Optimization in Machine Learning: Lecture 8

Duality

by Xiaolin Huang xiaolinhuang@sjtu.edu.cn SEIEE 2-429

Institute of Image Processing and Pattern Recognition

http://www.pami.sjtu.edu.cn/

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

目录 Contents

目录 Contents

Lagrangian

- consider the standard form

$$\begin{array}{ll} \min & f_0(x) \\ \mathrm{s.t.} & f_i(x) \leq 0, i = 1, \dots, m \\ & h_i(x) = 0, i = 1, \dots, p \end{array}$$

- its Lagrangian is

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

  - weighted sum of the objective and the constraint functions
  - $L: \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$
  - $\mathrm{dom}\, L: D \times \mathbf{R}^m \times \mathbf{R}^p$

Lagrange dual function

- Lagrange dual function $g: \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$

$$g(\lambda,\nu) = \inf_{x\in D} L(x,\lambda,\nu)$$

$$= \inf_{x\in D}\left(f_0(x) + \sum_{i=1}^{m}\lambda_i f_i(x) + \sum_{i=1}^{p}\nu_i h_i(x)\right)$$

- lower bound: if $\lambda \geq 0$, then $g(\lambda,\nu) \leq f^*$
  - proof : suppose $\tilde{x}$ is feasible, i.e., $f_i(\tilde{x}) \leq 0, h_i(\tilde{x}) = 0$

    then with $\lambda \geq 0$, we have

$$f_0(\tilde{x}) \geq f_0(\tilde{x}) + \sum_{i=1}^{m}\lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p}\nu_i h_i(\tilde{x})$$

$$\geq \inf_{x\in D}\left(f_0(x) + \sum_{i=1}^{m}\lambda_i f_i(x) + \sum_{i=1}^{p}\nu_i h_i(x)\right) = g(\lambda,\nu)$$

Lagrange dual and conjugate function

$$\begin{array}{ll} \min & f_0(x) \\ \text{s.t.} & x = 0 \end{array}$$

- conjugate function $f^*(y) = \sup_x (y^\top x - f(x))$
- Lagrangian: $L(x, \lambda, v) = f_0(x) + v^\top x$
- Lagrange dual function

$$g(v) = \inf_{x \in D} (f_0(x) + v^\top x) = -f_0^*(-v)$$

Lagrange dual and conjugate function

$$\begin{array}{ll} \min & f_0(x) \\ \text{s.t.} & Ax \le b \\ & Cx = d \end{array}$$

- conjugate function $f^*(y) = \sup_x (y^\top x - f(x))$
- Lagrangian: $L(x, \lambda, \nu) = f_0(x) + \lambda^\top (Ax - b) + \nu^\top (Cx - d)$
- Lagrange dual function

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in D} (f_0(x) + \lambda^\top (Ax - b) + \nu^\top (Cx - d)) \\ &= \inf_{x \in D} (f_0(x) + (\lambda^\top A + \nu^\top C)x - \lambda^\top b - \nu^\top d\ ) \\ &= -f_0^*(-A^\top \lambda - C^\top \nu) - \lambda^\top b - \nu^\top d \end{aligned}$$

Dual problem

- $g(\lambda, \nu)$ could give a lower bound for the primary problem

- could we get a better, or even tightest, lower bound?

$$\begin{array}{ll} \max & g(\lambda, \nu) \\ \text{s.t.} & \lambda \geq 0 \end{array}$$

  - it is convex, even the primal problem is not

- LP

$$f(x) = f_0^\top x \rightarrow f^*(y) = \sup_{x \in \mathbf{dom}\, f} (x^\top (y - f_0)) = \begin{cases} 0, & y = f_0 \\ +\infty, & y \neq f_0 \end{cases}$$

$$\boxed{\begin{array}{ll} \min & f_0^\top x \\ \text{s.t.} & Ax \leq b \\ & Cx = d \end{array} \quad \Rightarrow \quad \begin{array}{ll} \max & -\lambda^\top b - \nu^\top d \\ \text{s.t.} & A^\top \lambda + C^\top \nu = f_0 \\ & \lambda \geq 0 \end{array}}$$

$$g(\lambda, \nu) = -f_0^*(-A^\top \lambda - C^\top \nu) - \lambda^\top b - \nu^\top d$$

Dual problem example: SVM

- SVM in primal space

$$\begin{aligned} \min_{x,z} \quad & \frac{1}{2}\|x\|_2^2 + C\sum_i \rho_i \\ \text{s.t.} \quad & b_i(x^\top a_i + z) \geq 1 - \rho_i \\ & \rho_i \geq 0 \end{aligned}$$

- the corresponding Lagrangian

$$L(x, z, \rho; \lambda, \nu) = \frac{1}{2}\|x\|_2^2 + C\sum_i \rho_i$$

$$+ \sum_i \lambda_i(1 - \rho_i - b_i(x^\top a_i + z)) - \nu^T \rho$$

Dual problem example: SVM

- to obtain the $g(\lambda,\nu) = \inf_{x,z,\rho} L(x,z,\rho;\lambda,\nu)$

$$L(x,z,\rho;\lambda,\nu) = \frac{1}{2}\|x\|_2^2 + C\sum_i \rho_i + \sum_i \lambda_i(1-\rho_i - b_i(x^\top a_i + z)) - \nu^T\rho$$

$$\frac{\partial L}{\partial x} = x - \sum_i b_i\lambda_i a_i = 0$$

$$\frac{\partial L}{\partial z} = \sum_i \lambda_i b_i = 0$$

$$\frac{\partial L}{\partial \rho_i} = C - \lambda_i - \nu_i = 0$$

$$\lambda_i \geq 0, \nu_i \geq 0$$

$$\begin{aligned} \min_{\lambda}\ & \textstyle\sum_i \sum_j \lambda_i b_i a_i^\top a_j b_j \lambda_j - \sum_i \lambda_i \\ \text{s.t.}\ & \textstyle\sum_i \lambda_i b_i = 0 \\ & 0 \leq \lambda_i \leq C \end{aligned}$$

Dual problem example: LP

- two products I and II with different profit
- three resources A, B, and C with different inventories

	Product I	Product II	inventory
Resource A	0	5	15
Resource B	6	2	24
Resource C	1	1	5
Profit	**2**	**1**	

$$
\begin{aligned}
\max \quad & 2x_1 + x_2 \\
\text{s.t.} \quad & \phantom{6x_1 + {}} 5x_2 \le 15 \\
& 6x_1 + 2x_2 \le 24 \\
& x_1 + x_2 \le 5 \\
& x_1, x_2 \ge 0
\end{aligned}
$$

Weak duality

- **Primal** **Dual**

$$\begin{aligned} \max \quad & 2x_1 + x_2 \\ \text{s.t.} \quad & 5x_2 \le 15 \\ & 6x_1 + 2x_2 \le 24 \\ & x_1 + x_2 \le 5 \\ & x_1, x_2 \ge 0 \end{aligned}$$

$$\begin{aligned} \min \quad & 15y_1 + 24y_2 + 5y_3 \\ \text{s.t.} \quad & 6y_2 + y_3 \ge 2 \\ & 5y_1 + 2y_2 + y_3 \ge 1 \\ & y_1, y_2, y_3 \ge 0 \end{aligned}$$

- **Weak duality**

$$2x_1 + x_2 \le 15y_1 + 24y_2 + 5y_3$$

for any feasible solution

Strong duality

- **Primal**

$$\begin{aligned} \max \quad & 2x_1 + x_2 \\ \text{s.t.} \quad & 5x_2 \le 15 \\ & 6x_1 + 2x_2 \le 24 \\ & x_1 + x_2 \le 5 \\ & x_1, x_2 \ge 0 \end{aligned}$$

Dual

$$\begin{aligned} \min \quad & 15y_1 + 24y_2 + 5y_3 \\ \text{s.t.} \quad & 6y_2 + y_3 \ge 2 \\ & 5y_1 + 2y_2 + y_3 \ge 1 \\ & y_1, y_2, y_3 \ge 0 \end{aligned}$$

- **Strong duality**

$$2x_1^* + x_2^* = 15y_1^* + 24y_2^* + 5y_3^*$$

  - for the optimal solution

Complementary slackness

- **Primal** **Dual**

$$\begin{aligned} \max \quad & 2x_1 + x_2 \\ \text{s.t.} \quad & 5x_2 \le 15 \\ & 6x_1 + 2x_2 \le 24 \\ & x_1 + x_2 \le 5 \\ & x_1, x_2 \ge 0 \end{aligned}$$

$$\begin{aligned} \min \quad & 15y_1 + 24y_2 + 5y_3 \\ \text{s.t.} \quad & 6y_2 + y_3 \ge 2 \\ & 5y_1 + 2y_2 + y_3 \ge 1 \\ & y_1, y_2, y_3 \ge 0 \end{aligned}$$

- **Complementary slackness**

$$5x_2^* < 15 \ \rightarrow y_1^* = 0$$

  - if there are redundant resource A for the optimal product plan, giving up A will not affect our profile, so the shadow price is zero.
  - instrict inequality/inactive constraint correspond to zero dual variable

Primal-dual relationship

- **Primal**

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, i = 1, \dots, m \\ & h_i(x) = 0, i = 1, \dots, p \end{aligned}$$

$$f^* = f_0(x^*)$$

Dual

$$\begin{aligned} \max \quad & g(\lambda, \nu) = \inf_{x \in D} (f_0(x) + \lambda^\top f(x) + \nu^\top h(x)) \\ \text{s.t.} \quad & \lambda \geq 0 \end{aligned}$$

$$g^* = g(\lambda^*, \nu^*)$$

- weak duality $\quad f(x) \geq g(\lambda, \nu), \quad f^* \geq g^*$
- strong duality $\quad f^* = g^*$
- slackness condition $\quad \lambda_i^* f_i(x^*) = 0$

$$\lambda_i^* > 0 \rightarrow f_i(x^*) = 0 \qquad f_i(x^*) < 0 \rightarrow \lambda_i^* = 0$$

Duality gap:

$\min f(x) - g(\lambda, \nu)$

Slater's constraint qualification

- for the primal problem

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, i = 1, \dots, m \\ & h_i(x) = 0, i = 1, \dots, p \end{aligned}$$

Slater's constraint qualification requires the problem is strictly feasible:

$$\exists x \in \operatorname{int} D, f_i(x) < 0, h_i(x) = 0$$

- if the problem is convex and the Slater's qualification satisfied, then there is

strong duality

- there are many other qualifications

Strong duality proof

- convexity, int $D \neq \emptyset$, A is full ranked, and the optimal value is f^*

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, i = 1, \dots, m \\ & Ax - b = 0, i = 1, \dots, p \end{aligned}$$

- consider a set

$$C = \{(u, v, t) \colon \exists x \in D, f_i(x) \leq u_i, \forall i, Ax - b = v_i, \forall i, f_0(x) \leq t\}$$

- it is convex and do not have joint point with the following convex set

$$D = \{(0,0,s) \in \mathbf{R}^m \times \mathbf{R}^p \times \mathbf{R} \colon s < f^*\}$$

Strong duality proof

- using separating hyperplane theorem for

$$C = \{(u, v, t): \exists x \in D, f_i(x) \leq u_i, \forall i\ , Ax - b = v_i, \forall i, f_0(x) \leq t\}$$

$$E = \{(0,0, t) \in \mathbf{R}^m \times \mathbf{R}^p \times \mathbf{R}: t < f^*\}$$

