# *ECE 602 – Section 7*
# *Practical Applications*

- Approximation and fitting
- Norm minimization problems
- Signal reconstruction and regularization
- Statistical estimation and Maximum Likelihood
- Minimal and maximal volume ellipsoids
- Optimal experiment design
- Support Vector Machines

- The simplest *norm approximation problem* is

$$\min_x \|Ax - b\|$$

where $A \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^m$ are given.

- $x^*$ is called an *approximate solution* to $Ax \approx b$ in the norm $\|\cdot\|$.

- The vector

$$r = Ax - b$$

is called the *residual* for the problem.

- It is a *convex optimization problem*, and hence there is at least one optimal solution.

- **Geometry**

  By expressing $Ax = a_1x_1 + a_2x_2 + \ldots + a_nx_n$, we see that $x^*$ is the closest point in $\mathcal{R}(A)$ w.r.t. $b$ (the *regression problem*).

- **Estimation**

  Considering the linear measurement model $b = Ax + v$, $x^*$ is the most plausible guess for $x$, given $b$.

- **Optimal design**

  Considering $x_1, \ldots, x_n$ to be *design variables*, we can view $Ax$ as a vector of $m$ *results*. So, $x^*$ is the *best design* which approximates a *desired result or target* $b$.

- We can also consider the *weighted norm approximation problem*

$$\min_x \|W(Ax - b)\|$$

where $W \in \mathbf{R}^{m \times m}$ is a *weighting matrix* (usually, $W \succeq 0$).

- The *least-squares approximation problem* is defined as

$$\min_x \|Ax - b\|_2^2 = \sum_{i=1}^{m} |r_i|^2$$

- The problem can be solved analytically through differentiating

$$f(x) = \|Ax - b\|_2^2 = x^T A^T A x - 2 b^T A x + b^T b$$

and setting $\nabla f(x) = 2A^T A x - A^T b = 0$, which results in *a system of normal equations*

$$A^T A x = A^T b$$

with the unique solution $x^* = (A^T A)^{-1} A^T b = A^\dagger b$.

- The *minimax* (aka *Chebyshev*) *approximation problem* is defined as

$$\min_x \|Ax - b\|_\infty = \max\{|r_1|, \ldots, |r_m|\}$$

- The Chebyshev approximation problem can be cast as an LP

$$\begin{aligned} &\min_{x,t} \; t \\ &\text{s.t.} \; -t\mathbf{1} \preceq Ax - b \preceq t\mathbf{1} \end{aligned}$$

with variables $x \in \mathbf{R}^n$ and $t \in \mathbf{R}$.

- The *sum of (absolute) residuals approximation problem* is defined as

$$\min_x \|Ax - b\|_1 = |r_1| + |r_2| + \ldots + |r_m|$$

- In the context of estimation, it is also know as a *robust estimator*.
- The $\ell_1$-norm approximation problem can be cast as an LP

$$\begin{aligned} &\min_{x,t} \ \mathbf{1}^T t \\ &\text{s.t.} \ -t \preceq Ax - b \preceq t \end{aligned}$$

with variables $x \in \mathbf{R}^n$ and $t \in \mathbf{R}^m$.

- The *penalty function approximation problem* has the form

$$\min_x \; \phi(r_1) + \phi(r_2) + \ldots + \phi(r_m)$$
$$\text{s.t. } r = Ax - b$$

where $\phi : \mathbf{R} \to \mathbf{R}$ is called the *(residual) penalty function*.

- When $\phi$ is convex, the penalty function approximation problem is a convex optimization problem.
- In many cases, $\phi$ is symmetric, nonnegative, and satisfies $\phi(0) = 0$.
- Roughly speaking, $\phi(u)$ is a measure of our "dislike" of a residual.
- The shape of $\phi(u)$ has profound effect on the distribution of residuals.

- *Quadratic penalty*

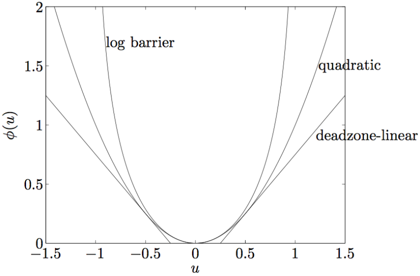
$$\phi(u) = |u|^2$$

- *Deadzone-linear penalty*

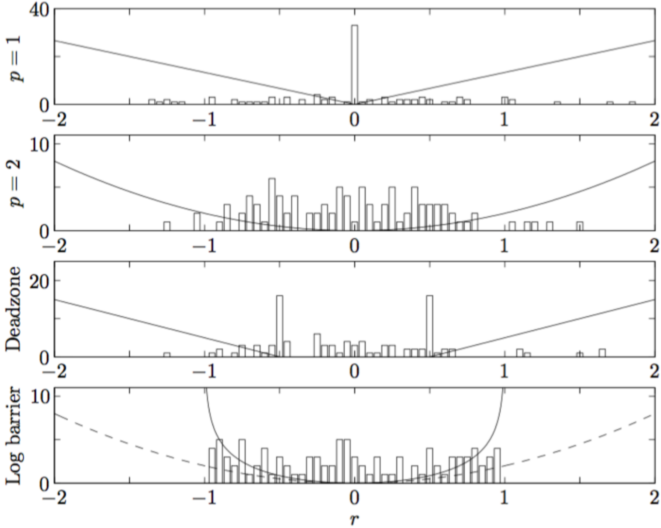
$$\phi(u) = \max\{u - a, 0\}, \;\; a > 0$$

- *Log-barier penalty*

$$\phi(u) = \begin{cases} -a^2 \log\left(1 - \frac{u^2}{a^2}\right), & |u| < a \\ \infty, & |u| \geq a \end{cases}$$



Consider a norm/penalty minimization problem of size $m = 100$ and $n = 30$.

