Biostatistics 615/815 Lecture 21: Simulated Annealing

Hyun Min Kang

April 5th, 2011

Recap - Dynamic Polymorphisms

```
class shape {                    // shape is an abstract class
public:
  virtual double area() = 0;  // shape object will never be created
}

class rectangle : public shape {
public:
  double x;
  double y;
  virtual double area() { return x*y; }
};

class circle : public shape {
public:
  double r;
  circle(double _r) : r(_r) {}
  virtual double area() { return M_PI*r*r; }
};
```

Recap : Function objects using dynamic polymorphisms

```
class optFunc {
 public:
  virtual double operator() (std::vector<double>& x) = 0;
};
class arbitraryOptFunc : public optFunc {
 public:
  virtual double operator() (std::vector<double>& x) {
    return 100*(x[1]-x[0]*x[0])*(x[1]-x[0]*x[0])+(1-x[0])*(1-x[0]);
  }
};
class mixLLKFunc : public optFunc {
  ... // many auxilrary functions
public:
  std::vector<double> data;
  virtual double operator() (std::vector<double>& x) {
    ...
  }
};
```

E-M algorithm : A Basic Strategy

- Complete data $(\mathbf{x}, \mathbf{z})$ - what we would like to have
  - Observed data $\mathbf{x}$ - individual observations
  - Missing data $\mathbf{z}$ - hidden / missing variables
- The algorithm
  - Use estimated parameters to infer $\mathbf{z}$
  - Update estimated parameters using $\mathbf{x}$
  - Repeat until convergence

Recap: The E-M algorithm

Expectation step (E-step)

- Given the current estimates of parameters $\theta^{(t)}$, calculate the conditional distribution of latent variable **z**.
- Then the expected log-likelihood of data given the conditional distribution of **z** can be obtained

$$Q(\theta|\theta^{(t)}) = \mathbf{E}_{\mathbf{z}|\mathbf{x},\theta^{(t)}} \left[\log p(\mathbf{x}, \mathbf{z}|\theta)\right]$$

Maximization step (M-step)

- Find the parameter that maximize the expected log-likelihood

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{t})$$

Summary : The E-M Algorithm

- Iterative procedure to find maximum likelihood estimate
  - E-step : Calculate the distribution of latent variables and the expected log-likelihood of the parameters given current set of parameters
  - M-step : Update the parameters based on the expected log-likelihood function
- The iteration does not decrese the marginal likelihood function
- But no guarantee that it will converge to the MLE
- Particularly useful when the likelihood is an exponential family
  - The E-step becomes the sum of expectations of sufficient statistics
  - The M-step involves maximizing a linear function, where closed form solution can often be found

Local and global optimization methods

Local optimization methods

- "Greedy" optimization methods
  - Can get trapped at local minima
  - Outcome might depend on starting point
- Examples
  - Golden Search
  - Nelder-Mead Simplex Method
  - E-M algorithm

Today

- Simulated Annealing
- Markov-Chain Monte-Carlo Method
- Designed to search for global minimum among many local minima

Local minimization methods

The problem

- Most minimization strategies find the *nearest* local minimum from the starting point
- Standard strategy
  - Generate trial point based on current estimates
  - Evaluate function at proposed location
  - Accept new value if it improves solution

The solution

- We need a strategy to find other minima
- To do so, we sometimes need to select new points that does not improve solution
- How?

Simulated Annealing

Annealing

- One manner in which crystals are formed
- Gradual cooling of liquid
  - At high temperatures, molecules move freely
  - At low temperatures, molecules are "stuck"
- If cooling is slow
  - Low energy, organized crystal lattice formed

