Biostatistics 615 - Statistical Computing

Topic 6
Multi-dimensional Optimization

Hyun Min Kang

Topic Overview

1. Coordinate descent algorithm
2. Nelder-Mead method
3. Gradient descent algorithm
4. Newton's method
5. Quasi-Newton methods
6. Advanced methods for multidimensional optimization

Multi-dimensional Optimization

What is multi-dimensional optimization?

- A function takes multiple variables (or a vector) as parameters.
- The function still returns a single-dimensional value.
- The objective is to find a combination of parameters that minimizes the objective function.

Why is multi-dimensional optimization challenging?

- The search space exponentially increases with the number of dimensions.
- A local minimum may not be the global minimum.

Topic Overview

1. Coordinate descent algorithm
2. Nelder-Mead method
3. Gradient descent algorithm
4. Newton's method
5. Quasi-Newton methods
6. Advanced methods for multidimensional optimization

Coordinate descent

Steps

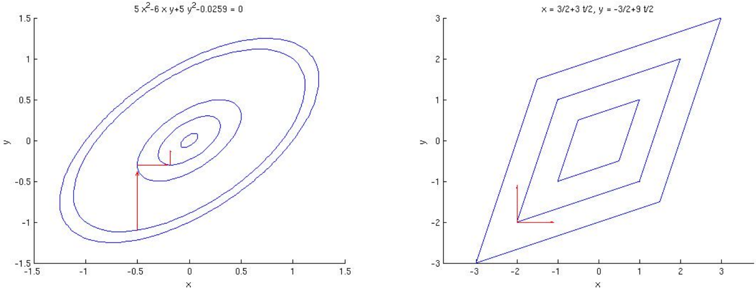
1. Start with initial values $\mathbf{x}^0 = (x_1^0, \ldots, x_n^0)$.
2. Minimize one variable at a time, fixing all the other variables.

$$x_i^{k+1} = \text{argmin}_y \; F(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, y, x_{i+1}^k, \ldots, x_n^k)$$

3. Repeat by selecting the next variable to minimize until convergence.

Key Features

- One-dimensional optimization algorithms may be used.
- Works well for additively separable functions.
- May converge slowly.
- May not converge with specific correlation between dimensions.



(http://en.wikipedia.org/wiki/Coordinate_descent)

Hands-on Session

Visit https://bit.ly/615top06r
Multi-dimensional coordinate descent

Topic Overview

1. Coordinate descent algorithm
2. Nelder-Mead method
3. Gradient descent algorithm
4. Newton's method
5. Quasi-Newton methods
6. Advanced methods for multidimensional optimization

The Nelder-Mead Method

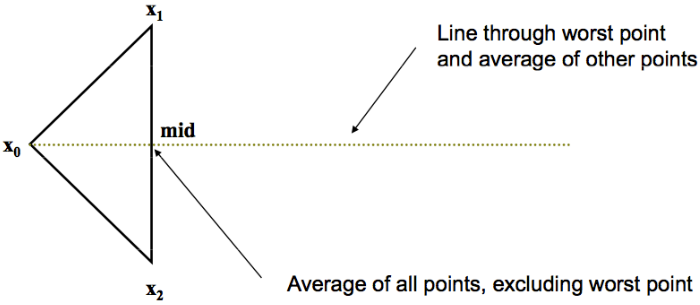
- a.k.a. downhill simplex method or amoeba method
- Calculate likelihoods at simplex vertexes
  - Geometric shape with $k+1$ corners
  - A triangle in $k=2$ dimensions
  - A tetrahedron in $k=3$ dimensions
- Simplex *crawls*
  - Towards minimum
  - Away from maximum
- A commonly used nonlinear optimization method, without using derivatives.
- May converge to non-stationary points.