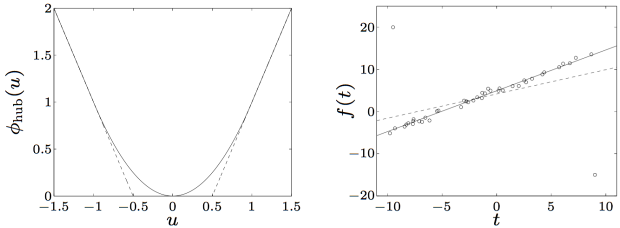


- The *robust least-squares* or *Huber penalty function* is defined as

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

- Given $\{(t_i, y_i)\}_{i=1}^{42}$, consider the following two regression problems

$$\min_{\alpha,\beta} \sum_{i=1}^{42} (y_i - \alpha - \beta t_i)^2 \quad \text{versus} \quad \min_{\alpha,\beta} \sum_{i=1}^{42} \phi_{\text{hub}}(y_i - \alpha - \beta t_i)$$




- We see that the *robust regression* (solid line) is much less sensitive to the effect of *outliers*.

- *Norm minimization with non-negativity constraints*

$$\min_x \|Ax - b\| \quad \text{s.t. } x \succeq 0$$

- *Norm minimization with "box" constraints*

$$\min_x \|Ax - b\| \quad \text{s.t. } l \preceq x \preceq u$$

- *Norm minimization over probability simplex*

$$\min_x \|Ax - b\| \quad \text{s.t. } x \succeq 0, \quad \mathbf{1}^T x = 1$$

- *Norm minimization with ball constraints*

$$\min_x \|Ax - b\| \quad \text{s.t. } \|x - x_0\| \leq d$$

- The basic *least-norm problem* has the form

$$\min_x \|x\| \quad \text{s.t. } Ax = b$$

where $b \in \mathbf{R}^m$ and $A \in \mathbf{R}^{m \times n}$ $(n > m)$ are problem data.

- **Design interpretation:** $x$ are *design variables*; $b$ are required results, $x^*$ is smallest ("most efficient") design that satisfies requirements.
- **Estimation interpretation:** $b = Ax$ are (perfect) measurements of $x$; $x^*$ is smallest ("most plausible") estimate consistent with $b$.
- **Geometric interpretation:** $x^*$ is a point in the affine set $\{x | Ax = b\}$ with minimum distance to 0.

- The most common least-norm problem is the *least $\ell_1$-norm problem*

$$\min_x \|x\|_2^2 \quad \text{s.t. } Ax = b$$

where the matrix $A$ is "wide".

- Introducing the dual variable $\nu \in \mathbf{R}^m$, the KKT conditions require

$$2x^* + A^T \nu^* = 0, \quad Ax^* = b$$

which can be concisely expressed as

$$\begin{bmatrix} 2I & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \nu^* \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix}$$

- Solving w.r.t. $x^*$ and $\nu^*$ results in

$$x^* = A^T \left(AA^T\right)^{-1} b$$
$$\nu^* = -2\left(AA^T\right)^{-1} b$$

- Note that, since $\text{rank}(A) = m < n$, $AA^T$ is invertible.

- The least $\ell_1$-norm problem is defined as

$$\begin{aligned} &\min_x \ \|x\|_1 \\ &\text{s.t. } Ax = b \end{aligned}$$

- The least $\ell_1$-norm problem tends to produce a solution $x$ *with a large number of components equal to zero*.

- We say it tends to produce *sparse solutions* to $Ax = b$.

- The problem can be solved as an LP of the form

$$\begin{aligned} &\min_{x,t} \ \mathbf{1}^T t \\ &\text{s.t. } -t \preceq x \preceq t, \quad Ax = b \end{aligned}$$

with variables $x \in \mathbf{R}^n$ and $y \in \mathbf{R}^n$.

- A common form of *regularization* is

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

where $\gamma > 0$ is a *regularization parameter*.

- In cases where $A$ is poorly conditioned, or even singular, regularization gives a compromise between solving $Ax = b$ and keeping $x$ *small*.
- The *Tikhonov regularization problem* has the form of

$$\min_x \ \|Ax - b\|_2^2 + \gamma\|x\|_2^2$$

- The problem has a closed form solution

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

- Note that $A^T A + \gamma I \succ 0$ for all $\gamma > 0$.

- Consider the following *observation model*

$$y(t) = \sum_{\tau=0}^{t} h(\tau)u(t-\tau) = \mathcal{H}\{u\}(t), \quad t = 0, 1, \ldots, T$$

with $\{y(t)\}_{t=0}^{N}$ given and the *impulse response* $\{h(t)\}_{t=0}^{N}$ known.

- Our goal is to estimate $\{u(t)\}_{t=0}^{N}$ through solving

$$\min_{u} \; f_1(u) + \delta\, f_2(u) + \eta\, f_3(u)$$

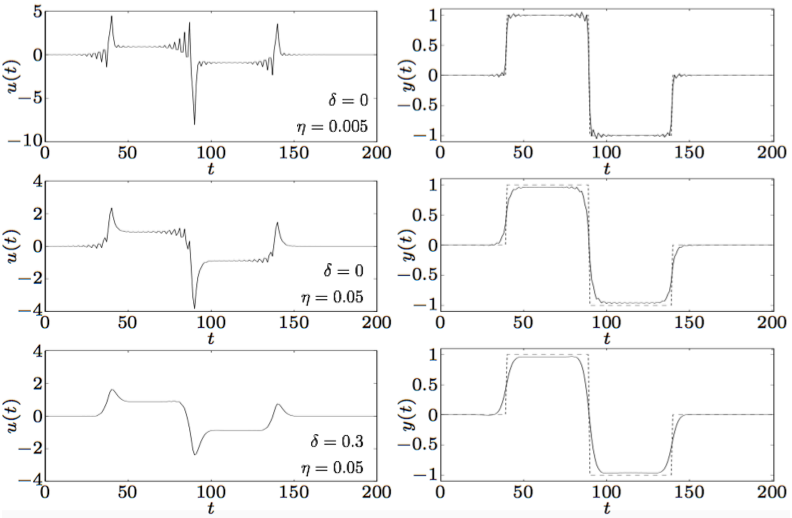
where

$$f_1(u) = \sum_{t=0}^{T} \left|\mathcal{H}\{u\}(t) - y(t)\right|^2 \quad \text{(data fidelity)}$$

$$f_2(u) = \sum_{t=0}^{T} \left|u(t)\right|^2 \quad \text{(smallness)}$$

$$f_3(u) = \sum_{t=0}^{T-1} \left|u(t+1) - u(t)\right|^2 \quad \text{(smoothness)}$$

