Simulated Annealing

What is annealing?

Annealing is the process in which we heat a metal to incredibly high temperatures in order to **break down its fine crystal structure**. We then cool the metal at a proper rate which allows us to reform the crystal structure to be optimally suited to a specific use case

Statmech 001

This is basically fundamental to statistical mechanics. The **boltzmann distribution** describes the distribution of energy in a system (in general). The form of the distribution which matters to us is shown below

$$f(E) = e^{-\frac{E}{kT}}$$

Where E is the energy of the system, k is boltzmann's constant, and T is the system temperature. In english, this means that the probability of a particle in the system having energy E is one divided by euler's e to the power of the energy divided by boltzmann's constant times temperature (which turns out to be the average energy of the system but that is just me being excitable)

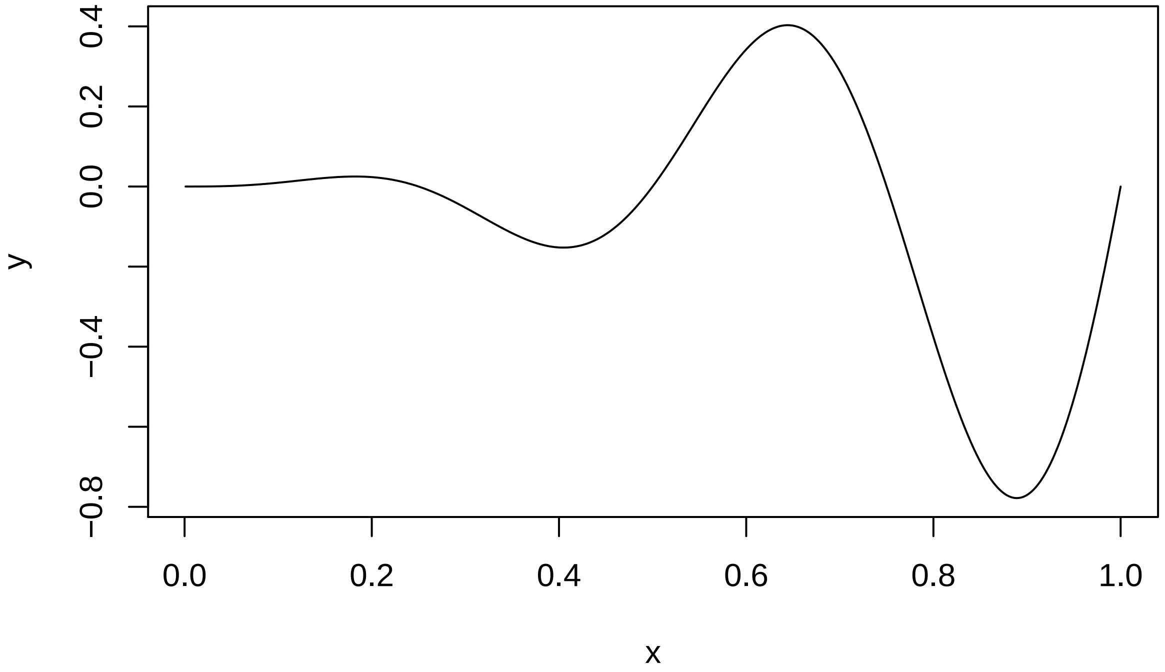
This suggests that we can define the probability of a change in energy of magnitude δE

$$\mathbb{P}(\delta E) = e^{-\frac{\delta E}{kT}}$$

What does this mean to us? This means a few things intuitively. First, the probability of a change in energy is overall greater at higher temperatures. This makes perfect sense, as hotter things are, in general, more volatile. Likewise, a significant change in energy is unlikely at a low temperature, which also is fairly intuitive. The other thing this suggests is that large jumps in energy are less likely (δE big), which is a reasonable claim as well, as we do not see a lot of fires freezing immediately, or icebergs suddenly boiling.

Loss as energy

What does any of this have to do with machine learning, optimization, or loss functions?? Let us imagine a non convex loss function, for example something like this:



For this function, a naive "ball simply rolling down a hill" approach will not work at all, as it will get stuck in the first local optima. One way to mediate this is to add a bit of stochasticity (randomness), and descend the loss gradient with some random jumps in the wrong direction to get out of the local optima. Another approach to solving this problem is to **model the loss function as system energy, and try to minimize the system energy**. This is essentially what simulated annealing is doing.

Simulated annealing: search pattern

With simulated annealing, we are going to first start with an initial guess, lets call it x_1, and the energy, or loss, at x_1 we will call $f(x_1)$. To generate a new guess, we are going to take a random step away from x_1, and call it x_2. If $f(x_2) < f(x_1)$, we are going to go ahead and say that x_2 is our new starting point for the next random step. However, if $f(x_2) > f(x_1)$, we are not just going to throw it away. This represents an increase in loss, or an increase in energy. Boltzmann tells us that there is a chance we will actually move to a higher energy state, described in the equations above. Therefore, we will calculate the probability of the increase in energy, and then we will produce a random number between 0 and 1. If that random number is less than the probability, we will accept x_2 as our new starting point (so if the probability of moving to a higher energy state is 0.9, and we produce 0.7, we are going to go to the higher energy state). If it is greater than the probability, we will stick with x_1

Simulated annealing: Cooling

This is where the real interesting part of simulated annealing lies. In the async, Professor Santerre discussed the problems of how to choose the step size, and exploration vs exploitation. In the annealing process, we first heat the metal to a very high temperature, which causes the underlying structure to break down. Therefore, in simulated annealing, we also start at a very high "temperature". Because the temperature is high, we are going to accept a lot of random points, which **ignore the loss landscape**. At a high temperature, we are essentially randomly sampling the loss function (*exploration*). Then, after each round of sampling, we let the temperature cool, essentially slowly refining our random samples, or shifting from exploration to exploitation (slowly). Finally, when the temperature is cold, we descend to the bottom of the optima we are at, (hopefully) ending up at the global optima.