06. Applications and Algorithms

By Yang Lin[1] (2024 秋季, @NJU)

Contents

[1]Institute: Nanjing University. Email: linyang@nju.edu.cn.

1 Approximation

1.1 Norm Approximation

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

($A \in \mathbf{R}^{m \times n}$ with $m \geq n$)

Interpretations of solution $x = \arg\min_x \|Ax - b\|$:

1 Approximation

1.1 Norm Approximation

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

$(A \in \mathbf{R}^{m\times n}$ with $m \geq n)$

Interpretations of solution $x = \arg\min_x \|Ax - b\|$:

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

2. **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^*

3. **Optimal design:** x are design variables(input), Ax is result(output)

x^* is design that best approximates desired result b

1 Approximation

Examples.

Least-squares approximation ($\|\cdot\|_2$): solution satisfies normal equations

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

($x^* = (A^T A)^{-1} A^T b$ if **rank** $A = n$)

Chebyshev approximation ($\|\cdot\|_\infty$): can be solved as an LP

$$\begin{array}{ll} \texttt{minimize} & t \\ \texttt{subject to} & -t\mathbf{1} \preceq Ax - b \preceq t\mathbf{1} \end{array}$$

Sum of absolute residuals approximation ($\|\cdot\|_1$): can be solved as an LP

$$\begin{array}{ll} \texttt{minimize} & \mathbf{1}^T y \\ \texttt{subject to} & -y \preceq Ax - b \preceq y \end{array}$$

1 Approximation

1.2 Penalty function approximation.

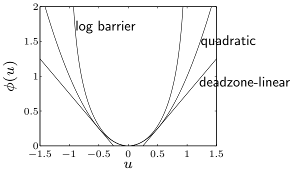
$$\begin{array}{ll} \texttt{minimize} & \phi(r_1) + \cdots \phi(r_m) \\ \texttt{subject to} & r = Ax - b \end{array}$$

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

Examples

1. quadratic: $\phi(u) = u^2$
2. deadzone-linear with width a: $\phi(u) = \max\{0, |u| - a\}$
3. log-barrier with limit a:

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

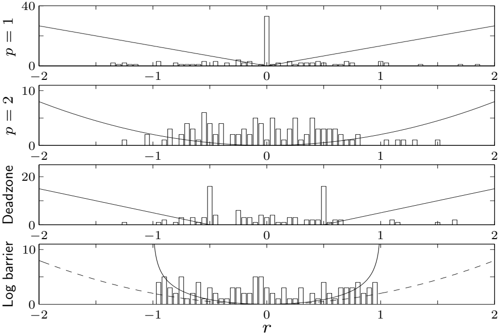


1 Approximation

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

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

Shape of penalty function has large effect on distribution of residuals

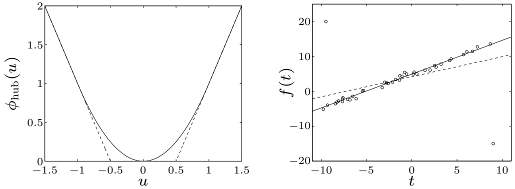


1 Approximation

Huber penalty function (with parameter M)

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

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

1 Approximation

1.3 Regularized approximation

$$\texttt{minimize (w.r.t. } \mathbf{R}_+^2\texttt{)} \quad (\|Ax - b\|, \|x\|)$$

$(A \in \mathbf{R}^{m \times n})$

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

1. **estimation**: linear measurement model $y = Ax + v$, with prior knowledge that $\|x\|$ is small
2. **optimal design**: small x is cheaper or more efficient, or the linear model $y = Ax$ is only valid for small x
3. **robust approximation**: good approximation $Ax \approx b$ with small x is less sensitive to errors in A than good approximation with large x

