An Introduction to MCMC Techniques for Optimization

Kyle Simek

Department of Computer Science, University of Arizona

November 19, 2010

MCMC Review

- Goal: To generate samples from an arbitrary probability distribution
- Problem: Most distributions can't be sampled from directly.
- Problem: Many distributions are only known up to a scaling constant
  - Example: Bayesian models, $p(D)$ is unknown.
- Solution: Markov Chain Monte Carlo
  - Simulate a Markov Chain that converges to the target distribution.
  - After converging, all simulated states of the chain are samples from the target distribution.
- **Canonical Example**: Metropolis-Hastings

Algorithm: Metropolis Hastings

1. Pick starting state, x_0
2. Pick a new state, x_1 $q(x|x_0)$

Issues with Metropolis Hastings

- Random Walk
  - Takes N^2 steps to move N steps away from initial state
  - Strategies:
    - Gibbs Sampling
    - Stochastic Dynamics
    - Hybrid Monte Carlo
  - Not covered today
- Quasi-ergodicity
  - Recall: Ergodicity $\triangleq$ every state x' is reachable from any state x in a finite number of steps.
  - Def: Quasi-ergodicity - probabilitiy of moving from x to x' is non-zero but so small as to be effectively impossible.
  - Many **deep** local maxima
  - Sampler is likely to get caught and won't escape.
  - Focus of today's talk.

Dealing with Quasi-ergodicity

- Simulated Annealing
  - Deterministic cooling schedules
  - Adaptive cooling schedules
  - Restarting
  - Simulated Tempering (a.k.a. Serial Tempering, a.k.a. Umbrella Sampling)
  - Replica Exchange (a.k.a. Parallel Tempering)
- Stochastic Tunneling
- Others (Not covered today)
  - Jump Walking
  - Smart Walking
  - Cool Walking
  - Smart Darting

Statistical Physics

- Most techniques discussed today have roots in statistical physics.
- **Traditional Formulation**: Given some probability distribution, $p(x)$, find the state x^* with the maximum probability.
- **Statistical Physics Formulation**: Given some potential energy function $U(x)$, find the state x^* with lowest energy.
- **Example**: *Molecular dynamics* - given a system of N molecules, find low-energy crystal configurations.

Traditional vs. Statistical Physics Formulations

- Physical formulation is analagous to Traditional formulation
  - Boltzman distribution for MD:

$$p(x) = Z(T)e^{-U(x)/T}$$

  - Each of these terms has an analogue in the traditional formulation.
  - $U(x) =$ "potential function".
    - Proportional to the negative log probability.
    - **Minimizing $U(x) =$ maximizing $p(x)$.**
    - When $U(x)$ is quadratic, $p(x)$ is a Gaussian distrubution.
  - $T =$ system's "temperature".
    - Informally, a "sharpness" parameter (more on this later).
    - Represents variance when $p(x)$ is Gaussian.
  - $Z(T) =$ "partition function", i.e. normalization factor.
    - Often unknown, but in most MCMC scenarios not needed (we'll see an exception).
    - In Bayesian models, it contains the unknown $p(D)$ in the denominator of Bayes' Rule.

Simulated Annealing - General Idea

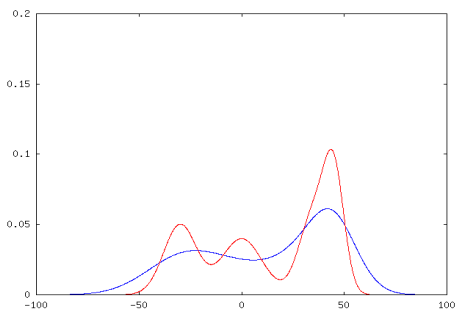
- Def: Annealing - a metalurgic process of cooling hot metal slowly to increase it's ductatility and relieve internal stresses.
  - Recall that temperature is a measure of the speed of molecules/atoms in the system.
  - During cooling, fast-moving atoms slow down and lose energy.
  - In **fast** cooling, motion of atoms is halted before they can find a low-energy state.
  - By decreasing temperature **slowly**, atoms have time to find low-energy states while they are still fairly mobile.
- We can model this system statstically.

Simulated Annealing (ctd.)

- Let X be the position and momentum of all molecules in a system.
- Let $H(x)$ be the the system's energy in configuration x.
- Let T be the system's temperature.
- The exact position and momentum of all molecules is unknowable.
- But we can model them probabilistically as

$$p(x) = Z(T)e^{(} - H(x)/T)$$

- Note that increasing the temperature "irons out" the peaks in the distribution.



Simulated Annealing (ctd.)

