Metaheuristic Optimization: Simulated Annealing

Adaptive and Cooperative Algorithms (ECE 457A)

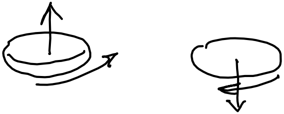
ECE, MME, and MSCI Departments,
University of Waterloo, ON, Canada

Course Instructor: Benyamin Ghojogh

Fall 2023

Boltzmann (Gibbs) Distribution and Statistical Physics

Boltzmann (Gibbs) Distribution



- Centuries ago, the **Boltzmann distribution** (1868) [1], also called the **Gibbs distribution** (1902) [2], was proposed.
- This energy-based distribution was found to be useful for modeling the **physical systems statistically** [3].
- One of these systems was the **Ising model** which modeled interacting particles with binary spins [4, 5].
- Assume we have several **particles** $\{x_i\}_{i=1}^{d}$ in **statistical physics**.
- These particles can be seen as random variables which can **randomly have a state**. For example, if the particles are **electrons**, they can have **states** $+1$ **and** -1 **for counterclockwise and clockwise spins**, respectively.

Boltzmann (Gibbs) Distribution

softmax function $\frac{e^{x_i}}{\sum e^{x_j}}$, $Z = \sum e^{x_j}$

$x_1, \dots, x_d \Rightarrow \frac{x_i}{\sum_{j=1} x_j}$

- The **Boltzmann distribution** (1868) [1], also called the **Gibbs distribution** (1902) [2], can show the **probability that a physical system can have a specific state**. i.e., every of the particles has a specific state. The probability mass function of this distribution is [3]:

$$\boxed{\mathbb{P}(x) = \frac{e^{-\beta E(x)}}{Z},} \tag{1}$$

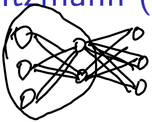
where $E(x)$ is the energy of variable x and Z is the normalization constant so that the probabilities sum to one.

- This normalization constant is called the **partition function** which is hard to compute as it sums over all possible configurations of states (values) that the particles can have. If we define $\mathbb{R}^d \ni \boldsymbol{x} := [x_1, \dots, x_d]^\top$, we have:

$$\boxed{Z := \underbrace{\sum_{\boldsymbol{x} \in \mathbb{R}^d}}\, e^{-\beta E(x)}.} \tag{2}$$

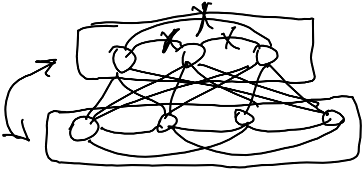
$x_1 \quad x_2 \quad x_3 \rightarrow 2 \times 2 \times 2 = 8$

Boltzmann (Gibbs) Distribution



BM

RBM



- We had:

$$\boxed{\mathbb{P}(x) = \frac{e^{-\beta E(x)}}{Z}.}$$

- The coefficient $\beta \geq 0$ is defined as:

$$\boxed{\beta := \frac{1}{k_\beta T} \propto \frac{1}{T},} \tag{3}$$

where k_β is the **Boltzmann constant** and $T \geq 0$ is the **absolute thermodynamic temperature in Kelvins**.

- If the temperature tends to absolute zero, $\boxed{T \to 0,}$ we have $\beta \to \infty$ and $\mathbb{P}(x) \to 0$, meaning that the **absolute zero temperature occurs extremely rarely in the universe**.

Boltzmann (Gibbs) Distribution

- Recall Eqs. (10) and (2):

$$\begin{cases} \mathbb{P}(x) = \dfrac{e^{-\beta E(x)}}{Z}, \\ Z := \displaystyle\sum_{\boldsymbol{x} \in \mathbb{R}^d} e^{-\beta E(x)}. \end{cases}$$

- The **free energy** is defined as:

$$\boxed{F(\beta) := \frac{-1}{\beta} \ln(Z),} \tag{4}$$

where ln(.) is the natural logarithm.

- The **internal energy** is defined as:

$$\boxed{U(\beta) := \frac{\partial}{\partial \beta}\big(\beta\, F(\beta)\big).} \tag{5}$$

- Therefore, we have:

$$U(\beta) = \frac{\partial}{\partial \beta}(-\ln(Z)) = \frac{-1}{Z}\frac{\partial Z}{\partial \beta} \overset{(2)}{=} \sum_{\boldsymbol{x} \in \mathbb{R}^d} E(x)\, \frac{e^{-\beta E(x)}}{Z} \overset{(10)}{=} \sum_{\boldsymbol{x} \in \mathbb{R}^d} \mathbb{P}(x) E(x). \tag{6}$$

Boltzmann (Gibbs) Distribution

- Recall Eqs. (10) and (4) and (6):

$$\left\{\begin{aligned} \mathbb{P}(x) &= \frac{e^{-\beta E(x)}}{Z}, \\ F(\beta) &:= \frac{-1}{\beta}\ln(Z), \\ U(\beta) &= \sum_{\boldsymbol{x}\in\mathbb{R}^d} \mathbb{P}(x)E(x). \end{aligned}\right.$$

