Introduction to Machine Learning
Lecture 15

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

Regression

Regression Problem

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

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

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

- **Loss function**: $L\colon Y \times Y \to \mathbb{R}_+$ 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 $h\colon X \to \mathbb{R}$ in H with small generalization error with respect to target f

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

Notes

- Empirical error:

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

- 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.

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 $1-\delta$,

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

- Proof: By the union bound,

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

By Hoeffding, since $L\big(h(x), f(x)\big) \in [0, M]$,

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

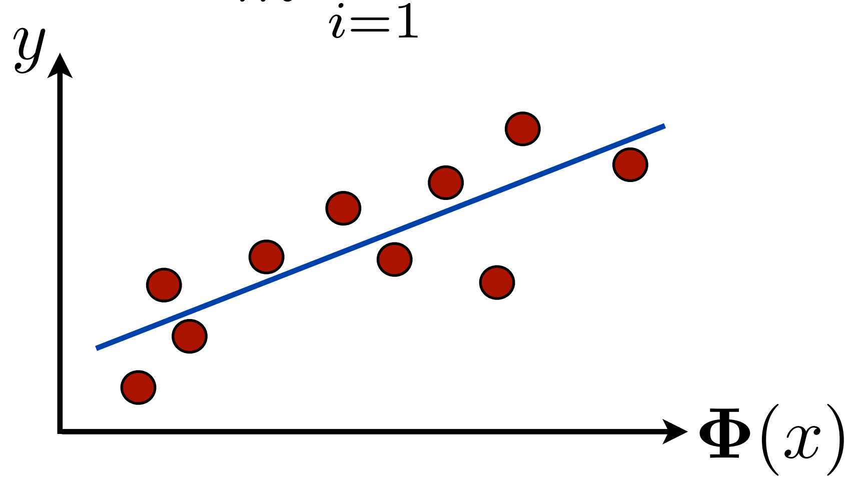
Linear Regression

- Feature mapping $\mathbf{\Phi}: X \to \mathbb{R}^N$.
- Hypothesis set: linear functions.

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

- Optimization problem: empirical risk minimization.

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



Linear Regression - Solution

- Rewrite objective function as $F(\mathbf{W}) = \frac{1}{m}\|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|^2$,
  $\mathbf{X} = \begin{bmatrix} \Phi(x_1) \ldots \Phi(x_m) \\ 1 \quad \ldots \quad 1 \end{bmatrix}$

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

- 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, no regularization.
  - For output labels in $\mathbb{R}^p$, $p > 1$, solve p distinct linear regression problems.

Ridge Regression

(Hoerl and Kennard, 1970)

- Optimization problem:

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

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 diferentiable 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:

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

Ridge Regression - Equivalent Formulations

- Optimization problem:

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

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

- Optimization problem:

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

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

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

Ridge Regression Equations

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

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

- KKT conditions:

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

$$\nabla_{\xi_i} L = 2\xi_i - \alpha'_i = 0 \iff \boxed{\xi_i = \alpha'_i/2.}$$

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

Moving to The Dual

- Plugging in the expression of $\mathbf{w}$ and ξ_is gives

$$L = \sum_{i=1}^{m} \frac{{\alpha'_i}^2}{4} + \sum_{i=1}^{m} \alpha'_i y_i - \sum_{i=1}^{m} \frac{{\alpha'_i}^2}{2} - \frac{1}{2\lambda} \sum_{i,j=1}^{m} \alpha'_i \alpha'_j \mathbf{\Phi}(x_i)^\top \mathbf{\Phi}(x_j) + \lambda \Big(\frac{1}{4\lambda^2} \| \sum_{i=1}^{m} \alpha'_i \mathbf{\Phi}(x_i) \|^2 - \Lambda^2 \Big).$$

- Thus,

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

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

RR - Dual Optimization Problem

- Optimization problem:

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

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

- Solution:

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

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

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:** $\boldsymbol{\alpha}$ such that

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

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

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:

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

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

- Solution:

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

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

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.

