Pattern Recognition (PR)

Prof. Dr.-Ing. Andreas Maier
Pattern Recognition Lab (CS 5), Friedrich-Alexander-Universität Erlangen-Nürnberg
Winter Term 2020/21

This is a printable version of the slides of the lecture

Pattern Recognition (PR)
Winter term 2020/21
Friedrich-Alexander University of Erlangen-Nuremberg.

Erlangen, January 8, 2021
Prof. Dr.-Ing. Andreas Maier

Norms and Norm Dependent Linear Regression

Motivation

- Different norms and similarity measures play an important role in machine learning and pattern recognition.
- In this chapter we summarize important definitions and facts on norms.
- We consider the problem of linear regression for different norms.
- We will briefly look into associated optimization problems.

Inner Product

Definition

The *inner product of vectors* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ is defined by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{d} x_i y_i \quad .$$

Example

The *Euclidean norm* (L_2-norm) can be written in terms of an inner product:

$$\|\boldsymbol{x}\|_2 = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} = \sqrt{\boldsymbol{x}^T \boldsymbol{x}} = \sqrt{\sum_{i=1}^{d} x_i^2} \quad .$$

Inner Product (cont.)

Definition

The *inner product of matrices* $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m \times n}$ is defined by

$$\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \mathrm{tr}(\boldsymbol{X}^T \boldsymbol{Y}) = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{i,j} y_{i,j} \quad .$$

Example

The *Frobenius norm* can be written in terms of an inner product:

$$\|\boldsymbol{X}\|_F = \sqrt{\langle \boldsymbol{X}, \boldsymbol{X} \rangle} = \sqrt{\mathrm{tr}(\boldsymbol{X}^T \boldsymbol{X})} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{i,j}^2} \quad .$$

Norms

Definition

The function $\|\cdot\|$ is called a *norm* if it

1. is nonnegative: $\forall \boldsymbol{x}: \ \|\boldsymbol{x}\| \geq 0$
2. is definite: $\|\boldsymbol{x}\| = 0$ only if $\boldsymbol{x} = 0$
3. is homogeneous: $\|a\boldsymbol{x}\| = |a| \cdot \|\boldsymbol{x}\|$ where $a \in \mathbb{R}$
4. fulfills the triangle inequality:

$$\forall \boldsymbol{x}, \boldsymbol{y}: \ \|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$$

Norms (cont.)

- The L_0-norm of a d-dimensional vector denotes the number of non-zero entries. Despite its name, the L_0-norm is not a norm because it is not homogeneous.
- The L_p-norm ($p \geq 1$) of a d-dimensional vector is defined as

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}$$

Norms (cont.)

- L_1-norm: sum of absolute values

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{d} |x_i|$$

- L_2-norm: sum of squared values

$$\|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{d} x_i^2 \right)^{\frac{1}{2}}$$

- L_∞-norm: maximum norm

$$\|\boldsymbol{x}\|_\infty = \lim_{p \to \infty} \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}} = \max_i \{ |x_i| \; ; \; i = 1, 2, \ldots, d \}$$

Norms (cont.)

Definition

Let $\boldsymbol{P}$ be a symmetric positive definite matrix.
The *quadratic $L_{\boldsymbol{P}}$-norm* is defined by

$$\|\boldsymbol{x}\|_{\boldsymbol{P}} = \sqrt{\boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x}} = \sqrt{(\boldsymbol{P}^{\frac{1}{2}} \boldsymbol{x})^T \boldsymbol{P}^{\frac{1}{2}} \boldsymbol{x}} = \|\boldsymbol{P}^{\frac{1}{2}} \boldsymbol{x}\|_2$$

Norms (cont.)

Note:

- The L_2-norm is the same as the quadratic L_1-norm.
- The Mahalanobis distance between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ based on the covariance matrix $\boldsymbol{\Sigma}$ is given by the quadratic $L_{\boldsymbol{\Sigma}^{-1}}$-norm:

$$\|\boldsymbol{x}-\boldsymbol{y}\|_{\boldsymbol{\Sigma}^{-1}} = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y})}$$

- A norm is a measure for the length of a vector. It can also be used to measure the distance between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\text{dist}(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|$$

Norms (cont.)

Norms of matrices can be defined by norms of vectors.

