Optimization Theory and Algorithms

Instructor: Prof. LIAO, Guocheng （廖国成）

Email: liaogch6@mail.sysu.edu.cn

School of Software Engineering
Sun Yat-sen University

Outline

- Approximation
- Estimation
- Classification

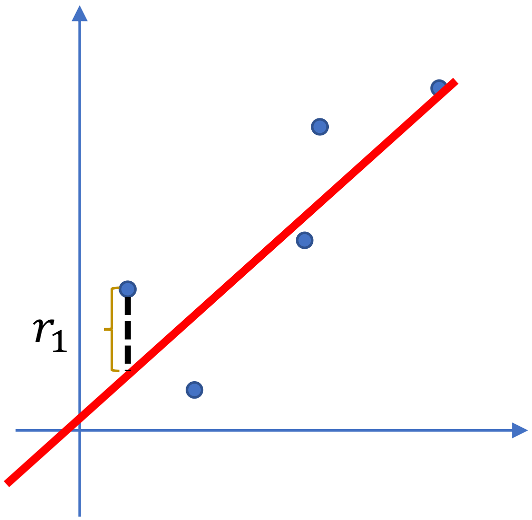
Norm approximation

$$\min ||Ax - b||$$

- $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$
- $r = Ax - b \in \mathbb{R}^m$ is **residual**
- Approximation solution of $Ax \approx b$, in $||\cdot||$
- Convex problem

➢ $b \in \mathcal{R}(A)$: the optimal value is zero; $r = 0$

➢ $b \notin \mathcal{R}(A)$



Interpretation of norm approximation

$$\min ||Ax - b||$$

Approximation interpretation (regression problem)

$$Ax = x_1 a_1 + \cdots + x_n a_n$$

- $a_1, \ldots, a_n \in \mathbb{R}^m$ are columns of A
- To approximate b by a linear combination of the columns of A

Design interpretation

- $x_1, \ldots, x_n$ are design variables (input); Ax is result (output); b is target
- To find the best design that makes the result as closed to the target as possible.

Examples of norm approximation

$$\min ||Ax - b||$$

L2-norm (least-squares approximation)

$$\min ||Ax - b||_2^2 = r_1^2 + r_2^2 + \cdots + r_m^2$$

KKT conditions: $x^\star = (A^T A)^{-1} A^T b$

L∞-norm (minmax approximation)

$$\min \|Ax - b\|_\infty = \max\{|r_1|, |r_2|, \ldots, |r_2|\}$$

Linear programming

$$\min t$$
$$\text{s.t. } -t\mathbf{1} \preccurlyeq Ax - b \preccurlyeq t\mathbf{1}$$

L1-norm (sum of absolute residuals approximation)

$$\min \|Ax - b\|_1 = |r_1| + |r_2| + \cdots + |r_m|$$

Linear programming

$$\min \mathbf{1}^T t$$
$$\text{s.t. } -t \preccurlyeq Ax - b \preccurlyeq t$$

Penalty function approximation

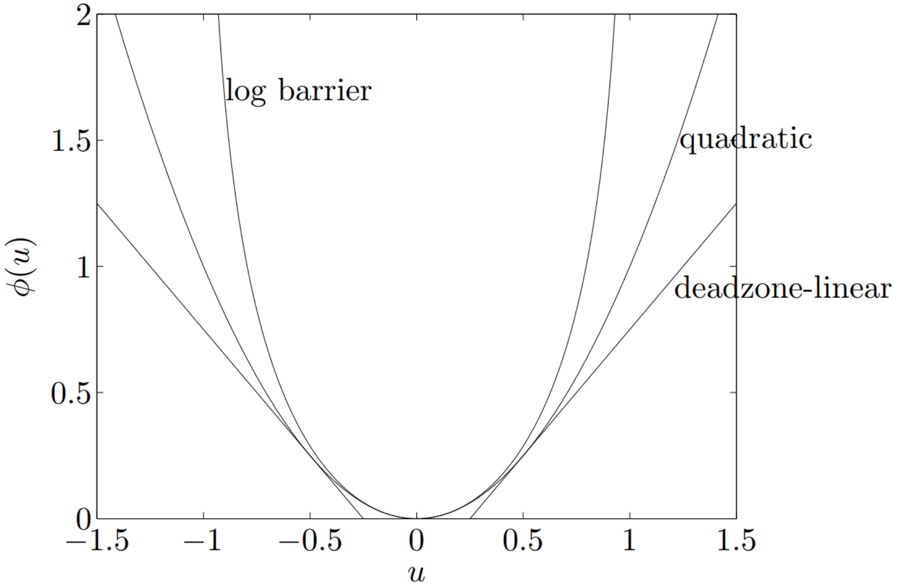
$$\begin{array}{ll} \text{min} & \phi(r_1) + \phi(r_2) + \cdots + \phi(r_m) \\ \text{s.t.} & r = Ax - b \end{array}$$