Support Vector Regression

(Vapnik, 1995)

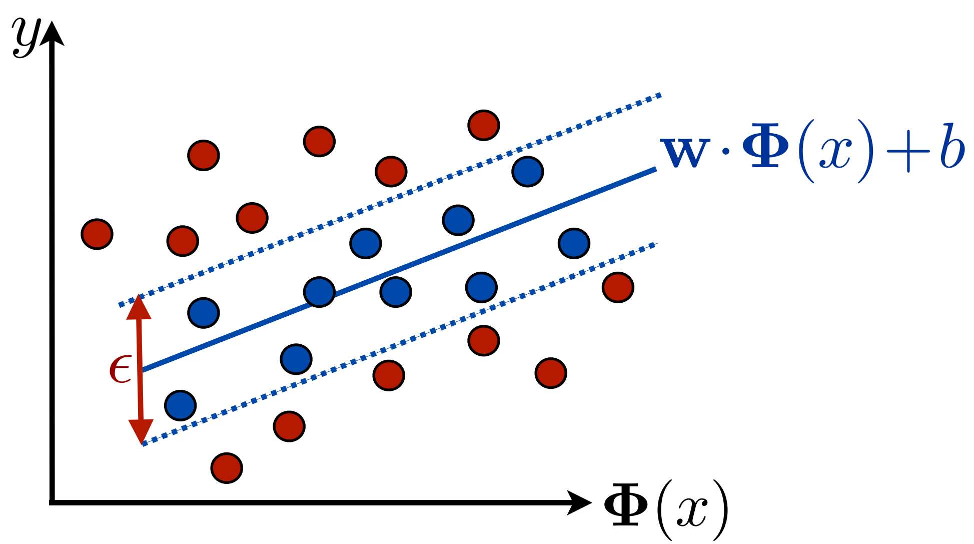
- Hypothesis set:

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

- Loss function: ϵ-insensitive loss.

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

Fit 'tube' with width ϵ to data.



Support Vector Regression (SVR)

(Vapnik, 1995)

- Optimization problem: similar to that of SVM.

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

- Equivalent formulation:

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

$$\begin{aligned}\text{subject to } (\mathbf{w}\cdot\mathbf{\Phi}(x_i)+b) - y_i &\leq \epsilon + \xi_i\\ y_i - (\mathbf{w}\cdot\mathbf{\Phi}(x_i)+b) &\leq \epsilon + \xi_i'\\ \xi_i \geq 0, \xi_i' &\geq 0.\end{aligned}$$

SVR - Dual Optimization Problem

- Optimization problem:

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

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

- Solution:

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

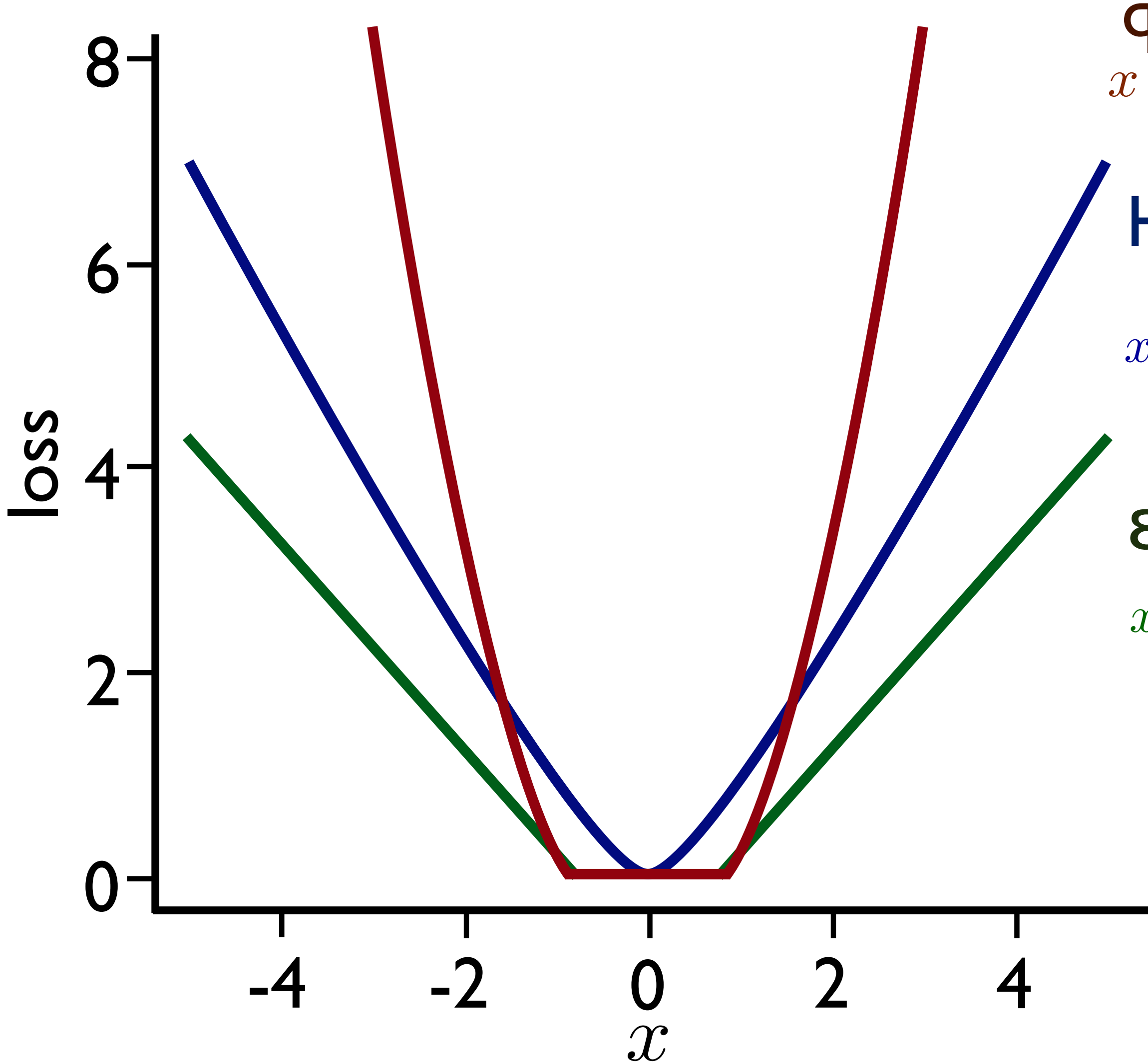
with $b = \begin{cases} -\sum_{i=1}^m (\alpha'_j - \alpha_j) K(x_j, x_i) + y_i + \epsilon & \text{when } 0 < \alpha_i < C \\ -\sum_{i=1}^m (\alpha'_j - \alpha_j) K(x_j, x_i) + y_i - \epsilon & \text{when } 0 < \alpha'_i < C. \end{cases}$