- We can trade off the objectives by solving for different $\delta > 0$ and $\eta > 0$.




- Different values of $\delta$ and $\eta$ yield solutions of various nature (leftmost subplots), which "track" $y(t)$ differently (rightmost subplots).

- In *reconstruction problems*, we start with a signal $x \in \mathbf{R}$ (which may be considered to be a function of time, for instance), which is assumed to be corrupted by *additive noise* $v$, *viz.*

$$y = x + v$$

- The goal of *signal reconstruction* is to find an estimate $x$ given $y$ which can be achieved through solving

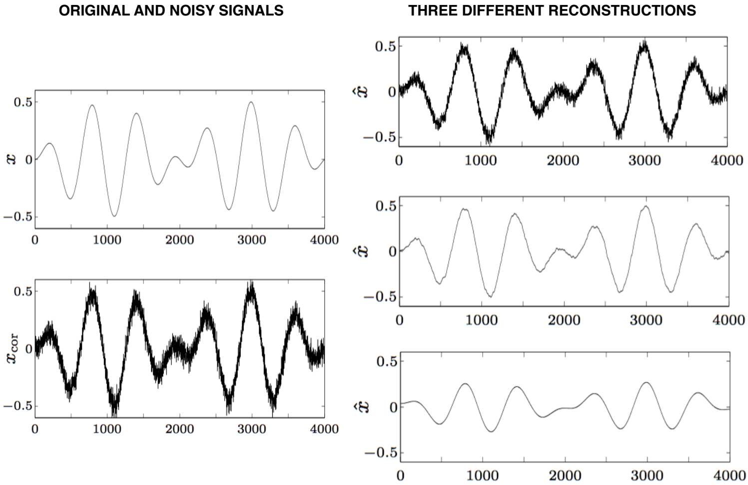
$$\min_x \; \|x - y\|_2^2 + \gamma\, \phi(x)$$

where $\phi : \mathbf{R}^n \to \mathbf{R}$ is a *regularization function*.

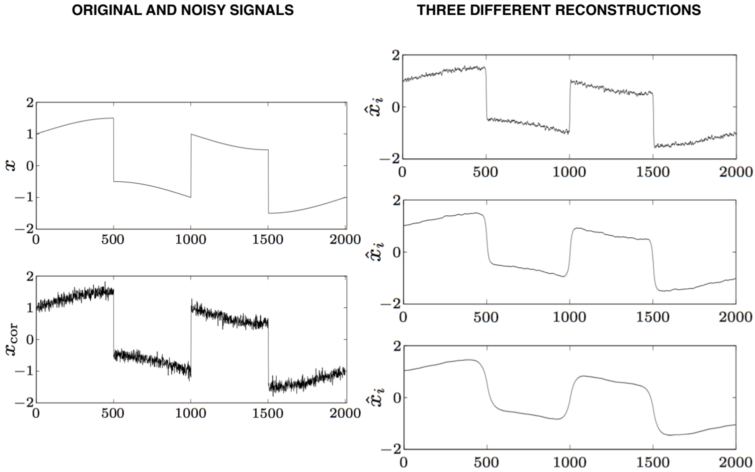
- There are many kinds of regularization functions.

| *Quadratic regularization* | *Total variation regularization* |
| --- | --- |
| $\phi_{\text{quad}}(x) = \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2$ | $\phi_{\text{tv}}(x) = \sum_{i=1}^{n-1} \lvert x_{i+1} - x_i \rvert$ |

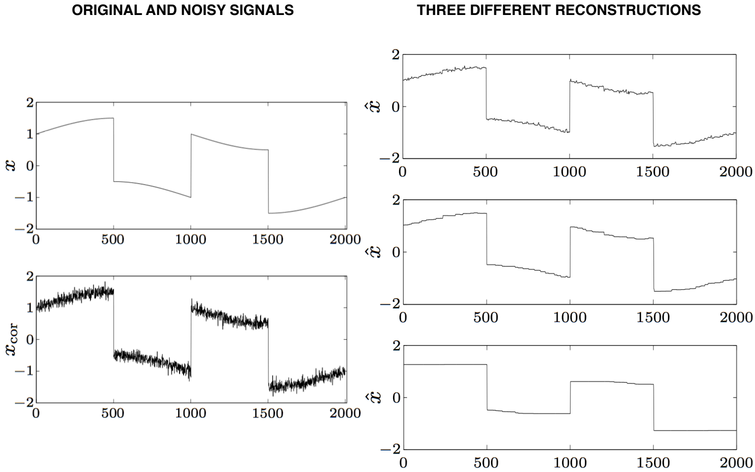
# Example: Quadratic smoothing




(**Left**) Original signal and its noisy measurements; (**Right**) Signal reconstructions obtained at *increasing* values of $\gamma$.




(**Left**) Original signal and its noisy measurements; (**Right**) Signal reconstructions obtained at *increasing* values of $\gamma$. Note how the “signal edges" get blurred.




(**Left**) Original signal and its noisy measurements; (**Right**) Signal reconstructions obtained at *increasing* values of $\gamma$. Note how the "signal edges" are preserved.