(http://en.wikipedia.org/wiki/Nelder%E2%80%93Mead_method)

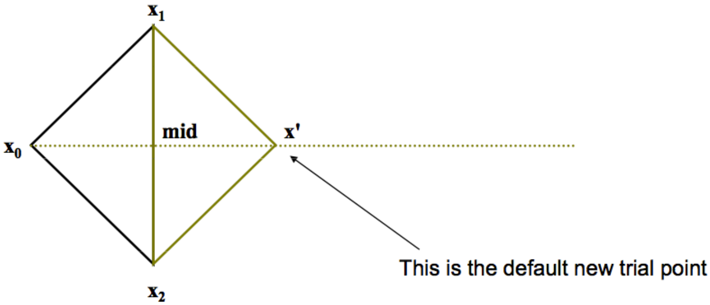
Nelder-Mead Method in Two Dimensions

- Evaluate functions at three vertexes
  - The highest (worst) point
  - The next highest point
  - The lowest (best) point
- Intuition
  - Move away from high point, towards low point

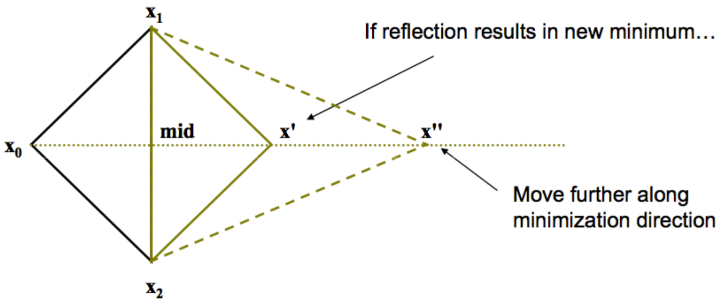
Direction for Optimization



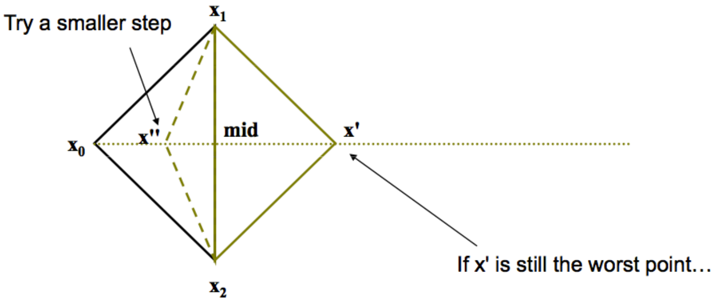
Reflection



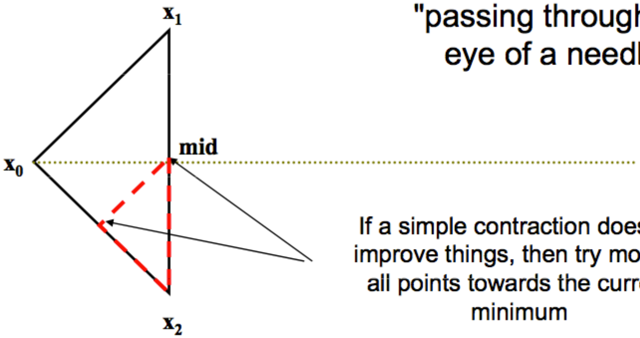
Reflection and Expansion



Contraction



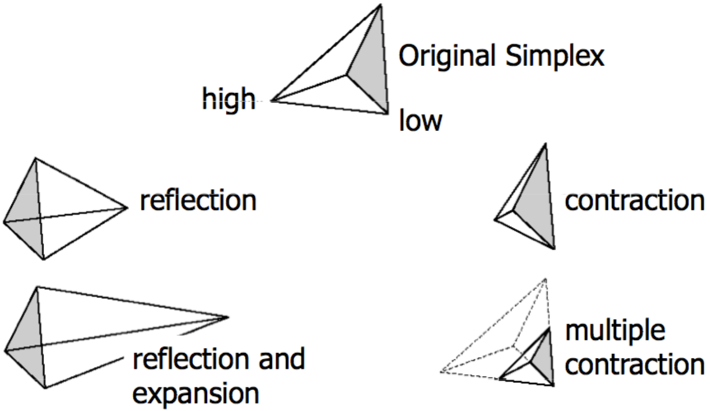
Multiple Contraction



"passing through the eye of a needle"

