2. Selected optimization problems in data science

Maryam Fazel, Univ. of Washington

A sampling of (convex) optimization problems in learning and statistics:

- regression, penalty fcts, and regularization
- classification and SVM
- maximum-likelihood and logistic regression
- optimal experiment design

Selected optimization problems in data science

A sampling of (convex) optimization problems in learning and statistics:

- regression, penalty fcts, and regularization
- classification and SVM
- maximum-likelihood and logistic regression
- optimal experiment design

For nonconvex examples, see deep learning lecture (also workshop next week)

References: Convex Optimization, Boyd & Vandenberghe, Cambridge Publishing, 2004. Chapters 6, 7, 8. Other refs cited throughout.

Data fitting and regression

$$\text{minimize} \quad \|Ax - b\|$$

where $A \in \mathbf{R}^{m \times n}$, $\|\cdot\|$ is a norm on $\mathbf{R}^m$. interpretations of solution $x^\star$:

- **approximation**: $Ax^\star$ is the best approximation of b (observations, labels) by a linear combination of columns of A (features)

- **geometric**: $Ax^\star$ is point in $\mathcal{R}(A)$ closest to b

- **estimation**: linear measurement model

$$y = Ax + v$$

  y are measurements, x is unknown, v is measurement error

  given $y = b$, best guess of x is $x^\star$

Penalty function approximation

$$\begin{array}{ll} \text{minimize} & \phi(r_1) + \cdots + \phi(r_m) \\ \text{subject to} & r = Ax - b \end{array}$$

($A \in \mathbf{R}^{m \times n}$, $\phi : \mathbf{R} \to \mathbf{R}$ is a convex penalty function)

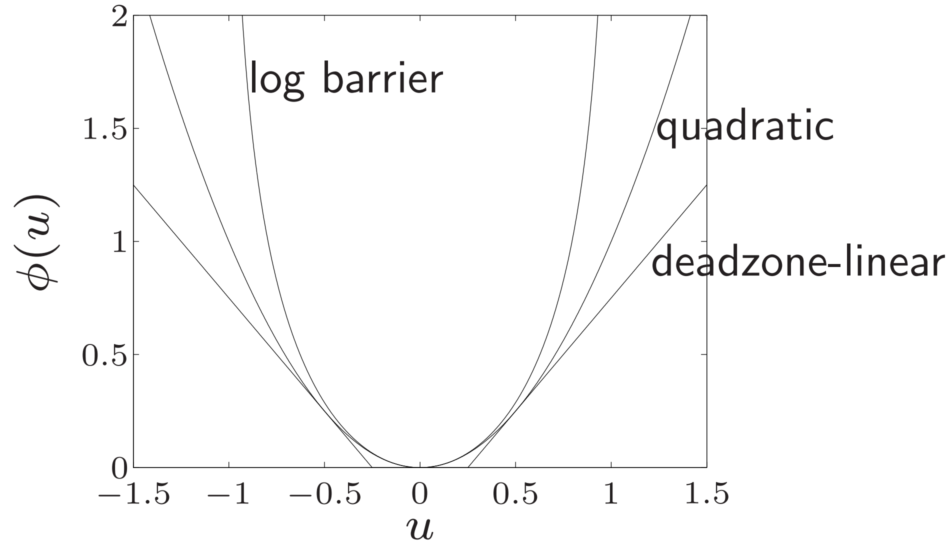
examples

- abs value: $\phi(u) = |u|$ (ℓ_1 norm)
- quadratic: $\phi(u) = u^2$
- deadzone-linear with width a:

$$\phi(u) = \max\{0, |u| - a\}$$

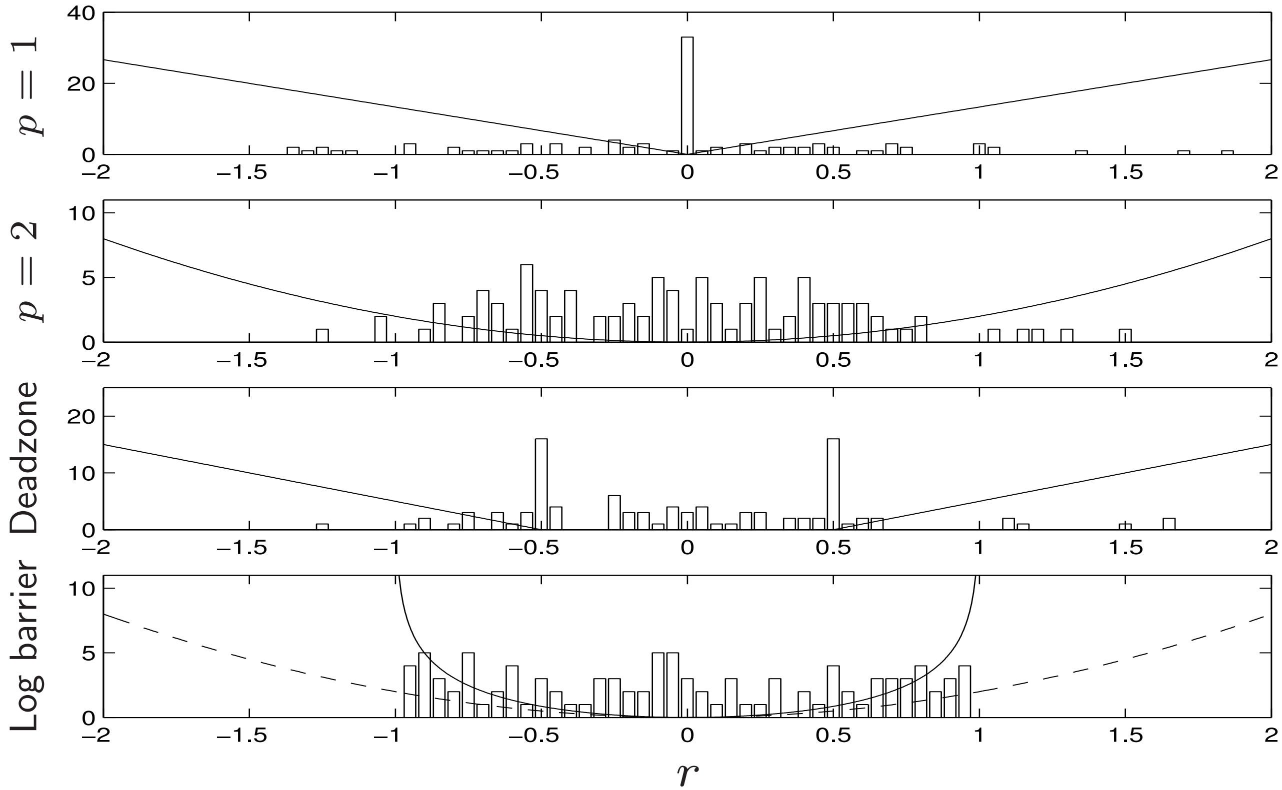
- log-barrier with limit a:

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



example ($m = 100$, $n = 30$): histogram of residuals for penalties

$$\phi(u) = |u|, \quad \phi(u) = u^2, \quad \phi(u) = \max\{0, |u| - a\}, \quad \phi(u) = -\log(1 - u^2)$$

