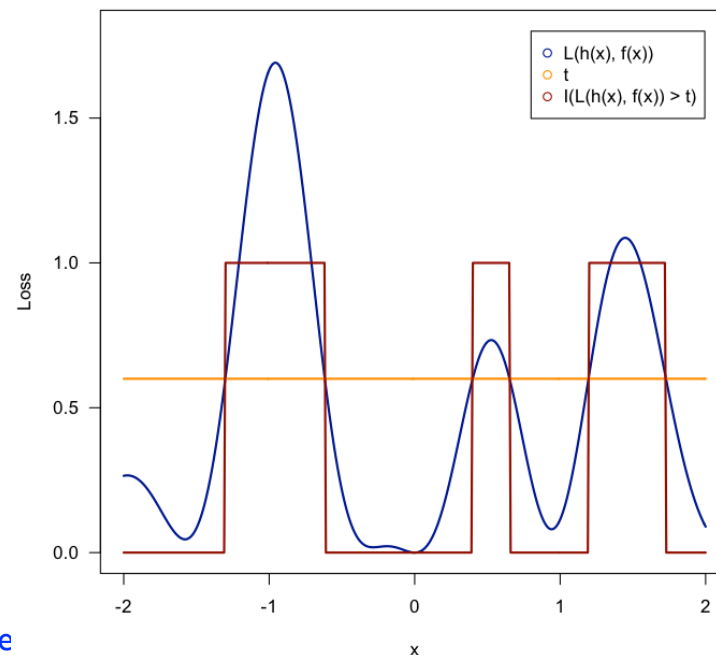
$$\left| \left\{ \begin{bmatrix} \text{sgn}(g(x_1) - t_1) \\ \vdots \\ \text{sgn}(g(x_m) - t_m) \end{bmatrix} : g \in G \right\} \right| = 2^m.$$



Pseudo-Dimension

(Pollard, 1984)

- **Definition:** Let G be a family of functions mapping from X to \mathbb{R} . The pseudo-dimension of G , $\text{Pdim}(G)$, is the size of the largest set shattered by G .
- **Definition** (equivalent, see also (Vapnik, 1995)):
$$\text{Pdim}(G) = \text{VCdim}\left(\{(x, t) \mapsto 1_{(g(x) - t) > 0} : g \in G\}\right).$$



Pseudo-Dimension - Properties

- **Theorem:** Pseudo-dimension of hyperplanes.

$$\text{Pdim}(\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} + b : \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}) = N + 1.$$

- **Theorem:** Pseudo-dimension of a vector space of real-valued functions H :

$$\text{Pdim}(H) = \dim(H).$$

Generalization Bounds

Classification \longrightarrow Regression

- **Lemma** (Lebesgue integral): for $f \geq 0$ measurable,

$$\mathbb{E}_D[f(x)] = \int_0^\infty \Pr_D[f(x) > t] dt.$$

- Assume that the loss function L is bounded by M .

$$\begin{aligned} |R(h) - \widehat{R}(h)| &= \left| \int_0^M \left(\Pr_{x \sim D}[L(h(x), f(x)) > t] - \Pr_{x \sim S}[L(h(x), f(x)) > t] \right) dt \right| \\ &\leq M \sup_{t \in [0, M]} \left| \Pr_{x \sim D}[L(h(x), f(x)) > t] - \Pr_{x \sim S}[L(h(x), f(x)) > t] \right| \\ &= M \sup_{t \in [0, M]} \left| \mathbb{E}_{x \sim D}[1_{L(h(x), f(x)) > t}] - \mathbb{E}_{x \sim S}[1_{L(h(x), f(x)) > t}] \right|. \end{aligned}$$

$$\Pr \left[\sup_{h \in H} |R(h) - \widehat{R}(h)| > \epsilon \right] \leq \Pr \left[\sup_{\substack{h \in H \\ t \in [0, M]}} \left| R(1_{L(h, f) > t}) - \widehat{R}(1_{L(h, f) > t}) \right| > \frac{\epsilon}{M} \right].$$

Standard classification generalization bound.

Generalization Bound - Pdim

- **Theorem:** Let H be a family of real-valued functions. Assume that $\text{Pdim}(\{L(h, f) : h \in H\}) = d < \infty$ and that the loss L is bounded by M . Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \hat{R}(h) + M \sqrt{\frac{2d \log \frac{em}{d}}{m}} + M \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

- **Proof:** follows observation of previous slide and VCDim bound for indicator functions of lecture 3.

Notes

- Pdim bounds in unbounded case modulo assumptions: existence of an envelope function or moment assumptions.
- Other relevant capacity measures:
 - covering numbers.
 - packing numbers.
 - fat-shattering dimension.