- there exists nonzero $(\tilde{\lambda}, \tilde{v}, \mu) \in \mathbf{R}^m \times \mathbf{R}^p \times \mathbf{R}$ and α, such that

$$\tilde{\lambda}^\top u + \tilde{v}^\top v + \mu t \geq \alpha, \qquad \forall (u, v, t) \in C$$

$$\Rightarrow \tilde{\lambda} \geq 0, \mu \geq 0$$

$$\tilde{\lambda}^\top u + \tilde{v}^\top v + \mu t \leq \alpha, \qquad \forall (u, v, t) \in E$$

$$\Rightarrow \mu t \leq \alpha, \forall t \leq f^*$$

$$\mu f^* \leq \alpha$$

Strong duality proof

- together with $\tilde{\lambda}^\top u + \tilde{\nu}^\top v + \mu t \geq \alpha, \forall (u, v, t) \in C$ and $\mu f^* \leq \alpha$, we have

$$C = \{(u, v, t) \colon \exists x \in D, f_i(x) \leq u_i, \forall i, Ax - b = v_i, \forall i, f_0(x) \leq t\}$$

$$\sum_{i=1}^{m} \tilde{\lambda}_i f_i(x) + \tilde{\nu}^\top (Ax - b) + \mu f_0(x) \geq \alpha \geq \mu f^*, \forall x \in D$$

- if $\mu \neq 0$

$$\sum_{i=1}^{m} \frac{\tilde{\lambda}_i}{\mu} f_i(x) + \frac{\tilde{\nu}^\top}{\mu} (Ax - b) + f_0(x) \geq f^*, \forall x \in D$$

$$g(\lambda, \nu) = \inf_{x \in D} (f_0(x) + v^\top f(x) + \nu^T h(x)) \geq f^*$$

weak duality: $g(\lambda, \nu) \leq f^*$

$$\left.\begin{array}{l} g(\lambda, \nu) \geq f^* \\ g(\lambda, \nu) \leq f^* \end{array}\right\} \exists \lambda, \nu \colon g(\lambda, \nu) = f^*$$

Strong duality proof

- together with $\tilde{\lambda}^{\top}u+\tilde{v}^{\top}v+\mu t\geq\alpha,\forall(u,v,t)\in C$ and $\mu f^*\leq\alpha$, we have

$$C=\{(u,v,t):\exists x\in D, f_i(x)\leq u_i,\forall i\ , Ax-b=v_i,\forall i, f_0(x)\leq t\}$$

$$\sum_{i=1}^{m}\tilde{\lambda}_i f_i(x)+\tilde{v}^{\top}(Ax-b)+\mu f_0(x)\geq\alpha\geq\mu f^*,\forall x\in D$$

- if $\mu=0$

$$\sum_{i=1}^{m}\tilde{\lambda}_i f_i(x)+\tilde{v}^{\top}(Ax-b)\geq 0,\forall x\in D$$

$$\tilde{v}^{\top}(Ax-b)\geq 0,\forall x\in D$$

$$\tilde{v}^{\top}(A\tilde{x}-b)=0$$

slater's condition: there exist strictly feasible solutions $\tilde{x}$

$$\Downarrow$$

$$\tilde{v}^{\top}A=0$$

$$\sum_{i=1}^{m}\tilde{\lambda}_i f_i(\tilde{x})\geq 0 \Rightarrow \tilde{\lambda}=0$$

$$\left.\begin{array}{r}\tilde{\lambda}=0\\ \mu=0\\ (\tilde{\lambda},\tilde{v},\mu)\text{ is nonzero}\end{array}\right\}\ \tilde{v}\neq 0 \longrightarrow$$

contradict A is full ranked

Primal-dual relationship

- **Primal**

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, i = 1, \dots, m \\ & h_i(x) = 0, i = 1, \dots, p \end{aligned}$$

$$f^* = f_0(x^*)$$

Dual

$$\begin{aligned} \max \quad & g(\lambda, \nu) = \inf_{x \in D} \left(f_0(x) + \nu^\top f(x) + \nu^T h(x)\right) \\ \text{s.t.} \quad & \lambda \ge 0 \end{aligned}$$

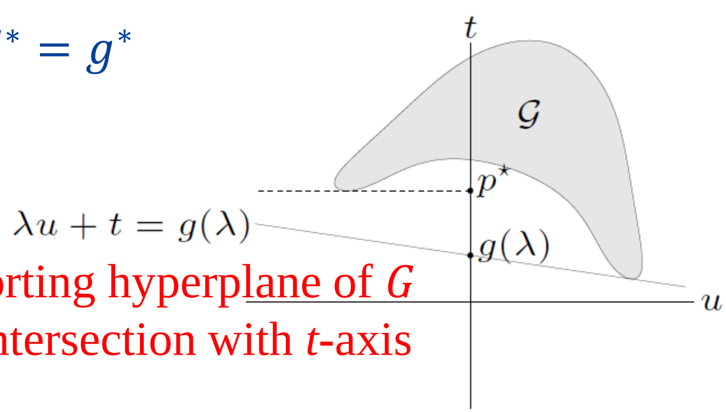
$$g^* = g(\lambda^*, \nu^*)$$

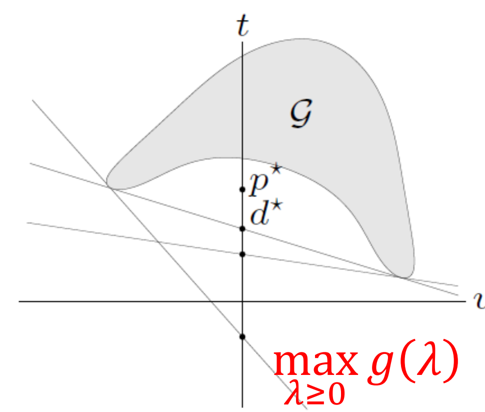
$\min f_0(x), \text{s.t.}, f_1(x) \le 0$

$G = \{(f_1(x), f_0(x)), x \in D\}$

$g(\lambda) = \inf_{(t,u) \in D} t + \lambda u$

- weak duality $\quad f(x) \ge g(\lambda, \nu), \quad f^* \ge g^*$
- strong duality $\quad f^* = g^*$





$\max_{\lambda \ge 0} g(\lambda)$

$t + \lambda u = g(\lambda)$ is a supporting hyperplane of G
$g(\lambda)$ is the value of the intersection with t-axis (since $\lambda \ge 0$)

Primal-dual relationship

- **Primal** **Dual**

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, i = 1, \dots, m \\ & h_i(x) = 0, i = 1, \dots, p \end{aligned}$$

$$\begin{aligned} \max \quad & g(\lambda, \nu) = \inf_{x \in D} (f_0(x) + \nu^\top f(x) + \nu^T h(x)) \\ \text{s.t.} \quad & \lambda \ge 0 \end{aligned}$$

$$g^* = g(\lambda^*, \nu^*)$$

$$f^* = f_0(x^*)$$

- weak duality $f(x) \ge g(\lambda, \nu)$, $f^* \ge g^*$
- strong duality $f^* = g^*$
  - convex problems with further condition
  - only convexity is not sufficient
  - convexity is not necessary: strong duality holds for some non-convex problems

Complementary slackness

- when the strong duality holds

$$f_0(x^*) = g(\lambda^*, \nu^*) = \inf_x \left(f_0(x) + {\lambda^*}^\top f(x) + {\nu^*}^\top h(x) \right)$$

$$\leq f_0(x^*) + {\lambda^*}^\top f(x^*) + {\nu^*}^\top h(x^*)$$

$$\leq f_0(x^*)$$

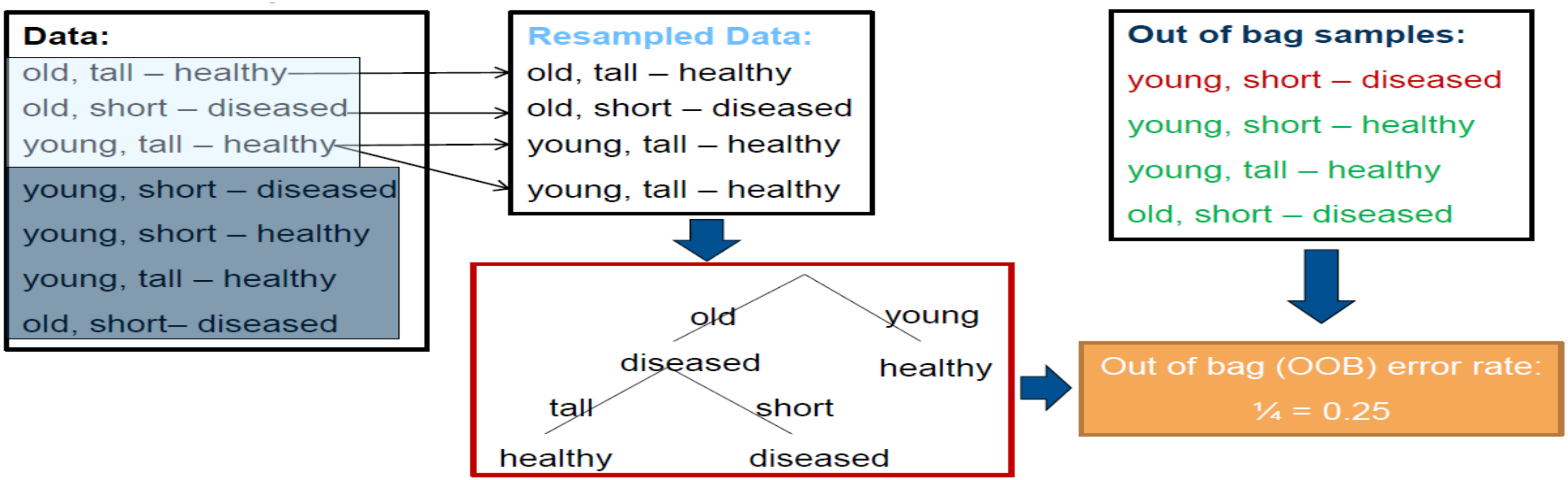
- "=" should be true for the last inequality

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0 \Rightarrow \lambda_i^* f_i(x^*) = 0$$

$$\lambda_i^* > 0 \rightarrow f_i(x^*) = 0 \qquad f_i(x^*) < 0 \rightarrow \lambda_i^* = 0$$