Definition

Let $\|.\|_p$ and $\|.\|_q$ be norms for vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$.
The *operator norm* of a matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ is defined by

$$\|\boldsymbol{X}\|_{p,q} = \sup\{\|\boldsymbol{X}\boldsymbol{u}\|_p;\ \|\boldsymbol{u}\|_q \leq 1\}$$

Example

If $p = q = 2$, i. e. we use the L_2-norm twice, the operator norm of $\boldsymbol{X}$ results in the maximum singular value:

$$\|\boldsymbol{X}\|_{2,2} = \|\boldsymbol{X}\|_2 = \sigma_{\max}(\boldsymbol{X}) = \sqrt{\lambda_{\max}(\boldsymbol{X}^T\boldsymbol{X})}$$

Unit Balls

Definition

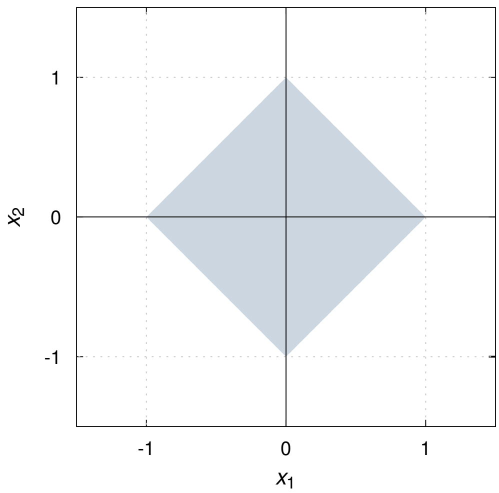
The set

$$\mathscr{B} = \{\boldsymbol{x}; \|\boldsymbol{x}\| \leq 1\}$$

of all vectors $\boldsymbol{x}$ of length less or equal to one according to the norm $\|.\|$ is called the *unit ball*.

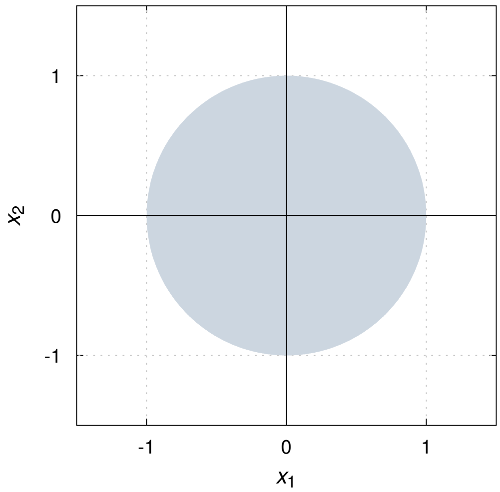
Unit Balls (cont.)

The unit ball for the L_1-norm:



Unit Balls (cont.)

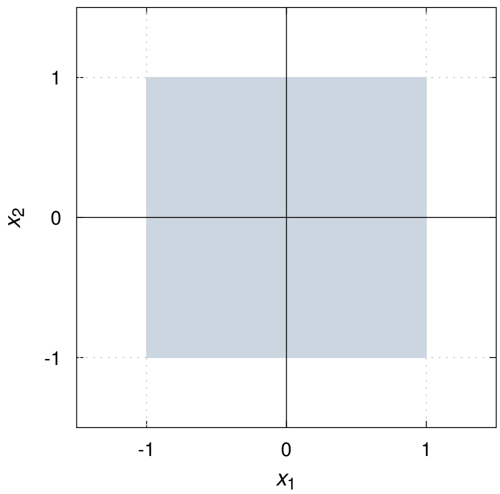
The unit ball for the L_2-norm:



Pattern Recognition Lab
FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
FACULTY OF ENGINEERING

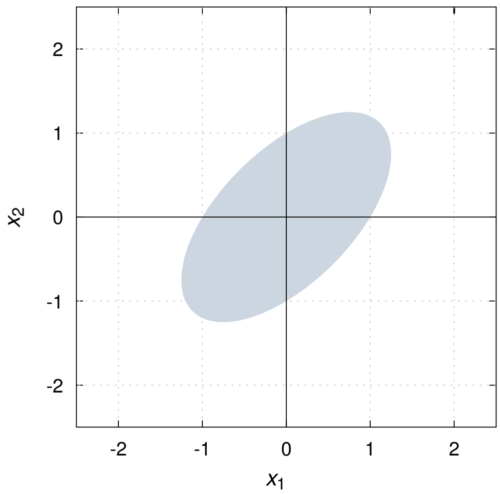
Unit Balls (cont.)