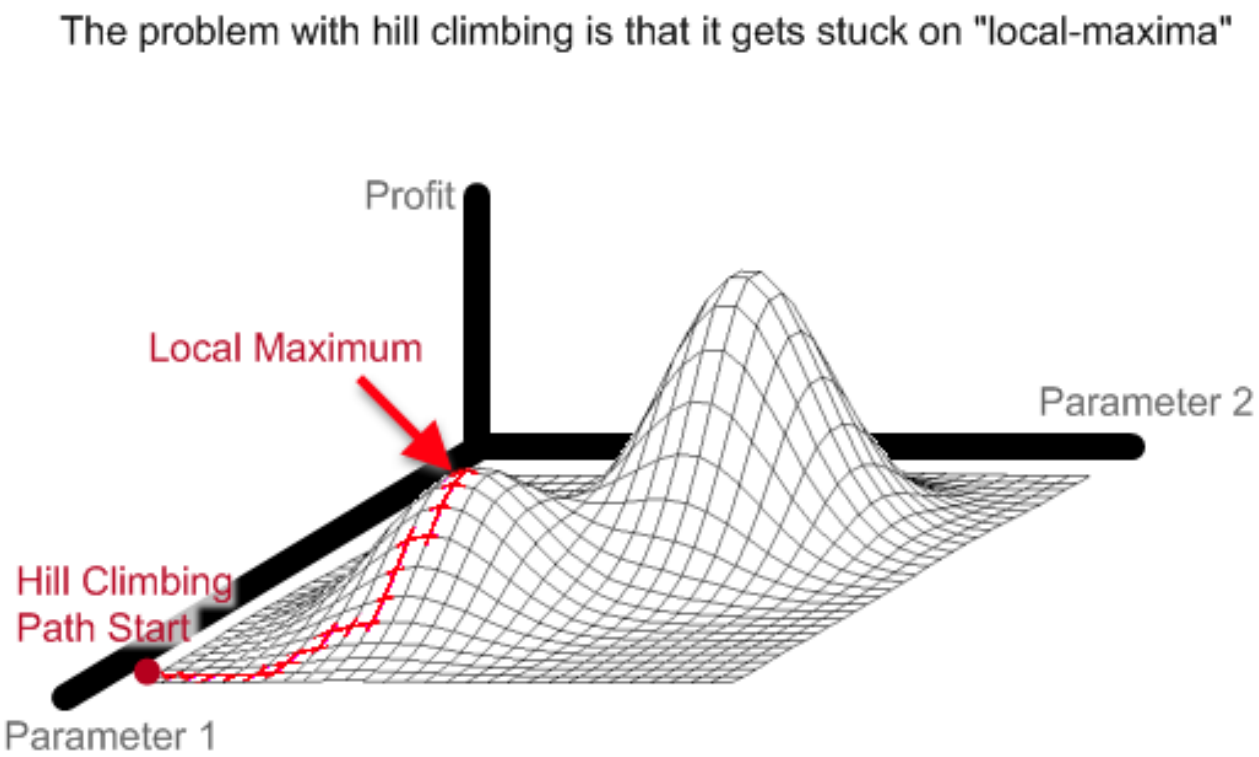
Simulated Annealing

- Analogy with thermodynamics
- Incorporate a temperature parameter into the minimization procedure
- At high temperatures, explore parameter space
- At lower temperatures, restrict exploration

Simulated Annealing Strategy

- Consider decreasing series of temperatures
- For each temperature, iterate these step
  - Propose an update and evaluation function
  - Accept updates that improve solution
  - Accept some updates that don't improve solution
    - Acceptance probability depends on "temperature" parameter
- If cooling is sufficiently slow, the global minimum will be reached