- Let $A \in \mathbf{R}^{m \times n}$ be a *random matrix*, with mean $\bar{A}$ (in which case $A$ is referred to as *uncertain*).

- In general, *robust approximation problems* are concerned with solving

$$\min_x \|Ax - b\|$$

under the conditions of uncertainty.

- In particular, the *stochastic robust approximation problem* is defined as

$$\min_x \mathbf{E}\{\|Ax - b\|\}$$

where $\mathbf{E}$ is the *operator of statistical expectation*.

- It is *always* a convex optimization problem, but usually *intractable*, except for a number of special cases.

- Consider the *statistical robust least-squares problem*

$$\min_x \mathbf{E}\{\|Ax-b\|_2^2\}$$

- The objective function can be expressed as

$$\mathbf{E}\{\|Ax-b\|_2^2\} = \mathbf{E}\{\|(\bar{A}+U)x-b\|_2^2\} =$$

$$= \mathbf{E}\Big\{\big((\bar{A}+U)x-b\big)^T\big((\bar{A}+U)x-b\big)\Big\} = \big(\bar{A}x-b\big)^T\big(\bar{A}x-b\big)+$$

$$+\mathbf{E}\{x^T U^T U x\} = \|\bar{A}x-b\|_2^2 + x^T \Sigma\, x$$

- Thus, the original problem is equivalent to

$$\min_x \|\bar{A}x-b\|_2^2 + \|\Sigma^{1/2}x\|_2^2$$

with solution

$$x^* = \big(\bar{A}^T\bar{A}+\Sigma\big)^{-1}\bar{A}^T b$$

- Note that, when $\Sigma = \gamma I$, we obtain a *Tikhonov regularized problem*.

- The uncertainly in $A$ can be described *deterministically* by assuming $A$ to be an *arbitrary element* of set $\mathcal{A}$, *viz.*

$$A \in \mathcal{A} \subseteq \mathbf{R}^{m \times n}$$

- Then, the *worst-case robust approximation problem* is then defined as

$$\min_{x} \sup_{A \in \mathcal{A}} \|Ax - b\|$$

- In this case, the objective function

$$e_{\mathrm{wc}}(x) = \sup_{A \in \mathcal{A}} \|Ax - b\|$$

is referred to as the *worst-case error*.

- The tractability of the problem depends on the norm and the structure of $\mathcal{A}$.

- Consider the following uncertainty model

$$A = A_0 + u\,A_1$$

where the matrices $A_0$ and $A_1$ are *fixed* and $u \in [-1,\,1]$.

- Under this model we explore solutions to

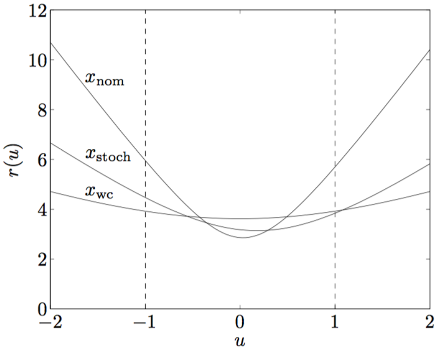
$$\min_x \left\|A\,x - b\right\|_2^2 = \left\|(A_0 + A_1 u)x - b\right\|_2^2$$

obtained using different approaches.

  - **Nominal optimal approach:** The optimal solution $x_{\text{nom}}$ is found under assumption $u = 0$ (i.e., $A = A_0$).
  - **Stochastic robust approximation:** The optimal solution $x_{\text{stoch}}$ is found assuming $u$ is *uniformly distributed* in $[-1,\,1]$.
  - **Worst-case robust approximation:** The optimal solution $x_{\text{wc}}$ is found by solving

$$\sup_{-1 \le u \le 1} \left\|(A_0 + A_1 u)x - b\right\|_2^2$$

- For $A_0 = 10$ and $A_1 = 1$, we analyze $r(u) = \left\|(A_0 + A_1 u)x - b\right\|_2^2$ as a function of $u$.




- $x_{\text{nom}}$ achieves the smallest residual when $u = 0$, but yields much larger residuals as $u$ approaches either $-1$ or $1$.
- $x_{\text{wc}}$ has the largest residual at $u = 0$, but its residuals stays nearly constant, when $u$ varies over $[-1, 1]$.

- Consider a family of *probability distributions* on $\mathbf{R}^m$, represented by a *probability density function (pdf)* $p_x$ that is *parameterized* $x$.

- Given a set of observed values from $p_x(\cdot)$, *our goal is to estimate $x$*.

- *Maximum likelihood (ML) estimation* searches $\hat{x}_{\mathrm{ML}}$ based on

$$\hat{x}_{\mathrm{ML}} = \arg\max_x \; \log p_x(y)$$
$$\text{s.t. } x \in \mathcal{C}$$

  where $y$ is the problem data and $\mathcal{C}$ describes the domain of $l = \log p_x$ that is called the *log-likelihood function*.

- Alternatively, instead of $x \in \mathcal{C}$, we can use $p_x(y) = 0$, for all $x \notin \mathcal{C}$.

- ML estimation is a *convex* optimization problem if $\log p_x(y)$ is *concave* in $x$ for fixed $y$.

- We consider a linear measurement model

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

where $v_i \sim p$ are *independent and identically distributed (i.i.d.)*.

- The probability density is then equal to

$$p_x(y) = \prod_{i=1}^{m} p(y_i - a_i^T x)$$

and, hence, the log-likelihood function is given by

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

- Consequently, the ML estimate $\hat{x}_{\mathrm{ML}}$ is any optimal solution to the problem

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

- **Gaussian noise** with $p(z) = (2\pi\sigma^2)^{-1/2} \exp(-z^2/2\sigma^2)$

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

Thus, the ML estimation is the solution of an LS problem.