The unit ball for the L_∞-norm:



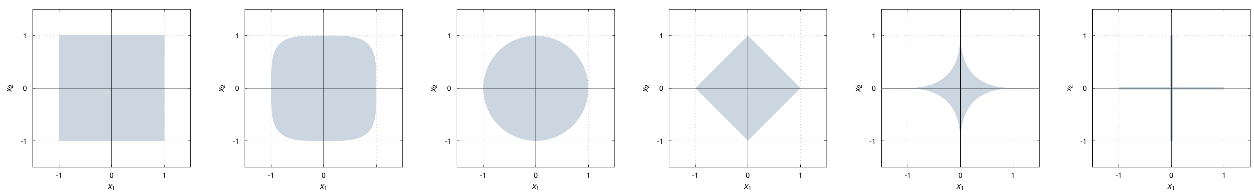
Unit Balls (cont.)

The unit ball for the $L_{\boldsymbol{P}}$-norm:



Unit Balls (cont.)

Summary: unit balls for the L_∞-, L_4-, L_2-, L_1-, $L_{0.5}$- and L_0-norm



The $L_{0.5}$- and the L_0-norm are not norms

Pattern Recognition Lab

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Next Time in Pattern Recognition

Norm Dependent Linear Regression

In pattern recognition and pattern analysis (as in many other fields) one of the most important norm dependent linear regression problems is:

$$\text{minimize} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|$$

or alternatively

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|$$

Norm Dependent Linear Regression (cont.)

- Different norms will lead to different results.
- The estimation error $\varepsilon \in \mathbb{R}$ is defined by $\varepsilon = \|\boldsymbol{x}^* - \hat{\boldsymbol{x}}\|$, where $\boldsymbol{x}^*$ denotes the correct value.
- The residual $\boldsymbol{r} = (r_1, r_2, \ldots, r_m)^T$ is defined by $\boldsymbol{r} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$.
- If $\boldsymbol{b}$ is in the range of $\boldsymbol{A}$, the residual will be the zero vector.

Least-Squares Linear Regression

Minimization of the residual using the L_2-norm:

$$
\begin{aligned}
\hat{\boldsymbol{x}} &= \operatorname*{argmin}_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2 \\
&= \operatorname*{argmin}_{\boldsymbol{x}} \sum_{i=1}^{m} r_i^2 \\
&= \operatorname*{argmin}_{\boldsymbol{x}} (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \\
&= \operatorname*{argmin}_{\boldsymbol{x}} \left(\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{b} - \boldsymbol{b}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{b}\right) \\
&= \operatorname*{argmin}_{\boldsymbol{x}} \left(\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{b}\right)
\end{aligned}
$$

Least-Squares Linear Regression (cont.)

The partial derivatives are:

$$\frac{\partial}{\partial \boldsymbol{x}} \left(\boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x} - 2 \boldsymbol{b}^T \boldsymbol{A} \boldsymbol{x} + \boldsymbol{b}^T \boldsymbol{b} \right) = 2 \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x} - 2 \boldsymbol{A}^T \boldsymbol{b} = 0$$

Using the partial derivatives we get a closed form solution for the L_2-norm:

$$\hat{\boldsymbol{x}} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{b}$$

if the columns of $\boldsymbol{A}$ are mutually independent.

Chebyshev Linear Regression

Minimization of the residual using the L_∞-norm:

$$\text{minimize} \quad \left\{ \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_\infty = \max \{ |r_1|, |r_2|, \ldots, |r_m| \} \right\}$$

This optimization problem can be rewritten in terms of a LP-problem:

$$\begin{array}{ll} \text{minimize} & r \\ \text{subject to} & -r \cdot 1 \preceq \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \preceq r \cdot 1 \end{array}$$

where $r \in \mathbb{R}$ and $1 \in \{1\}^m$.