If a simple contraction doesn't improve things, then try moving all points towards the current minimum

Summary : The Nelder-Mead Method (in 3-D)



Hands-on Session

Visit https://bit.ly/615top06r
Nelder-Mead algorithm

Topic Overview

1. Coordinate descent algorithm
2. Nelder-Mead method
3. Gradient descent algorithm
4. Newton's method
5. Quasi-Newton methods
6. Advanced methods for multidimensional optimization

Gradient descent

- a.k.a. steepest descent.
- First order optimization algorithm (uses gradient).

$$x_{n+1} = x_n - \gamma_n \nabla F(x_n)$$

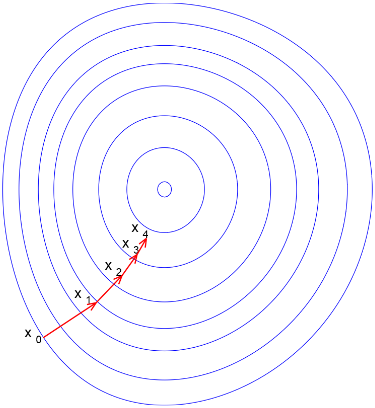
- The step size γ_n can be set as constant, or determined using a line search (e.g., golden section search).

$$\gamma_n = \frac{|(x_n - x_{n-1})^\top \{\nabla F(x_n) - \nabla F(x_{n-1})\}|}{\|\nabla F(x_n) - \nabla F(x_{n-1})\|^2}$$

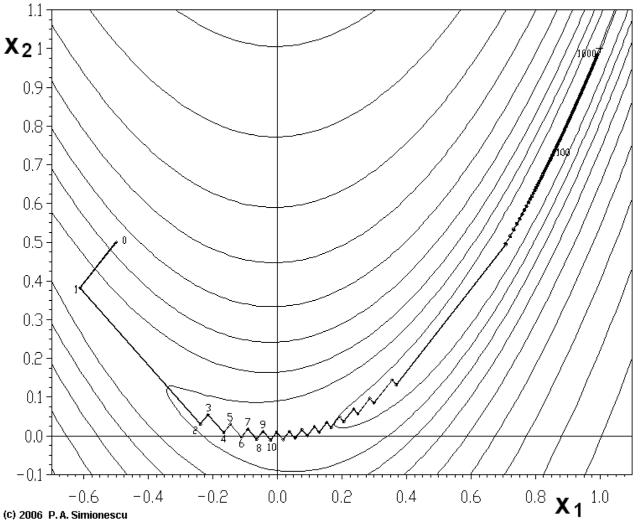
- May converge slow.

(http://en.wikipedia.org/wiki/Gradient_descent)

Gradient descent



(http://en.wikipedia.org/wiki/Gradient_descent)



(http://en.wikipedia.org/wiki/Gradient_descent)

Visit https://bit.ly/615top06r
Gradient Descent

Topic Overview

1. Coordinate descent algorithm
2. Nelder-Mead method
3. Gradient descent algorithm
4. Newton's method
5. Quasi-Newton methods
6. Advanced methods for multidimensional optimization

Newton's method

- Second order optimization algorithm (uses both gradient and Hessian).

$$x_{n+1} = x_n - [H_F(x_n)]^{-1} \nabla F(x_n)$$

- Converge much faster than gradient descent (quadratic convergence).
- Hessian is expensive to compute and store if dimension is high.
- The Hessian may be close to singular.