- Support vectors: points strictly outside the tube.

Notes

- Advantages:
  - strong theoretical guarantees.
  - sparser solution.
  - use of kernels.
- Disadvantages:
  - selection of two parameters: C and ϵ. Heuristics:
    - search C near maximum y, ϵ near average difference of ys, measure of no. of SVs.
  - large matrices: low-rank approximations of kernel matrix.

Alternative Loss Functions



quadratic ε-insensitive

$x \mapsto \max(0, |x| - \epsilon)^2$

Huber

$$x \mapsto \begin{cases} x^2 & \text{if } |x| \leq c \\ 2c|x| - c^2 & \text{otherwise.} \end{cases}$$

ε-insensitive

$x \mapsto \max(0, |x| - \epsilon)$

SVR - Quadratic Loss

- Optimization problem:

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

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

- 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.

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)

$\textsc{WidrowHoff}(\mathbf{w}_0)$

$\mathbf{w}_1 \leftarrow \mathbf{w}_0 \qquad \triangleright \text{typically } \mathbf{w}_0 = \mathbf{0}$
$\textbf{for } t \leftarrow 1 \textbf{ to } T \textbf{ do}$
$\quad \textsc{Receive}(\mathbf{x}_t)$
$\quad \widehat{y}_t \leftarrow \mathbf{w}_t \cdot \mathbf{x}_t$
$\quad \textsc{Receive}(y_t)$
$\quad \mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + 2\eta(\mathbf{w}_t \cdot \mathbf{x}_t - y_t)\mathbf{x}_t \qquad \triangleright\, \eta > 0$
$\textbf{return } \mathbf{w}_{T+1}$

Dual On-Line SVR

(Vijayakumar and Wu, 1988)

$(b = 0)$

$\textsc{DualSVR}()$

$\boldsymbol{\alpha} \leftarrow \mathbf{0}$
$\boldsymbol{\alpha}' \leftarrow \mathbf{0}$
for $t \leftarrow 1$ **to** T **do**
    $\textsc{Receive}(x_t)$
    $\widehat{y}_t \leftarrow \sum_{s=1}^{T} (\alpha'_s - \alpha_s) K(x_s, x_t)$
    $\textsc{Receive}(y_t)$
    $\alpha'_{t+1} \leftarrow \alpha'_t + \min(\max(\eta(y_t - \widehat{y}_t - \epsilon), -\alpha'_t), C - \alpha'_t)$
    $\alpha_{t+1} \leftarrow \alpha_t + \min(\max(\eta(\widehat{y}_t - y_t - \epsilon), -\alpha_t), C - \alpha_t)$
return $\sum_{t=1}^{T} \alpha_t K(x_t, \cdot)$