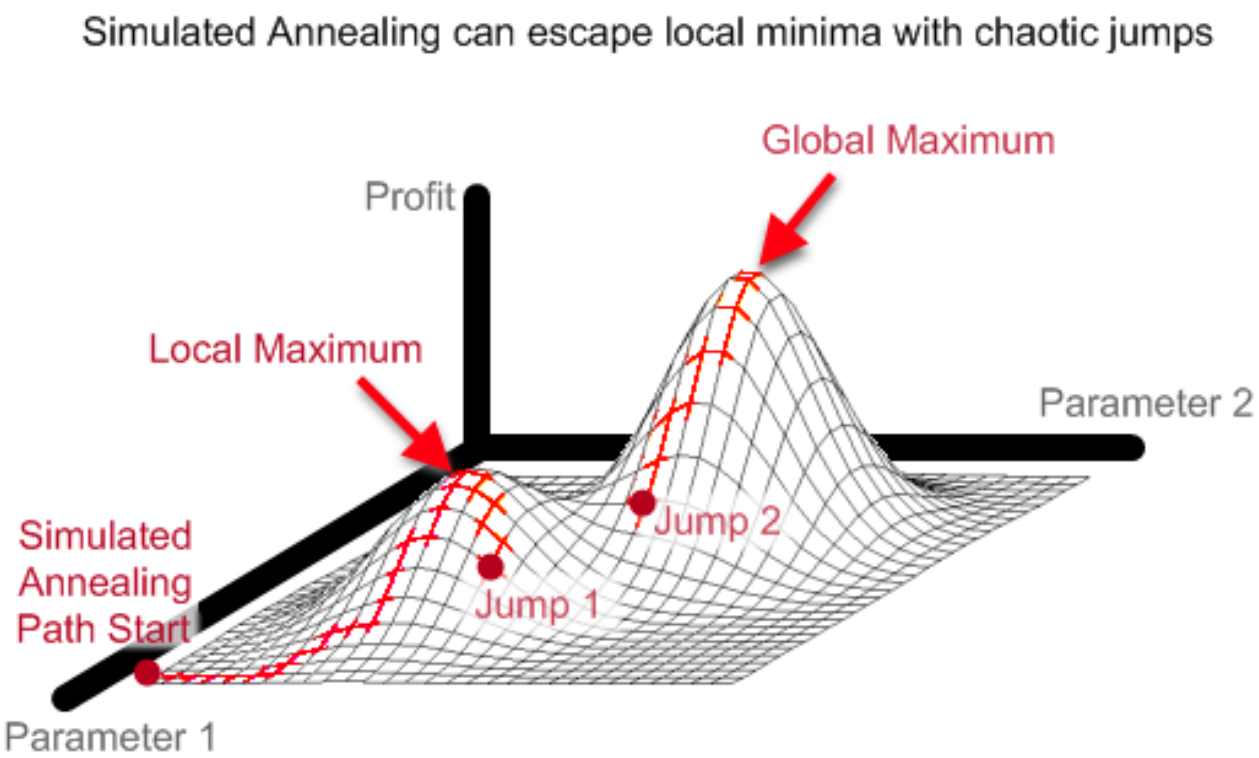
Local minimization methods



Images by Max Dama from

http://maxdama.blogspot.com/2008/07/trading-optimization-simulated.html

Global minimization with Simulated Annealing



Images by Max Dama from

http://maxdama.blogspot.com/2008/07/trading-optimization-simulated.html

Example Applications

- The traveling salesman problem (TSP)
  - Salesman must visit every city in a set
  - Given distances between pairs of cities
  - Find the shortest route through the set
- No polynomial time algorithm is known for finding optimal solution
- Simulated annealing or other stochastic optmization methods often provide near-optimal solutions.

Simulated Annealing TSP : Update Scheme

- A good scheme should be able to
  - Connect any two possible paths
  - Propose improvements to good solutions
- Some possible update schemes
  - Swap a pair of cities in current path
  - Invert a segment in current path

