Presentation

constrained optimization

Wenda Chen

Speech Data and Constrained Optimization Models

Part 1: Speech Signal data (continuous):

- Adaptive filtering and LMS
- ICA with Negentropy criteria for source separation

Part 2: Transcription data (discrete) (*Present next time*):

- Dynamic programming for confusion network
- Linear regression and MMSE for feature analysis

Constrained Optimization

Suppose we have a cost function (or objective function)

$$f(\mathbf{x}) : \mathbb{R}^N \longrightarrow \mathbb{R}$$

Our aim is to find values of the parameters (decision variables) **x** that minimize this function

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$$

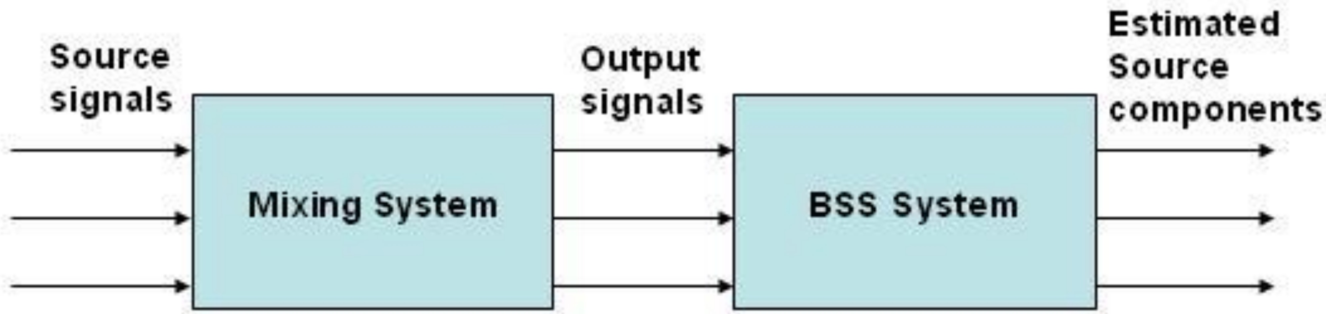
Subject to the following constraints:

- equality: $a_i(\mathbf{x}) = 0$
- nonequality: $c_j(\mathbf{x}) \geq 0$

If we seek a maximum of $f(\mathbf{x})$ (profit function) it is equivalent to seeking a minimum of $-f(\mathbf{x})$

Blind Source Separation

- Input: Source Signals
- Output: Estimated Source Components



Signals received and collected are convolutive mixtures

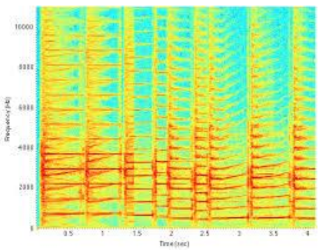
$$x_k(t) = \sum_{i=1}^{N} \sum_{l=1}^{R} a_{ki}(l) s_i(t-l) \text{ for } k = 1, 2, \ldots N$$

Pre-whitening: $E\{xx^H\} = I$

Adaptive Filter to Independent Component Analysis (ICA)

- Work on the signals multiplication in frequency domain and in discrete frequency bands by taking short time FFT

$$y(f,t) = W(f)x(f,t)$$



- Adaptive filter framework with LMS method

$$J_{\text{mse}}(\mathrm{w}) = E\{|e(n)|^2\} = E\{|x(n) - \hat{x}(n)|^2\}$$

- Negentropy maximization criteria from information theory, instead of target signal difference [2]

$$J(y) = H(y_{guass}) - H(y)$$

- In practice, due to the robustness to outliers, the cost function can be chosen as

$$J(w) = E\{G(|w^H x|^2)\} \quad G(y) = \log(y + \eta)$$

Newton method

Fit a quadratic approximation to $f(x)$ using both gradient and curvature information at x.

- Expand $f(x)$ locally using a Taylor series.

$$f(x+\delta x) = f(x) + f'(x)\delta x + \frac{1}{2} f''(x)\delta x^2 + o(\delta x^2)$$

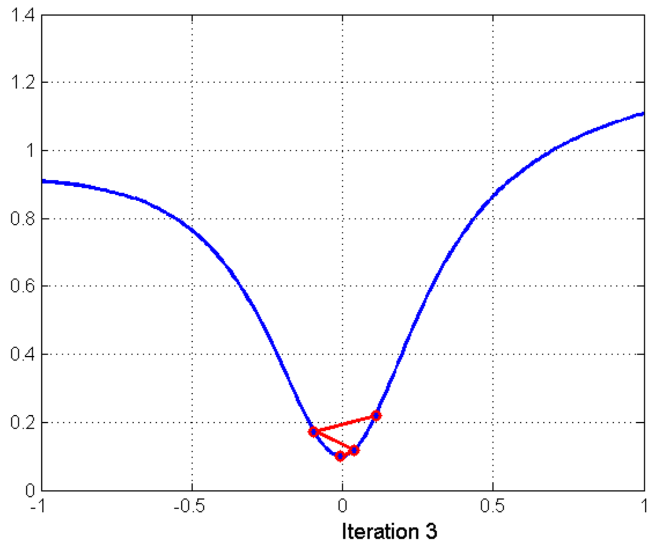
- Find the δx which minimizes this local quadratic approximation.