1 Approximation

1.3.1 Scalarized problem

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

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

other common method: minimize $\|Ax - b\|^2 + \delta\|x\|^2$ with $\delta > 0$

Tikhonov regularization

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

can be solved as a least-squares problem

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

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

1 Approximation

1.3.2 Optimal input design

linear dynamical system with impulse response h:

$$y(t) = \sum_{\tau=0}^{t} h(\tau)u(t-\tau), \;\; t = 0, 1, \ldots, N$$

Input design problem: multicriterion problem with 3 objectives

1. tracking error with desired output y_{des} : $J_{\text{track}} = \sum_{t=0}^{N}(y(t) - y_{\text{des}}(t))^2$
2. input magnitude: $J_{\text{mag}} = \sum_{t=0}^{N} u(t)^2$
3. input variation: $J_{\text{der}} = \sum_{t=0}^{N-1}(u(t+1) - u(t))2$

track desired output using a small and slowly varying input signal

regularized least-squares formulation

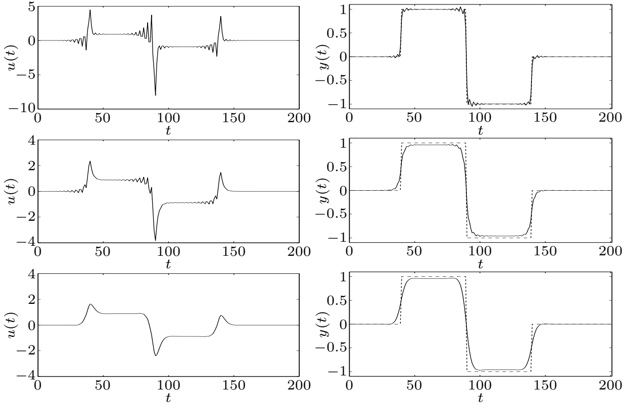
$$\texttt{minimize} \quad J_{\text{track}} + \delta J_{\text{der}} + \eta J_{\text{mag}}$$

for fixed δ, η, a least-squares problem in $\mu(0), \ldots, \mu(N)$

1 Approximation

example: 3 solutions on optimal trade-off curve

(top) $\delta = 0$, small η; (middle) $\delta = 0$,larger η; (bottom) large δ, η



1.3.3 Signal reconstruction

$$\texttt{minimize (w.r.t. } \mathtt{R}_{+}^{2}\texttt{)} \quad (\|\hat{x} - x_{\mathrm{cor}}\|_2, \phi(\hat{x}))$$

$x \in \mathbf{R}^n$ is unknown signal

$x_{\mathrm{cor}} = x + v$ is (known) corrupted version of x, with additive noise v

variable $\hat{x}$ (reconstructed signal) is estimate of x

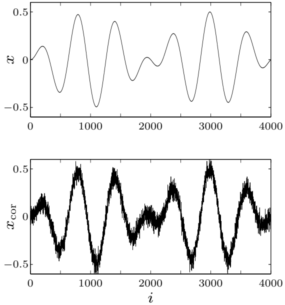
$\phi : \mathbf{R}^n \to \mathbf{R}$ is regularization function or smoothing objective

examples: quadratic smoothing, total variation smoothing:

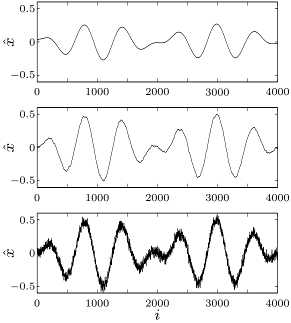
$$\phi_{\mathrm{quad}}(\hat{x}) = \sum_{i=1}^{n-1} (\hat{x}_{i+1} - \hat{x}_i)^2, \;\; \phi_{\mathrm{tv}}(\hat{x}) = \sum_{i=1}^{n-1} |\hat{x}_{i+1} - \hat{x}_i|$$

1 Approximation

quadratic smoothing example



original signal x and noisy signal x_{cor}

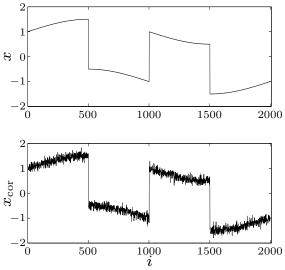


three solutions on trade-off curve $\|\hat{x} - x_{\mathrm{cor}}\|_2$ versus $\phi_{\mathrm{quad}}(\hat{x})$