Examples of simulated annealing results

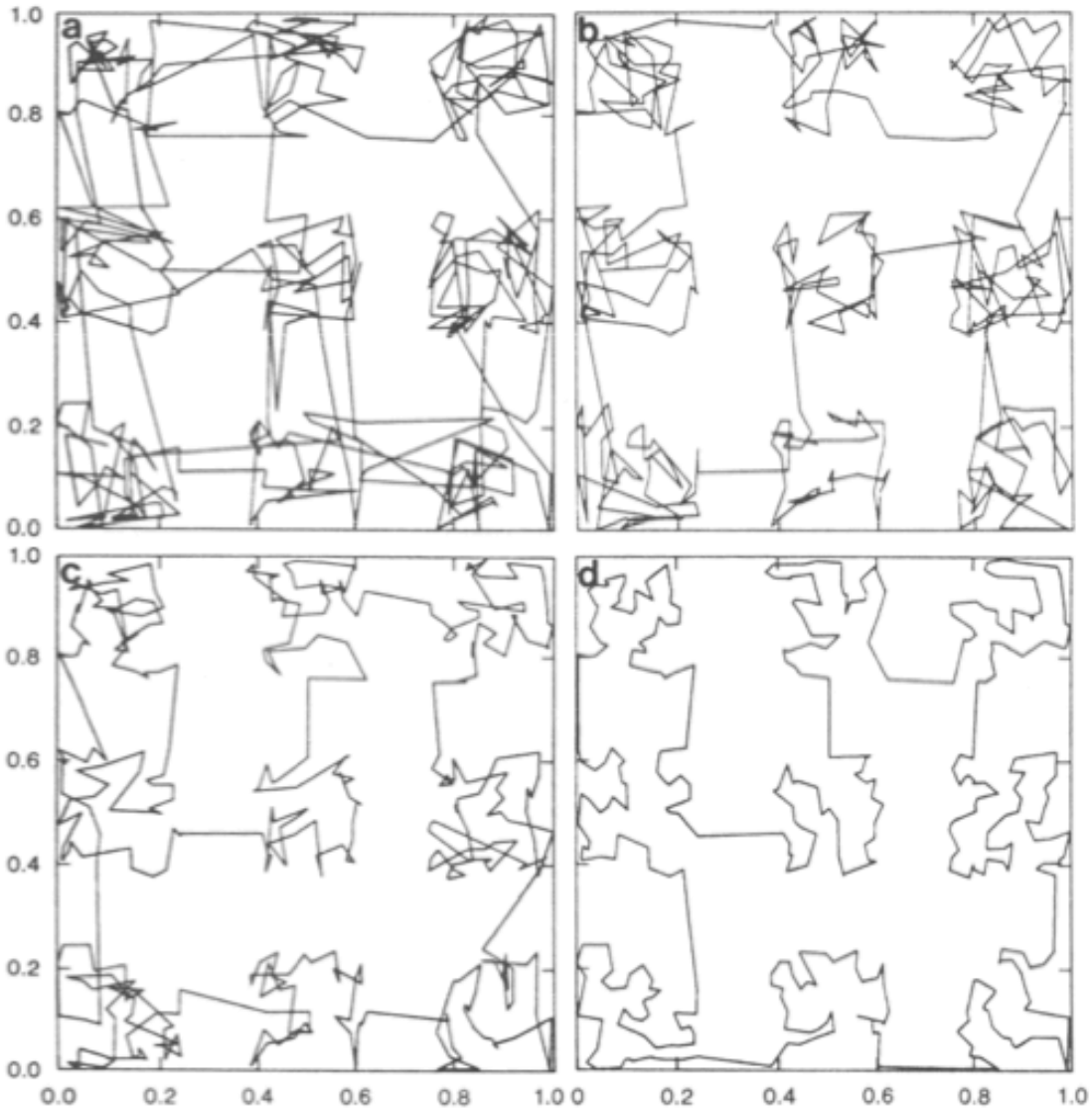


Fig. 9. Results at four temperatures for a clustered 400-city traveling salesman problem. The

Update scheme in Simulated Annealing

- Random walk by Metropolis criterion (1953)
- Given a configuration, assume a probability proportional to the Boltzmann factor

$$P_A = e^{-E_A/T}$$

- Changes from E_0 to E_1 with probability

$$\min\left(1, \frac{P_1}{P_0}\right) = \min\left(1, \exp\left(-\frac{E_1 - E_0}{T}\right)\right)$$