- $\phi: \mathbb{R} \to \mathbb{R}$: convex penalty function; evaluate a cost or penalty for a residual
- Examples of penalty functions
  - Quadratic penalty function $\phi(u) = u^2$
  - Absolute value penalty function $\phi(u) = |u|$
  - Deadzone-linear penalty function

$$\phi(u) = \begin{cases} 0 & |u| \le a \\ |u| - a & |u| > a. \end{cases}$$

  - Log barrier penalty function

$$\phi(u) = \begin{cases} -a^2 \log(1 - (u/a)^2) & |u| < a \\ \infty & |u| \ge a. \end{cases}$$



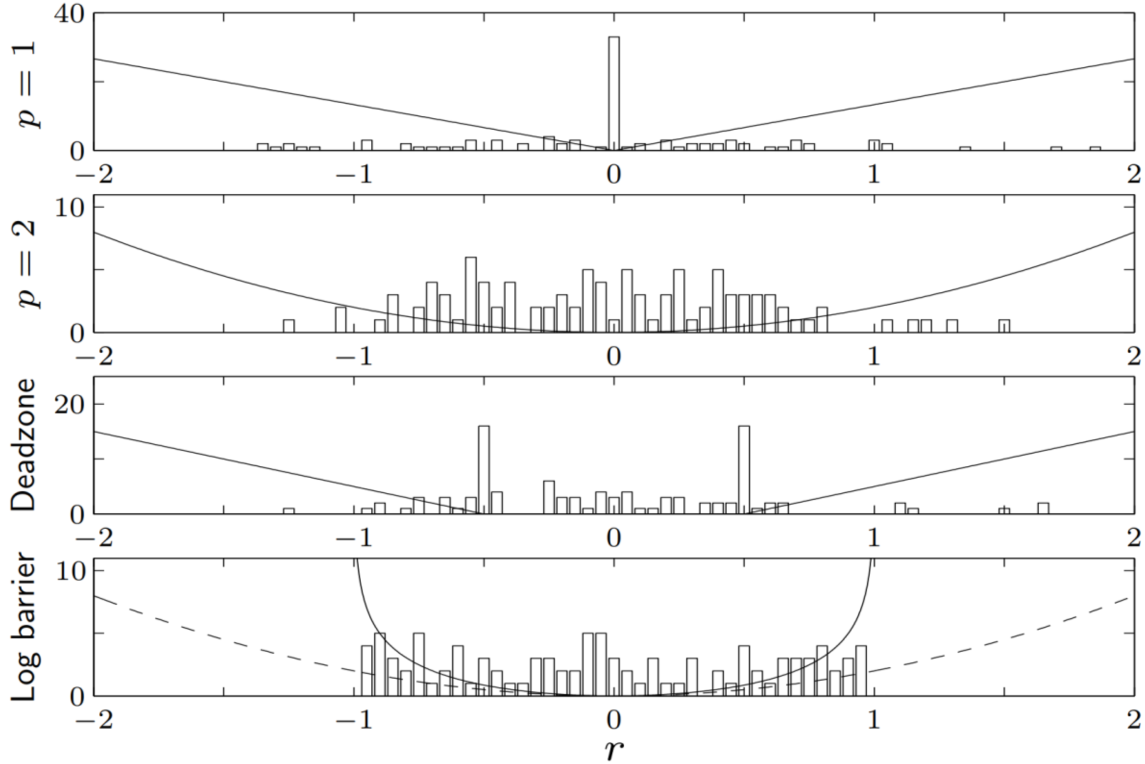
Penalty functions

$\phi(u) = |u|$

$\phi(u) = u^2$

$$\phi(u) = \begin{cases} 0 & |u| \leq a \\ |u| - a & |u| > a. \end{cases}$$

$\phi(u) = -\log(1 - u^2)$



- $\phi(u) = |u|$: more weight on small residuals; less weight on large residuals
- $\phi(u) = u^2$: more weight on large residuals; less weight on small residuals
- Dead-zone: no weight on small residuals; relatively small weight on large residuals
- Log barrier: less weight on small residuals; significant weight on large residuals

Approximation with constraints

Non-negative constraints on variables

$$\begin{aligned} &\min\ ||Ax - b|| \\ &\text{s.t.}\quad x \succcurlyeq 0 \end{aligned}$$

- x is known to be non-negative, e.g., prices, powers, area

Variable bounds

$$\begin{aligned} &\min\ ||Ax - b|| \\ &\text{s.t.}\quad l \preccurlyeq x \preccurlyeq u \end{aligned}$$

- x is known to lie in some bounded intervals.