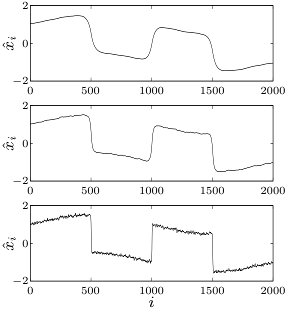
1 Approximation

total variation reconstruction example

quadratic smoothing smooths out noise and sharp transitions in signal



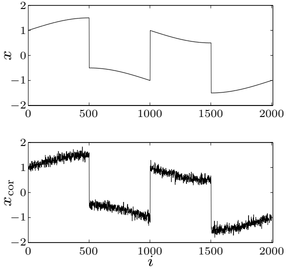
original signal x and noisy signal x_{cor}



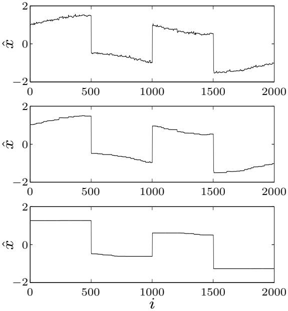
three solutions on trade-off curve $\|\hat{x} - x_{\mathrm{cor}}\|_2$ versus $\phi_{\mathrm{quad}}(\hat{x})$

1 Approximation

total variation smoothing preserves sharp transitions in signal



original signal x and noisy signal x_{cor}



three solutions on trade-off curve $\|\hat{x} - x_{\mathrm{cor}}\|_2$ versus $\phi_{\mathrm{tv}}(\hat{x})$

1 Approximation

1.4 Robust approximation

Minimize $\|Ax-b\|$ with uncertain A

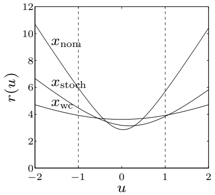
two approaches:

1. Stochastic: assume A is random, minimize $\mathbf{E}\|Ax-b\|$
2. Worst-case: set $\mathcal{A}$ of possible values of A, minimize $\sup_{A\in\mathcal{A}}\|Ax-b\|$

tractable only in special cases (certain norms $\|\cdot\|$, distributions, sets $\mathcal{A}$)

example: $A(u) = A_0 + uA_1$

1. x_{norm} minimizes $\|A_0x-b\|_2^2$
2. x_{stoch} minimizes $\mathbf{E}\|A(u)x-b\|_2^2$ with u uniform on $[-1,1]$
3. x_{wc} minimizes $\sup_{-1\leq u\leq 1}\|A(u)x-b\|_2^2$ figure shows $r(u) = \|A(u)x-b\|_2$



1 Approximation

Stochastic robust LS with $A = \bar{A} + U$, U random, $\mathbf{E}\, U = 0$, $\mathbf{E}\, U^T U = P$

$$\texttt{minimize} \quad \mathbf{E}\|(\bar{A} + U)x - b\|_2^2$$

1. explicit expression for objective:

$$\begin{aligned} \mathbf{E}\|Ax - b\|_2^2 &= \mathbf{E}\|\bar{A}x - b + Ux\|_2^2 \\ &= \|\bar{A}x - b\|_2^2 + \mathbf{E}\, x^T U^T U x \\ &= \|\bar{A}x - b\|_2^2 + x^T P x \end{aligned}$$