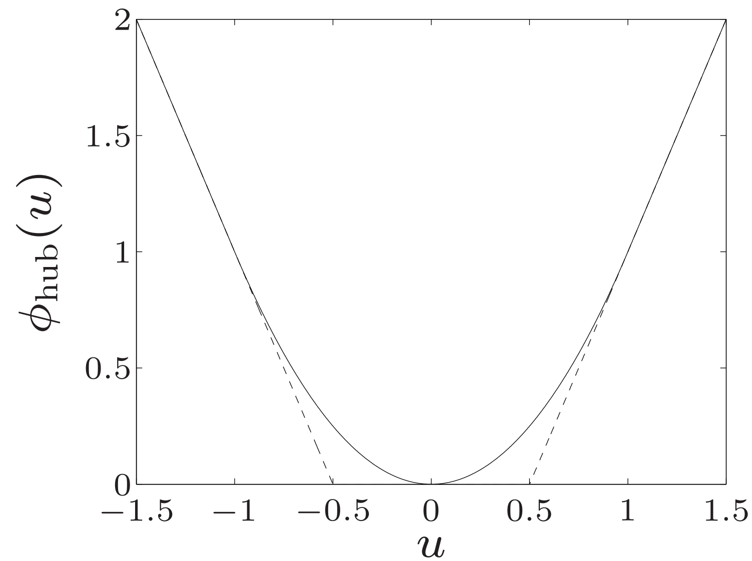


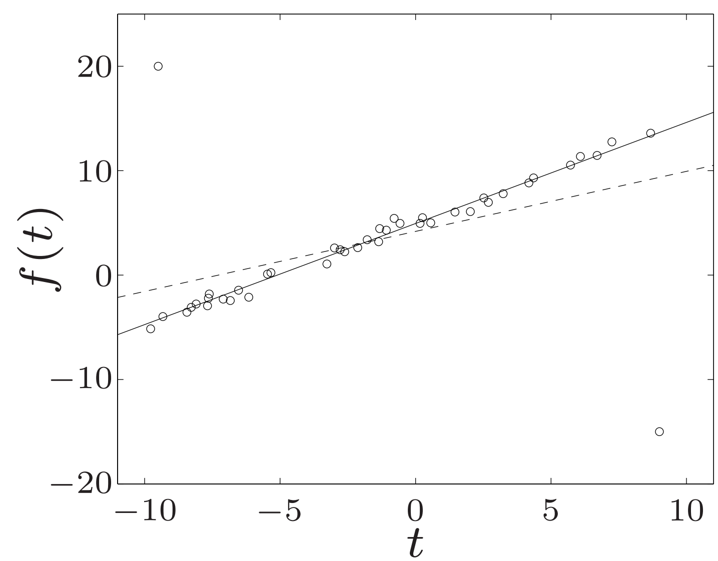
shape of penalty function has large effect on distribution of residuals

Huber penalty function (with parameter M)

$$\phi_{\mathrm{hub}}(u) = \begin{cases} u^2 & |u| \leq M \\ M(2|u| - M) & |u| > M \end{cases}$$

linear growth for large u makes approximation less sensitive to outliers





- left: Huber penalty for $M = 1$
- right: affine function $f(t) = \alpha + \beta t$ fitted to 42 points t_i, y_i (circles) using quadratic (dashed) and Huber (solid) penalty

Regularized regression

$$\text{minimize} \quad \|Ax - b\| + \gamma\|x\|$$

$A \in \mathbf{R}^{m \times n}$, norms on $\mathbf{R}^m$ and $\mathbf{R}^n$ can be different

interpretation: find good approximation $Ax \approx b$ with small x

solution for $\gamma > 0$ traces out optimal trade-off curve

Tikhonov regularization

$$\text{minimize} \quad \|Ax - b\|_2^2 + \delta\|x\|_2^2$$

can be solved as a least-squares problem

$$\text{minimize} \quad \left\| \begin{bmatrix} A \\ \sqrt{\delta} I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2$$

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

Regularization for sparsity

$$\text{minimize} \quad \|Ax - b\|_2^2 + \gamma \|x\|_1$$

- ℓ_1-norm encourages sparsity [Logan '65; Claerbout,Muir '73; Rudin,Osher,Fatemi '92,...]

- very broadly used: compressed sensing; LASSO, group-LASSO, and variations; Graphical LASSO;. . .

- statistical guarantees (under some assumptions on A) [Candes,Romberg,Tao '04; Donoho '04],. . .

(see Po-Ling's lecture)

Regularization for low-rank

nuclear norm encourages low-rank $X \in \mathbf{R}^{m \times n}$ [F.,Hindi,Boyd '01; F. '02]:

$$\|X\|_* = \sum_i \sigma_i(X)$$

where $\sigma_i(X)$ is the ith singular value

$$\text{minimize} \quad \|\mathcal{A}(X) - b\|_2^2 + \gamma \|X\|_*$$

where $\mathcal{A} : \mathbf{R}^{m \times n} \mapsto \mathbf{R}^p$ is a linear map

- **matrix completion:** observe a subset of entries
- *Netflix prize* (2009), 17770 movies x 480189 users
- theoretical guarantees (under suitable assumptions): [Recht, F., Parrilo '10; Candes,Recht '09,. . .]

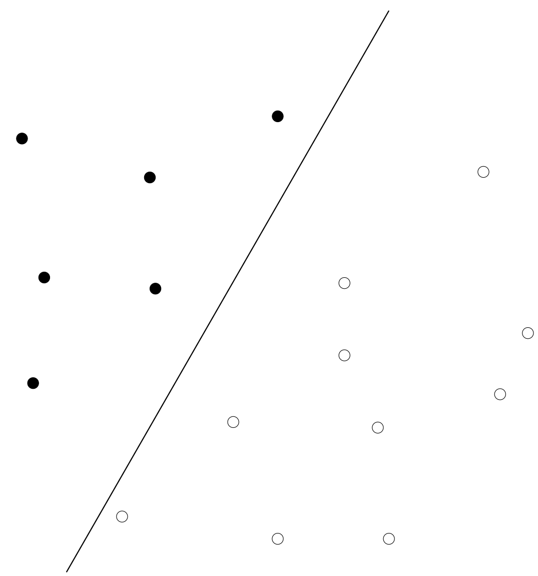
many applications: recommender systems, low-rank subspace tracking, low-dim Euclidean embedding, dynamical system identification, cluster detection,. . .