Sum of Absolute Residuals

Minimization of the residual using the L_1-norm:

$$\text{minimize} \quad \left\{ \|\boldsymbol{Ax} - \boldsymbol{b}\|_1 = \sum_{i=1}^{m} |r_i| \right\}$$

This optimization problem can be rewritten in terms of a LP-problem:

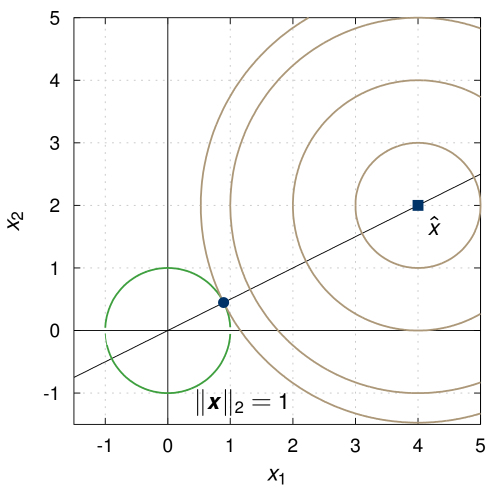
$$\begin{array}{ll} \text{minimize} & 1^T \boldsymbol{r} \\ \text{subject to} & -\boldsymbol{r} \preceq \boldsymbol{Ax} - \boldsymbol{b} \preceq \boldsymbol{r} \end{array}$$

where $\boldsymbol{r} \in \mathbb{R}^m$ and $1 \in \{1\}^m$.

Ridge Regression and Unit Balls

Ridge regression is defined via the optimization problem

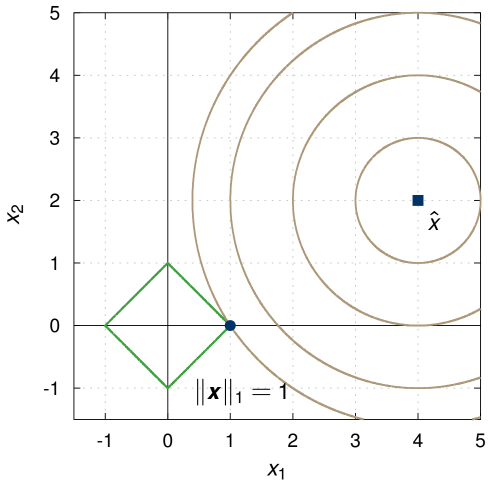
$$\text{minimize} \quad \|A\boldsymbol{x} - \boldsymbol{b}\|_2 + \lambda \cdot \|\boldsymbol{x}\|_2$$



Lasso and Unit Balls

The lasso (Tibshirani 1996) is defined via the optimization problem

$$\text{minimize} \quad \|\boldsymbol{Ax} - \boldsymbol{b}\|_2 + \lambda \cdot \|\boldsymbol{x}\|_1$$



Compressed Sensing

- In the previous chapter we motivated regularized linear regression.
- Assume we have fewer measurements than required to estimate the parameter vector $\boldsymbol{x}$.
- Solution of the underdetermined case required.
- We call a vector S-sparse if its support, i. e. the number of non-zero entries, is less or equal to S
- The vector $\boldsymbol{x}$ can be recovered mostly always by solving the convex optimization problem (quadratic programming):

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \boldsymbol{Ax} = \boldsymbol{b}. \end{array}$$

Penalty Function

Motivated by the discussion of different norms, we now introduce and study penalty functions.

Definition

The *penalty function approximation problem* is defined as follows:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m} \phi(r_i) \\ \text{subject to} & \boldsymbol{r} = (r_1, r_2, \ldots, r_m)^T = \boldsymbol{Ax} - \boldsymbol{b}, \end{array}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the penalty function for the components of the residual vector.

Penalty Function (cont.)

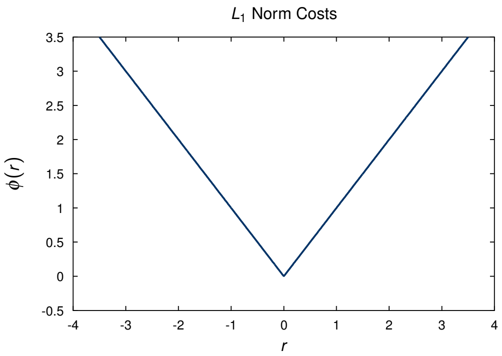
Note:

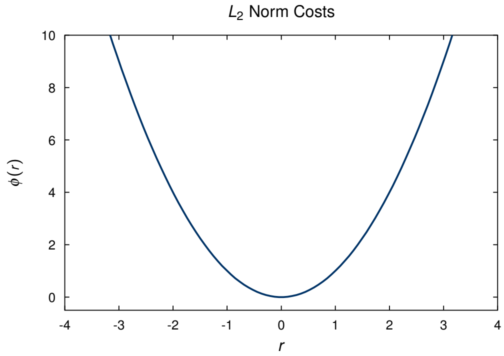
- The penalty function ϕ assigns costs to residuals.
- If ϕ is a convex function, the penalty function approximation problem is a convex optimization problem.