- The **entropy** is defined as:

$$\begin{aligned} H(\beta) &:= -\sum_{\boldsymbol{x}\in\mathbb{R}^d} \mathbb{P}(x)\ln\big(\mathbb{P}(x)\big) \overset{(10)}{=} -\sum_{\boldsymbol{x}\in\mathbb{R}^d} \mathbb{P}(x)\big(-\beta E(x) - \ln(Z)\big) \\ &= \beta \sum_{\boldsymbol{x}\in\mathbb{R}^d} \mathbb{P}(x)E(x) + \ln(Z) \underbrace{\sum_{\boldsymbol{x}\in\mathbb{R}^d} \mathbb{P}(x)}_{=1} \overset{(a)}{=} -\beta\, F(\beta) + \beta\, U(\beta), \end{aligned} \qquad (7)$$

where (a) is because of Eqs. (6) and (4).

Boltzmann (Gibbs) Distribution

Lemma

A physical system prefers to be in low energy; hence, the system always loses energy to have less energy.

Proof.

On the one hand, according to the second law of thermodynamics, entropy of a physical system always increases by passing time [6]. Entropy is a measure of randomness and disorder in system. On the other hand, when a system loses energy to its surrounding, it becomes less ordered. Hence, by passing time, the energy of system decreases to have more entropy. Q.E.D. □

Corollary

According to Eq. (10):

$E(x) \downarrow$

$$\mathbb{P}(x) = \frac{e^{-\beta E(x)}}{Z}, \uparrow$$

and Lemma 1, the probability $\mathbb{P}(x)$ of states in a system tend to increase by passing time.

Boltzmann (Gibbs) Distribution

- This corollary makes sense because **systems tend to become more probable**. This idea is also used in **simulated annealing** [7] where the **temperature of system is cooled down gradually**.
- Simulated annealing is a metaheuristic optimization algorithm in which a **temperature parameter** controls the amount of **global search** versus **local search**. It **reduces the temperature gradually** to decrease the exploration and increase the exploitation of the search space, gradually.
- Recall Eqs. (10) and (3):

$$\begin{cases} \mathbb{P}(x) = \dfrac{e^{-\beta E(x)}}{Z}, & (8) \\ \beta := \dfrac{1}{k_\beta T} \propto \dfrac{1}{T}. & (9) \end{cases}$$

- Therefore, the probability mass function of the Boltzmann distribution or Gibbs distribution can be written as:

$$\mathbb{P}(x) = \frac{e^{-\frac{E(x)}{T}}}{Z}, \tag{10}$$

where $E(x)$ is the energy of variable x, and T is the Kelvin temperature, and Z is the normalization constant so that the probabilities sum to one. We can write it as:

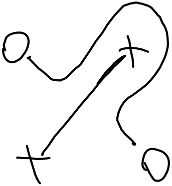
$$\boxed{\mathbb{P}(\Delta E) = \frac{e^{-\frac{\Delta E}{T}}}{Z},} \tag{11}$$

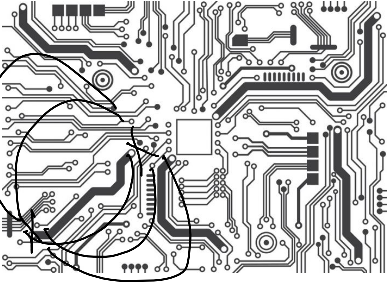
where ΔE is the difference of energy.

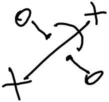
Simulated Annealing

Simulated Annealing: Idea

- **Simulated annealing** was proposed in 1983 [7] and is inspired by the **annealing schedule** in high-energy physics for forming the shape of materials.
- It is used in various applications such as **VLSI (Very Large Scale Integration) and circuit routing**.







HSIC

Simulated Annealing: algorithm

$(\Delta E/T \uparrow \Rightarrow \Delta E/T \uparrow \Rightarrow \frac{\Delta E}{T} \downarrow \Rightarrow T \uparrow)$

- step 1: choose some random initial candidates and an initial temperature
- step 2: in every iteration, do a local search in a neighborhood of candidates and choose a neighbor point for every candidate.
  - for every candidate, if the fitness of the neighbor solution is better than the candidate: accept it and replace the candidate with that.
  - otherwise, accept it with some Boltzmann probability:

$\Delta E = \text{cost(neighbor)} - \text{cost(point)}$

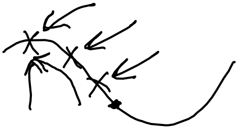
$$\mathbb{P}(\Delta E) = \begin{cases} 1 & \text{if } \boxed{\Delta E \leq 0} \\ e^{-\frac{\Delta E}{T}} & \text{if } \Delta E > 0, \end{cases} \tag{12}$$

0.7

where ΔE is the change of cost (cost of neighbor minus cost of candidate) (or fitness of candidate minus fitness of neighbor).