Classification and Support Vector Machines

Linear classification

separate two sets of points $\{x_1, \ldots, x_N\}$, $\{y_1, \ldots, y_M\}$ by a hyperplane:

$$a^T x_i + b > 0, \quad i = 1, \ldots, N, \qquad a^T y_i + b < 0, \quad i = 1, \ldots, M$$



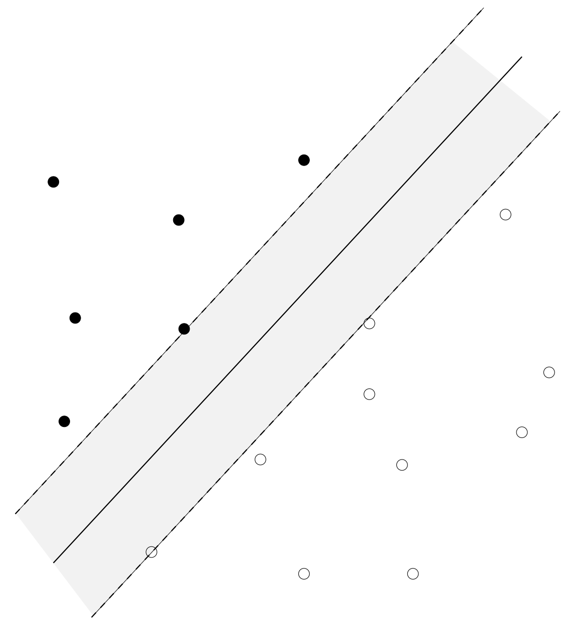
homogeneous in a, b, hence equivalent to

$$a^T x_i + b \geq 1, \quad i = 1, \ldots, N, \qquad a^T y_i + b \leq -1, \quad i = 1, \ldots, M$$

a set of linear inequalities in a, b

Robust linear classification

(Euclidean) distance between hyperplanes



$$\begin{aligned}
\mathcal{H}_1 &= \{z \mid a^T z + b = 1\} \\
\mathcal{H}_2 &= \{z \mid a^T z + b = -1\}
\end{aligned}$$

is $\mathbf{dist}(\mathcal{H}_1, \mathcal{H}_2) = 2/\|a\|_2$

to separate two sets of points by maximum margin,

$$\begin{array}{ll}
\text{minimize} & (1/2)\|a\|_2 \\
\text{subject to} & a^T x_i + b \geq 1, \quad i = 1, \ldots, N \\
& a^T y_i + b \leq -1, \quad i = 1, \ldots, M
\end{array} \tag{1}$$

(after squaring objective) a QP in a, b

Lagrange dual of maximum margin separation problem (1)

$$
\begin{array}{ll}
\text{maximize} & \mathbf{1}^T\lambda + \mathbf{1}^T\mu \\
\text{subject to} & 2\left\|\sum_{i=1}^N \lambda_i x_i - \sum_{i=1}^M \mu_i y_i\right\|_2 \leq 1 \\
& \mathbf{1}^T\lambda = \mathbf{1}^T\mu, \quad \lambda \succeq 0, \quad \mu \succeq 0
\end{array}
\qquad (2)
$$

from duality, optimal value is inverse of maximum margin of separation

interpretation

- change variables to $\theta_i = \lambda_i/\mathbf{1}^T\lambda$, $\gamma_i = \mu_i/\mathbf{1}^T\mu$, $t = 1/(\mathbf{1}^T\lambda + \mathbf{1}^T\mu)$
- invert objective to minimize $1/(\mathbf{1}^T\lambda + \mathbf{1}^T\mu) = t$

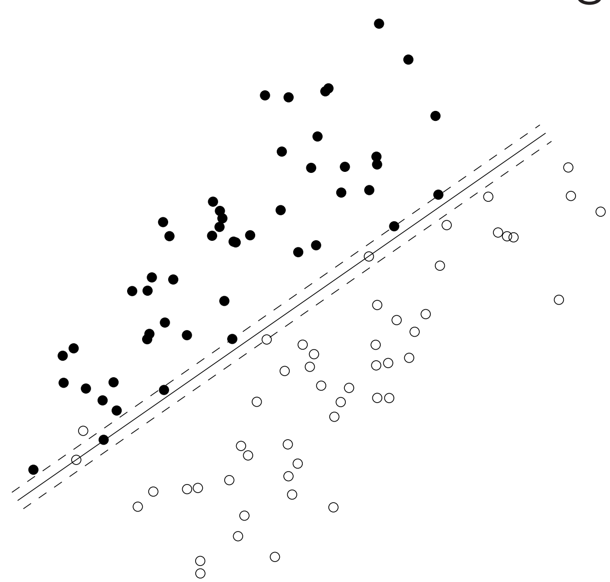
$$
\begin{array}{ll}
\text{minimize} & t \\
\text{subject to} & \left\|\sum_{i=1}^N \theta_i x_i - \sum_{i=1}^M \gamma_i y_i\right\|_2 \leq t \\
& \theta \succeq 0, \quad \mathbf{1}^T\theta = 1, \quad \gamma \succeq 0, \quad \mathbf{1}^T\gamma = 1
\end{array}
$$