This Lecture

- Generalization bounds
- Linear regression
- Kernel ridge regression
- Support vector regression
- Lasso

Linear Regression

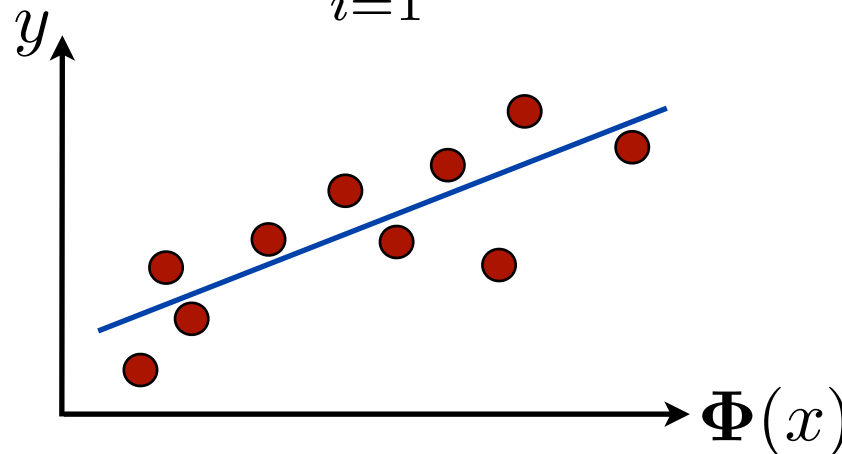
■ Feature mapping $\Phi : X \rightarrow \mathbb{R}^N$.

■ Hypothesis set: linear functions.

$$\{x \mapsto \mathbf{w} \cdot \Phi(x) + b : \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}.$$

■ **Optimization problem:** empirical risk minimization.

$$\min_{\mathbf{w}, b} F(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^m (\mathbf{w} \cdot \Phi(x_i) + b - y_i)^2.$$



Linear Regression - Solution

- Rewrite objective function as $F(\mathbf{W}) = \frac{1}{m} \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|^2$,
 $\mathbf{X} = \begin{bmatrix} \Phi(x_1) & \dots & \Phi(x_m) \\ 1 & \dots & 1 \end{bmatrix} \in \mathbb{R}^{(N+1) \times m}$

$$\text{with } \mathbf{X}^\top = \begin{bmatrix} \Phi(x_1)^\top & 1 \\ \vdots & \\ \Phi(x_m)^\top & 1 \end{bmatrix} \quad \mathbf{W} = \begin{bmatrix} w_1 \\ \vdots \\ w_N \\ b \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} .$$

- Convex and differentiable function.

$$\nabla F(\mathbf{W}) = \frac{2}{m} \mathbf{X}(\mathbf{X}^\top \mathbf{W} - \mathbf{Y}).$$

$$\nabla F(\mathbf{W}) = 0 \Leftrightarrow \mathbf{X}(\mathbf{X}^\top \mathbf{W} - \mathbf{Y}) = 0 \Leftrightarrow \mathbf{X}\mathbf{X}^\top \mathbf{W} = \mathbf{X}\mathbf{Y}.$$

Linear Regression - Solution

■ Solution:

$$\mathbf{w} = \begin{cases} (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{Y} & \text{if } \mathbf{X}\mathbf{X}^\top \text{ invertible.} \\ (\mathbf{X}\mathbf{X}^\top)^\dagger\mathbf{X}\mathbf{Y} & \text{in general.} \end{cases}$$

- **Computational complexity:** $O(mN + N^3)$ if matrix inversion in $O(N^3)$.
- Poor guarantees in general, no regularization.
- For output labels in \mathbb{R}^p , $p > 1$, solve p distinct linear regression problems.

This Lecture

- Generalization bounds
- Linear regression
- Kernel ridge regression
- Support vector regression
- Lasso

Mean Square Bound - Kernel-Based Hypotheses

■ **Theorem:** Let $K: X \times X \rightarrow \mathbb{R}$ be a PDS kernel and let $\Phi: X \rightarrow \mathbb{H}$ be a feature mapping associated to K . Let $H = \left\{ \mathbf{x} \mapsto \mathbf{w} \cdot \Phi(x) : \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda \right\}$. Assume $K(x, x) \leq R^2$ and $|f(x)| \leq \Lambda R$ for all $x \in X$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}(h) + \frac{8R^2\Lambda^2}{\sqrt{m}} \left(1 + \frac{1}{2} \sqrt{\frac{\log \frac{1}{\delta}}{2}} \right)$$

$$R(h) \leq \widehat{R}(h) + \frac{8R^2\Lambda^2}{\sqrt{m}} \left(\sqrt{\frac{\text{Tr}[\mathbf{K}]}{mR^2}} + \frac{3}{4} \sqrt{\frac{\log \frac{2}{\delta}}{2}} \right).$$

Mean Square Bound - Kernel-Based Hypotheses

- **Proof:** direct application of the Rademacher Complexity Regression Bound (this lecture) and bound on the Rademacher complexity of kernel-based hypotheses (lecture 5):

$$\widehat{\mathfrak{R}}_S(H) \leq \frac{\Lambda \sqrt{\text{Tr}[\mathbf{K}]}}{m} \leq \sqrt{\frac{R^2 \Lambda^2}{m}}.$$

Ridge Regression

(Hoerl and Kennard, 1970)

■ Optimization problem:

$$\min_{\mathbf{w}} F(\mathbf{w}, b) = \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^m (\mathbf{w} \cdot \Phi(x_i) + b - y_i)^2,$$

where $\lambda \geq 0$ is a (regularization) parameter.