2. hence, robust LS problem is equivalent to LS problem

$$\texttt{minimize} \quad \|\bar{A}x - b\|_2^2 + \|P^{1/2}x\|_2^2$$

3. for $P = \delta I$, get Tikhonov regularized problem

$$\texttt{minimize} \quad \|\bar{A}x - b\|_2^2 + \delta\|x\|_2^2$$

1 Approximation

Worst-case robust LS with $\mathcal{A} = \{\bar{A} + u_1 A_1 + \cdots + u_p A_p \mid \|u\|_2 \leq 1\}$

$$\texttt{minimize} \quad \sup_{A \in \mathcal{A}} \|\bar{A}x - b\|_2^2 = \sup_{\|u\|_2 \leq 1} \|P(x)u + q(x)\|_2^2$$

where $P(x) = [A_1 x \; A_2 x \; \cdots \; A_p x]$, $q(x) = \bar{A}x - b$

1. strong duality holds between the following problems

$$\begin{array}{ll} \texttt{maximize} & \|Pu + q\|_2^2 \\ \texttt{subject to} & \|u\|_2^2 \leq 1 \end{array}$$

$$\begin{array}{ll} \texttt{minimize} & t + \lambda \\ \texttt{subject to} & \begin{bmatrix} I & P & q \\ P^T & \lambda I & 0 \\ q^T & 0 & t \end{bmatrix} \succeq 0 \end{array}$$

2. hence, robust LS problem is equivalent to SDP

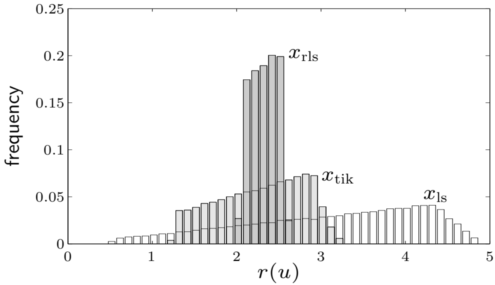
$$\begin{array}{ll} \texttt{minimize} & t + \lambda \\ \texttt{subject to} & \begin{bmatrix} I & P(x) & q(x) \\ P(x)^T & \lambda I & 0 \\ q(x)^T & 0 & t \end{bmatrix} \succeq 0 \end{array}$$

1 Approximation

example

$$r(u) = \|(A_0 + u_1 A_1 + u_2 A_2)x - b\|2$$

with u uniformly distributed on unit disk, for three values of x



1. x_{ls} minimizes $\|A_0 x - b\|_2$
2. x_{tik} minimizes $\|A_0 x - b\|_2^2 + \delta\|x\|_2^2$ (Tikhonov solution)
3. x_{wc} minimizes $\sup_{\|u\|_2 \le 1} \|A_0 x - b\|_2^2 + \|x\|_2^2$

2 Statistical Estimation

2.1 Maximum Likelihood Estimation

Parametric distribution estimation

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

Maximum Likelihood Estimation

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

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

2 Statistical Estimation

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

maximum likelihood estimate: any solution x of

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

y is observed value

2 Statistical Estimation

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

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

ML estimate is LS solution

2. **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 1-norm solution

3. **uniform noise on** $[-a, a]$**:**

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

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

2 Statistical Estimation

Logistic regression

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

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

1. a, b are parameters; $u \in \mathbf{R}^n$ are (observable) explanatory variables

2. Estimation problem: estimate a, b from m observations (u_i, y_i)

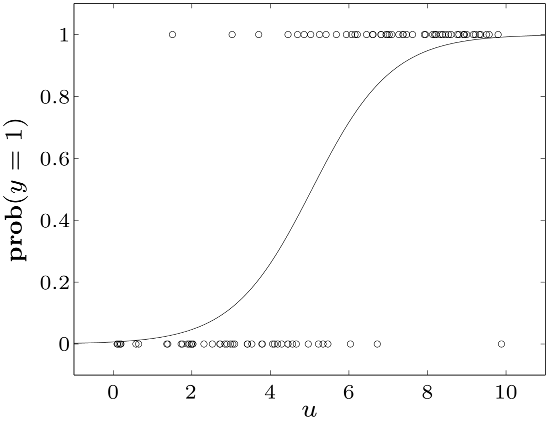
log-likelihood function (for $y_1 = \cdots = y_k = 1, y_{k+1} = \cdots = y_m = 0$):