optimal value is distance between convex hulls

Approximate linear separation of non-separable sets

$$
\begin{array}{ll}
\text{minimize} & \mathbf{1}^T u + \mathbf{1}^T v \\
\text{subject to} & a^T x_i + b \geq 1 - u_i, \quad i = 1, \ldots, N \\
& a^T y_i + b \leq -1 + v_i, \quad i = 1, \ldots, M \\
& u \succeq 0, \quad v \succeq 0
\end{array}
$$

- an LP in a, b, u, v
- at optimum, $u_i = \max\{0, 1 - a^T x_i - b\}$, $v_i = \max\{0, 1 + a^T y_i + b\}$
- can be interpreted as a heuristic for minimizing #misclassified points

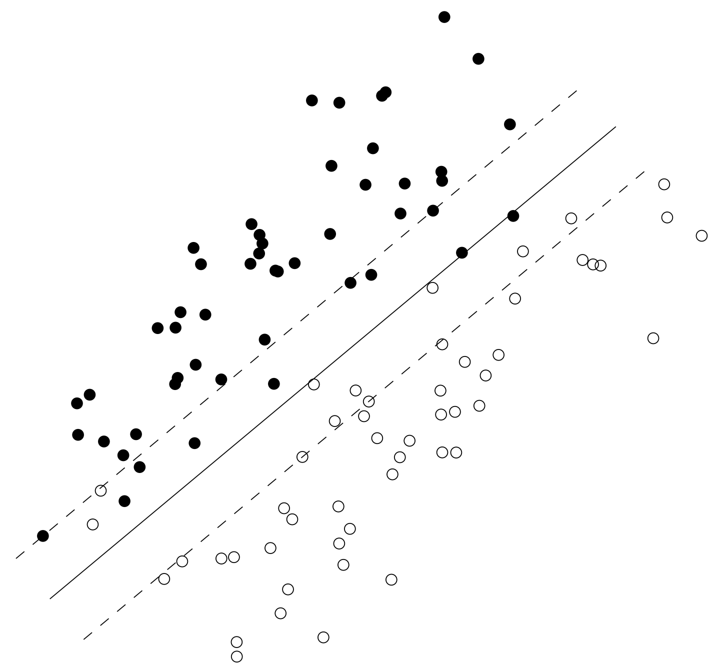


Support vector machine

$$\begin{array}{ll} \text{minimize} & \|a\|_2 + \gamma(\mathbf{1}^T u + \mathbf{1}^T v) \\ \text{subject to} & a^T x_i + b \geq 1 - u_i, \quad i = 1, \ldots, N \\ & a^T y_i + b \leq -1 + v_i, \quad i = 1, \ldots, M \\ & u \succeq 0, \quad v \succeq 0 \end{array}$$

produces point on trade-off curve between inverse of margin $2/\|a\|_2$ and classification error, measured by total slack $\mathbf{1}^T u + \mathbf{1}^T v$

same example as previous page, with $\gamma = 0.1$:



Nonlinear classification

separate two sets of points by a nonlinear function:

$$f(x_i) > 0, \quad i = 1, \dots, N, \qquad f(y_i) < 0, \quad i = 1, \dots, M$$

- choose a linearly parametrized family of functions

$$f(z) = \theta^T F(z)$$

  $F = (F_1, \dots, F_k) : \mathbf{R}^n \to \mathbf{R}^k$ are basis functions

- solve a set of linear inequalities in θ:

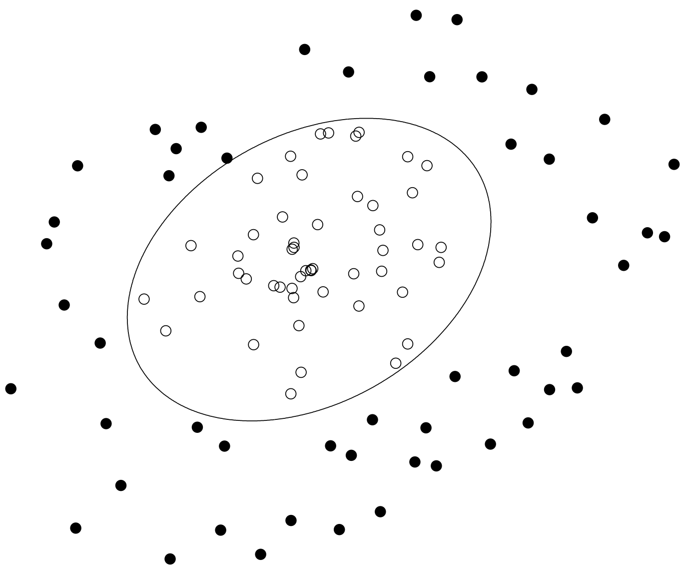
$$\theta^T F(x_i) \geq 1, \quad i = 1, \dots, N, \qquad \theta^T F(y_i) \leq -1, \quad i = 1, \dots, M$$

quadratic discrimination: $f(z) = z^T P z + q^T z + r$

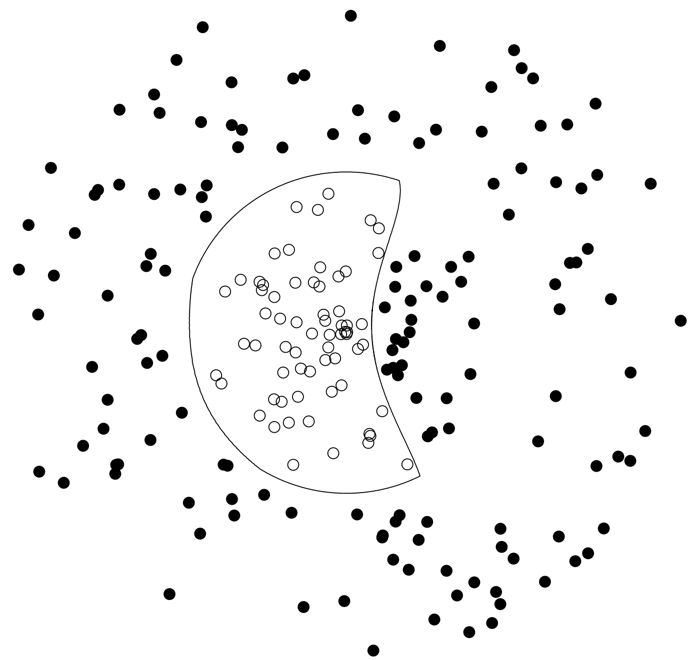
$$x_i^T P x_i + q^T x_i + r \geq 1, \qquad y_i^T P y_i + q^T y_i + r \leq -1$$

can add additional constraints ($e.g.$, $P \preceq -I$ to separate by an ellipsoid)

polynomial discrimination: $F(z)$ are all monomials up to a given degree



separation by ellipsoid



separation by 4th degree polynomial

Maximum-likelihood estimation, logistic regression

Parametric distribution estimation

- distribution estimation problem: estimate probability density $p(y)$ of a random variable from observed values
- parametric distribution estimation: choose from a family of densities $p_x(y)$, indexed by a parameter x

maximum likelihood estimation

$$\text{maximize (over } x) \quad \log p_x(y)$$

- y is observed value
- $l(x) = \log p_x(y)$ is called log-likelihood function
- can add constraints $x \in C$ explicitly, or define $p_x(y) = 0$ for $x \notin C$
- a convex optimization problem if $\log p_x(y)$ is concave in x for fixed y

Linear measurements with IID noise

linear measurement model

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

- $x \in \mathbf{R}^n$ is vector of unknown parameters
- v_i is IID measurement noise, with density $p(z)$
- y_i is measurement: $y \in \mathbf{R}^m$ has density $p_x(y) = \prod_{i=1}^m p(y_i - a_i^T x)$

maximum likelihood estimate: any solution x of

$$\text{maximize} \quad l(x) = \sum_{i=1}^m \log p(y_i - a_i^T x)$$

(y is observed value)

examples

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

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

  ML estimate is LS solution

- Laplacian noise: $p(z) = (1/(2a))e^{-|z|/a}$,

$$l(x) = -m\log(2a) - \frac{1}{a}\sum_{i=1}^{m}|a_i^T x - y_i|$$

  ML estimate is ℓ_1-norm solution

- uniform noise on $[-a, a]$:

$$l(x) = \begin{cases} -m\log(2a) & |a_i^T x - y_i| \le a, \quad i = 1, \ldots, m \\ -\infty & \text{otherwise} \end{cases}$$