- directly based on generalization bound.
- generalization of linear regression.
- closed-form solution.
- can be used with kernels.

Ridge Regression - Solution

- Assume $b=0$: often constant feature used (but not equivalent to the use of original offset!).

- Rewrite objective function as

$$F(\mathbf{W}) = \lambda \|\mathbf{W}\|^2 + \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|^2.$$

- Convex and differentiable function.

$$\nabla F(\mathbf{W}) = 2\lambda \mathbf{W} + 2\mathbf{X}(\mathbf{X}^\top \mathbf{W} - \mathbf{Y}).$$

$$\nabla F(\mathbf{W}) = 0 \Leftrightarrow (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})\mathbf{W} = \mathbf{X}\mathbf{Y}.$$

- **Solution:**

$$\mathbf{W} = \underbrace{(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}}_{\text{always invertible.}} \mathbf{X}\mathbf{Y}.$$

Ridge Regression - Equivalent Formulations

■ Optimization problem:

$$\min_{\mathbf{w}, b} \sum_{i=1}^m (\mathbf{w} \cdot \Phi(x_i) + b - y_i)^2$$

$$\text{subject to: } \|\mathbf{w}\|^2 \leq \Lambda^2.$$

■ Optimization problem:

$$\min_{\mathbf{w}, b} \sum_{i=1}^m \xi_i^2$$

$$\text{subject to: } \xi_i = \mathbf{w} \cdot \Phi(x_i) + b - y_i$$

$$\|\mathbf{w}\|^2 \leq \Lambda^2.$$

Ridge Regression Equations

■ **Lagrangian:** assume $b=0$. For all $\xi, \mathbf{w}, \alpha', \lambda \geq 0$,

$$L(\xi, \mathbf{w}, \alpha', \lambda) = \sum_{i=1}^m \xi_i^2 + \sum_{i=1}^m \alpha'_i (y_i - \xi_i - \mathbf{w} \cdot \Phi(x_i)) + \lambda (\|\mathbf{w}\|^2 - \Lambda^2).$$

■ **KKT conditions:**

$$\begin{aligned} \nabla_{\mathbf{w}} L &= - \sum_{i=1}^m \alpha'_i \Phi(x_i) + 2\lambda \mathbf{w} = 0 & \iff & \mathbf{w} = \frac{1}{2\lambda} \sum_{i=1}^m \alpha'_i \Phi(x_i). \\ \nabla_{\xi_i} L &= 2\xi_i - \alpha'_i = 0 & \iff & \xi_i = \alpha'_i / 2. \end{aligned}$$

$$\begin{aligned} \forall i \in [1, m], \alpha'_i (y_i - \xi_i - \mathbf{w} \cdot \Phi(x_i)) &= 0 \\ \lambda (\|\mathbf{w}\|^2 - \Lambda^2) &= 0. \end{aligned}$$

Moving to The Dual

- Plugging in the expression of w and ξ_i s gives

$$L = \sum_{i=1}^m \frac{\alpha_i'^2}{4} + \sum_{i=1}^m \alpha_i' y_i - \sum_{i=1}^m \frac{\alpha_i'^2}{2} - \frac{1}{2\lambda} \sum_{i,j=1}^m \alpha_i' \alpha_j' \Phi(x_i)^\top \Phi(x_j) + \lambda \left(\frac{1}{4\lambda^2} \left\| \sum_{i=1}^m \alpha_i' \Phi(x_i) \right\|^2 - \Lambda^2 \right).$$

- Thus,

$$\begin{aligned} L &= -\frac{1}{4} \sum_{i=1}^m \alpha_i'^2 + \sum_{i=1}^m \alpha_i' y_i - \frac{1}{4\lambda} \sum_{i,j=1}^m \alpha_i' \alpha_j' \Phi(x_i)^\top \Phi(x_j) - \lambda \Lambda^2 \\ &= -\lambda \sum_{i=1}^m \alpha_i^2 + 2 \sum_{i=1}^m \alpha_i y_i - \sum_{i,j=1}^m \alpha_i \alpha_j \Phi(x_i)^\top \Phi(x_j) - \lambda \Lambda^2, \end{aligned}$$

with $\alpha_i' = 2\lambda \alpha_i$.

RR - Dual Optimization Problem

■ Optimization problem:

$$\max_{\alpha \in \mathbb{R}^m} -\lambda \alpha^\top \alpha + 2\alpha^\top \mathbf{y} - \alpha^\top (\mathbf{X}^\top \mathbf{X}) \alpha$$

or
$$\max_{\alpha \in \mathbb{R}^m} -\alpha^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}) \alpha + 2\alpha^\top \mathbf{y}.$$

■ Solution:

$$h(x) = \sum_{i=1}^m \alpha_i \Phi(\mathbf{x}_i) \cdot \Phi(x),$$

with
$$\alpha = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{y}.$$

Direct Dual Solution

- **Lemma:** The following matrix identity always holds.

$$(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}.$$

- **Proof:** Observe that $(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})\mathbf{X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})$. Left-multiplying by $(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}$ and right-multiplying by $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}$ yields the statement.