Perturbation

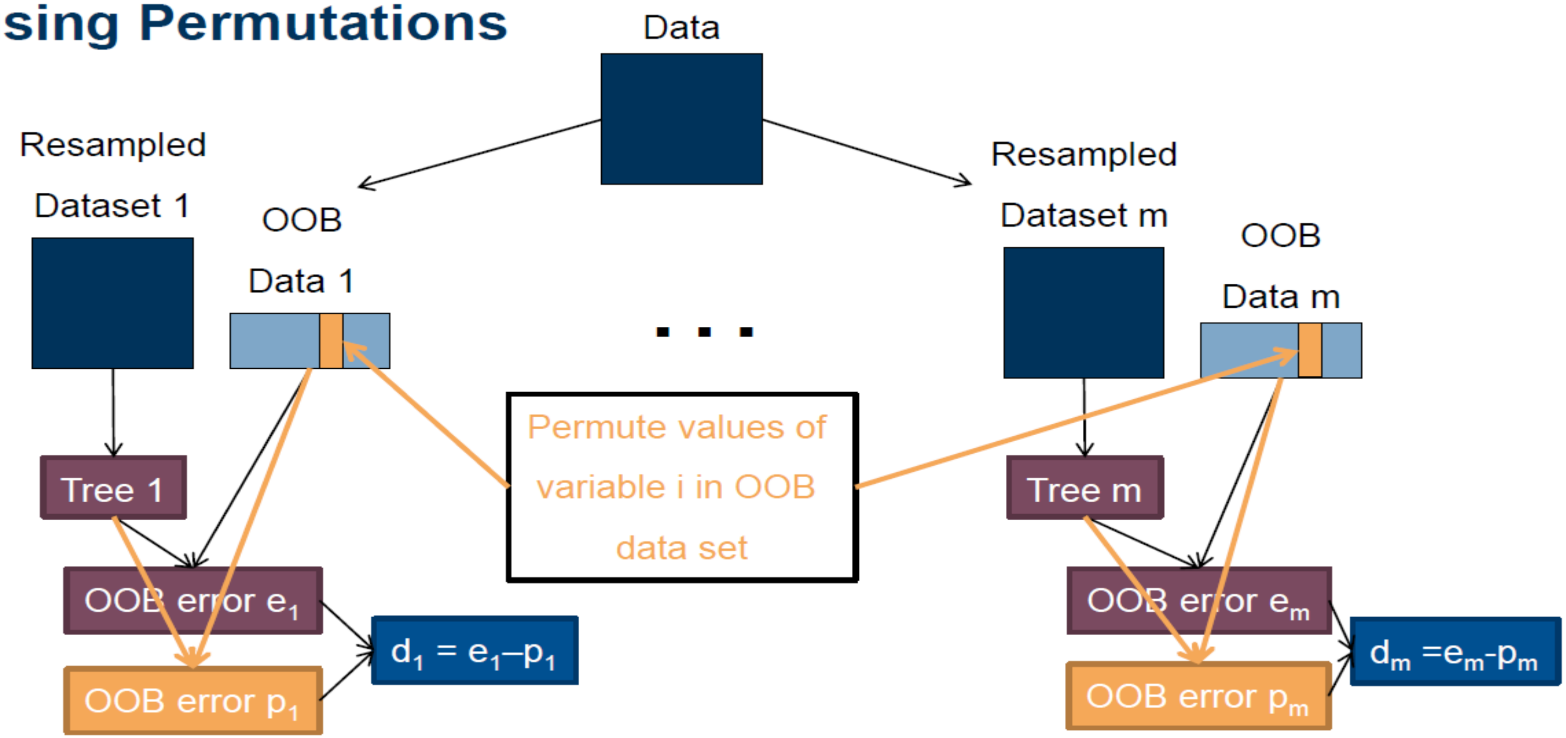
- perturbation: an intuitive way for sensitivity analysis
  - e.g., in random forest, one can permute one contribution and see the difference on the error, e.g., out-of-bag (OOB) error



Perturbation

- perturbation: an intuitive way for sensitivity analysis

using Permutations



$$\left.\begin{aligned} \overline{d} &= \frac{1}{m}\sum_{i=1}^{m} d_i \\ s_d^2 &= \frac{1}{m-1}\sum_{i=1}^{m}(d_i - \overline{d})^2 \end{aligned}\right\} \quad v_i = \frac{\overline{d}}{s_d}$$

Perturbation

- unperturbed optimization problem and its dual

$$\begin{array}{ll} \min & f_0(x) \\ \text{s.t.} & f_i(x) \le 0, i = 1, \dots, m \\ & h_i(x) = 0, i = 1, \dots, p \end{array} \qquad \begin{array}{ll} \max & g(\lambda, \nu) \\ \text{s.t.} & \lambda \ge 0 \end{array}$$

- perturbed optimization problem and its dual

$$\begin{array}{ll} \min & f_0(x) \\ \text{s.t.} & f_i(x) \le u_i, i = 1, \dots, m \\ & h_i(x) = v_i, i = 1, \dots, p \end{array} \qquad \begin{array}{ll} \max & g(\lambda, \nu) - u^\top \lambda - v^\top \nu \\ \text{s.t.} & \lambda \ge 0 \end{array}$$

$f^*(u, v)$ how about the optimal value changes as a function of u and v

Global sensitivity

- apply weak duality to the perturbed problem

$$f^*(u,v) \geq \underbrace{g(\lambda^*,\nu^*)}_{\text{assume the strong duality holds}} - u^\top\lambda^* - v^\top\nu^* = f^*(0,0) - u^\top\lambda^* - v^\top\nu^*$$

- perturbed optimization problem and its dual

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq u_i, i = 1, \dots, m \\ & h_i(x) = v_i, i = 1, \dots, p \end{aligned}$$

$$\begin{aligned} \max \quad & g(\lambda,\nu) - u^\top\lambda - v^\top v \\ \text{s.t.} \quad & \lambda \geq 0 \end{aligned}$$

$f^*(u,v)$ how about the optimal value changes as a function of u and v

Global sensitivity

- apply weak duality to the perturbed problem

$$f^*(u,v) \geq \underbrace{g(\lambda^*,\nu^*)}_{\text{assume the strong duality holds}} - u^\top\lambda^* - v^\top\nu^* = f^*(0,0) - u^\top\lambda^* - v^\top\nu^*$$

- sensitivity interpretation
  - if λ_i^* is large, f^*increases (beacomes worese)greatly when we tighten constraint ($u_i < 0$)
  - if λ_i^* is large and positive, ….

$$\begin{array}{ll} \min & f_0(x) \\ \text{s.t.} & f_i(x) \leq u_i, i = 1, \dots, m \\ & h_i(x) = v_i, i = 1, \dots, p \end{array}$$

Local sensitivity

- apply weak duality to the perturbed problem

$$f^*(u,v) \geq g(\lambda^*,\nu^*) - u^\top\lambda^* - v^\top\nu^* = f^*(0,0) - u^\top\lambda^* - v^\top\nu^*$$

- if $f^*(u,v)$ is differentiable at the original, then

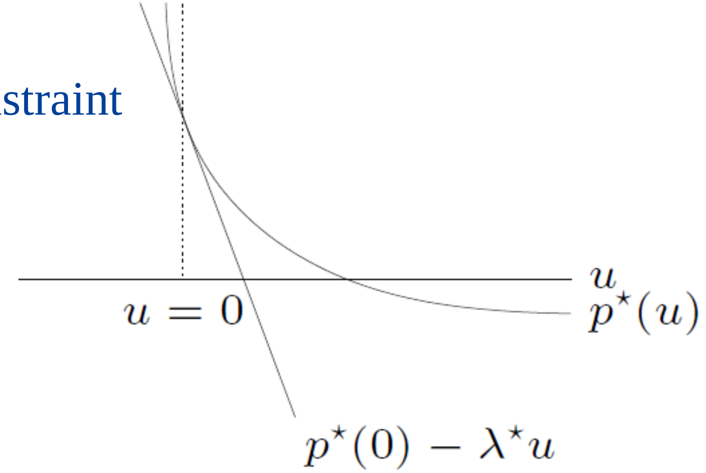
$$\lambda_i^* = -\frac{df^*(u,v)}{du_i}\Big|_{u=0,v=0} \quad \nu_i^* = -\frac{df^*(u,v)}{dv_i}\Big|_{u=0,v=0}$$

$f^*(u)$ for a problem with one inequality constraint

⬇

if the inequality constraint is linear

if the inequality becomes equation



Problem reformulation

- equivalent formulations may have very different dual
- reformulation can be useful when the dual is difficult or uninterestin
- for example

$$\min_x f_0(Ax+b) \Rightarrow g = \inf_x L(x) = \inf_x f_0(Ax+b)$$

$$\Downarrow$$

$$\min_{x,z} f_0(z) \quad \text{s.t. } Ax+b-z=0 \Rightarrow g(v) = \inf_{x,z} f_0(z) - v^\top(Ax+b-z)$$

$$= \begin{cases} -f_0^*(v) + b^\top v & A^\top v = 0 \\ -\infty & \text{otherwise} \end{cases}$$

目录 Contents

Karush-Kuhn-Tucker conditions

suppose all these functions are differentiable

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, i = 1, \dots, m \\ & h_i(x) = 0, i = 1, \dots, p \end{aligned}$$

- x^* is optimal to the primal problem and u^*, v^* to the dual problem

primal feasible $\quad f_i(x^*) \le 0$

$\quad h_i(x^*) = 0$

dual feasible $\quad \lambda_i^* \ge 0$

complementary slackness $\quad \lambda_i^* f_i(x^*) = 0$

$$\nabla f_0(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) + \sum_i v_i^* h(x^*) = 0$$

gradient of Lagrangian

Optimality

- for convex problems

$$L(x,\lambda,\nu) = f_0(x) + \lambda^\top f(x) + \nu^T h(x) \text{ is convex}$$

- $L(x,\lambda,\nu)$ achieves the minimum when x^* satisfying

$$\nabla f_0(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) + \sum_i \nu_i^* h(x^*) = 0$$

$$\Downarrow$$

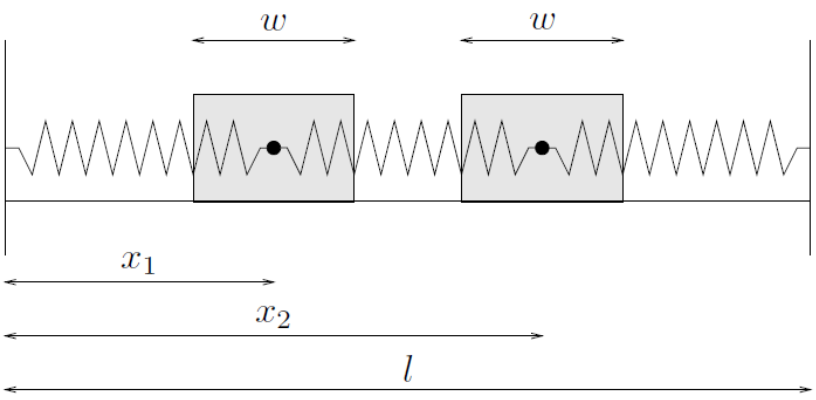
$$g(\lambda^*,\nu^*) = f_0(x^*) + {\lambda^*}^\top f(x^*) + {\nu^*}^\top h(x^*) = f_0(x^*)$$

duality gap is zero, and they are optimal

Optimality

- for convex problems
  - if Slater's condition is satisfied, KKT is sufficient and necessary
  - if not, KKT is necessary, but not sufficient
- for non-convex problems
  - KKT is necessary

Interpretation



- the energy

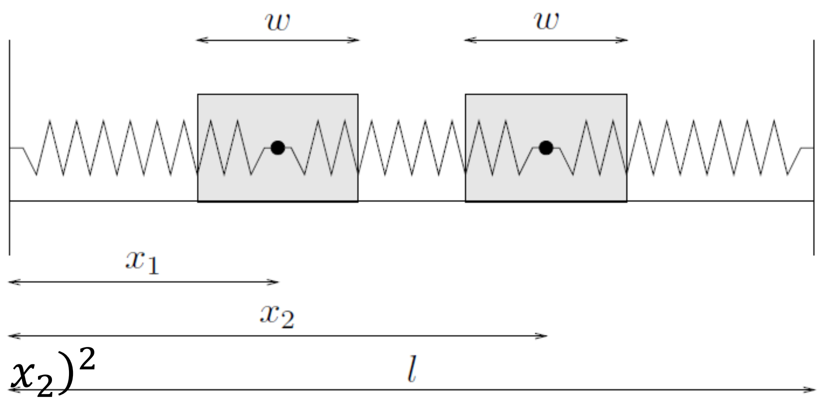
$$f_0(x_1,x_2)=\frac{1}{2}k_1x_1^2+\frac{1}{2}k_2(x_2-x_1)^2+\frac{1}{2}k_3(l-x_2)^2$$

- physical constraints: $\frac{w}{2}-x_1\le 0, w+x_1-x_2\le 0, \frac{w}{2}-l+x_2\le 0$

- the equilibrium could be achieved

$$\begin{aligned}\min\quad & \frac{1}{2}k_1x_1^2+\frac{1}{2}k_2(x_2-x_1)^2+\frac{1}{2}k_3(l-x_2)^2\\ \text{s.t.}\quad & \frac{w}{2}-x_1\le 0\\ & w+x_1-x_2\le 0\\ & \frac{w}{2}-l+x_2\le 0\end{aligned}$$