- Given sufficient time, leads to equilibrium state

Using Markov Chains

Markov Chain Revisited

- The Markovian property

$$\Pr(q_t|q_{t-1}, q_{t-2}, \cdots, q_0) = \Pr(q_t|q_{t-1})$$

- Transition probability

$$\theta_{ij} = \Pr(q_t = j|q_{t-1} = i)$$

Simulated Annealing using Markov Chain

- Start with some state q_t.
- Propose a changed q_{t+1} given q_t
- Decide whether to accept change based on $\theta_{q_t q_{t+1}}$
  - Decision is based on relative probabilities of two outcomes

Key requirements

- Irreducibility : it is possible to get any state from any state
  - There exist t where $\Pr(q_t = j | q_o = i) > 0$ for all (i, j).
- Aperiodicity : return to the original state can occur at irregular times

$$\gcd\{t : \Pr(q_t = i | q_0 = i) > 0\} = 1$$

- These two conditions guarantee the existence of a unique equilibrium distribution

Equilibrium distribution

- Starting point does not affect results
- The marginal distribution of resulting state does not change
- Equlibrium distribution π satisfies

$$\begin{aligned} \pi &= \lim_{t\to\infty}\Theta^{t+1} \\ &= (\lim_{t\to\infty}\Theta^{t})\Theta \\ &= \pi\Theta \end{aligned}$$

- In Simulated Annealing, $\Pr(E) \propto e^{-E/T}$

Simulated Annealing Recipes

1. Select starting temperature and initial parameter values
2. Randomly select a new point in the neighborhood of the original
3. Compare the two points using the *Metropolis criterion*
4. Repeat steps 2 and 3 until system reaches equilibrium state
   - In practice, repeat the process N times for large N.
5. Decrease temperature and repeat the above steps, stop when system reaches frozen state

Practical issues

- The maximum temperature
- Scheme for decreasing temperature
- Strategy for proposing updates
  - For mixture of normals, suggestion of Brooks and Morgan (1995) works well
  - Select a component to update, and sample from within plausible range

Implementing Simulated Annealing

```
class normMixSA {
public:
  int k;                        // # of components
  int n;                        // # of data
  std::vector<double> data;     // observed data
  std::vector<double> pis;      // pis
  std::vector<double> means;    // means
  std::vector<double> sigmas;   // sds
  double llk;                   // current likelihood
  normMixSA(std::vector<double>& _data, int _k);  // constructor
  void initParams();                              // initialize paemetrs
  double updatePis(double temperature);
  double updateMeans(double temperature, double lo, double hi);
  double updateSigmas(double temperature, double sdlo, double sdhi);
  double runSA(double eps);                       // run Simulated Annealing
  static int acceptProposal(double current, double proposal, double temperature);
};
```

Evaluating Proposals in Simulated Annealing

```
int normMixSA::acceptProposal(double current, double proposal,
                              double temperature) {
  if ( proposal < current )  return 1;  // return 1 if likelihood decreased
  if ( temperature == 0.0 )  return 0;  // return 0 if frozen
  double prob = exp(0-(proposal-current)/temperature);
  return (randu(0.,1.) < prob); // otherwise, probablistically accept proposal
}
```

Updating Means and Variances

- Select component to update at random
- Sample a new mean (or variance) within plausible range for parameter
- Decide whether to accept proposal or not

Updating Means

```
double normMixSA::updateMeans(double temperature, double min, double max) {
  int c = randn(0,k)           // select a random integer between 0..(k-1)
  double old = means[c];       // save the old mean for recovery
  means[c] = randu(min, max); // update mean and evaluate the likelihood
  double proposal = 0-mixLLKFunc::mixLLK(data, pis, means, sigmas);
  if ( acceptProposal(llk, proposal, temperature) ) {
    llk = proposal;            // if accepted, keep the changes
  }
  else {
    means[c] = old;            // if rejected, rollback the changes
  }
  return llk;
}
```

```
double normMixSA::updateSigmas(double temperature, double min, double max) {
  int c = randn(0,k)            // select a random integer between 0..(k-1)
  double old = sigmas[c];       // save the old mean for recovery
  sigmas[c] = randu(min, max); // update a component and evaluate the likelihood
  double proposal = 0-mixLLKFunc::mixLLK(data, pis, means, sigmas);
  if ( acceptProposal(llk, proposal, temperature) ) {
    llk = proposal;             // if accepted, keep the changes
  }
  else {
    sigmas[c] = old;            // if rejected, rollback the changes
  }
  return llk;
}
```

