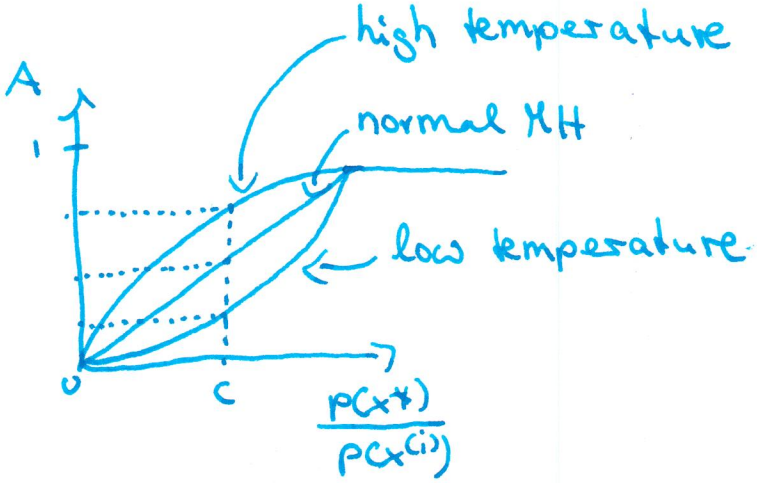
Lecture 15 : Simulated Annealing

- In the previous lecture, Pavlos has introduced the notion of a temperature parameter to alter the acceptance probability of MH:



- As you can see from the dotted lines, we accept with a higher probability for the high temperature, than for the low temperature.

  ⇒ high temperature encourages our walker to explore the space.

- So where do we put our temperature into the acceptance probability?

- We define $p'(x) = e^{-(-\log(p(x)))}$

- So far $p'(x) = p(x)$

- Now we introduce the temperature:

$$p'(x) = e^{-\frac{(-\log(p(x)))}{T}}$$

We can see that we got a Boltzmann distribution by defining the energy $E = -\log(p(x))$

$$\Rightarrow p'(x) = e^{-\frac{1}{T}E(x)}$$

Note that for $T \neq 1 \Rightarrow p'(x) \neq p(x)$, which is why in parallel tempering we only sample from the chain with $T = 1$.

Now we look at the resulting acceptance probability

$$A = \frac{p'(x^*)}{p'(x)} = \frac{e^{-\frac{1}{T}E(x^*)}}{e^{-\frac{1}{T}E(x)}}$$

$$= e^{-\frac{1}{T}E(x^*) + \frac{1}{T}E(x)} = e^{-\frac{1}{T}\Delta E}$$

with $\Delta E = E(x^*) - E(x)$

- So for $T \to 0$ our acceptance probability also goes down $A \to 0$

⇒ For small temperatures we become more greedy and only accept steps up the hill.

- In parallel tempering we said $T<1$ is not useful, but now we are changing our goal from sampling to optimization.

  ⇒ we don't care about the whole distribution, we only want the maximum / minimum of $p(x)$.

  ⇒ we can cast this into the problem of minimizing $E(x)$, the energy of the system.

Simulated annealing follows the idea of annealing in metallurgy, where the material is first heated and then cooled slowly, so that all molekules have time to settle into a low energy structure.

Let's look at the actual algorithm:

- set initial x_0
- set initial temperature
- set proposal step size
- run MH for n iterations
- cool down
- loop until convergence.

Open questions:

- the usual x_0 and step size
- initial temperature
- number of iterations to run MH
- cooling schedule
- convergence / stopping criteria.

$\Rightarrow$ The ART of simulated annealing.

The following are some practical choices / heuristics, it depends on your problem what will work well.

Step-size: An idea is to skew the proposals towards states with an energy similar to the current state. The idea is that this heuristic tends to exclude a few very good candidates, but also many more bad candidates.

Initial temperature: The temperature is working against ΔE in a sense in the beginning, so should be a bit higher than ΔE.

number of iterations to run MH:

- Ideally until the chain for the current temperature converges, Pavlos often uses 100.

Cooling Schedule:

- linear: $T_{k+1} = \alpha \cdot T_k \quad 0.8 < \alpha < 0.99$
- logarithmic: $T_{k+1} = \frac{1}{\log(T_k)}$
- exponential: $0.95^k \cdot T_k$

stopping criteria:

- fixed number of iterations
- changes in energy are very small
- objective limit: run until solution is good enough.

In addition: reheating sometimes it can be useful to reheat the system to explore more and find different local minima.

TSP example

Why does simulated annealing work?

- It not only works well in praxis, if we would have an infinite amount of time we are actually guaranteed to find the global optimum!

- Let's look again at $p'(x)$, the distribution our MH in the inner loop samples:

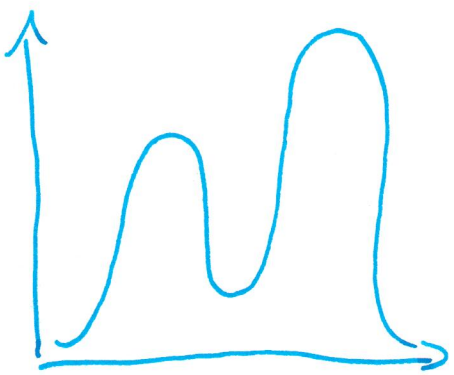
$$p'(x) = e^{-\frac{[-\log(p(x))]}{T}}$$

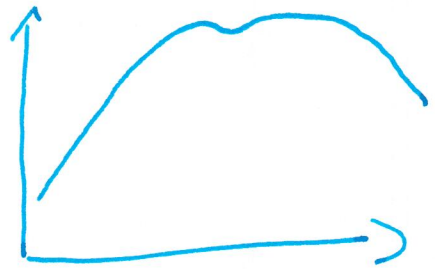
$$= e^{-\frac{1}{T}[-\log(p(x))]}$$

$$= \left[e^{\log(p(x))}\right]^{\frac{1}{T}}$$

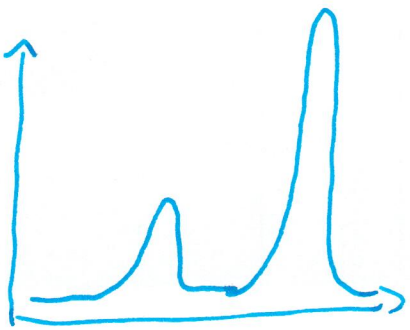
$$= [p(x)]^{\frac{1}{T}}$$

So what does the temperature do to our $p(x)$?





high temperature



low temperature

What we are doing in simulated annealing is running several MH chains (one for each temperature). And by slowly changing the temperature we make $p'(x)$ more and more peaked, trapping the MH at the global maximum.

$\Rightarrow$ If we cool too fast, the MH is getting trapped in a local optimum $\Rightarrow$ this is where restarting comes in handy.