Interpretation

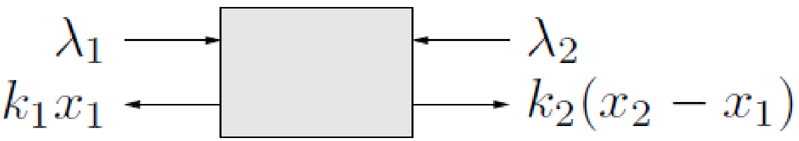


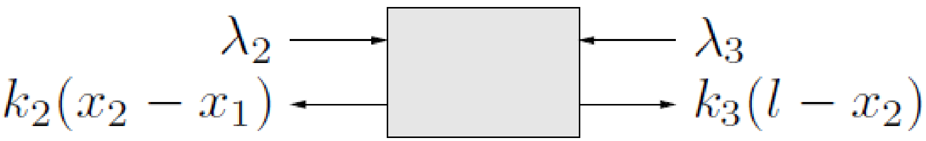
$$
\begin{aligned}
\min \quad & \frac{1}{2}k_1x_1^2 + \frac{1}{2}k_2(x_2-x_1)^2 + \frac{1}{2}k_3(l-x_2)^2 \\
\text{s.t.} \quad & \frac{w}{2} - x_1 \le 0 \\
& w + x_1 - x_2 \le 0 \\
& \frac{w}{2} - l + x_2 \le 0
\end{aligned}
$$

$$\lambda_1, \lambda_2, \lambda_3 \ge 0$$

- KKT $\quad \lambda_1\left(\frac{w}{2} - x_1\right) = 0, \lambda_2\,(w + x_1 - x_2) = 0, \lambda_3\left(\frac{w}{2} - l + x_2\right) = 0$

$$\begin{bmatrix} k_1x_1 - k_2(x_2-x_1) \\ k_2(x_2-x_1) - k_3(l-x_2) \end{bmatrix} + \lambda_1\begin{bmatrix} -1 \\ 0 \end{bmatrix} + \lambda_2\begin{bmatrix} 1 \\ -1 \end{bmatrix} + \lambda_3\begin{bmatrix} 0 \\ 1 \end{bmatrix} = 0$$





Newton's method with equality constraints

$$\min_x f(x) \quad \text{s.t. } Ax = b$$

- if the x^k is feasible, we need to guarantee that $A\Delta x_{\text{nt}} = 0$
- optimality condition

$$Ax^* = b, \nabla f(x^*) + A^\top v^* = 0$$

$$\Downarrow$$

$$A\left(x^k + \Delta x_{\text{nt}}\right) = b, \nabla f\left(x^k + \Delta x_{\text{nt}}\right) + A^\top v \approx \nabla f\left(x^k\right) + \nabla^2 f\left(x^k\right)\Delta x_{\text{nt}} + A^\top v = 0$$

$$\Downarrow \; Ax^k = b$$

$$A\Delta x_{\text{nt}} = 0, \;\; \nabla^2 f\left(x^k\right)\Delta x_{\text{nt}} + A^\top v = -\nabla f\left(x^k\right)$$

Newton's method with equality constraints

$$\min_x f(x) \quad \text{s.t. } Ax = b$$

- the Newton's direction is obtained by

$$\begin{bmatrix} \nabla^2 f(x^k) & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\text{nt}} \\ v \end{bmatrix} = \begin{bmatrix} -\nabla f(x^k) \\ 0 \end{bmatrix}$$

---

given starting point $x \in \textbf{dom}\, f$ with $Ax = b$, tolerance $\epsilon > 0$.

repeat

1. Compute the Newton step and decrement Δx_{nt}, $\lambda(x)$.
2. *Stopping criterion.* **quit** if $\lambda^2/2 \leq \epsilon$.
3. *Line search.* Choose step size t by backtracking line search.
4. *Update.* $x := x + t\Delta x_{\text{nt}}$.

Newton's method with equality constraints

$$\min_{x} f(x) \quad \text{s.t. } Ax = b$$

- if the x^k is infeasible, we need to first let the solution goes to feasible set

$$A\left(x^k + \Delta x_{\text{nt}}\right) = b, \nabla f\left(x^k + \Delta x_{\text{nt}}\right) + A^\top v \approx \nabla f\left(x^k\right) + \nabla^2 f\left(x^k\right)\Delta x_{\text{nt}} + A^\top v = 0$$

$$\Big\Downarrow \quad Ax^k \neq b$$

$$A\Delta x_{\text{nt}} = -(Ax^k - b), \;\; \nabla^2 f\left(x^k\right)\Delta x_{\text{nt}} + A^\top v = -\nabla f\left(x^k\right)$$

$$\Big\Downarrow$$

$$\begin{bmatrix} \nabla^2 f\left(x^k\right) & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\text{nt}} \\ v \end{bmatrix} = -\begin{bmatrix} \nabla f\left(x^k\right) \\ Ax^k - b \end{bmatrix}$$

primal-dual interpretation

Inequality constrained problem

$$\begin{array}{ll} \min & f_0(x) \\ \text{s.t.} & f_i(x) \le 0, i = 1, \dots, m \\ & Ax = b \end{array}$$

suppose all these functions are twice continuously differentiable

- convex problem, A is full-ranked, f^* is finite and attained
- we assume the problem is strictly feasible

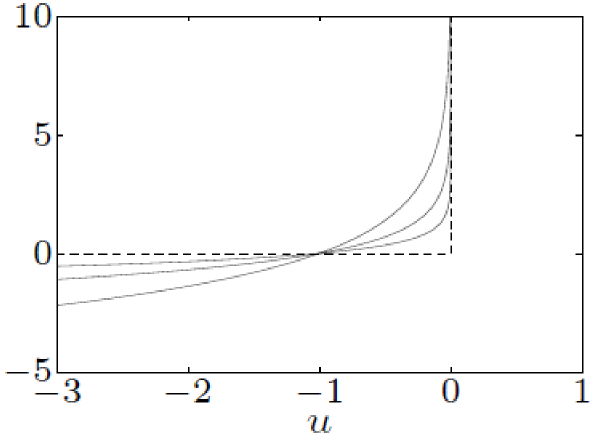
  (and so strong duality holds an dual optimum is attained)
- using indicator function to reformulate

$$\begin{array}{ll} \min & f_0(x) + \Sigma_{i=1}^m I_-(f_i(x)) \\ \text{s.t.} & Ax = b \end{array}$$

$$I_-(u) = \begin{cases} 0, & u \le 0 \\ \infty, & u > 0 \end{cases}$$

Logarithmic barrier

$$\begin{aligned} \min \quad & f_0(x) + \Sigma_{i=1}^{m} I_-(f_i(x)) \\ \text{s.t.} \quad & Ax = b \end{aligned}$$



- the indicator function is not continuous
- use logarithmic function to approach the indicator

$$I_-(u) \approx -\frac{1}{t}\log(-u)\,, t \to \infty$$

- logarithmic barrier function

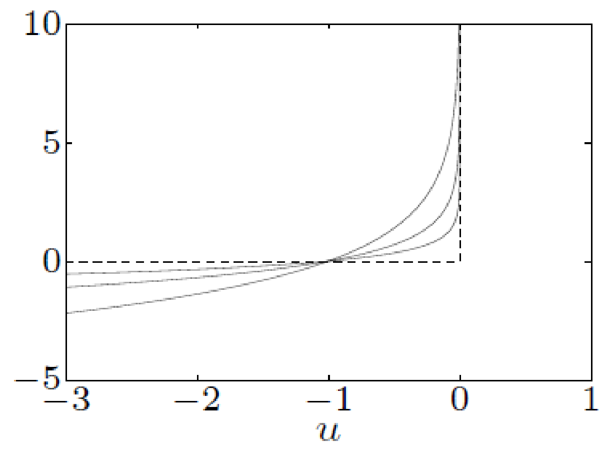
$$\phi(x) = -\sum_{i=1}^{m} \log\left(-f_i(x)\right)$$

$$\nabla\phi(x) = \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla f_i(x)$$

$$\nabla^2\phi(x) = \sum_{i=1}^{m} \frac{1}{f_i(x)^2} \nabla f_i(x) \nabla f_i(x)^\top + \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla^2 f_i(x)$$

Logarithmic barrier

$$\begin{array}{ll} \min & f_0(x) + \sum_{i=1}^{m} I_-(f_i(x)) \\ \text{s.t.} & Ax = b \end{array}$$



- the indicator function is not continuous
- use logarithmic function to approach the indicator

$$I_-(u) \approx -\frac{1}{t}\log(-u), t \to \infty$$

- logarithmic barrier function

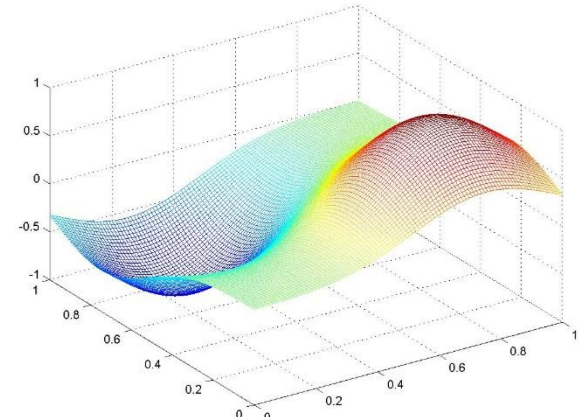
$$\phi(x) = -\sum_{i=1}^{m} \log(-f_i(x))$$

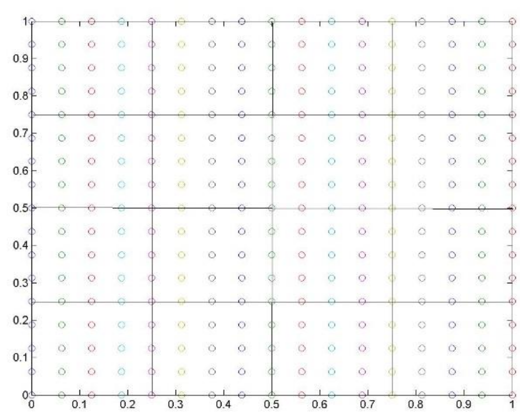
- with increasing t, we can approach the original problem

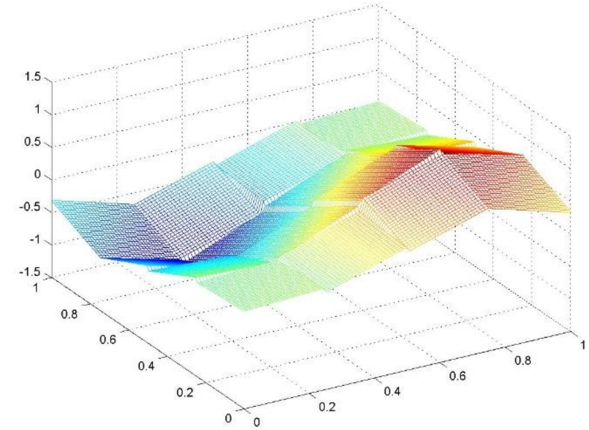
$$\min_x t f_0(x) + \phi(x), \text{s.t. } Ax = b$$

Nonlinearity: subregion

- Takagi-Sugeno Model:
  - divide the domain into subregions
  - locally train a linear model