Probability distribution

$$\begin{aligned} &\min\ ||Ax - b|| \\ &\text{s.t.}\quad x \succcurlyeq 0 \\ &\qquad\ \ \mathbf{1}^T x = 1 \end{aligned}$$

- x is frequency or probability distribution

Regularized approximation

$$\min ||Ax - b|| + \gamma||x||$$

- $\gamma > 0$
- Interpretation: $||x||$ should not be too large
- Trade-off between $||Ax - b||$ and $||x||$

L2-norm regularization $\min ||Ax - b||_2^2 + \gamma||x||_2^2$

- $x = (A^T A + \gamma I)^{-1} A^T b$

L1-norm regularization $\min \|Ax - b\|_1 + \gamma\|x\|_1$

- Solution is sparse (there are many zeros in x)
- Absolute value puts more weight on small x

Maximum likelihood estimation

- Parametric distribution estimation: given some observed values y, to estimate the probability density function $\boxed{p_x(y)}$ with parameter x
- **Maximum likelihood estimation**:

Likelihood function

$$\max_x \log p_x(y)$$

- $\log p_x(y)$ log-likelihood function
- To find the parameter that maximizes the probability that the observed values y are generated

Linear measurements with IID noise

Linear measurement model

$$y_i = a_i^T x + v_i, i = 1, \dots, m$$

- $x \in \mathbb{R}^m$ is a vector of unknow parameters
- a_i is known data
- v_i is noise, with probability density function $p(v)$
- y_i is measurement with density $p_x(y_i) = p(y_i - a_i^T x)$

Maximum likelihood estimation

$$\max_x l(x) = \log \prod_{i=1}^{m} p(y_i - a_i^T x) = \sum_{i=1}^{m} \log(p_x(y_i))$$

Examples of noise

$$\max_x l(x) = \log \prod_{i=1}^{m} p(y_i - a_i^T x) = \sum_{i=1}^{m} \log(p_x(y_i))$$

Gaussian noise $\mathcal{N}(0, \sigma^2)$: $p(v) = (2\pi\sigma^2)^{-1/2} e^{-\frac{v^2}{2\sigma^2}}$

$$l(x) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - a_i^T x)^2 \Rightarrow \min ||Ax - y||_2^2$$

ML estimation with Gaussian noise is equivalent to least-squares approximation

Laplacian noise : $p(v) = (1/2b) e^{-\frac{|v|}{b}}$

$$l(x) = m \log 1/2b - \frac{1}{b}\sum_{i=1}^{m}|y_i - a_i^T x| \Rightarrow \min ||Ax - y||_1$$

ML estimation with Laplacian noise is equivalent to L1-norm approximation

Maximum a posterior probability (MAP) estimation

- Posterior probability = prior probability + likelihood

$$p(x,y) = \underbrace{\boxed{p(x)}}_{\text{prior probability}}\underbrace{\boxed{p(y|x)}}_{\text{conditional probability (likelihood)}}$$

$$\max_x \log p(x) + \log p_x(y)$$

- Comparison between ML estimation and MAP estimation: incorporating extra prior probability of parameter x

Gaussian prior

$$\max_x \log p(x) + \log p_x(y)$$

Gaussian prior $\mathcal{N}(0, \delta^2)$: $\quad p(x) = (2\pi\delta^2)^{-1/2} e^{-\frac{x^2}{2\delta^2}}$

$$\max_x -\frac{1}{2}\log(2\pi\delta^2) - \frac{x^2}{2\delta^2} - \frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - a_i^T x)^2$$

$$\text{min } ||Ax - y||_2^2 + \gamma||x||_2^2$$

MAP estimation with Gaussian prior is equivalent to least-squares approximation with quadratic regularization.