Updating Mixture Proportions

- Mixture proportions must sum to 1.0
- When updating one proportion, must take others into account
- Select a component at tandom
  - Increase or decrease probability by up to 25%
  - Rescale all proportions so they sum to 1.0

Updating Mixture Proportions

```
double normMixSA::updatePis(double temperature) {
  std::vector<double> pisCopy = pis; // make a copy of pi
  int c = randn(0,k);                // select a component to update
  pisCopy[c] *= randu(0.8,1.25);  // update the component
  // normalize pis
  double sum = 0.0;
  for(int i=0; i < k; ++i)
    sum += pisCopy[i];
  for(int i=0; i < k; ++i)
    pisCopy[i] /= sum;
  double proposal = 0-mixLLKFunc::mixLLK(data, pisCopy, means, sigmas);
  if ( acceptProposal(llk, proposal, temperature) ) {
    llk = proposal;
    pis = pisCopy;   // if accepted, update pis to pisCopy
  }
  return llk;
}
```

Initializing parameters

```
void normMixSA::initParams() {
  double sum = 0, sqsum = 0;
  for(int i=0; i < n; ++i) {
    sum += data[i];
    sqsum += (data[i]*data[i]);
  }
  double mean = sum/n;
  double sigma = sqrt(sqsum/n - sum*sum/n/n);
  for(int i=0; i < k; ++i) {
    pis[i] = 1./k;                // uniform priors
    means[i] = data[rand() % n]; // pick random data points
    sigmas[i] = sigma;            // pick uniform variance
  }
}
```

Putting things together

```
double normMixSA::runSA(double eps) {
  initParams();                              // initialize parameter
  llk = 0-mixLLKFunc::mixLLK(data, pis, means, sigmas); // initial likelihood
  double temperature = MAX_TEMP;             // initialize temperature
  double lo = min(data), hi = max(data); // min(), max() can be implemented
  double sd = stdev(data);                   // stdev() can also be implemented
  double sdhi = 10.0 * sd, sdlo = 0.1 * sd;
  while( temperature > eps ) {
    for(int i=0; i < 1000; ++i) {
      switch( randn(0,3) ) {  // generate a random number between 0 and 2
      case 0:                 // update one of the 3*k components
        llk = updatePis(temperature);  break;
      case 1:
        llk = updateMeans(temperature, lo, hi); break;
      case 2:
        llk = updateSigmas(temperature, sdlo, sdhi); break;
      }
    }
    temperature *= 0.90;      // cool down slowly
  }
  return llk;
}
```

Running examples

```
user@host:~/> ./mixSimplex ./mix.dat
Minimim = 3043.46, at pi = 0.667271,
between N(-0.0304604,1.00326) and N(5.01226,0.956009)

user@host:~/> ./mixEM ./mix.dat
Minimim = -3043.46, at pi = 0.667842,
between N(-0.0299457,1.00791) and N(5.0128,0.913825)

user@host:~/> ./mixSA ./mix.dat
Minimim = 3043.46, at pi = 0.667793,
between N(-0.030148,1.00478) and N(5.01245,0.91296)
```

Comparisons

2-component Gaussian mixtures

- Simplex Method : 306 Evaluations
- E-M Algorithm : 12 Evaluations
- Simulated Annealing : $\sim 100,000$ Evaluations

For higher dimensional problems

- Simplex Method may not converge, or converge very slowly
- E-M Algorithm may stuck at local maxima when likelihood function is multimodal
- Simulated Annealing scale robustly with the number of dimensions.

Summary

Today - Simulated Annealing

- Simulated Annealing
- Markov-Chain Monte-Carlo method
- Searching for global minimum among local minima

Next lecture

- More on MCMC Method
- A simple Gibbs Sampler