  ML estimate is any x with $|a_i^T x - y_i| \le a$

Logistic regression

random variable $y \in \{0, 1\}$ with distribution

$$p = \mathbf{prob}(y = 1) = \frac{\exp(a^T u + b)}{1 + \exp(a^T u + b)}$$

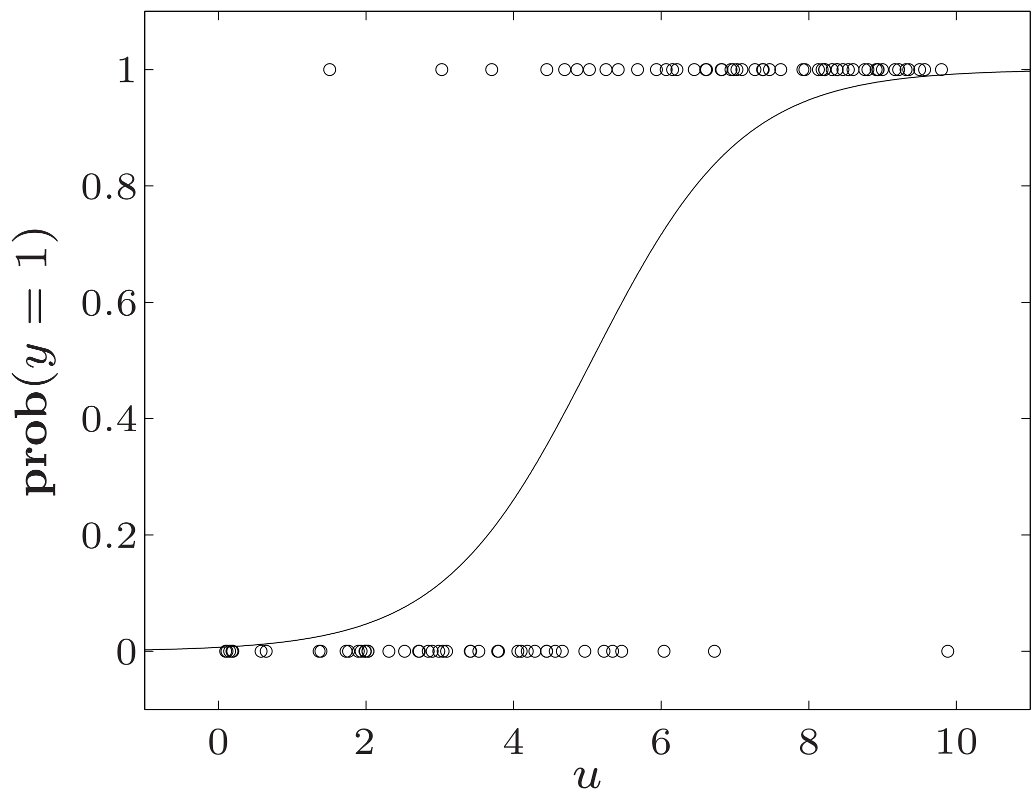
- a, b are parameters; $u \in \mathbf{R}^n$ are (observable) explanatory variables
- example: p is probability of acquiring a certain disease; variables u include weight, age, blood pressure, . . .
- estimation problem: estimate a, b from m observations (u_i, y_i)

log-likelihood function (for $y_1 = \cdots = y_k = 1$, $y_{k+1} = \cdots = y_m = 0$):

$$
\begin{aligned}
l(a,b) &= \log \left(\prod_{i=1}^{k} \frac{\exp(a^T u_i + b)}{1 + \exp(a^T u_i + b)} \prod_{i=k+1}^{m} \frac{1}{1 + \exp(a^T u_i + b)} \right) \\
&= \sum_{i=1}^{k} (a^T u_i + b) - \sum_{i=1}^{m} \log(1 + \exp(a^T u_i + b))
\end{aligned}
$$

concave in a, b

example ($n = 1$, $m = 50$ measurements)



- circles show 50 points (u_i, y_i)
- solid curve is ML estimate of $p = \exp(au + b)/(1 + \exp(au + b))$

regularization on a, b, constraints are also common

Optimal experiment design

Experiment design

m linear measurements $y_i = a_i^T x + w_i$, $i = 1, \ldots, m$ of unknown $x \in \mathbf{R}^n$