- **Laplacian noise** with $p(z) = (1/2a)\exp(-|z|/a)$

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

In this case, the ML estimation is an $\ell_1$-norm approximation.

- **Uniform noise** with $p(z) = 1/2a$, for $z \in [-a, a]$

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

Thus, the ML solution is any $x$ satisfying $\|Ax - y\|_\infty \le a$.

- We can interpret any *penalty function approximation problem*

$$\min_x \sum_{i=1}^m \phi(b_i - a_i^T x)$$

as an ML estimation problem, with the noise probability density defined as

$$p(z) = \frac{\exp(-\phi(z))}{\int \exp(-\phi(u))du}$$

and measurements $b$.

- If $\phi(x)$ grows very rapidly as $|x| \to \infty$, the corresponding *pdf* will have relatively "light" tails.
- This allows us to understand the robustness of $\ell_1$-norm approximation to large errors.

- Consider a random variable $y \in \{0, 1\}$, with

$$\textbf{prob}\,(y = 1) = \rho$$
$$\textbf{prob}\,(y = 0) = 1 - \rho$$

where $\rho \in [0, 1]$ is assumed to depend on some *explanatory variables* $u \in \mathbf{R}^n$ (e.g., weight, age, height, blood pressure, etc.)

- In the *logistic model*, the probability $\rho$ is defined as

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

where $a \in \mathbf{R}^n$ and $b \in \mathbf{R}$ are *model parameters*.

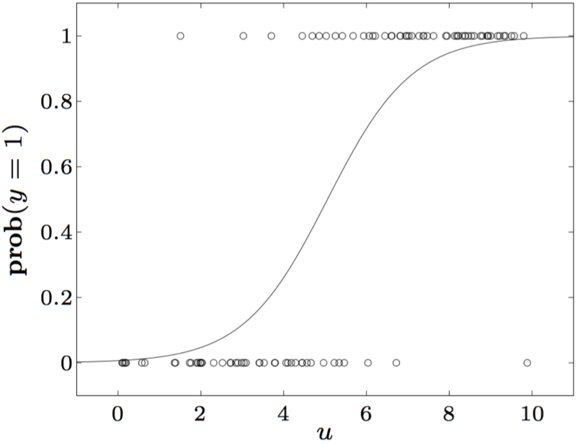
- The problem of finding the ML estimates of $a$ and $b$ is called *logistic regression*.

- Suppose we are given of a set of $m$ *training samples* $\{u_i, y_i\}_{i=1}^m$, where for each explanatory variable $u_i \in \mathbf{R}$, its associated *label* $y_i \in \{0, 1\}$ is provided as well.
- Without loss of generality, let us assume that $y_i = 1$, for $i = 1, 2, ..., k$, and $y_i = 0$, for $i = k+1, k+2, \ldots, m$.
- In this case, the log-likelihood function is defined as

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

- Note that $l(a, b)$ is concave in $a$ and $b$ concurrently.

- The following example is based on $m = 50$ training samples.




- The circled points depict the training data $(u_i, y_i)$.
- The solid curve is $p(u) = \exp(\hat{a}_{\mathrm{ML}} u + \hat{b}_{\mathrm{ML}})/(1 + \exp(\hat{a}_{\mathrm{ML}} u + \hat{b}_{\mathrm{ML}}))$.

- Consider the problem of estimating $x \in \mathbf{R}^n$ from its measurements

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

where $v_i$ are the samples of *measurement noise*, which is assumed to be *i.i.d.* with $\mathcal{N}(0, 1)$.

- The ML estimate of $x$ is given by

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

which is a linear function of the measurements.

- The *covariance* of the *estimation error* $e = \hat{x} - x$ is given by

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

that characterizes the *experiment's informativeness*.

- The problem of *experimental design* consists in finding such *test vectors* $a_i$ that minimize the *size* of $E$.

- The test vectors $\{a_i\}_{i=1}^m$ are assumed to be selected from a smaller set of *prototype vectors* $\{v_j\}_{j=1}^p$, in which case each $v_j$ can be picked more than one time (usually, $m \gg p$).

- Let $m_j$ be the number of vectors $a_i$ equal to $v_j$ (or, by the same token, the number of times the vector $v_j$ has been selected), with

$$m_1 + m_2 + \ldots + m_p = M$$

- The error covariance matrix can then be expressed as

$$E = \Big(\sum_{i=1}^{m} a_i a_i^T\Big)^{-1} = \Big(\sum_{j=1}^{p} m_j v_j v_j^T\Big)^{-1}$$

which now becomes a function of $(m_1, \ldots, m_p)$, thereby allowing us to consider

$$\min_{m_1,\ldots,m_p} \Big(\sum_{j=1}^{p} m_j v_j v_j^T\Big)^{-1} \text{ s.t. } \sum_{i=1}^{p} m_i = M, \; m_i \geq 0, \; i = 1, 2, \ldots, p$$

- Unfortunately, this is a difficult *integer* problem.

- Define $\lambda \in \mathbf{R}^p$ with $\lambda_i = m_i/M$ be the *relative frequency* of the $i$th experiment. The error covariance can then be expressed in terms of $\lambda$ as

$$E = \frac{1}{M}\left(\sum_{j=1}^{p} \lambda_j v_j v_j^T\right)^{-1}$$

where all $\lambda_i$ are positive and sum up to one (i.e., $\lambda \succeq 0$ and $\mathbf{1}^T\lambda = 1$).

- Consequently, the *relaxed experiment design problem* is defined as

$$\min_{\lambda} \ (\text{w.r.t.}\, \mathbf{S}_+^n) \ \ E = \frac{1}{M}\left(\sum_{j=1}^{p} \lambda_j v_j v_j^T\right)^{-1}$$
$$\text{s.t. } \lambda \succeq 0, \ \mathbf{1}^T\lambda = 1$$

which is a convex optimization problem.