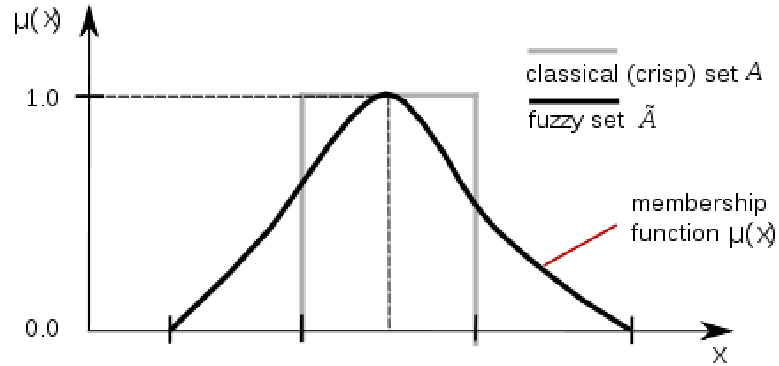




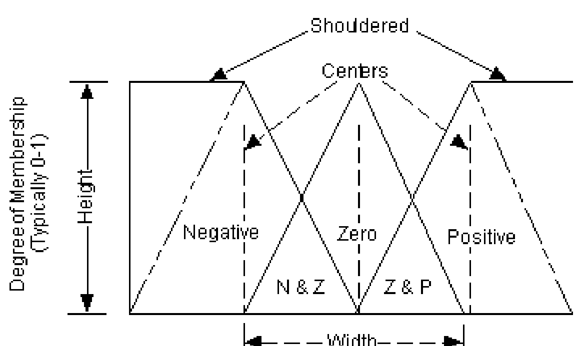


Fuzzy systems

- how to deal with the discontinuity
  - local linear functions $f_i(x), \forall x \in S_i$
  - simply sum them with an indicator function $I(A) = 1$ iff A is true



$$F(x) = \sum_{i=1} I(x \in S_i) f_i(x)$$

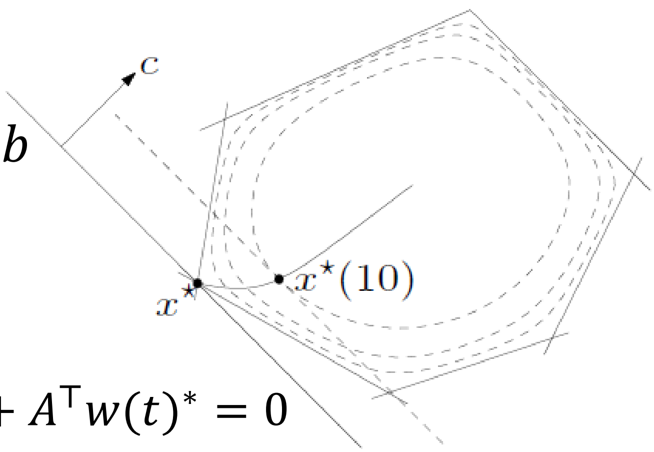


  - we could replace the indicator function by a *membership* functions
  - ANFIS (Adaptive neuro fuzzy inference system)

Central path

- central path

$$x^*(t) = \underset{x}{\operatorname{argmin}}\, t f_0(x) + \phi(x), \text{s.t. } Ax = b$$



- for a given t, there exists a $w(t)^*$ such that

$$t\nabla f_0\big(x^*(t)\big) + \sum_{i=1}^{m} \frac{1}{-f_i\big(x^*(t)\big)} \nabla f_i\big(x^*(t)\big) + A^\top w(t)^* = 0$$

$$\lambda^*(t) = 1/\big(-t f_i\big(x^*(t)\big)\big) \qquad \nu^*(t) = w(t)^*/t$$

- the following Lagrangian is minimized at $x^*(t)$

$$L(x, \lambda_i^*(t), \nu^*(t)) = f_0(x) + \Sigma_{i=1}^{m} \lambda_i^*(t) f_i(x) + \nu^*(t)^\top (Ax - b)$$

$$f^* \geq g\big(\lambda_i^*(t), \nu^*(t)\big) = L\big(x^*(t), \lambda^*(t), \nu^*(t)\big) = f_0\big(x^*(t)\big) - m/t$$

KKT conditions

- consider $x^*(t), \lambda^*(t), \nu^*(t)$

$$x^*(t) = \underset{x}{\operatorname{argmin}}\, tf_0(x) + \phi(x), \text{s.t. } Ax = b$$

$$\lambda^*(t) = 1/(-tf_i(x^*(t)))$$

$$\nu^*(t) = w(t)^*/t$$

$$\begin{array}{ll} \min & f_0(x) \\ \text{s.t.} & f_i(x) \le 0, i = 1, \dots, m \\ & Ax = b \end{array}$$

primal feasible

$$f_i(x^*(t)) \le 0$$

$$h_i(x^*(t)) = 0$$

dual feasible

$$\lambda_i^*(t) \ge 0$$

complementary slackness

$$\lambda_i^*(t) f_i(x^*(t)) = -1/t$$

$$\nabla f_0(x^*(t)) + \sum_i \lambda_i^* \nabla f_i(x^*(t)) + \sum_i \nu_i^* h(x^*(t)) = 0$$

gradient of Lagrangian

Barrier method

given strictly feasible x, $t := t^{(0)} > 0$, $\mu > 1$, tolerance $\epsilon > 0$.

repeat

1. *Centering step.* Compute $x^\star(t)$ by minimizing $tf_0 + \phi$, subject to $Ax = b$.
2. *Update.* $x := x^\star(t)$.
3. *Stopping criterion.* **quit** if $m/t < \epsilon$.
4. *Increase* t. $t := \mu t$.

- two loops: outer *iteration*, and *centering*
- outer iteration
  - a larger μ leads to faster convergence
- centering
  - standard analysis for unconstrained problems
  - a larger μ leads to slower convergence

Feasibility and phase I

- interior-point method needs a strictly feasible solution
- if not, solve a feasibility problem

$$
\begin{array}{ll}
\text{find} & x \\
\text{s.t.} & f_i(x) \le 0, i = 1, \dots, m \\
& Ax = b
\end{array}
$$

$$
\begin{array}{ll}
\min & s \\
\text{s.t.} & f_i(x) \le s, i = 1, \dots, m \\
& Ax = b, s \ge 0
\end{array}
$$

$$
\begin{array}{ll}
\min & \sum s_i \\
\text{s.t.} & f_i(x) \le s_i, i = 1, \dots, m \\
& Ax = b, s_i \ge 0
\end{array}
$$

目录 Contents

Dual problem example of SVM

- SVM in primal space

$$\begin{aligned}\min_{x,z}\quad & \tfrac{1}{2}\|x\|_2^2 + C\textstyle\sum_i \rho_i\\ \text{s.t.}\quad & b_i(x^\top a_i + z) \geq 1-\rho_i\\ & \rho_i \geq 0\end{aligned}$$

- the corresponding Lagrangian

$$L(x,z,\rho;\lambda,\nu) = \frac{1}{2}\|x\|_2^2 + C\sum_i \rho_i$$

$$+\sum_i \lambda_i(1-\rho_i - b_i(x^\top a_i + z)) - \nu^T\rho$$

Dual problem example of SVM

- to obtain the $g(\lambda, \nu) = \inf_{x,z,\rho} L(x, z, \rho; \lambda, \nu)$

$$L(x, z, \rho; \lambda, \nu) = \frac{1}{2}\|x\|_2^2 + C\sum_i \rho_i + \sum_i \lambda_i(1 - \rho_i - b_i(x^\top a_i + z)) - \nu^T \rho$$

$$\frac{\partial L}{\partial x} = x - \sum_i b_i \lambda_i a_i = 0$$

$$\frac{\partial L}{\partial z} = \sum_i \lambda_i b_i = 0$$

$$\frac{\partial L}{\partial \rho_i} = C - \lambda_i - \nu_i = 0$$

$$\lambda_i \geq 0, \nu_i \geq 0$$

$$\begin{aligned} \min_{\lambda} \; & \textstyle\sum_i \sum_j \lambda_i b_i a_i^\top a_j b_j \lambda_j - \sum_i \lambda_i \\ \text{s.t.} \; & \textstyle\sum_i \lambda_i b_i = 0 \\ & 0 \leq \lambda_i \leq C \end{aligned}$$

Primal-dual relation

- SVM in primal space

$$\begin{aligned} \min_{x,z} \quad & \frac{1}{2}\|x\|_2^2 + C\textstyle\sum_i \rho_i \\ \text{s.t.} \quad & b_i(x^\top a_i + z) \geq 1 - \rho_i \\ & \rho_i \geq 0 \end{aligned}$$

- SVM in dual space

$$\begin{aligned} \min_{\lambda} \quad & \textstyle\sum_i \sum_j \lambda_i b_i a_i^\top a_j b_j \lambda_j - \sum_i \lambda_i \\ \text{s.t.} \quad & \textstyle\sum_i \lambda_i b_i = 0 \\ & 0 \leq \lambda_i \leq C \end{aligned}$$

- strong duality and

$$f(x) = x^\top a + z = \sum_i \lambda_i b_i a_i^\top a + z$$

Support vector

- SVM in dual space $$f(x) = \sum_i \lambda_i b_i a_i^\top a + z$$

$$\min_\lambda \sum_i \sum_j \lambda_i b_i a_i^\top a_j b_j \lambda_j - \sum_i \lambda_i$$

$$\text{s.t.} \quad \sum_i \lambda_i b_i = 0$$

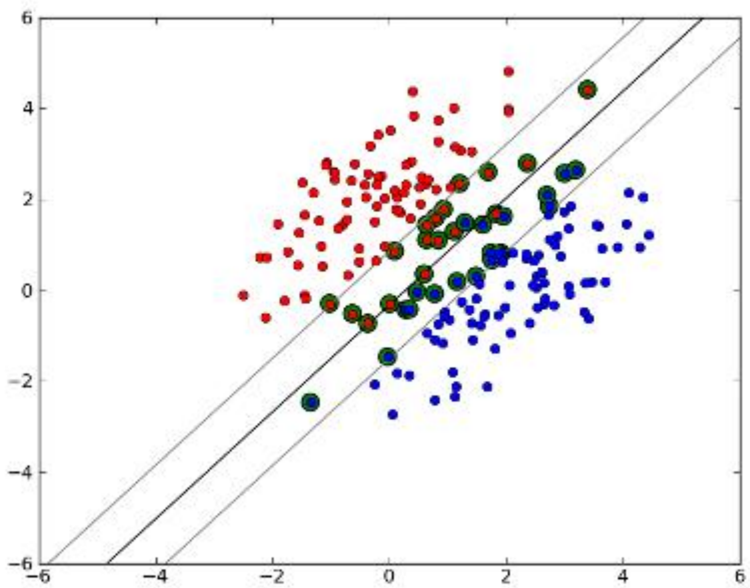
$$0 \leq \lambda_i \leq C$$

- there are only a part of samples

$$\lambda_i \neq 0$$

"Support Vector"

which one is support vector?