Penalty Function (cont.)

Penalty functions of the L_1-, L_2-norms:

$$\phi_{L_1}(r) = |r|; \qquad \phi_{L_2}(r) = r^2$$



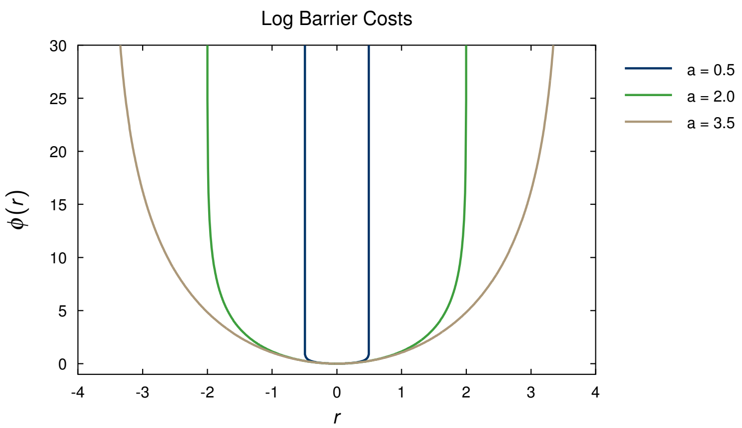


- In L_1 small deviations are weighted higher than using L_2.
- In L_1 large deviations are weighted lower than using L_2.

Penalty Function (cont.)

Log barrier function

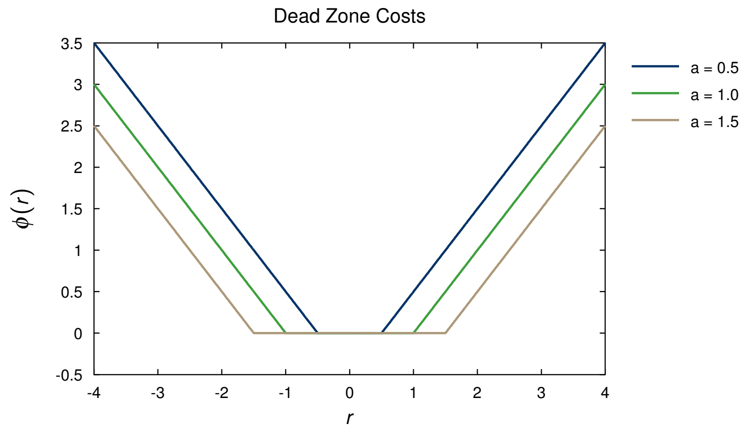
$$\phi_{\text{barrier}}(r) = \begin{cases} -a^2 \log\left(1 - \left(\frac{r}{a}\right)^2\right), & \text{if } \quad |r| < a \\ \infty, & \text{otherwise} \end{cases}$$



Penalty Function (cont.)

Dead zone linear penalty function

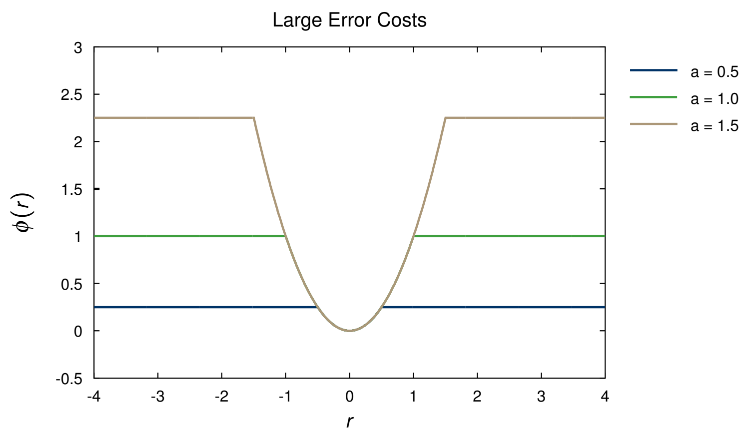
$$\phi_{\mathrm{dz}}(r) = \begin{cases} 0, & \text{if} \quad |r| \le a \\ |r| - a, & \text{otherwise} \end{cases}$$



Penalty Function (cont.)