- Given an optimal $\lambda^*$, the related $m_i^*$ can be recovered as

$$m_i^* = \text{round}\,(\lambda_i^* M), \quad i = 1, \ldots, p$$

- Given an estimate $\hat{x}$ of $x$, the error covariance $E$ can be used to define the *$\alpha$-confidence level ellipsoid* as

$$\mathcal{E} = \left\{x \mid (x - \hat{x})^T E^{-1} (x - \hat{x}) \leq T\right\}$$

where $T$ is a function of both $\alpha$ and $n$.

- There are different ways of *scalarization* of $E$ which lead to optimization of various geometric properties of $\mathcal{E}$.

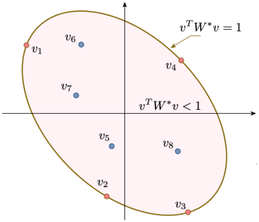
| Design type | Optimization problem | Geometric meaning |
|---|---|---|
| ***D*-optimal design** | $\min_\lambda \ \log\det\left(\sum_{j=1}^p \lambda_j v_j v_j^T\right)^{-1}$ s.t. $\lambda \succeq 0, \ \mathbf{1}^T\lambda = 1$ | minimizes the *volume* of $\mathcal{E}$ |
| ***E*-optimal design** | $\min_\lambda \ \left\|\left(\sum_{j=1}^p \lambda_j v_j v_j^T\right)^{-1}\right\|_2$ s.t. $\lambda \succeq 0, \ \mathbf{1}^T\lambda = 1$ | minimizes the *diameter* of $\mathcal{E}$ |
| ***A*-optimal design** | $\min_\lambda \ \mathbf{tr}\left(\sum_{j=1}^p \lambda_j v_j v_j^T\right)^{-1}$ s.t. $\lambda \succeq 0, \ \mathbf{1}^T\lambda = 1$ | minimizes the error *variance* |

- The *dual* of the $D$-optimal experiment design problem can be expressed as

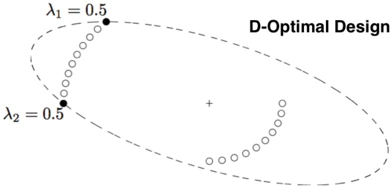
$$\max_{W \in \mathbf{S}^n_{++}} \quad \log\det W$$
$$\text{s.t. } v_i^T W v_i \le 1,\ i = 1, \ldots, p$$

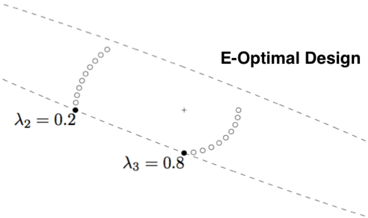
- The optimal $W^*$ defines the *minimum volume ellipsoid* $\{x \,|\, x^T W x \le 1\}$ that is centred at zero and contains all the points $v_1, \ldots, v_p$.

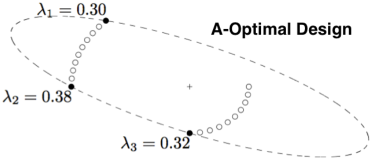
- The optimal design only uses experiments $v_i$ that lie on the surface of the ellipsoid defined by $W^*$.
- In the figure, only the experiments $v_1$, ..., $v_4$ (red points) define the *optimal $D$-design*.



Consider a problem with $x \in \mathbf{R}^2$ and $p = 20$.



- In a normed (linear) space with $\|\cdot\|$, the *distance* of $x_0 \in \mathbf{R}^n$ to a closed set $\mathcal{C} \subseteq \mathbf{R}^n$, is defined as

$$\mathbf{dist}(x_0, C) = \inf \{\|x - x_0\| \mid x \in C\}$$

- If for some $z \in \mathcal{C}$, $\|z - x_0\| = \mathbf{dist}(x_0, C)$, then $z$ is called a *projection of $x_0$ onto $\mathcal{C}$*.

- In general, there could be more than one projection, unless $C$ is both *closed and convex*.

- Let $P_C : \mathbf{R}^n \to \mathbf{R}^n$ be such that

$$P_C(x_0) \in \mathcal{C} \quad \text{and} \quad \|P_C(x_0) - x_0\| = \mathbf{dist}(x_0, \mathcal{C})$$

which implies that

$$P_C(x_0) = \arg\min_{x \in \mathcal{C}} \{\|x - x_0\|\}$$

- We refer to $P_C$ as the *operator of projection on $C$*.

- **Projection on the unit $\ell_2$-norm ball in $\mathbf{R}^n$**

$$P_C(x) = \begin{cases} x, & \|x\|_2 \leq 1 \\ x/\|x\|_2, & \|x\|_2 > 1 \end{cases}$$

- **Projection on $\mathbf{R}^n_+$**

$$P_C(x) = (x)_+$$

- **Projection on $\mathcal{C} = \{x \mid Ax = b\}$** $(A \in \mathbf{R}^{m\times n}, \operatorname{rank}(A) = m > n\ )$

$$P_C(x) = x + A^T(AA^T)^{-1}(b - Ax)$$

- **Projection on $\mathbf{S}^n_+$:**

$$P_C(X) = \sum_{i=1}^{n} \max\{\lambda_i, 0\} q_i q_i^T$$

where $X = \sum_{i=1}^n \lambda_i q_i q_i^T$ is the *eigenvalue decomposition* of $X$.

- Suppose $\mathcal{C} \subseteq \mathbf{R}^n$ is bounded, with $\mathbf{int}\,\mathcal{C} \neq \emptyset$.
- The *Löwner-John ellipsoid $\mathcal{E}_{\text{LJ}}$ of the set $C$* is the *minimum volume* ellipsoid that contains $\mathcal{C}$.
- Recall that a general ellipsoid can be *implicitly* represented as

$$\mathcal{E} = \{v \mid \|Av + b\|_2 \leq 1\}$$

for some $A \in \mathbf{S}^n_{++}$ and $b \in \mathbf{R}^n$.