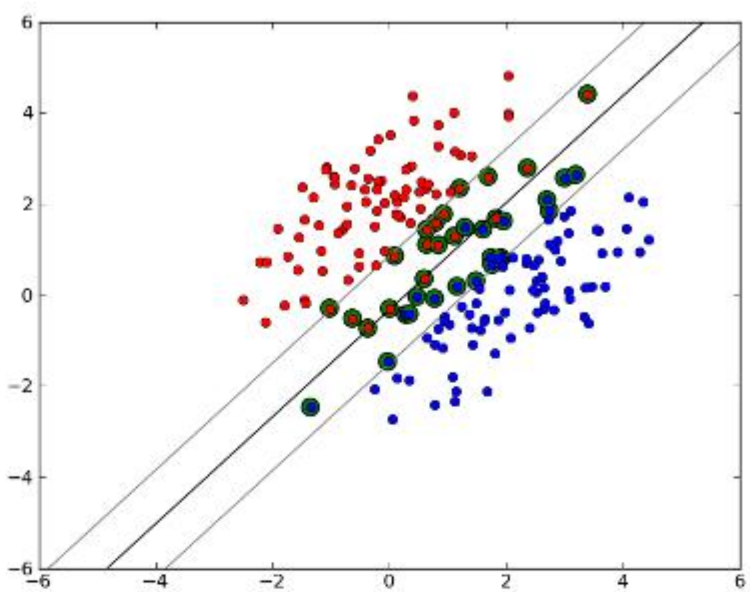
Support vector

- complementary slackness

$$\lambda_i = C \quad \begin{cases} \lambda_i > 0 \longrightarrow 1 - \rho_i - b_i(x^\top a_i + z) = 0 \\ \nu_i = 0 \longrightarrow \rho_i \geq 0 \end{cases}$$

$$\Downarrow$$

$$b_i(x^\top a_i + z) \leq 1$$



Support vector

- complementary slackness

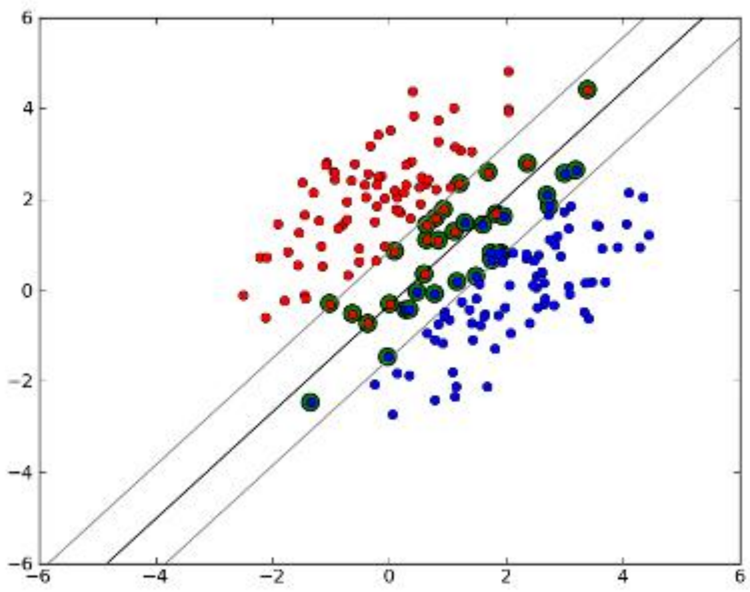
$$0 < \lambda_i < C \quad \begin{cases} \lambda_i > 0 \longrightarrow 1 - \rho_i - b_i(x^\top a_i + z) = 0 \\ \nu_i > 0 \longrightarrow \rho_i = 0 \end{cases}$$

$$\Downarrow$$

$$b_i(x^\top a_i + z) = 1$$

$$\Downarrow$$

determine z



Kernel trick

- to introduce non-linearity, usually a non-linear mapping is needed:

$$\phi(a)\colon\ \mathbf{R}^n \to \mathbf{R}^d$$

- SVM in primal space

$$\begin{aligned}\min_{x,z}\ & \frac{1}{2}\|x\|_2^2 + C\sum_i \rho_i \\ \text{s.t.}\ & b_i(x^\top\phi(a_i) + z) \geq 1 - \rho_i, \\ & \rho_i \geq 0\end{aligned}$$

- SVM in dual space

$$\begin{aligned}\min_{\lambda}\ & \sum_i\sum_j \lambda_i b \boxed{\phi(a_i)^\top\phi(a_j)} b_j \lambda - \sum_i \lambda_i \\ \text{s.t.}\ & \textstyle\sum_i \lambda_i b_i = 0 \\ & 0 \leq \lambda_i \leq C\end{aligned}$$

Kernel trick

- the discriminant function is :

$$f(x) = \sum_i \lambda_i b \boxed{\phi(a_i)^\top \phi(a)} - z$$

- kernel trick: we do not need to know the formulation of $\phi(x)$, instead, we only need to know the inner product

- kernel functions: $K(u, v)$: $\mathbf{R}^n \times \mathbf{R}^n \to R$

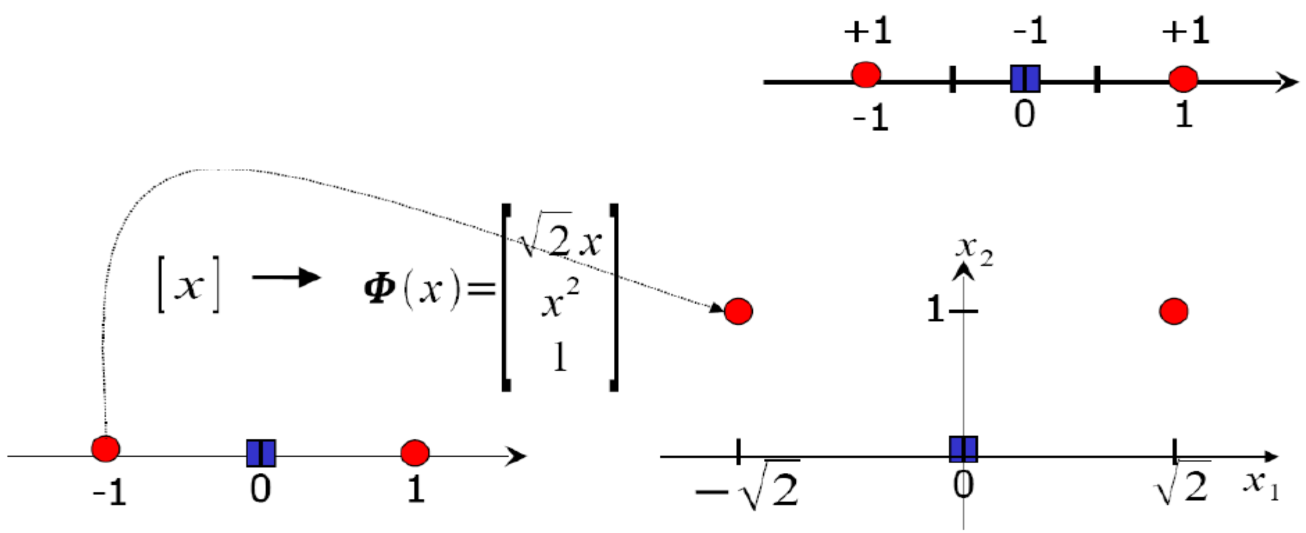
describe the relationship of the two samples

Polynomial kernel

- polynomial kernel:

$$K(u, v) = (u^T v + c)^d$$

- when $c = 0, d = 1$, it reduces to *linear kernel*
- for a one-dimensional, two-order polynomial kernel



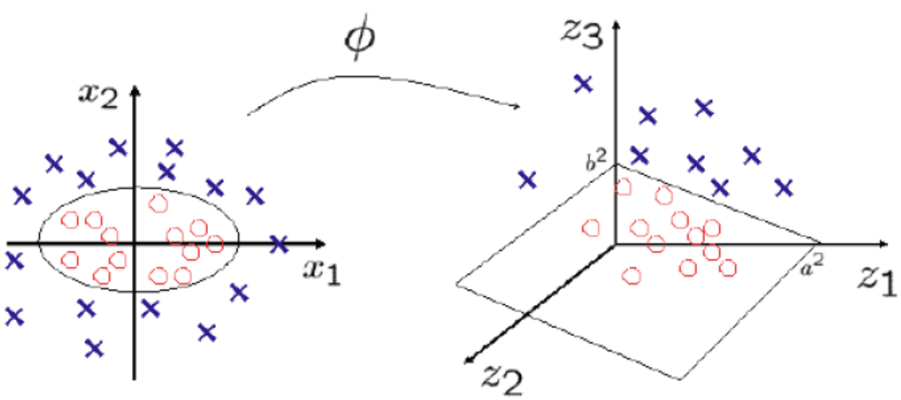
Polynomial kernel

- polynomial kernel:

$$K(u, v) = (u^T v + c)^d$$

- when $c = 0, d = 1$, it reduces to *linear kernel*
- for a two-dimensional, two-order polynomial kernel

$$\phi(\mathbf{u}) = [u_1^2, \sqrt{2}u_1 u_2, u_2^2];$$



RBF kernel

- Radial basis function (Gaussian) kernel

$$\mathcal{K}(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{u}-\mathbf{v}\|_2^2}{2\sigma^2}\right)$$

- even for one-dimensional case

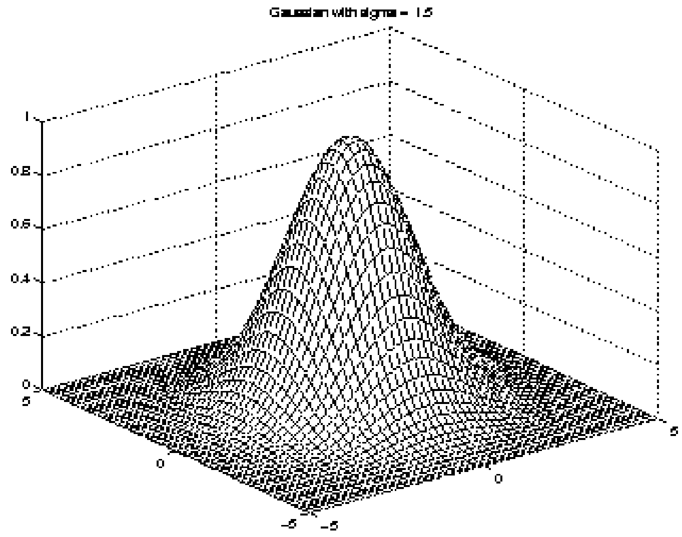
$$\phi(u) = \exp\left(-\frac{u^2}{2}\right)\left[1, \sqrt{2}u, \sqrt{\frac{1}{2!}}u^2, \sqrt{\frac{1}{3!}}u^3, \ldots\right]^T$$

an indefinite dimensional mapping

Mercer kernel

- radial basis function (Gaussian) kernel

$$\mathcal{K}(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{\|\mathbf{u}-\mathbf{v}\|_2^2}{2\sigma^2}\right)$$



- it is a similarity/dissimilarity measure
- many similarity matrix can be used if:
- Mercer's Theorem:

the matrix introduced by K is positive-semidefinite

$$K_{ij} = K(a_i, a_j) = \phi(a_i)^T \phi(a_j)$$

Solving algorithms

- solve SVM from primal or dual?

$$\begin{aligned} \min_{x,z}\quad & \tfrac{1}{2}\|x\|_2^2 + C\textstyle\sum_i \rho_i \\ \text{s.t.}\quad & b_i(x^\top a_i + z) \ge 1 - \rho_i \\ & \rho_i \ge 0 \end{aligned}$$

$$\begin{aligned} \min_{\lambda}\quad & \textstyle\sum_i \sum_j \lambda_i b_i a_i^\top a_j b_j \lambda - \sum_i \lambda_i \\ \text{s.t.}\quad & \textstyle\sum_i \lambda_i b_i = 0 \\ & 0 \le \lambda_i \le C \end{aligned}$$

independent of m

- primal space: $n + 1$ unknown variables ($2m + 1$ constraints)
- dual space: m unknown variables and $m + 1$ constraints
- big data problem usually solved from primal

independent of n

SMO

- consider the dual problem

$$
\begin{aligned}
\min_{\lambda} \ & \sum_i \sum_j \lambda_i b_i K_{ij} b_j \lambda_j - \sum_i \lambda_i \\
\text{s.t.} \ & \textstyle\sum_i \lambda_i b_i = 0 \\
& 0 \le \lambda_i \le C
\end{aligned}
$$

- choose only a small number of variables
- the smallest number is 2
- Sequential Minimization Optimization (SMO)