(http://en.wikipedia.org/wiki/Newton%27s_method_in_optimization)

Hands-on Session

Visit https://bit.ly/615top06r

Newton's method

Topic Overview

1. Coordinate descent algorithm
2. Nelder-Mead method
3. Gradient descent algorithm
4. Newton's method
5. Quasi-Newton methods
6. Advanced methods for multidimensional optimization

Quasi-Newton methods

- Approximation of the Newton method without requiring Hessian.
- The Hessian is updated by analyzing successive gradient vectors.
- A number of methods.
    - Davidon, 1959
    - DFP, 1963
    - Broyden, 1965
    - BFGS, 1970
    - L-BFGS, 1979
    - L-BFGS-B, 1995
- R function `optim` implements Nelder-Mead, BFGS and L-BFGS-B, as well as Brent, Conjugate Gradient and Simulated Annealing.

(http://en.wikipedia.org/wiki/Quasi-Newton_method)

Hands-on Session

Visit https://bit.ly/615top06r

- Utilizing the `optim()` function built-in R
- Evaluation of optimization functions

Topic Overview

1. Coordinate descent algorithm
2. Nelder-Mead method
3. Gradient descent algorithm
4. Newton's method
5. Quasi-Newton methods
6. Advanced methods for multidimensional optimization

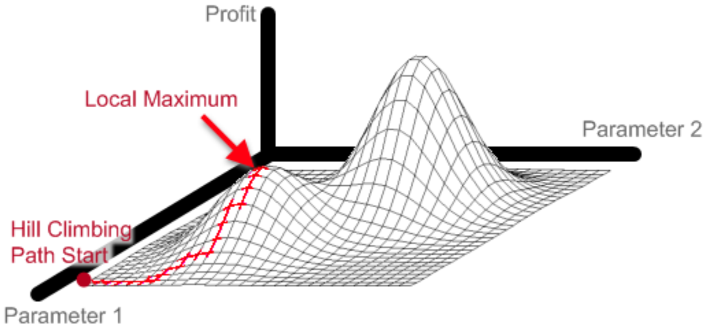
Advanced methods for multidimensional optimization

Hill-climbing algorithms

Aims to find local optimum that minimize/maximize the objective function.

- Coordinate descent
- Nelder-Mead algorithm
- Gradient descent
- Newton's method
- Quasi-Newton methods (e.g. L-BFGS-B)
- E-M algorithm (covered later)
- Quadratic programming (not covered in 615)
- ADMM, or Alternating Direction Method of Multipliers (not covered in 615)

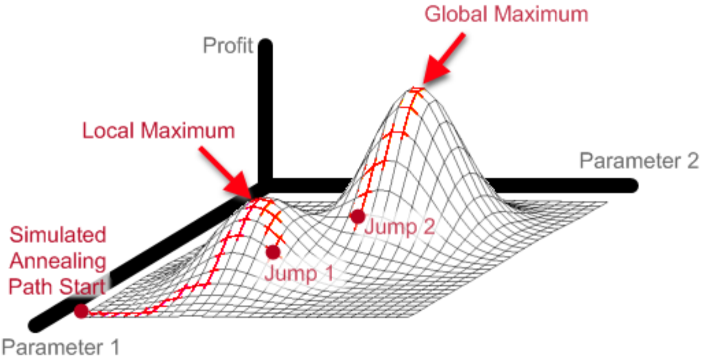
What are the limitations of the hill-climbing algorithms?



Hill-climbing algorithms may get stuck at a local optima and cannot move to the global optima.

Image credit: Max Dama

Methods for global optimization



To achieve global optimum, *chaotic jumps* may be necessary (i.e. do not always climb up).

Image credit: Max Dama

Global optimization methods

Monte Carlo methods

- Simulated Annealing
- Monte-Carlo Markov Chain (MCMC)
- Gibbs Sampler

Stochastic optimization algorithms

- Stochastic gradient descent
- Root Mean Squared Propagation (RMSProp)
- Adam (Adaptive moment estimation)

Some of these algorithms may be covered in BIOSTAT815