- **Dual solution:** α such that

$$\mathbf{W} = \sum_{i=1}^m \alpha_i K(x_i, \cdot) = \sum_{i=1}^m \alpha_i \Phi(x_i) = \mathbf{X}\alpha.$$

By lemma, $\mathbf{W} = (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{Y} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{Y}$.

This gives

$$\alpha = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{Y}.$$

Computational Complexity

	Solution	Prediction
Primal	$O(mN^2 + N^3)$	$O(N)$
Dual	$O(\kappa m^2 + m^3)$	$O(\kappa m)$

Kernel Ridge Regression

(Saunders et al., 1998)

■ Optimization problem:

$$\max_{\alpha \in \mathbb{R}^m} -\lambda \alpha^\top \alpha + 2\alpha^\top \mathbf{y} - \alpha^\top \mathbf{K} \alpha$$

or
$$\max_{\alpha \in \mathbb{R}^m} -\alpha^\top (\mathbf{K} + \lambda \mathbf{I}) \alpha + 2\alpha^\top \mathbf{y}.$$

■ Solution:

$$h(x) = \sum_{i=1}^m \alpha_i K(x_i, x),$$

with
$$\alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y}.$$

Notes

■ Advantages:

- strong theoretical guarantees.
- generalization to outputs in \mathbb{R}^p : single matrix inversion (Cortes et al., 2007).
- use of kernels.

■ Disadvantages:

- solution not sparse.
- training time for large matrices: low-rank approximations of kernel matrix, e.g., Nyström approx., partial Cholesky decomposition.

This Lecture

- Generalization bounds
- Linear regression
- Kernel ridge regression
- Support vector regression
- Lasso

Support Vector Regression

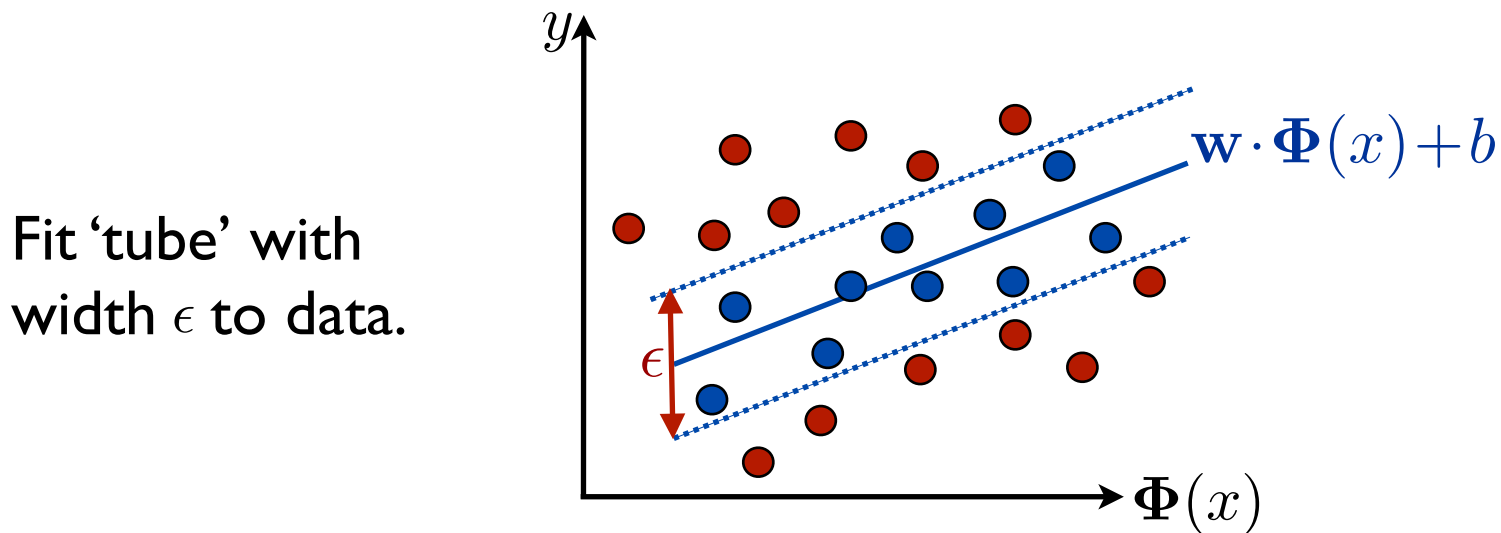
(Vapnik, 1995)

- Hypothesis set:

$$\{x \mapsto \mathbf{w} \cdot \Phi(x) + b : \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}.$$

- Loss function: ϵ -insensitive loss.

$$L(y, y') = |y' - y|_{\epsilon} = \max(0, |y' - y| - \epsilon).$$



Support Vector Regression (SVR)

(Vapnik, 1995)

- **Optimization problem:** similar to that of SVM.