—— this idea is not restricted to SVM

SMO

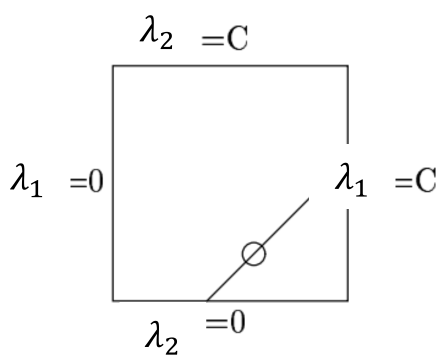
- consider the dual problem

$$\min_{\lambda} \sum_i \sum_j \lambda_i b_i K_{ij} b_j \lambda_j - \sum_i \lambda_i$$
$$\text{s.t.} \quad \sum_i \lambda_i b_i = 0$$
$$0 \le \lambda_i \le C$$

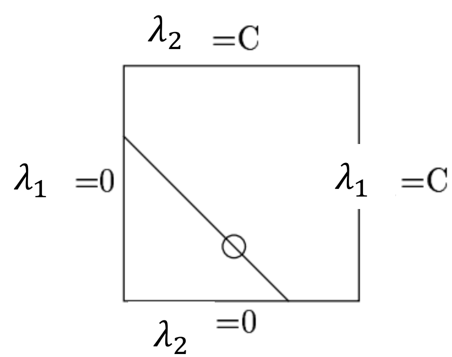
$$\Rightarrow$$

$$\min_{\lambda_1, \lambda_2} \quad K_{11}\lambda_1^2 + 2b_1 b_2 K_{12}\lambda_1\lambda_2 + K_{22}\lambda_2^2 - \lambda_1 - \lambda_2$$
$$\text{s.t.} \quad b_1\lambda_1 + b_2\lambda_2 = 0$$
$$0 \le \lambda_1, \lambda_2 \le C$$

- according to different b_1, b_2



$a_1 \neq a_2 \rightarrow \lambda_1 - \lambda_2 = 0$



$a_1 = a_2 \rightarrow \lambda_1 + \lambda_2 = 0$

SMO

- λ_1, λ_2 can be optimally updated, actually with analytic expressions

$$\min_{\lambda_1 \lambda_2} \begin{array}{c} K_{11}\lambda_1^2 + 2b_1b_2K_{12}\lambda_1\lambda_2 + K_{22}\lambda_2^2 \\ -\lambda_1 - \lambda_2 \end{array}$$

$$\text{s.t.} \quad \begin{array}{l} b_1\lambda_1 + b_2\lambda_2 = 0 \\ 0 \le \lambda_1, \lambda_2 \le C \end{array}$$

- the remaining question is how to select the variables to be update
  - purely random: the improvement may be small, but no time required
  - the largest improvement pair: a 2-D loop
  - the largest two variables violating optimality condition: 1-D loop

Another example: one-bit CS

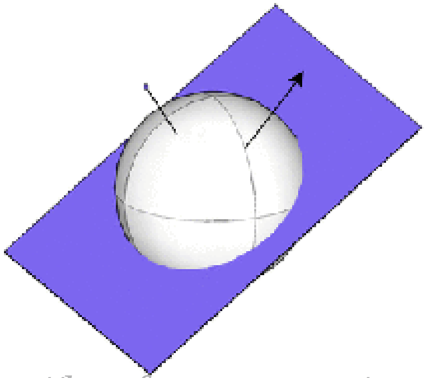
- compressive sensing

$$\min_{x,z} \quad \mu \|x\|_1 + \frac{1}{m} \sum \left(b_i - a_i^\top x\right)^2$$

- in real situation, the observations (actually all variables) are quantized
- the extreme case, we only have one-bit information

$$b_i = \mathrm{sign}(a_i^\top x + \varepsilon)$$

- is that possible to also recover the signal?
- but norm information is needed



https://boufounos.com/research/1-bit-compressive-sensing/

Another example: one-bit CS

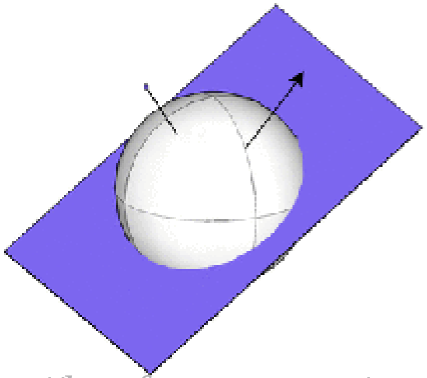
- one-bit compressive sensing

$$\begin{array}{ll} \min_{x,z} & \mu\|x\|_1 + \frac{1}{m}\sum \max\{0, 1 - b_i(a_i^\top x)\} \\ \text{s.t.} & \|x\|_2 = 1 \end{array}$$

- in real situation, the observations (actually all variables) are quantized
- the extreme case, we only have one-bit information

$$b_i = \operatorname{sign}(a_i^\top x + \varepsilon)$$

- is that possible to also recover the signal?
- but norm information is needed



https://boufounos.com/research/1-bit-compressive-sensing/

Another example: one-bit CS

- one-bit compressive sensing

$$\begin{aligned} \min_{x,z} \quad & \mu \|x\|_1 + \frac{1}{m} \sum \max\{0, 1 - b_i(a_i^\top x)\} \\ \text{s.t.} \quad & \|x\|_2 = 1 \end{aligned}$$

- relaxation

$$\begin{aligned} \min_{x,z} \quad & \mu \|x\|_1 + \frac{1}{m} \sum \max\{0, 1 - b_i(a_i^\top x)\} \\ \text{s.t.} \quad & \|x\|_2 \leq 1 \end{aligned}$$

- reformulation (could have different dual)

$$\begin{aligned} \min_{x,y,z} \quad & \mu \|y\|_1 + \frac{1}{m} \sum \max\{0, 1 + z_i\} + I_{\|x\|_2 \leq 1}(x) \\ \text{s.t.} \quad & x = y, z_i = -b_i\left(a_i^\top x\right) \end{aligned}$$

Another example: one-bit CS

$$\begin{aligned} \min_{x,y,z} \quad & \mu\,\|y\|_1 + \frac{1}{m}\sum \max\{0, 1+z_i\} + I_{\|x\|_2\le 1}(x) \\ \text{s.t.} \quad & x = y, z_i = -b_i\left(a_i^\top x\right) \end{aligned}$$

- Langrangian

$$L(x,y,z;\lambda,\nu) = \mu\|y\|_1 + \frac{1}{m}\sum \max\{0, 1+z_i\} + I_{\|x\|_2\le 1}(x) + \lambda^\top(x-y) + \nu^\top(-b\cdot Ax - z)$$

- minimization over primal variables

$$\min_x L(x,y,z;\lambda,\nu) = \min_x I_{\|x\|_2\le 1}(x) + \lambda^\top x - \nu^\top(b\cdot Ax) = -\|\sum \nu_i b_i a_i - \lambda\|_2$$

$$\min_y L(x,y,z;\lambda,\nu) = \min_y \mu\|y\|_1 - \lambda^\top y = \begin{cases} 0, & \|\lambda\|_\infty \le \mu \\ -\infty, & \text{otherwise} \end{cases}$$

$$\min_{z_i} L(x,y,z;\lambda,\nu) = \min_{z_i} \frac{1}{m}\max\{0, 1+z_i\} - \nu_i z_i = \begin{cases} \nu_i, & |\nu_i| \le 1/m \\ -\infty, & \text{otherwise} \end{cases}$$

Another example: one-bit CS

$$\max_{\lambda,\nu} \ \sum \nu_i - \|\sum \nu_i b_i a_i - \lambda\|_2$$
$$\text{s.t.} \ \ \|\lambda\|_\infty \le \mu, \ \ \|\nu\|_\infty \le 1/m$$

separable
dual coordinate ascent

- Langrangian

$$L(x,y,z;\lambda,\nu) = \mu\|y\|_1 + \frac{1}{m}\sum \max\{0, 1+z_i\} + I_{\|x\|_2\le 1}(x) + \lambda^\top(x-y) + \nu^\top(-b\cdot Ax - z)$$

- minimization over primal variables

$$\min_x L(x,y,z;\lambda,\nu) = \min_x I_{\|x\|_2\le 1}(x) + \lambda^\top x - \nu^\top(b\cdot Ax) = -\|\sum \nu_i b_i a_i - \lambda\|_2$$

$$\min_y L(x,y,z;\lambda,\nu) = \min_y \mu\|y\|_1 - \lambda^\top y = \begin{cases} 0, & \|\lambda\|_\infty \le \mu \\ -\infty, & \text{otherwise} \end{cases}$$

$$\min_{z_i} L(x,y,z;\lambda,\nu) = \min_{z_i} \frac{1}{m}\max\{0, 1+z_i\} - \nu_i z_i = \begin{cases} \nu_i, & |\nu_i| \le 1/m \\ -\infty, & \text{otherwise} \end{cases}$$

Principle component analysis

- for finding the principle axes of a dataset

$$\max_{x^\top x=1} \quad x^\top C x \qquad\qquad C = \frac{1}{n-1}\boxed{A^\top A}$$

it should be zero-mean, otherwise, the covariance matrix will be ?

- this optimization problem can be solved as the following:
- 1. compute the mean
- 2. compute the covariance
- 3. find the principle axes
- 4. project data onto the eigenvectors

Principle component analysis

- for finding the principle axes of a dataset

$$\max_{x^\top x=1} \quad x^\top C x \qquad\qquad C = \frac{1}{n-1}A^\top A$$

- the Lagrangian is

$$L(x,\lambda) = x^\top C x - \lambda(x^\top x - 1)$$

- from the KKT condition

$$\frac{\partial L(x,\lambda)}{\partial x} = Cx - \lambda x = 0$$

$$\Downarrow$$

$$Cx = \lambda x$$

Other principle components

- for the second principle axis:
  - maximize the variance
  - uncorrelated (orthogonal) with $x_1^\top A$

$$\mathrm{cov}(x_1^\top A, x_2^\top A) = x_1^\top A x_2 = x_2^\top A x_1 = \lambda x_1^\top x_2 = 0$$

- then we are going to solve

$$\max_{x^\top x = 1, x_1^\top x_2 = 0} \quad x^\top A x$$

- similarly, consider its Langrangian

$$L(x, \lambda) = x^\top A x - \lambda(x^\top x - 1) - \beta x_1^\top w$$

Other principle components

- consider the derivatives of $L(x, \lambda) = x^\top A x - \lambda(x^\top x - 1) - \beta x_1^\top x$

$$\frac{\partial L(x, \lambda)}{\partial x} = Cx - \lambda x - \beta x_1 = 0$$

- multiply by $x_1^\top$ on the left, we have

$$\boxed{x_1^\top C x - \lambda x_1^\top x} - \beta \boxed{x_1^\top x_1} = 0 \quad \rightarrow \quad \beta = 0$$

$$Cx = \lambda x$$

both equal zero, because of the un-correlation

nonzero

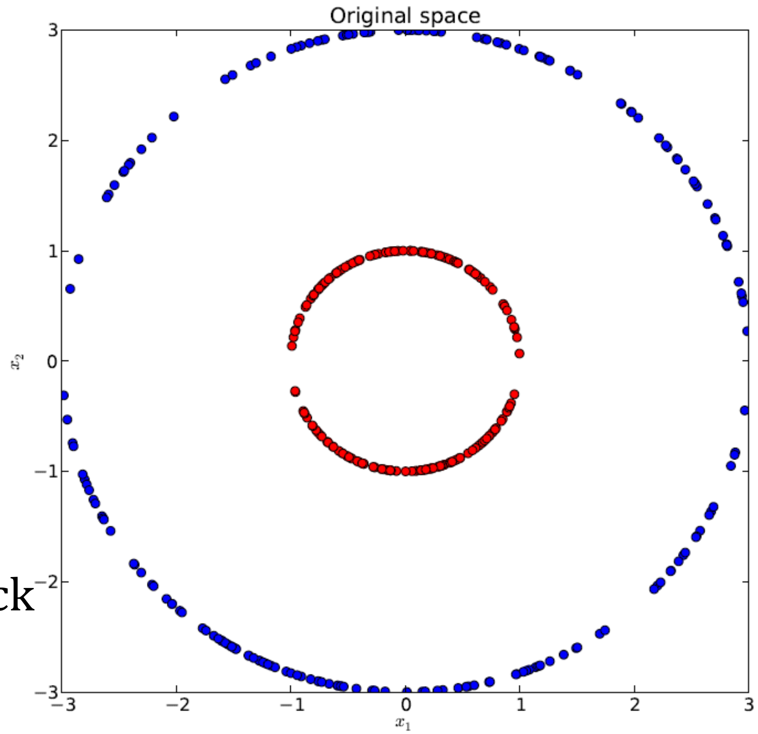
Nonlinear PCA

- PCA is to find linear subspace

- can we extend PCA to nonlinear problems?
  - the previous PCA is in primal
  - can we go to dual space and use *kernel trick*?