LASSO

(Tibshirani, 1996)

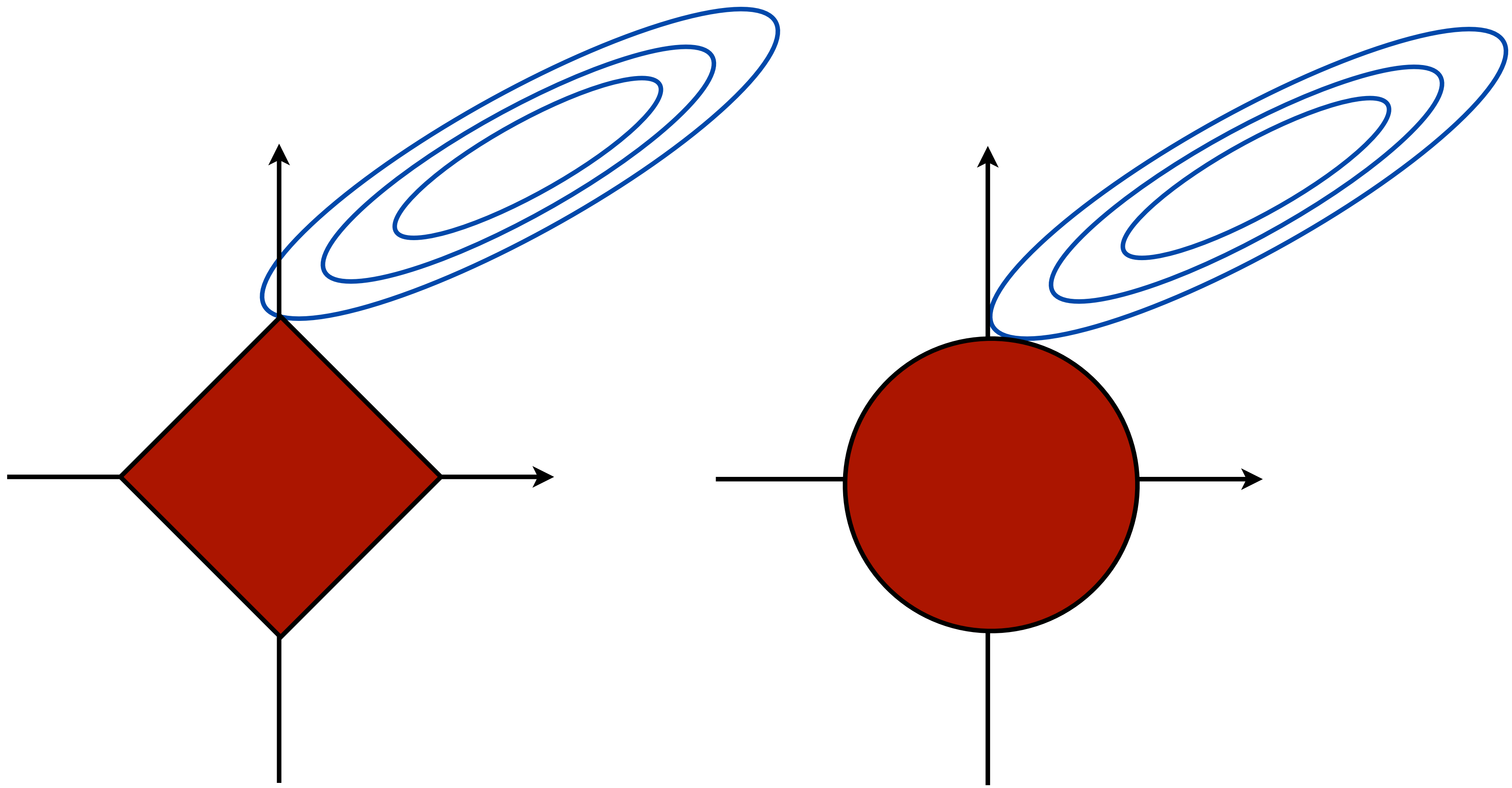
- Optimization problem: 'least absolute shrinkage and selection operator'.

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

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

- Solution: 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

Notes

- Advantages:
  - theoretical guarantees.
  - sparse solution.
  - feature selection.
- Drawbacks:
  - no natural use of kernels.
  - no closed-form solution (not necessary, but can be nice for theoretical analysis).

Regression

- Many other families of algorithms: including
  - neural networks.
  - decision trees.
  - 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.