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m |y_i - (\mathbf{w} \cdot \Phi(x_i) + b)|_{\epsilon}.$$

- **Equivalent formulation:**

$$\min_{\mathbf{w}, \xi, \xi'} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \xi'_i)$$

subject to $(\mathbf{w} \cdot \Phi(x_i) + b) - y_i \leq \epsilon + \xi_i$

$y_i - (\mathbf{w} \cdot \Phi(x_i) + b) \leq \epsilon + \xi'_i$

$\xi_i \geq 0, \xi'_i \geq 0.$

SVR - Dual Optimization Problem

■ Optimization problem:

$$\max_{\alpha, \alpha'} -\epsilon(\alpha' + \alpha)^\top \mathbf{1} + (\alpha' - \alpha)^\top \mathbf{y} - \frac{1}{2}(\alpha' - \alpha)^\top \mathbf{K}(\alpha' - \alpha)$$

subject to: $(\mathbf{0} \leq \alpha \leq \mathbf{C}) \wedge (\mathbf{0} \leq \alpha' \leq \mathbf{C}) \wedge ((\alpha' - \alpha)^\top \mathbf{1} = 0)$.

■ Solution:

$$h(x) = \sum_{i=1}^m (\alpha'_i - \alpha_i) K(\mathbf{x}_i, \mathbf{x}) + b$$

$$\text{with } b = \begin{cases} -\sum_{i=1}^m (\alpha'_j - \alpha_j) K(x_j, x_i) + y_i + \epsilon & \text{when } 0 < \alpha_i < C \\ -\sum_{i=1}^m (\alpha'_j - \alpha_j) K(x_j, x_i) + y_i - \epsilon & \text{when } 0 < \alpha'_i < C. \end{cases}$$

■ Support vectors: points strictly outside the tube.

Notes

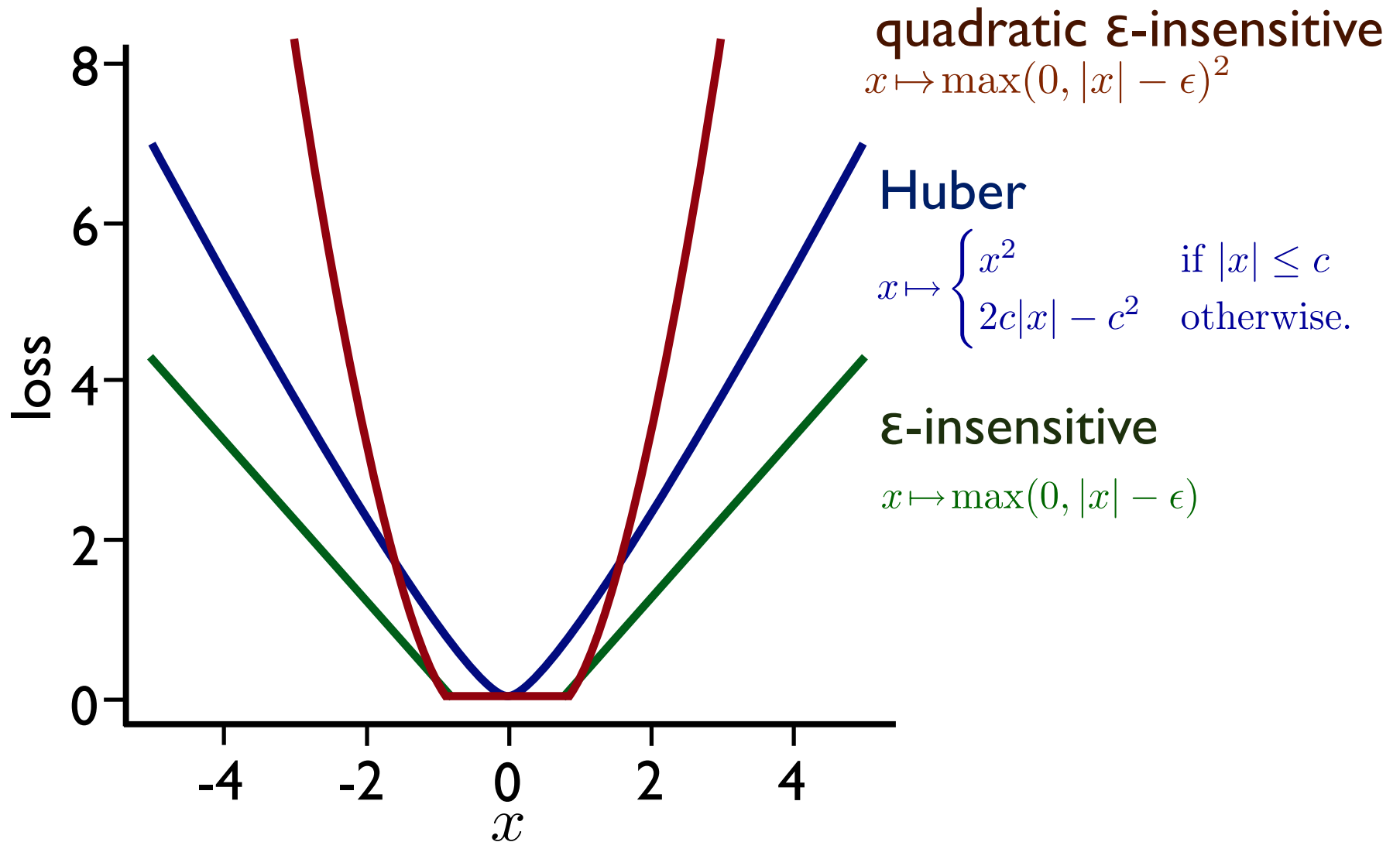
■ Advantages:

- strong theoretical guarantees (for that loss).
- sparser solution.
- use of kernels.