- Since the volume of $\mathcal{E}$ is proportional to $\det A^{-1}$, $\mathcal{E}_{\text{LJ}}$ can be found by solving

$$\begin{aligned} &\min_{A,b} \ \log\det A^{-1} \\ &\text{s.t. } \sup_{v \in \mathcal{C}} \|Av + b\|_2 \leq 1 \end{aligned}$$

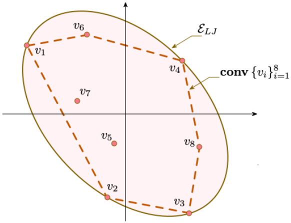
- Unfortunately, this (convex optimization) problem is *tractable* only in certain special cases.

- As a special case, suppose $C = \{x_1, \ldots, x_m\} \in \mathbf{R}^n$. Then, $\mathcal{E}_{\mathrm{LJ}}$ can be found by solving:

$$\begin{aligned} &\min_{A,x} \ \log\det A^{-1} \\ &\text{s.t. } \|Ax_i + b\|_2 \leq 1, \ i = 1, \ldots, m \end{aligned}$$

which is a convex optimization problem with quadratic constraints.

- The solution also gives the Löwner-John ellipsoid for $\mathbf{conv}\,\{x_1, \ldots, x_m\}$.




- When shrunk by a factor $n$, the Löwner-John ellipsoid is guaranteed to lie inside of $\mathbf{conv}\,\mathcal{C}$. Moreover, if $\mathcal{C}$ is symmetric, then the factor $1/n$ can be tightened to $1/\sqrt{n}$.

- Consider the problem of finding the ellipsoid of maximum volume that lies inside a convex set $\mathcal{C}$.

- The ellipsoid can be *explicitly* parametrized as

$$\mathcal{E} = \{Bu + d \mid \|u\|_2 \leq 1\}$$

for some $B \in \mathbf{S}^n_{++}$ and $d \in \mathbf{R}^n$.

- Hence, the *maximum-volume ellipsoid inscribed in $\mathcal{C}$* can be found via solving

$$\max_{B \succ 0, d} \ \log \det B$$
$$\text{s.t.} \ \sup_{\|u\|_2 \leq 1} I_{\mathcal{C}}(Bu + d) \leq 0$$

where $I_{\mathcal{C}}$ stands for the *indicator function* of $\mathcal{C}$.

- Again, this (convex optimization) problem is tractable only in certain special cases.

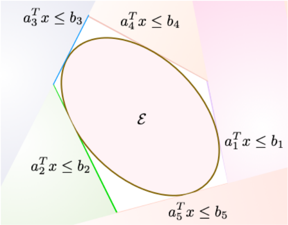
- As a special case, consider $C = \{x \mid a_i^T x \leq b_i,\ i = 1, \ldots, m\}$.

- In this case, the constraint is reduced to

$$\sup_{\|u\|_2 \leq 1} a_i^T(Bu + d) \leq b_i \quad \Longleftrightarrow \quad \|Ba_i\|_2 + a_i^T d \leq b_i, \quad i = 1, \ldots, m$$

- The maximum volume ellipsoid can now be found by solving

$$\min_{B \succ 0, d} \ \log\det B^{-1}$$
$$\text{s.t. } \|Ba_i\|_2 + a_i^T d \leq b_i, \quad i = 1, \ldots, m$$




- The maximum volume inscribed ellipsoid, expanded by a factor of $n$, covers $\mathcal{C}$. Again, this factor can be tightened to $\sqrt{n}$, if $\mathcal{C}$ is symmetric.

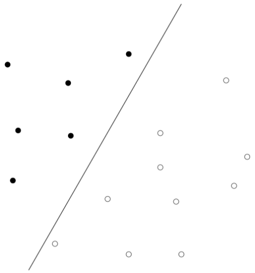
- In pattern recognition and classification problems we are given a set of *training samples*

$$\{x_1, \ldots, x_N\} \subset \mathbf{R}^n \quad \text{and} \quad \{y_1, \ldots, y_M\} \subset \mathbf{R}^n$$

and wish to find a function $f : \mathbf{R}^n \to \mathbf{R}$ such that

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

- If found, $\{x \mid f(x) = 0\}$ is said to *separate or classify* the two sets.



- In *linear discrimination*, we set $f(x) = a^T x - b$ so that

$$\begin{cases} a^T x_i - b > 0, & i = 1, \ldots, N \\ a^T y_i - b < 0, & i = 1, \ldots, M \end{cases}$$

- Geometrically, we seek a *hyperplane* that separates the two sets.
- Alternatively, the above strict inequalities are *feasible* if and only if

$$\begin{cases} a^T x_i - b \geq 1, & i = 1, \ldots, N \\ a^T y_i - b \leq -1, & i = 1, \ldots, M \end{cases}$$

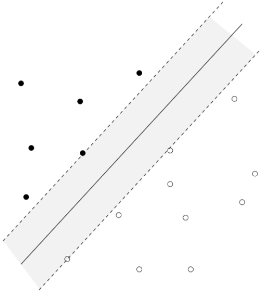
  are *feasible*.
- In general, the two sets of points can be linearly discriminated if and only if *their convex hulls do not intersect*.

- In *robust liner discrimination*, we seek $f(x) = a^T x - b$ that gives the maximum possible *"gap"* between the two sets, *viz.*

$$\begin{aligned} \max_{a,b,t} \; & t \\ \text{s.t.} \; & a^T x_i - b \geq t, \quad i = 1, \ldots, N \\ & a^T x_i - b \leq -t, \quad i = 1, \ldots, M \\ & \|a\|_2 \leq 1 \end{aligned}$$

- If the sets are linearly separable, then $t^* > 0$ and $\|a^*\|_2 = 1$.