$$C = \frac{1}{n-1}\boxed{A^{\top}A}$$

it is possible
to use kernel trick



PCA in terms of dot products

- the eigenvectors lie in the span of $a_1, a_2, \dots, a_m$
- ***Proof.***

$$Cx = \frac{1}{m}\sum_{j=1}^{m} a_j a_j^T x = \lambda x$$

Therefore,

$$x = \frac{1}{\lambda x}\sum_{j=1}^{m} a_j a_j^\top x$$

$$= \frac{1}{\lambda x}\sum_{j=1}^{m} \boxed{(a_j \cdot x) a_j}$$

$$(aa^\top)x = (a \cdot x)a$$

Show that $(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{v} = (\boldsymbol{x} \cdot \boldsymbol{v})\boldsymbol{x}$

$$
\begin{aligned}
(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{v} &= \begin{pmatrix} x_1x_1 & x_1x_2 & \dots & x_1x_M \\ x_2x_1 & x_2x_2 & \dots & x_2x_M \\ \vdots & \vdots & \ddots & \vdots \\ x_Mx_1 & x_Mx_2 & \dots & x_Mx_M \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_M \end{pmatrix} \\
&= \begin{pmatrix} x_1x_1v_1 + x_1x_2v_2 + \ldots + x_1x_Mv_M \\ x_2x_1v_1 + x_2x_2v_2 + \ldots + x_2x_Mv_M \\ \vdots \\ x_Mx_1v_1 + x_Mx_2v_2 + \ldots + x_Mx_Mv_M \end{pmatrix}
\end{aligned}
$$

$$= \begin{pmatrix} (x_1 v_1 + x_2 v_2 + \ldots + x_M v_M)\, x_1 \\ (x_1 v_1 + x_2 v_2 + \ldots + x_M v_M)\, x_2 \\ \vdots \\ (x_1 v_1 + x_2 v_2 + \ldots + x_M v_M)\, x_M \end{pmatrix}$$

$$= \begin{pmatrix} x_1 v_1 + x_2 v_2 + \ldots + x_M v_M \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{pmatrix}$$

$$= (\boldsymbol{x} \cdot \boldsymbol{v})\boldsymbol{x} \qquad \square$$

Nonlinear feature mapping

- we now can apply a nonlinear feature mapping $\phi(a)$
- then matrix $\Phi = \phi(A)$ (and assume it is centered)
- its principle component can be calculated as linear PCA:

$$x = \sum_{i=1}^{m} \alpha_i \phi(a_i)$$

$$Cx = \frac{1}{m}\sum_{j=1}^{m} \phi(a_j)\phi(a_j)^{\top} x = \lambda x$$

as showed previously, the solutions lie in **the span** of $\phi(a_i)$:

$$\frac{1}{m}\sum_{j=1}^{m} \phi(a_j)\phi(a_j)^{\top} \sum_{i=1}^{m} \alpha_i \phi(a_i) = \lambda \sum_{i=1}^{m} \alpha_i \phi(a_i)$$

Nonlinear feature mapping

- Kernel trick:

$$\boxed{\sum_{j=1}^{m} \phi(a_j)\phi(a_j)^T} \sum_{i=1}^{m} \alpha_i \phi(a_i) = m\lambda \boxed{\sum_{i=1}^{m} \alpha_i \phi(a_i)}$$

kernel trick

- Again, we do not need to know the feature mapping:

eigenvector

$$\phi(a)^\top x = \phi(a)^\top \sum_{i=1}^{m} \alpha_i \phi(a_i) = \sum_{i=1}^{m} \alpha_i K(a, a_i)$$

Kernel PCA

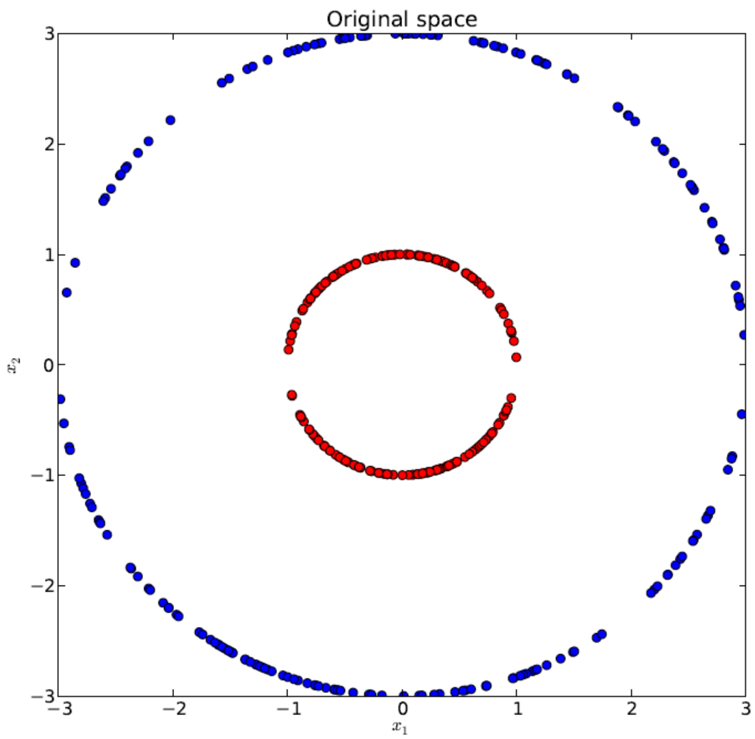
- calculate the kernel matrix K
- centralize the kernel matrix

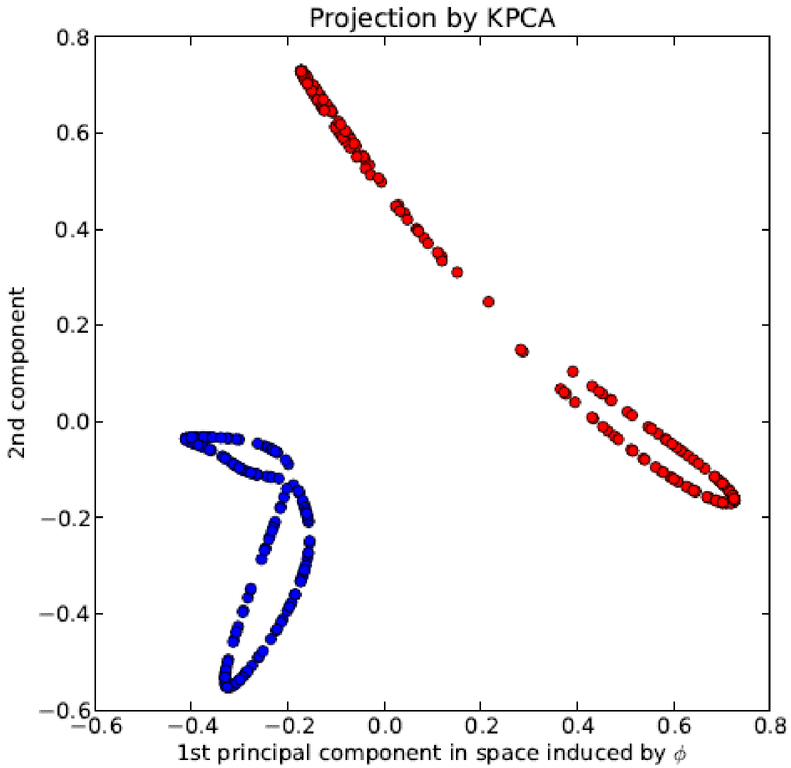
$$\widehat{K} = K - \frac{1}{n}11^T K - \frac{1}{n}K11^T + \frac{1^T K1}{n^2}11^T$$

- egien-value decomposition: $[U, V] = \text{eig}(\widehat{K})$
- find dual variables: $\alpha_i = \lambda_j^{-\frac{1}{2}} v_j$
- projection onto subspace:

$$\sum \alpha_{ji} K(a_i, a)$$

Kernel PCA


Original space
x_1
x_2


Projection by KPCA
1st principal component in space induced by ϕ
2nd component

Kernel PCA and PCA

- If we choose linear kernel: $C = A^\top A,\ A = AA^\top$
- PCA:

$$\boxed{C}w = \lambda w$$

$n \times n$

- KPCA:

$$\boxed{K}\beta = \mu\beta$$

$m \times m$

- the projected data

$$Ax = \sum \alpha_{ji} K(a_i, a)$$

Recall the PSD condition

- before we always requires the kernel matrix is PSD
- are the previous algorithm applicable?
  - SMO for SVM

$$\min_{\alpha} \sum_i \sum_j \alpha_i b_i K_{ij} b_j \alpha_j - \sum_i \alpha_i$$

$$\text{s.t.} \quad \sum_i \alpha_i b_i = 0$$

$$0 \le \alpha_i \le C$$

  - inverse problem for LS-SVM
  - eigenvalue for kPCA

$$K\alpha = \mu\alpha$$

Recall the PSD condition

- before we always requires the kernel matrix is PSD
- are the previous algorithm applicable?
  - SMO for SVM

$$\min_{\alpha} \sum_i \sum_j \alpha_i b_i K_{ij} b_j \alpha_j - \sum_i \alpha_i$$

$$\text{s.t.} \quad \sum_i \alpha_i b_i = 0$$

$$0 \leq \alpha_i \leq C$$

yes, but local optimality

  - inverse problem for LS-SVM — yes, solvable
  - eigenvalue for kPCA — yes, solvable

$$K\alpha = \mu\alpha$$

Recall the PSD condition

- before we always requires the kernel matrix is PSD
- does the primal-dual relationship exist?
  - SVM

$$\begin{aligned} \min_{x,z} \quad & \tfrac{1}{2}\|x\|_2^2 + C\textstyle\sum_i \rho_i \\ \text{s.t.} \quad & b_i(x^\top \phi(a_i) + z) \geq 1 - \rho_i, \\ & \rho_i \geq 0 \end{aligned}$$

$$\begin{aligned} \min_{\alpha} \quad & \sum_i \sum_j \alpha_i b_i K_{ij} b_j \alpha_j - \sum_i \alpha_i \\ \text{s.t.} \quad & \textstyle\sum_i \alpha_i b_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

  - if not, what is the relationship between them?

目录 Contents

Homework

- 5.29
- Consider the following problem

$$\begin{aligned} \min_{x,z} \quad & \frac{1}{2}\|x\|_2^2 - \nu\xi + C\sum_i \rho_i \\ \text{s.t.} \quad & b_i(x^\top \phi(a_i) + z) \geq \xi - \rho_i, \forall i, \\ & \xi \geq 0, \rho_i \geq 0, \forall i \end{aligned}$$

you are asked to prove that ν is an upper bound on the fraction of margin errors, i.e., the number of samples falling in the margin is less than νm, i.e.,

$$\#\{i: y_i f(x_i) < \rho\} \leq \nu m$$

THANKS