■ Disadvantages:

- selection of two parameters: C and ϵ . Heuristics:
 - search C near maximum y , ϵ near average difference of y s, measure of no. of SVs.
- large matrices: low-rank approximations of kernel matrix.

Alternative Loss Functions



SVR - Quadratic Loss

■ Optimization problem:

$$\max_{\alpha, \alpha'} -\epsilon(\alpha' + \alpha)^\top \mathbf{1} + (\alpha' - \alpha)^\top \mathbf{y} - \frac{1}{2}(\alpha' - \alpha)^\top \left(\mathbf{K} + \frac{1}{C} \mathbf{I} \right) (\alpha' - \alpha)$$

subject to: $(\alpha \geq \mathbf{0}) \wedge (\alpha' \geq \mathbf{0}) \wedge (\alpha' - \alpha)^\top \mathbf{1} = 0$.

■ Solution:

$$h(x) = \sum_{i=1}^m (\alpha'_i - \alpha_i) K(\mathbf{x}_i, \mathbf{x}) + b$$

with $b = \begin{cases} -\sum_{i=1}^m (\alpha'_j - \alpha_j) K(x_j, x_i) + y_i + \epsilon & \text{when } 0 < \alpha_i \wedge \xi_i = 0 \\ -\sum_{i=1}^m (\alpha'_j - \alpha_j) K(x_j, x_i) + y_i - \epsilon & \text{when } 0 < \alpha'_i \wedge \xi'_i = 0. \end{cases}$

■ Support vectors: points strictly outside the tube.

■ For $\epsilon=0$, coincides with KRR.

ϵ -Insensitive Bound - Kernel-Based Hypotheses

■ **Theorem:** Let $K: X \times X \rightarrow \mathbb{R}$ be a PDS kernel and let $\Phi: X \rightarrow H$ be a feature mapping associated to K . Let $H = \{\mathbf{x} \mapsto \mathbf{w} \cdot \Phi(x) : \|\mathbf{w}\|_H \leq \Lambda\}$. Assume $K(x, x) \leq R^2$ and $|f(x)| \leq \Gamma R$ for all $x \in X$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$\mathbb{E}[|h(x) - f(x)|_\epsilon] \leq \widehat{\mathbb{E}}[|h(x) - f(x)|_\epsilon] + \frac{R\Lambda}{\sqrt{m}} \left[2 + \left(\frac{\Gamma}{\Lambda} + 1 \right) \sqrt{\frac{\log \frac{1}{\delta}}{2}} \right].$$

$$\mathbb{E}[|h(x) - f(x)|_\epsilon] \leq \widehat{\mathbb{E}}[|h(x) - f(x)|_\epsilon] + \frac{\Lambda R}{\sqrt{m}} \left[2 \sqrt{\frac{\text{Tr}[\mathbf{K}]/R^2}{m}} + 3 \left(\frac{\Gamma}{\Lambda} + 1 \right) \sqrt{\frac{\log \frac{2}{\delta}}{2}} \right].$$

ϵ -Insensitive Bound - Kernel-Based Hypotheses

- **Proof:** Let $H_\epsilon = \{x \mapsto |h(x) - f(x)|_\epsilon : h \in H\}$ and let H' be defined by $H' = \{x \mapsto h(x) - f(x) : h \in H\}$.
- The function $\Phi_\epsilon : x \mapsto |x|_\epsilon$ is 1-Lipschitz and $\Phi_\epsilon(0) = 0$. Thus, by the contraction lemma,
$$\hat{\mathfrak{R}}_S(H_\epsilon) \leq \hat{\mathfrak{R}}_S(H').$$
- Since $\hat{\mathfrak{R}}_S(H') = \hat{\mathfrak{R}}_S(H)$ (see proof for Rademacher Complexity of L_p Loss), this shows that $\hat{\mathfrak{R}}_S(H_\epsilon) \leq \hat{\mathfrak{R}}_S(H)$.
- The rest is a direct application of the Rademacher Complexity Regression Bound (this lecture).

On-line Regression

- On-line version of batch algorithms:
 - stochastic gradient descent.
 - primal or dual.
- Examples:
 - Mean squared error function: **Widrow-Hoff** (or **LMS**) **algorithm** (Widrow and Hoff, 1995).
 - SVR ε -insensitive (dual) linear or quadratic function: **on-line SVR**.