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

concave in a, b

2 Statistical Estimation

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



circles show 50 points (u_i, y_i)

solid curve is ML estimate of $p = \exp(au + b)/(1 + \exp(au + b))$

*2.2 Optimal Detector Design

(Binary) hypothesis testing

detection (hypothesis testing) problem

given observation of a random variable $X \in \{1, \dots, n\}$, choose between:

1. hypothesis 1: X was generated by distribution $p = (p_1, \dots, p_n)$
2. hypothesis 2: X was generated by distribution $q = (q_1, \dots, q_n)$

Randomized detector

a nonnegative matrix $T \in \mathbf{R}^{2\times n}$, with $\mathbf{1}^T T = \mathbf{1}^T$

if we observe $X = k$, we choose hypothesis 1 with probability t_{1k}, hypothesis 2 with probability t_{2k}

if all elements of T are 0 or 1, it is called a deterministic detector

2 Statistical Estimation

detection probability matrix:

$$D = [Tp\ \ Tq] = \left[\begin{array}{cc} 1 - P_{\rm fp} & P_{\rm fn} \\ P_{\rm fp} & 1 - P_{\rm fn} \end{array}\right]$$

1. $P_{\rm fp}$ is probability of selecting hypothesis 2 if X is generated by distribution 1 (false positive)
2. $P_{\rm fn}$ is probability of selecting hypothesis 1 if X is generated by distribution 2 (false negative)

multicriterion formulation of detector design

$$\begin{array}{ll} \texttt{minimize (w.r.t. } \mathbf{R}_+^2\texttt{)} & (P_{\rm fp}, P_{\rm fn}) = ((Tp)_2, (Tq)_1) \\ \texttt{subject to} & t_{1k} + t_{2k} = 1, \ \ k = 1, \ldots, n \\ & t_{ik} \geq 0, \quad i = 1, 2, \quad k = 1, \ldots, n \end{array}$$

variable $T \in \mathbf{R}^{2 \times n}$

2 Statistical Estimation

scalarization (with weight $\lambda > 0$)

$$\begin{array}{ll} \texttt{minimize} & (Tp)_2 + \lambda(Tq)_1 \\ \texttt{subject to} & t_{1k} + t_{2k} = 1,\ \ t_{ik} \geq 0, i = 1, 2, \quad k = 1, \dots, n \end{array}$$

an LP with a simple analytical solution

$$(t_{1k}, t_{2k}) = \begin{cases} (1, 0) & p_k \geq \lambda q_k \\ (0, 1) & p_k < \lambda q_k \end{cases}$$

A deterministic detector, given by a likelihood ratio test

If $p_k = \lambda q_k$ for some k, any value $0 \leq t_{1k} \leq 1$, $t_{1k} = 1 - t_{2k}$ is optimal (i.e., Pareto-optimal detectors include non-deterministic detectors)

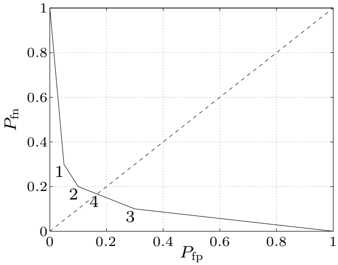
minimax detector

$$\begin{array}{ll} \texttt{minimize} & \max\{P_{\rm fp}, P_{\rm fn}\} = \max\{(Tp)_2, (Tq)_1\} \\ \texttt{subject to} & t_{1k} + t_{2k} = 1,\ \ t_{ik} \geq 0, i = 1, 2, \quad k = 1, \dots, n \end{array}$$

an LP; solution is usually not deterministic

example

$$P = \begin{bmatrix} 0.70 & 0.10 \\ 0.20 & 0.10 \\ 0.05 & 0.70 \\ 0.05 & 0.10 \end{bmatrix}$$



solutions 1, 2, 3 (and endpoints) are deterministic; 4 is minimax detector

3 Algorithms: Unconstrained minimization

$$\texttt{maximize} \quad f(x)$$

1. f convex, twice continuously differentiable (hence $\mathbf{dom}\, f$ open)
2. We assume optimal value $p^* = \inf_x f(x)$ is attained (and finite)