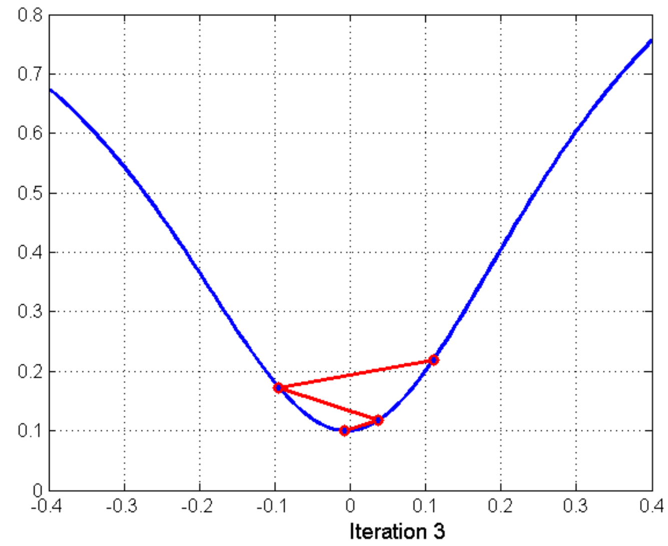
$$\delta x = -\frac{f'(x)}{f''(x)}$$

- Update x.

$$x_{n+1} = x_n - \delta x = x_n - \frac{f'(x)}{f''(x)}$$

Newton method

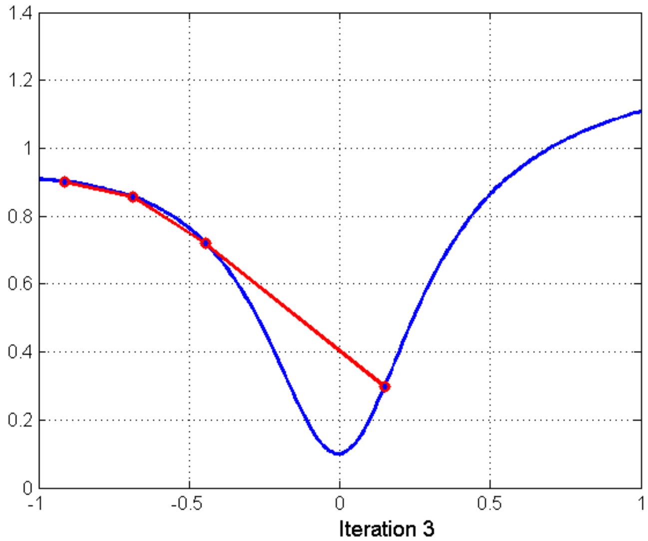


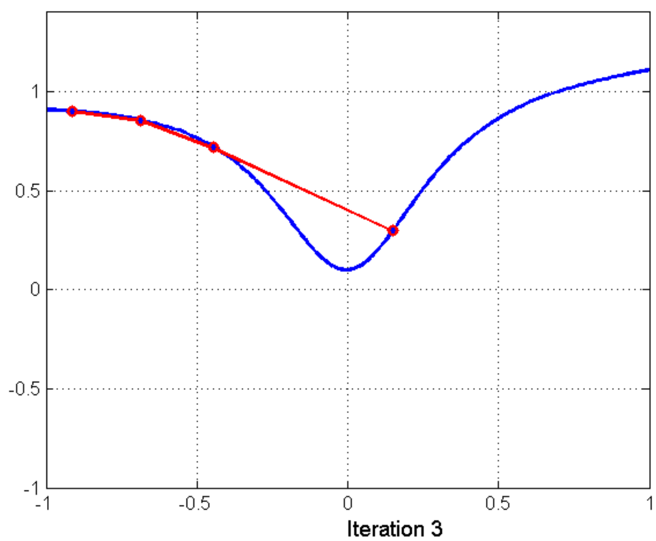


- avoids the need to bracket the root
- quadratic convergence (decimal accuracy doubles at every iteration)

Newton method

- Global convergence of Newton's method is poor.
- Often fails if the starting point is too far from the minimum.