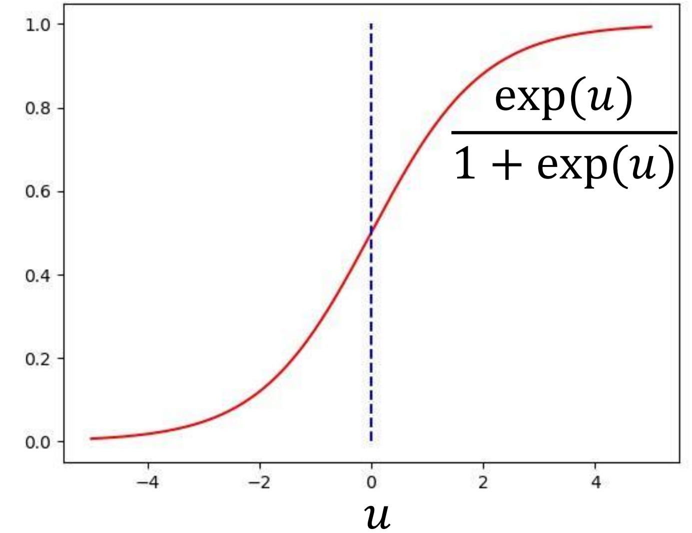
Logistic regression

- Binary classification: to label a sample with feature data x with $y \in \{0,1\}$
- Assume the probability:

$$p_1(x; a, b) = \mathbf{Prob}(y = 1) = \frac{\exp(a^T x + b)}{1 + \exp(a^T x + b)}$$

$$p_0(x; a, b) = \mathbf{Prob}(y = 0) = \frac{1}{1 + \exp(a^T x + b)}$$

- a and b are parameters to be estimated



Log-likelihood function $\quad l(a, b) = \sum_{i=1}^{m} y_i \log p_1(x_i; a, b) + (1 - y_i) \log p_0(x_i; a, b)$

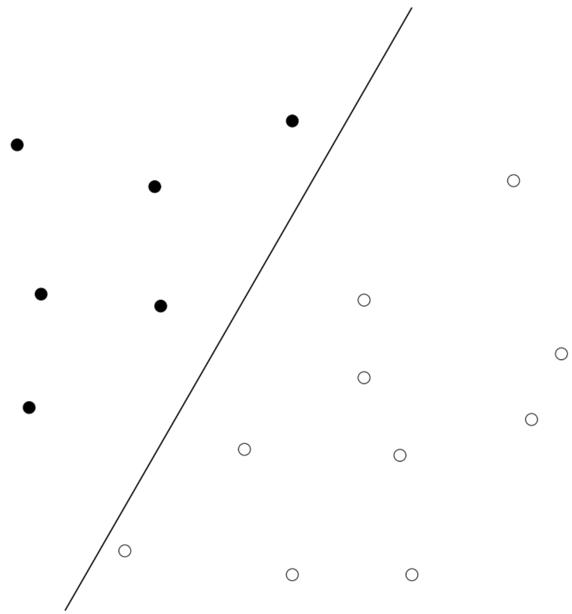
Solved via gradient descent or Newton's method

Linear classification

Separate two sets of points $\{x_1, \ldots, x_N\}$, $\{z_1, \ldots, z_N\}$ by a hyperplane $a^T x + b = 0$

$\{x_1, \ldots, x_N\}$

$a^T x_i + b > 0$



$\{z_1, \ldots, z_N\}$

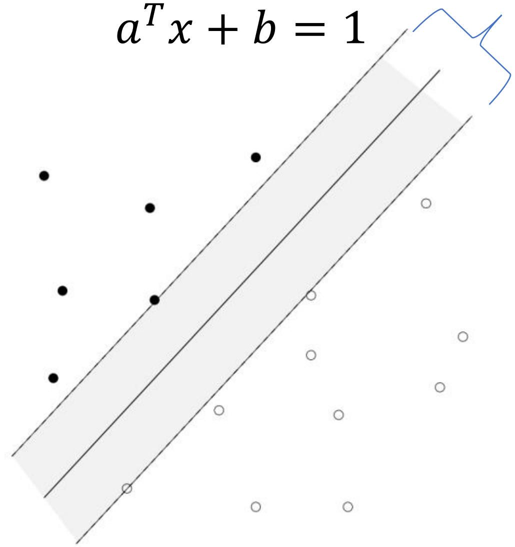
$a^T z_i + b < 0$

To find a and b such that:

$$a^T x_i + b \geq 1$$

$$a^T z_i + b \leq -1$$

Support vector machine



Separate two sets of points by maximum margin

$$\min \frac{1}{2}\|a\|_2$$
$$\text{s.t.} \quad a^T x_i + b \geq 1, i = 1, \dots, M$$
$$a^T z_i + b \leq -1, i = 1, \dots, N$$