Unconstrained Minimization Methods

1. produce sequence of points $x^{(k)} \in \mathbf{dom}\, f,\ k = 0, 1, \dots$ with

$$f(x^{(k)}) \to p^*$$

2. can be interpreted as iterative methods for solving optimality condition

$$\nabla f(x^*) = 0$$

3 Algorithms: Unconstrained minimization

3.1 Descent methods

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)} \quad \text{with} \quad f(x^{(k+1)}) < f(x^{(k)})$$

1. other notations: $x^+ = x + t\Delta x, x := x + t\Delta x$
2. Δx is the step, or search direction; t is the step size, or step length
3. from convexity, $f(x^+) < f(x)$ implies $\nabla f(x)^T \Delta x < 0$ (i.e., Δx is a descent direction)

General descent method.

given a starting point $x \in \mathbf{dom}\, f$.
repeat
1. Determine a descent direction Δx.
2. *Line search.* Choose a step size $t > 0$.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

---

3 Algorithms: Unconstrained minimization

Line search types

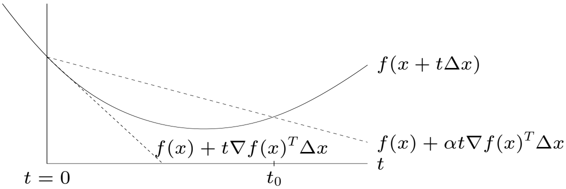
exact line search: $t = \arg\min_{t>0} f(x + t\delta x)$

backtracking line search (with parameters $\alpha \in (0, 1/2), \beta \in (0, 1)$)

1. starting at $t = 1$, repeat $t := \beta t$ until

$$f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^T \Delta x$$

2. graphical interpretation: backtrack until $t \leq t_0$



3 Algorithms: Unconstrained minimization

general descent method with $\Delta x = -\nabla f(x)$

---

given a starting point $x \in \mathbf{dom}\, f$.
repeat

1. $\Delta x := -\nabla f(x)$.
2. *Line search.* Choose step size t via exact or backtracking line search.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

---

1. stopping criterion usually of the form $\|\nabla f(x)\|_2 \leq \epsilon$
2. convergence result: for strongly convex f,

$$f(x^{(k)}) - p^* \leq c^k(f(x^{(0)} - p^*)$$

$c \in (0,1)$ depends on m, $x^{(0)}$, line search type

3. very simple, but often very slow; rarely used in practice

3 Algorithms: Unconstrained minimization

quadratic problem in $\mathbf{R}^2$

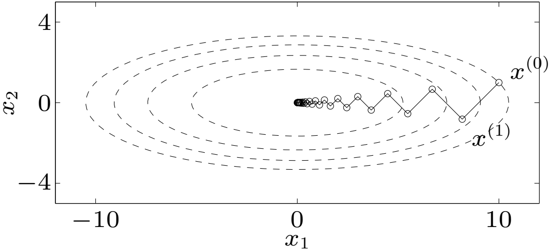
$$f(x) = (1/2)(x_1^2 + \gamma x_2^2) \quad (\gamma > 0)$$

with exact line search, starting at $x^{(0)} = (\gamma, 1)$:

$$x_1^{(k)} = \gamma \left(\frac{\gamma - 1}{\gamma + 1}\right)^k, \quad x_2^{(k)} = \left(-\frac{\gamma - 1}{\gamma + 1}\right)^k$$

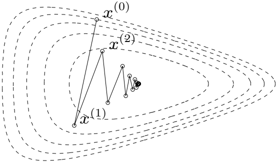
very slow if $\gamma \gg 1$ or $\gamma \ll 1$

example for $\gamma = 10$:

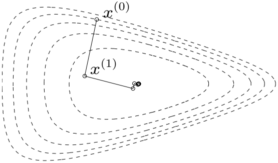


nonquadratic example

$$f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$$



backtracking line search

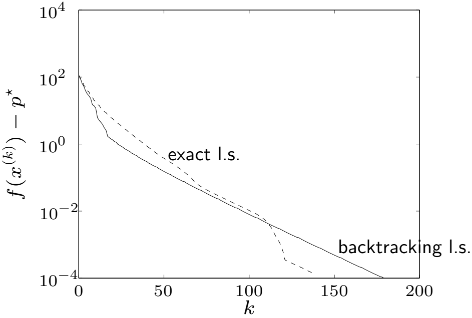


exact line search

3 Algorithms: Unconstrained minimization

a problem in $\mathbf{R}^{100}$

$$f(x) = c^T x - \sum_{i=1}^{500} \log(b_i - a_i^T x)$$



'linear' convergence, i.e., a straight line on a semilog plot

3 Algorithms: Unconstrained minimization

normalized steepest descent direction (at x, for norm $\|\cdot\|$):

$$\Delta x_{\text{nsd}} = \arg\min\{\nabla f(x)^T v \,|\, \|v\| = 1\}$$

interpretation: for small v, $f(x+v) \approx f(x) + \nabla f(x)^T v$

direction Δx_{nsd} is unit-norm step with most negative directional derivative

(unnormalized) steepest descent direction

$$\Delta x_{\text{sd}} = \|\nabla f(x)^T\|_* \Delta x_{\text{nsd}}$$

satisfies $\nabla f(x)^T \Delta x_{\text{sd}} = -\|\nabla f(x)^T\|_*^2$

steepest descent method

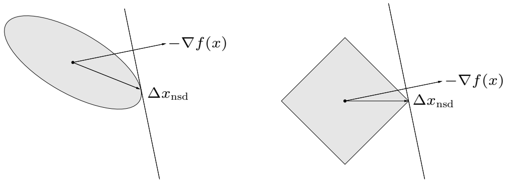
1. general descent method with $\Delta x = \Delta x_{\text{sd}}$
2. convergence properties similar to gradient descent

3 Algorithms: Unconstrained minimization

Examples

1. Euclidean norm: $\Delta x_{\text{sd}} = -\nabla f(x)$
2. Quadratic norm $\|x\|_P = (x^T P x)^{1/2} (P \in \mathbf{S}^n_{++})$: $\Delta x_{\text{sd}} = -P^{-1} \nabla f(x)$
3. l_1-norm: $\Delta x_{\text{sd}} = -(\partial f(x)/\partial x_i) e_i$, where $|\partial f(x)/\partial x_i| = \|\nabla f(x)\|_\infty$

unit balls and normalized steepest descent directions for a quadratic norm and the l_1-norm:



3 Algorithms: Unconstrained minimization

3.2 Newton step

$$\Delta x_{\mathrm{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

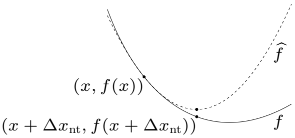
interpretations

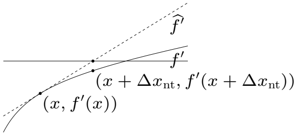
1. $x + \Delta x_{\mathrm{nt}}$ minimizes second order approximation

$$\hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

2. $x + \Delta x_{\mathrm{nt}}$ solves linearized optimality condition

$$\nabla f(x+v) \approx \nabla \hat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0$$

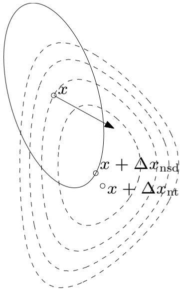




3 Algorithms: Unconstrained minimization

Δx_{nt} is steepest descent direction at x in local Hessian norm

$$\|u\|_{\nabla^2 f(x)} = \left(u^T \nabla^2 f(x) u\right)^{1/2}$$



dashed lines are contour lines of f; ellipse is $\{x + v \mid v^T \nabla^2 f(x) v = 1\}$
arrow shows $-\nabla f(x)$