- If "hotter" systems move around more quickly, we can use this to improve sampling.
  - **To explore**: Increase the temperature T to smooth out wells.
  - **To find minima**: Decrease T to make wells deeper.
  - (Demo)
- **In traditional formulation**: raise all probabilities to $1/T$:

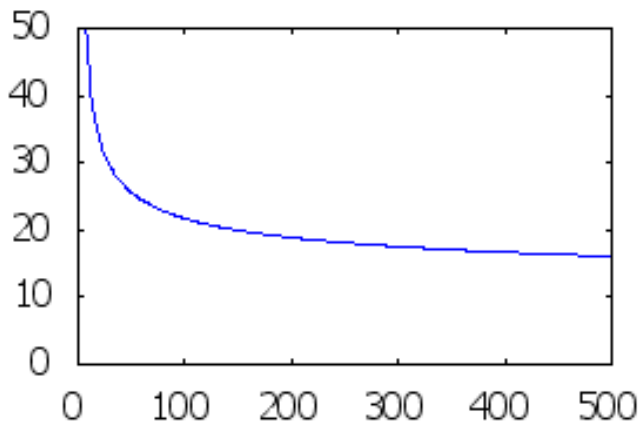
$$p(x) \to p(x)^{1/T}$$

.

- General algorithm: Start with high temperature T_0 and decrease throughout sampling.
- Def: **Cooling Schedule** - Defines how temperature changes as a function of time.
- **Performance depends heavilly on a good cooling schedule.**

Simulated Annealing - Deterministic Cooling schedules

- **Trivial Schedule**: if $T_0 = \inf$ and $T_1 = 0$, algorithm becomes gradient descent.
- **Optimal Schedule**: (under certain conditions) $T_t = T_0 / \log(t)$, for $t > 0$
  - Takes much too long to cool to $T = 0$.
  - Usually worse than exhaustive search (which is, itself, intractible).



- Another commonly used schedule: $T_{t+1} = \alpha T_t$.
  - α usually chosen by hand.

Simulated Annealing - Adaptive Cooling Schedule

- **Idea**: Adapt the cooling schedule based on the state of the Markov chain.
- **Constant Rate of Entropy Reduction**
  - "Peakiness" of a distribution is measured by entropy.
  - **Idea**: Entropy should change smoothly and linearly.
  - Rate of entropy reduction: $\frac{dS}{dT} = \frac{C}{T}$.
  - C: the heat capacity of the system; $C = \text{var}(E)/T^2$
  - $\text{var}(E)$: the variance of the energies at the current temperature.
  - **Cooling Schedule**: $T_t - T_{t+1} \propto T/C = \mathbf{T^3/\text{var}(E)}$

Algorithm: Cooling with Constant Rate of Entropy Reduction

1. Sample for a while at fixed temperature T_t.
2. Find variance of sampled energies.
3. Reduce temperature by $T^3/\text{var}(E)$

Restarting

- Problem: In simulated annealing, bad samples can get caught it deep wells if temperature drops too fast.
- **Idea**: "If you hit a dead end, start over"

Simulated Annealing with Restarting

1. Perform simulated annealing as usual.
2. Keep track of best model and its energy throughout sampling.
3. If restarting criterion is met,
   1. Replace current model and energy with the best
   2. Restart the annealing schedule

- "Restarting criterion" – various options
  - Restart when quality of samples has dropped too far below the best.
  - Restart randomly.

Simulated Tempering

- **Simulated Tempering**: continuously jump between higher and lower temperature states.
  - Samples don't get caught in deep wells.
  - Obviates the need for hand-built cooling schedule.
- Temperature is now part of the model; it is a parameter that is sampled over.
- A new temperature is proposed, and accepted with Metropolis probability:

$$min\left\{1, \frac{Z(T_j)}{Z(T_i)} \exp\{-\Delta H\}\right\}$$

  - $Z(T_i)$ is the normalizing constant for temperature T_i
  - $\Delta H = E(1/T_j - 1/T_i)$
- Problem: $Z(T_i)$ is usually unknown
  - "These constants can be preliminarily estimated by an iteration procedure." (Cong, et. al., 2002)
  - Could assume it doesn't change much, and ignore it?

Parallel Tempering (Replica Exchange)

- **Parallel Tempering**: Simulate multiple chains in parallel, and allow them to exchange temperatures.
- Probability of accepting a temperature swap:

$$min\left\{1, e^{(E_i - E_j)\left(\frac{1}{kT_i} - \frac{1}{kT_j}\right)}\right\}$$

- A popular extension to simulated tempering (serial tempering).
- Unlike serial tempering, no need for normalization factor (but it's unclear why).
- Mixes faster than just running two chains in isolation.

Stochastic Tunneling

- Instead of flattening out **all** wells, why not just flatten the ones shallower than the deepest one found so far?
- Transform the potential function:

$$f^*(x) = 1 - exp(-\gamma(f(x) - f_0))$$

- Note that the location of minima are unchanged.
- Depth of shallow wells are even shallower.
- Deep wells become deeper.
- Prevents re-exploring areas, avoids random walk somewhat.