- $t^*$ is equal to 1/2 of the "slab" thickness.



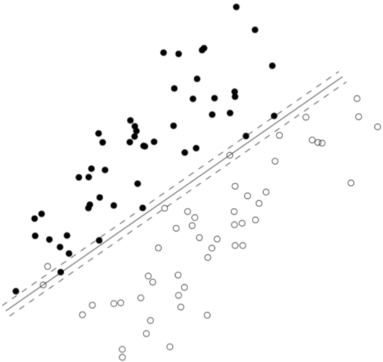
- Suppose $\{x_1, \ldots, x_N\}$ and $\{y_1, \ldots, y_M\}$ *cannot be linearly separated*.
- In this case we introduce $u \in \mathbf{R}_+^n$ and $v \in \mathbf{R}_+^m$ such that

$$\begin{cases} a^T x_i - b \geq 1 - u_i, \;\; i = 1, \ldots, N \\ a^T y_i - b \leq -(1 - v_i), \;\; i = 1, \ldots, M \end{cases}$$

- By making $u$ and $v$ large enough, the inequalities can always be made *feasible*.
- One can *maximize the sparsity* of $u$ and $v$ through

$$\begin{aligned} \min_{a,b,u,v} \;\; & \mathbf{1}^T u + \mathbf{1}^T v \\ \text{s.t. } & a^T x_i - b \geq 1 - u_i, \quad i = 1, \ldots, N \\ & a^T x_i - b \leq -(1 - v_i), \quad i = 1, \ldots, M \\ & u \succeq 0, \;\; v \succeq 0 \end{aligned}$$

- In fact, this problem minimizes the *number of points* that violate either $a_i^T - b \geq 1$ or $a_i^T - b \leq -1$.

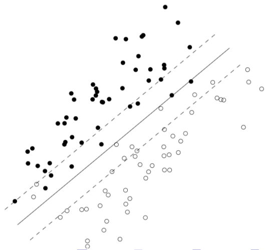


- In this example, $a^T z - b$ misclassifies 1 out of 100 points.
- The dashed lines are the hyperplanes $a^T z - b = \pm 1$.
- Four points are correctly classified, but lie within the slab.

- The width of the slab $\{z \mid -1 \leq a^T z - b \leq 1\}$ is equal to $2/\|a\|_2$.
- The *standard support vector classifier* is defined as the solution of

$$\begin{aligned} \min_{a,b,u,v} \quad & \|a\|_2 + \gamma \left(\mathbf{1}^T u + \mathbf{1}^T v\right) \\ \text{s.t.} \quad & a^T x_i - b \geq 1 - u_i, \quad i = 1, \ldots, N \\ & a^T x_i - b \leq -(1 - v_i), \quad i = 1, \ldots, M \\ & u \succeq 0, \; v \succeq 0 \end{aligned}$$

- Here $\gamma > 0$ gives the relative weight of the number of misclassified points compared to the width of the slab.



- In *non-linear discrimination*, we seek a nonlinear function $f : \mathbf{R}^n \to \mathbf{R}$ such that

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

- In particular, in the case of *quadratic discrimination*, the feasibility constraints are

$$\begin{aligned} x_i^T P x_i + q^T x_i + r > 0, &\quad i = 1, \ldots, N \\ y_i^T P y_i + q^T y_i + r < 0, &\quad i = 1, \ldots, M \end{aligned}$$

for some (variables) $P \in \mathbf{S}^n$, $q \in \mathbf{R}^n$, and $r \in \mathbf{R}$.

- Alternatively, one can solve a *nonstrict* feasibility problem of the form

$$\begin{aligned} x_i^T P x_i + q^T x_i + r \geq 1, &\quad i = 1, \ldots, N \\ y_i^T P y_i + q^T y_i + r \leq -1, &\quad i = 1, \ldots, M \end{aligned}$$

- The separating surface $\{z \mid z^T P z + q^T z + r = 0\}$ defines two *classification regions*, *viz.*

$$\{z \mid z^T P z + q^T z + r \geq 0\} \quad \text{and} \quad \{z \mid z^T P z + q^T z + r \leq 0\}$$

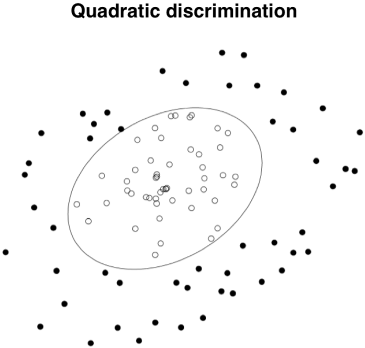
- We can impose conditions on the shape of the separating surface. For example, requiring $P \prec 0$ will make the separating surface *ellipsoidal*.
- The resulting problem can be solved as an SDP feasibility problem

$$\begin{aligned} \text{find} \;\; & P,\, q,\, r \\ & x_i^T P x_i + q^T x_i + r \geq 1, \quad i = 1, \ldots, N \\ & y_i^T P y_i + q^T y_i + r \leq -1, \quad i = 1, \ldots, M \\ & P \preceq -I \end{aligned}$$

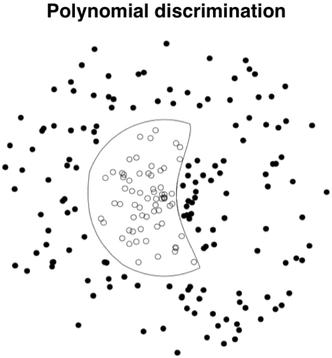
- Another example of nonlinear discrimination corresponds to $f$ defined as a polynomial of the form

$$f(x) = \sum_{i_1 + \ldots + i_n \leq d} a_{i_1, \ldots, i_n} x_1^{i_1} \cdots x_n^{i_n}$$

where $d$ is the degree of $f$.




$P \prec 0$




$d = 4$