- measurement errors w_i are IID $\mathcal{N}(0, 1)$
- ML (least-squares) estimate is

$$\hat{x} = \left(\sum_{i=1}^{m} a_i a_i^T \right)^{-1} \sum_{i=1}^{m} y_i a_i$$

- error $e = \hat{x} - x$ has zero mean and covariance

$$E = \mathbf{E}\, e e^T = \left(\sum_{i=1}^{m} a_i a_i^T \right)^{-1}$$

confidence ellipsoids are given by $\{x \mid (x - \hat{x})^T E^{-1} (x - \hat{x}) \leq \beta\}$

experiment design: choose $a_i \in \{v_1, \ldots, v_p\}$ (a set of possible test vectors) to make E 'small' optimization over the positive-semidefinite cone:

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{S}_+^n) & E = \left(\sum_{k=1}^p m_k v_k v_k^T\right)^{-1} \\ \text{subject to} & m_k \geq 0, \quad m_1 + \cdots + m_p = m \\ & m_k \in \mathbf{Z} \end{array}$$

- variables are m_k ($\#$ vectors a_i equal to v_k)
- difficult in general, due to integer constraint

relaxed experiment design

assume $m \gg p$, use $\lambda_k = m_k/m$ as (continuous) real variable

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{S}_+^n) & E = (1/m)\left(\sum_{k=1}^p \lambda_k v_k v_k^T\right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1 \end{array}$$

- common objectives to minimize: $\log\det E$, $\mathbf{tr}\, E$, $\lambda_{\max}(E)$, . . .
- can add other convex constraints, *e.g.*, bound experiment cost $c^T\lambda \leq B$

D-optimal design

$$\begin{array}{ll} \text{minimize} & \log\det\left(\sum_{k=1}^p \lambda_k v_k v_k^T\right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1 \end{array}$$

interpretation: minimizes volume of confidence ellipsoids

dual problem

$$\begin{array}{ll} \text{maximize} & \log\det W + n\log n \\ \text{subject to} & v_k^T W v_k \le 1, \quad k = 1,\ldots,p \end{array}$$

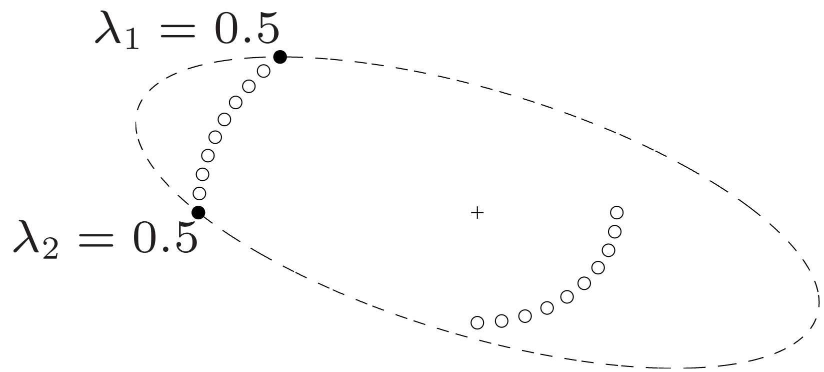
interpretation: $\{x \mid x^T W x \le 1\}$ is minimum volume ellipsoid centered at origin, that includes all test vectors v_k

complementary slackness: for λ, W primal and dual optimal

$$\lambda_k(1 - v_k^T W v_k) = 0, \quad k = 1,\ldots,p$$

optimal experiment uses vectors v_k on boundary of ellipsoid defined by W

example $(p = 20)$



design uses two vectors, on boundary of ellipse defined by optimal W

derivation of dual

first reformulate primal problem with new variable X:

$$
\begin{array}{ll}
\text{minimize} & \log\det X^{-1} \\
\text{subject to} & X = \sum_{k=1}^{p} \lambda_k v_k v_k^T, \quad \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1
\end{array}
$$

$$
L(X,\lambda,Z,z,\nu) = \log\det X^{-1} + \mathbf{tr}\left(Z\left(X - \sum_{k=1}^{p} \lambda_k v_k v_k^T\right)\right) - z^T\lambda + \nu(\mathbf{1}^T\lambda - 1)
$$

- minimize over X by setting gradient to zero: $-X^{-1} + Z = 0$
- minimum over λ_k is $-\infty$ unless $-v_k^T Z v_k - z_k + \nu = 0$

dual problem

$$
\begin{array}{ll}
\text{maximize} & n + \log\det Z - \nu \\
\text{subject to} & v_k^T Z v_k \leq \nu, \quad k = 1,\ldots,p
\end{array}
$$

change variable $W = Z/\nu$, and optimize over ν to get dual of page 2–28

Summary

- covered a few classes of problems, there are many more
- also many nonconvex (see, e.g., deep learning lecture)
- optimization problems are at the center of machine learning and data science

Questions?