Large error penalty function

$$\phi_{\mathrm{e}}(r) = \begin{cases} r^2, & \text{if} \quad |r| \le a \\ a^2, & \text{otherwise} \end{cases}$$



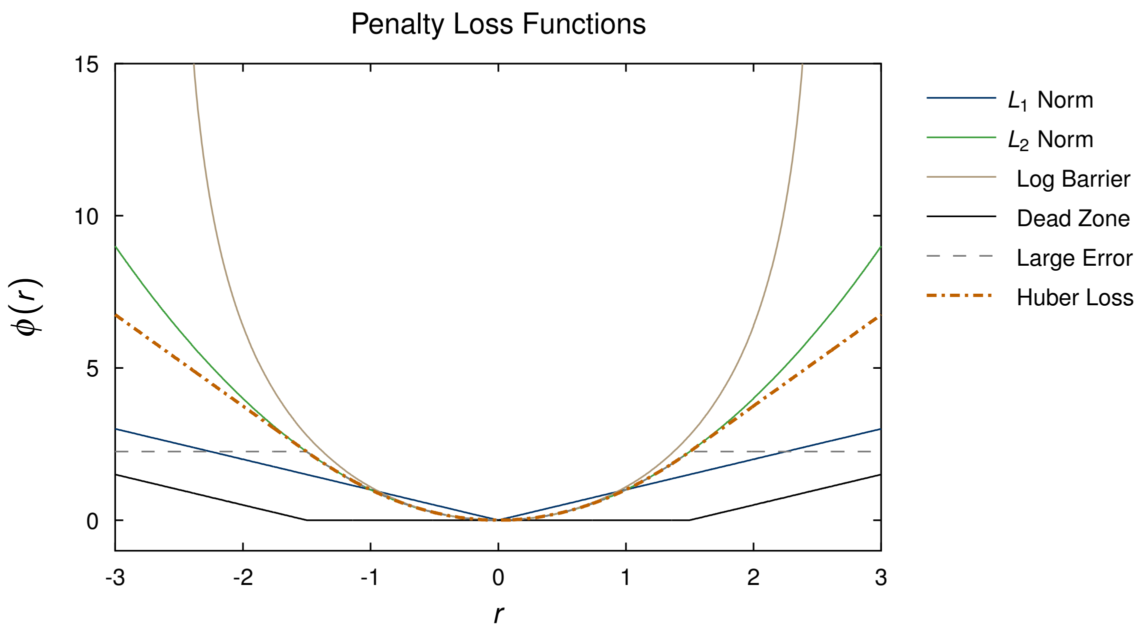
Penalty Function (cont.)

Huber function

$$\phi_{\text{Huber}}(r) = \begin{cases} r^2, & \text{if } |r| \leq a \\ a \cdot (2|r| - a), & \text{otherwise} \end{cases}$$

Pattern Recognition Lab
FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
FACULTY OF ENGINEERING

Penalty Functions (cont.)



Lessons Learned

- We have considered vector and matrix norms in more detail.
- Important vector norms: L_1, L_2, L_∞, and $L_{\boldsymbol{p}}$.
- Unit balls
- Linear regression for different norms: range from closed form solution to LP-problem.
- Regularized linear regression: range from closed form solution through QP-problem up to combinatorial optimization.
- We need to know the basics of algorithms for unconstrained and constrained optimization as well as convex optimization.

Pattern Recognition Lab

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Next Time in

Pattern Recognition



Further Readings

- G. Golub, C. F. Van Loan:
  Matrix Computations, 3rd Edition,
  The Johns Hopkins University Press, Baltimore, 1996.
- Lloyd N. Trefethen, David Bau III:
  Numerical Linear Algebra,
  SIAM, Philadelphia, 1997.
- S. Boyd, L. Vandenberghe:
  Convex Optimization,
  Cambridge University Press, 2004.
  ☞ http://www.stanford.edu/~boyd/cvxbook/

Further Readings (cont.)

- Compressed sensing is one of the most recent hot topics in pattern recognition and image processing. An excellent source is:

  http://www.dsp.ece.rice.edu/cs

  or the recent workshop on compressed sensing at Duke University:

  http://people.ee.duke.edu/%7Elcarin/compressive-sensing-workshop.html.

Comprehensive Questions

- What is the difference between the L_p- (p $\geq$ 1) and the $\boldsymbol{L_P}$-norm?

- How do the unit balls look like for L_∞-, L_4-, L_2-, L_1- and L_0-norm?

- What is the benefit of using the L_1- over the L_2-norm for sparse, underdetermined problems?

- What specific property of penalty functions is of special interest and why do we need different penalty functions at all?