- This gives a chance to even worse candidates for **exploration** (not to get stuck in local optimum).
- It starts with high temperature and cools down the temperature gradually in the iterations:
  - **linear reduction rule**: $T = T - \alpha$
  - **geometric reduction rule**: $T = T \times \alpha$, where $\alpha \in (0, 1)$
  - **slow-decrease rule**: $\boxed{T = \frac{T}{1+\beta T}}$, where β is a hyper-parameter

Simulated Annealing: algorithm



Algorithm Simulated annealing

Initialize the solution $\boldsymbol{x}$
while *not converged* **do**
  $\boldsymbol{x} \leftarrow$ get a point from the neighborhood $\mathcal{N}(\boldsymbol{x})$
  Evaluate fitness function and calculate ΔE
  if $\Delta E \leq 0$ **then**
    Update the solution
    ~~Break the loop~~
  else
    $u \leftarrow U(0,1)$
    if $u \leq e^{-\frac{\Delta E}{T}}$ **then**
      Update the solution
      ~~Break the loop~~
  $T \leftarrow$ Decrement the temperature T
Return the solution $\boldsymbol{x}$

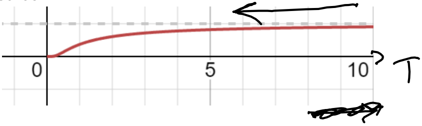
$e^{-\frac{\Delta E}{T}}$

$T\downarrow$ $\Delta E\uparrow$

Simulated Annealing: Analysis of temperature

$$\mathbb{P}(\Delta E) = \begin{cases} 1 & \text{if } \Delta E \leq 0 \\ e^{-\frac{\Delta E}{T}} & \text{if } \Delta E > 0 \end{cases}$$

The $e^{-\frac{1}{T}}$ graph with respect to T:



Analysis of temperature:

- In initial iterations, the temperature T is high so $e^{-\frac{\Delta E}{T}}$ is large (closer to one) so we give more chance to worse candidates so we have **more exploration**.
- In the end iterations, the temperature T is low so $e^{-\frac{\Delta E}{T}}$ is small (closer to zero) so we give less chance to worse candidates so we have **more exploitation**.
- It is like starting with large learning rate in gradient descent initially and then decrease the learning rate gradually.

$$P(\text{accepting } \underline{\text{worse point}}) = e^{-\frac{\Delta E}{T}}$$

initial iterations $T\uparrow$ $\Longrightarrow$ later $T\downarrow$

start & algorithm
- T large, ΔE large $\to$ neighbor very worse $\Rightarrow$ $P\downarrow\downarrow\downarrow$ $\leftarrow$
- T large, ΔE small $\to$ neighbor's cost close to mine $\Rightarrow$ $P\uparrow$ $\leftarrow$

end of algorithm
- T small, ΔE large $\to$ neigh very worse $\Rightarrow$ $P\downarrow\downarrow\downarrow\downarrow\downarrow\downarrow\downarrow\downarrow$ $\leftarrow$
- T small, ΔE small $\to$ neighbor's cost close to mine $\Rightarrow$ $P\downarrow$ $\leftarrow$

Simulated Annealing: Threshold accepting

- In some applications, it is time-consuming and resource-consuming to calculate the Boltzmann probability. In these cases, we can relax the Eq. (12) using a threshold q:

$$\boxed{\mathbb{P}(\Delta E) = \begin{cases} 1 & \text{if } \Delta E \leq q \\ 0 & \text{if } \Delta E > q, \end{cases}} \tag{13}$$

  where this threshold $q \geq 0$ is a decreasing function with respect to iteration index.
- This technique is called **threshold accepting** in simulated annealing.

Acknowledgment

- Some slides of this slide deck are inspired by teachings of Prof. Saeed Sharifian at the Amirkabir University of Technology, Department of Electrical Engineering.
- Some slides of this slide deck (the Boltzmann distribution and statistical physics) are inspired by teachings of Prof. Mehdi Molkaraie at University of Waterloo, Department of Statistics.
- This slide deck is based on our tutorial paper "Restricted boltzmann machine and deep belief network: Tutorial and survey" [8].

References

[1] L. Boltzmann, “Studien uber das gleichgewicht der lebenden kraft,” *Wissenschafiliche Abhandlungen*, vol. 1, pp. 49–96, 1868.

[2] J. W. Gibbs, *Elementary principles in statistical mechanics*.
Courier Corporation, 1902.

[3] K. Huang, *Statistical Mechanics*.
John Wiley & Sons, 1987.

[4] W. Lenz, “Beitršge zum verstšndnis der magnetischen eigenschaften in festen kšrpern,” *Physikalische Z*, vol. 21, pp. 613–615, 1920.

[5] E. Ising, “Beitrag zur theorie des ferromagnetismus,” *Zeitschrift für Physik*, vol. 31, no. 1, pp. 253–258, 1925.

[6] S. Carroll, *From eternity to here: the quest for the ultimate theory of time*.
Penguin, 2010.

[7] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi, “Optimization by simulated annealing,” *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[8] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, “Restricted boltzmann machine and deep belief network: Tutorial and survey,” *arXiv preprint arXiv:2107.12521*, 2021.