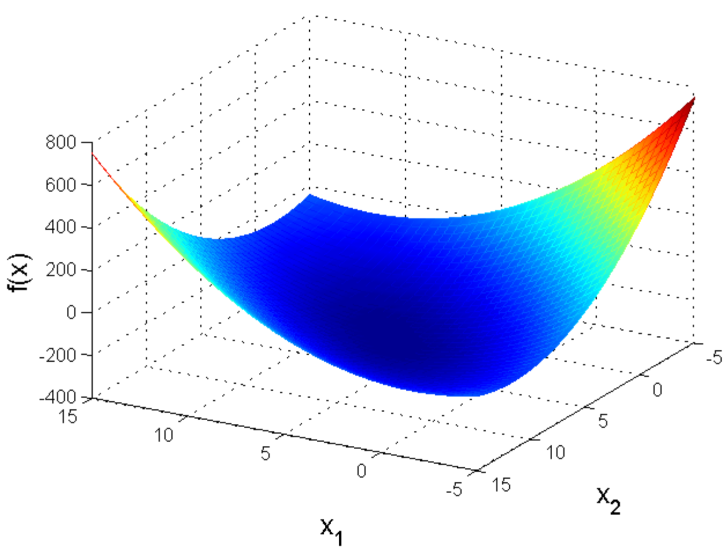


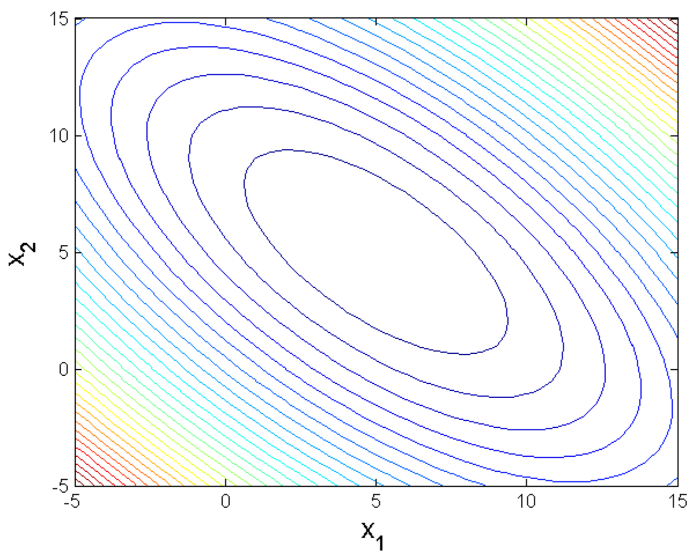


- in practice, must be used with a globalization strategy which reduces the step length until function decrease is assured

Extension to N (multivariate) dimensions

- How big N can be?
  - problem sizes can vary from a handful of parameters to many thousands





Taylor expansion

A function may be approximated locally by its Taylor series expansion about a point **x***

$$f(\mathbf{x}^* + \mathbf{x}) \approx f(\mathbf{x}^*) + \nabla f^T \mathbf{x} + \frac{1}{2}\mathbf{x}^T \mathbf{H} \mathbf{x}$$

where the gradient $\nabla f(\mathbf{x}^*)$ is the vector

$$\nabla f(\mathbf{x}^*) = \left[\frac{\partial f}{x_1} \dots \frac{\partial f}{x_N}\right]^T$$

and the Hessian **H(x*)** is the symmetric matrix

$$\mathbf{H}(\mathbf{x}^*) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_N \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_N^2} \end{bmatrix}$$

Equality constraints

- Minimize $f(\mathbf{x})$ subject to: $a_i(\mathbf{x}) = 0 \quad \text{for} \quad i = 1, 2, \ldots, p$

- The gradient of $f(\mathbf{x})$ at a local minimizer is equal to the linear combination of the gradients of $a_i(\mathbf{x})$ with

  Lagrange multipliers as the coefficients.

$$\nabla f(\mathbf{x}^*) = \sum_{i=1}^{p} \lambda_i^* \nabla a_i(\mathbf{x}^*)$$

In the BSS problem,

$$f(w) \; = \; E\{|w^H x|^2\} - 1 \; = \; ||w||^2 - 1 \; = \; 0$$

Inequality constraints

- Minimize $f(\mathbf{x})$ subject to: $c_j(\mathbf{x}) \geq 0$ for $j = 1, 2, \ldots, q$
- The gradient of $f(\mathbf{x})$ at a local minimizer is equal to the linear combination of the gradients of $c_j(\mathbf{x})$, which are active ($c_j(\mathbf{x}) = 0$)
- and Lagrange multipliers must be positive, $\mu_j \geq 0,\ j \in A$

$$\nabla f(\mathbf{x}^*) = \sum_{j \in A} \mu_j^* \nabla c_j(\mathbf{x}^*)$$

In the BSS problem, $g(w) = \varepsilon(y, r) - \xi \leq 0$ and $y = w^H x$

$$\begin{aligned} g(w) &= \frac{1}{N}(y-r)(y-r)^H - \xi \\ &= \frac{1}{N}(w^H x - r)(w^H x - r)^H - \xi \end{aligned}$$

Lagrangien

- We can introduce the function (Lagrangien)

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) - \sum_{i=1}^{p} \lambda_i a_i(\mathbf{x}) - \sum_{j=1}^{q} \mu_j c_j(\mathbf{x})$$

- The necessary condition for the local minimizer is

$$\nabla_x L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{0}$$

and it must be a feasible point (i.e. constraints are satisfied).