Widrow-Hoff

(Widrow and Hoff, 1988)

WIDROWHOFF(\mathbf{w}_0)

```
1   $\mathbf{w}_1 \leftarrow \mathbf{w}_0$     ▷ typically  $\mathbf{w}_0 = \mathbf{0}$ 
2  for  $t \leftarrow 1$  to  $T$  do
3      RECEIVE( $\mathbf{x}_t$ )
4       $\hat{y}_t \leftarrow \mathbf{w}_t \cdot \mathbf{x}_t$ 
5      RECEIVE( $y_t$ )
6       $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + 2\eta(\mathbf{w}_t \cdot \mathbf{x}_t - y_t)\mathbf{x}_t$     ▷  $\eta > 0$ 
7  return  $\mathbf{w}_{T+1}$ 
```

Dual On-Line SVR

(Vijayakumar and Wu, 1988)

($b=0$)

DUALSVR()

1 $\alpha \leftarrow \mathbf{0}$

2 $\alpha' \leftarrow \mathbf{0}$

3 **for** $t \leftarrow 1$ **to** T **do**

4 RECEIVE(x_t)

5 $\hat{y}_t \leftarrow \sum_{s=1}^T (\alpha'_s - \alpha_s) K(x_s, x_t)$

6 RECEIVE(y_t)

7 $\alpha'_{t+1} \leftarrow \alpha'_t + \min(\max(\eta(y_t - \hat{y}_t - \epsilon), -\alpha'_t), C - \alpha'_t)$

8 $\alpha_{t+1} \leftarrow \alpha_t + \min(\max(\eta(\hat{y}_t - y_t - \epsilon), -\alpha_t), C - \alpha_t)$

9 **return** $\sum_{t=1}^T \alpha_t K(x_t, \cdot)$

This Lecture

- Generalization bounds
- Linear regression
- Kernel ridge regression
- Support vector regression
- Lasso

LASSO

(Tibshirani, 1996)

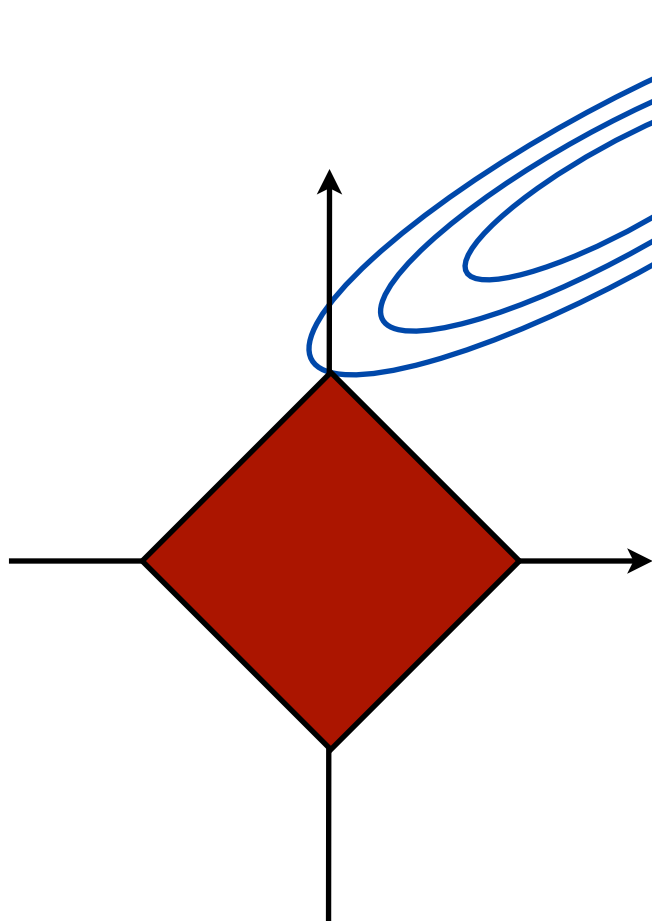
- **Optimization problem:** ‘least absolute shrinkage and selection operator’.

$$\min_{\mathbf{w}} F(\mathbf{w}, b) = \lambda \|\mathbf{w}\|_1 + \sum_{i=1}^m (\mathbf{w} \cdot \mathbf{x}_i + b - y_i)^2,$$