- These are Karush-Kuhn-Tucker conditions

Algorithm and Analysis

- For adaptive filtering, it is a MIMO optimization problem
- ICA with reference

$$\max_{y_i} C(y) = \Sigma_{i=0}^{m} J(y_i)$$

$$\text{subject to } \varepsilon(y_i, r_i) - \xi_l \leq 0, \quad \forall i = 1, 2, \ldots l$$

- Reference signals are chosen when very limited information is available about the source signals
- E.g. Use autocorrelation signal as reference for speech
- Optimization cost function (Lagrange function) for frequency domain ICA with reference:

$$L(w,\mu) = E\{G(|w^H x|^2)\} - \frac{1}{2\gamma}[\max\{\mu + \gamma g(w), 0\}^2 - \mu^2] - \lambda(f(w)) - \frac{1}{2}\gamma||f(w)||^2$$

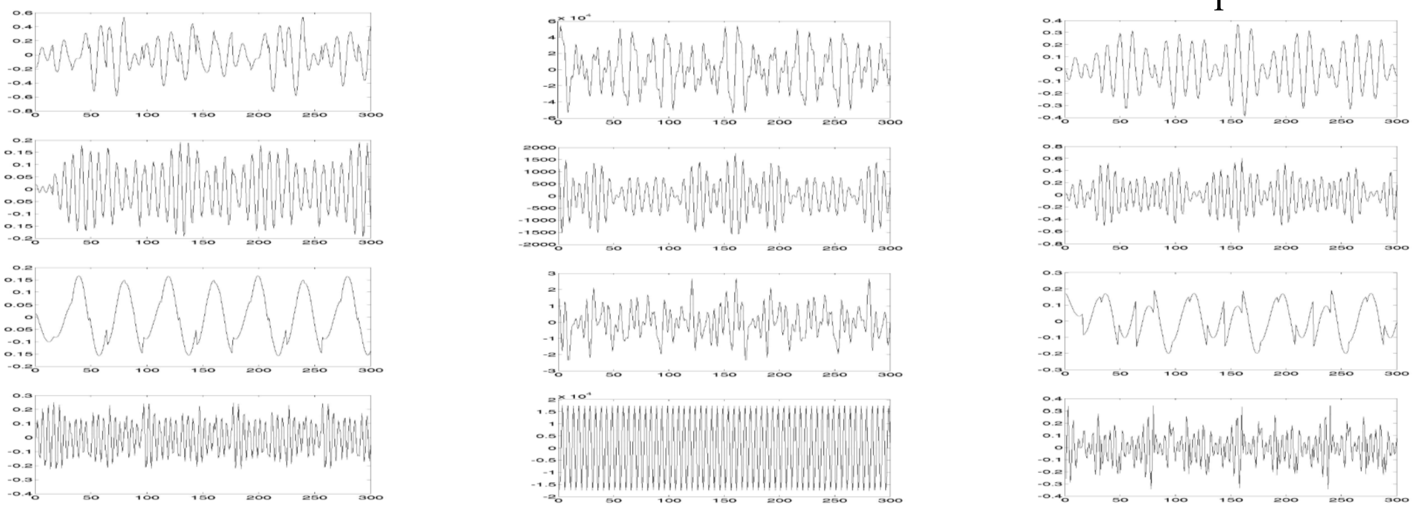
- Update weight and parameter using Newton's method:

$$w^{\text{new}} = w - \frac{\nabla L(w,\mu)}{\nabla^2 L(w,\mu)} \qquad w^{new} = \frac{w^{new}}{||w^{new}||} \qquad \mu^{new} = \max\{\mu + \gamma g(w), 0\}$$

Results and Reconstruction of Time-domain Signals

- Collection of data: BSS SP package and complex valued speech data
- Hermitian symmetric signal property for inverse Fourier Transform
- Reconstruction of the speech signals in time domain with selected frequency bands and overlap add method
- Speed of the approaches: time domain method converges faster

Source Mixtures Output



- Synthetic data SNR:

source	ICA-R	FICA
s_1	8.24(2.94)	10.25(5.06)
s_2	10.52(3.66)	11.35(1.05)
s_3	7.31(4.08)	10.13(3.40)
s_4	4.29(1.05)	6.49(2.28)