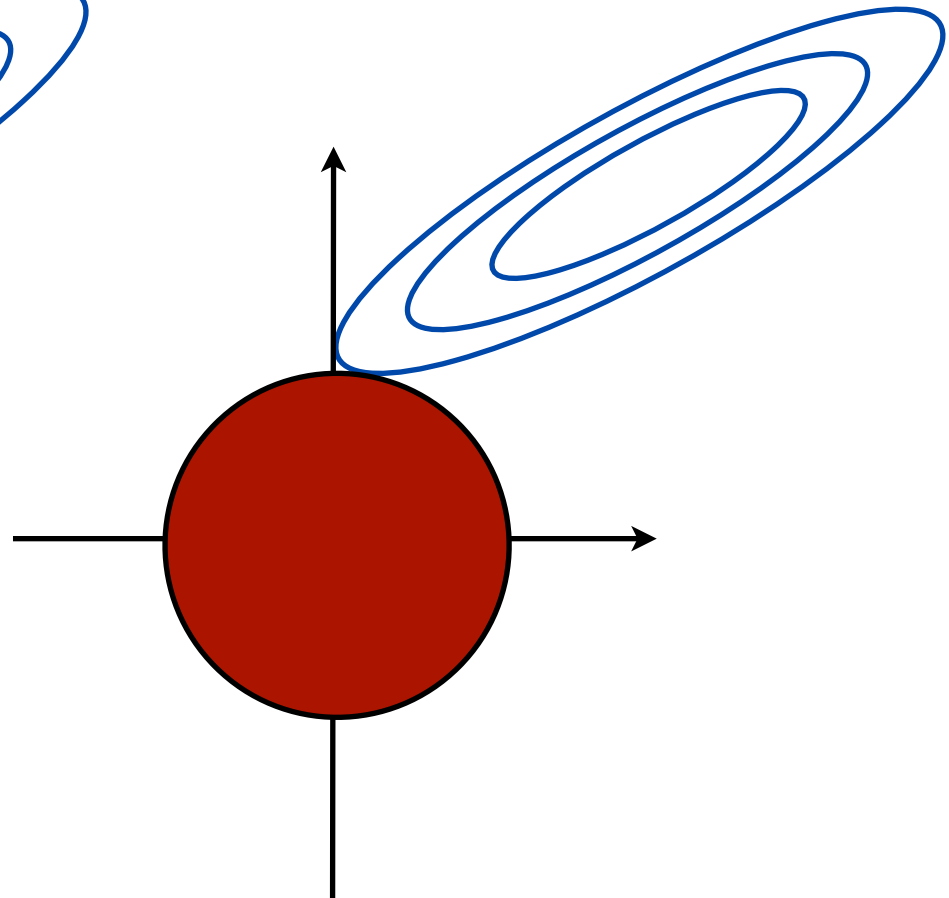
where $\lambda \geq 0$ is a (regularization) parameter.

- **Solution:** equiv. convex quadratic program (QP).
 - general: standard QP solvers.
 - specific algorithm: LARS (least angle regression procedure), entire path of solutions.

Sparsity of L1 regularization



L1 regularization



L2 regularization

Sparsity Guarantee

- Rademacher complexity of L1-norm bounded linear hypotheses:

$$\begin{aligned}\widehat{\mathfrak{R}}_S(H) &= \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\|\mathbf{w}\|_1 \leq \Lambda_1} \sum_{i=1}^m \sigma_i \mathbf{w} \cdot \mathbf{x}_i \right] \\ &= \frac{\Lambda_1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\left\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \right\|_{\infty} \right] && \text{(by definition of the dual norm)} \\ &= \frac{\Lambda_1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\max_{j \in [1, N]} \left| \sum_{i=1}^m \sigma_i x_{ij} \right| \right] && \text{(by definition of } \|\cdot\|_{\infty}\text{)} \\ &= \frac{\Lambda_1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\max_{j \in [1, N]} \max_{s \in \{-1, +1\}} s \sum_{i=1}^m \sigma_i x_{ij} \right] && \text{(by definition of } |\cdot| \text{)} \\ &= \frac{\Lambda_1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\mathbf{z} \in A} \sum_{i=1}^m \sigma_i z_i \right] \leq r_{\infty} \Lambda_1 \sqrt{\frac{2 \log(2N)}{m}}. && \text{(Massart's lemma)}\end{aligned}$$

Notes

■ Advantages:

- theoretical guarantees.
- sparse solution.
- feature selection.

■ Drawbacks:

- no natural use of kernels.
- no closed-form solution (not necessary, but can be convenient for theoretical analysis).

Regression

- Many other families of algorithms: including
 - neural networks.
 - decision trees (see multi-class lecture).
 - boosting trees for regression.

References

- Corinna Cortes, Mehryar Mohri, and Jason Weston. A General Regression Framework for Learning String-to-String Mappings. In *Predicting Structured Data*. The MIT Press, 2007.
- Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2002). Least angle regression. *Annals of Statistics* 2003.
- Arthur Hoerl and Robert Kennard. Ridge Regression: biased estimation of nonorthogonal problems. *Technometrics*, 12:55-67, 1970.
- C. Saunders and A. Gammerman and V.Vovk, Ridge Regression Learning Algorithm in Dual Variables, In *ICML '98*, pages 515--521, 1998.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society*, pages B. 58:267-288, 1996.
- David Pollard. *Convergence of Stochastic Processes*. Springer, New York, 1984.
- David Pollard. *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics, 1990.

References

- Sethu Vijayakumar and Si Wu. Sequential support vector classifiers and regression. In Proceedings of the International Conference on Soft Computing (SOCO'99), 1999.
- Vladimir N.Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, Basederlin, 1982.
- Vladimir N.Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
- Vladimir N.Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.
- Bernard Widrow and Ted Hoff. Adaptive Switching Circuits. *Neurocomputing: foundations of research*, pages 123-134, MIT Press, 1988.