Lam Schedule

Haochun Wang

06/30/2023

Outline

1. Minimize the Final Average Energy with respect to the Annealing Schedule (Assuming the Move Generation is Fixed)

2. Minimize the Final Average Energy with respect to the Move Generation

Outline

1. Minimize the Final Average Energy with respect to the Annealing Schedule (Assuming the Move Generation is Fixed)

2. Minimize the Final Average Energy with respect to the Move Generation

Basics

The energy of a system $X(s_n)$ is a stochastic process whose values depends both on step n and the inverse temperature at step n, denoted by s_n. Now consider a special case that s_n is a constant. As $n \to \infty$, the process $X(s_n)$ becomes stationary. We denote this kind of of stationary process by $\underline{X}(s_n)$ and we also denote its mean and variance by $\mu(s_n)$ and $\sigma^2(s_n)$.

About the Stationary Process

- **Why we should care about the stationary process:** it provides information on how far away we are from equilibrium at any given inverse temperature.
- **How we make use of this:** quasi-equilibrium and quasi-stationarity.
- **Quasi-equilibrium at a fixed inverse temperature**: $|\overline{X}(s) - \mu(s)| \leq \epsilon$. Here $\overline{X}(s)$ is the average energy. The potential problem of it is that it is not invariant upon scaling of energy.
- A process $X(s_{n-1})$ is **quasi-stationary** at inverse temperature s_n if $|\overline{X}(s_{n-1}) - \mu(s_n)| \leq \lambda\sigma(s_n)$.

About Quasi-stationarity

- **Quasi-stationary at an inverse temperature:** a process $X(s_{n-1})$ is **quasi-stationary at inverse temperature s_n** if $|\overline{X}(s_{n-1}) - \mu(s_n)| \leq \lambda\sigma(s_n)$.
- **By Lam:** $\overline{X}(s_{n-1})$ is the average energy of the system at step (n-1) (after a proposed move is either accepted or rejected). However, this is not quite clear. **My current definition** for this term is like this: consider a fixed (inverse) temperature schedule, then $\overline{X}(s_{n-1}) = lim_{m\to\infty} \frac{\sum_{i=1}^{m} X_i(s_{n-1})}{m}$, where X_i are chains that follows this specific temperature schedule.
- **A possible discussion: why Lam uses (n-1)-th step instead of n-th step?**

About λ-schedule and efficient λ-schedule

- **Quasi-stationary of a process:** a process is quasi-stationary if at all of the inverse temperatures, the quasi-stationary criterion is satisfied.
- λ-**schedule**: a schedule that gives rise to such a process.
- A λ-schedule is **n-step efficient** if among all the λ-schedules, it minimizes the final energy at step n.
- If a λ-schedule is **n-step efficient** for all $n \geq 1$, we call it efficient λ-schedule. Generally, the smaller λ is, the better the final average energy will become, but also the longer the computation time required.
- **To do next: derive the efficient λ-schedule.**

Derive the efficient λ-schedule

- Before deriving the efficient λ-schedule, we need to know how average energy of a system evolves. Here we consider the first-order autoregressive model.
- $X(s_n) = r(s_n)(X(s_{n-1} - \mu(s_n)) + \mu(s_n) + N_n$, where N_n is a white noise. Notice this formula is not given on the stationary process $\underline{X}(s_n)$. Instead, it is given on the quasi-stationary process $X(s_{n-1})$. The reason we can do that is when λ is sufficiently small, the properties between $X(s_{n-1})$ and $\underline{X}(s_n)$ are similar.

Derive the efficient λ-schedule

- For $X(s_n) = r(s_n)(X(s_{n-1} - \mu(s_n)) + \mu(s_n) + N_n$, we take expectation on both sides, we have the evolution of average energy characterized by: $\overline{X}(s_n) = r(s_n)(\overline{X}(s_{n-1}) - \mu(s_n)) + \mu(s_n)$.
- Also the quasi-stationary criterion can be further simplified.

Outline

1. Minimize the Final Average Energy with respect to the Annealing Schedule (Assuming the Move Generation is Fixed)

2. Minimize the Final Average Energy with respect to the Move Generation

Basics

- As we discussed last time, we have $\rho_2 = \mathbf{E}([\underline{X}(s_n) - \underline{X}(s_{n-1})]^2)$ and $s_{n+1} = s_n + \lambda \frac{\rho_2(s_n)}{2\sigma^3(s_n)}$. Also from previous derivations, we see that under certain constraints, decrement in average energy is a non-decreasing function of s_n and $\overline{X}(s_{n-1}) - \overline{X}(s_n) = \lambda \frac{\rho_2(s_n)}{2\sigma(s_n)}$.
- Maximizing the decrement in average energy is equivalent to maximizing ρ_2.
- For the section, I skip some tedious derivations since the intuition is more important. Some derivations can be found in the handout.

Basics

- In this chapter, one of the most important term is "energy density function" denoted by P(x). It is defined in this way: dividing the energy values into non-overlapping intervals of length Δx and group states according to their corresponding energy values. Then $P(x) = \frac{number\ of\ states\ with\ energy\ in\ the\ interval\ containing\ x}{\Delta x \cdot total\ number\ of\ states}$. Then the stationary probability density function at inverse temperature s is given by $P_s(x) = \frac{P(x) \cdot e^{-sx}}{Z(s)}$ where $Z(s) = \int_{-\infty}^{\infty} P(x) e^{-sx}\, dx$.
- For this chapter there are several new notations to be introduced. Ont thing to notice is that the subscripts are very important since different subscripts produce very different meanings.
- Two elements of the model: 1. Move Generation Model 2. Energy Density Model

Structure of the Move Generation Model

- A simple model to be considered is the conditional probability density model (model proposed energy given the current energy).
- One important thing to notice here: here we model proposed energy conditionally. This proposed energy may or may not be accepted. As a clarification, all the notations with a subscript p stand for "proposed".

Separable Model

- Intuitively, the more states with a given energy, the more likely this state is proposed. Also suppose two proposed energy values are within about the same distance from the current energy (disregarding the sign of the distance), they should be proposed with similar frequencies (by Lam).
- Apart from those two, we also need to think about the impact brought by the energy density function.

Separable Model

- A rough model is given like this: $f_p(x_p|x) \propto G(|x_p - x|)Q(x_p)$.
- $f_p(x_p|x)$ is the conditional probability density function of x_p given current energy x.
- $G(|x_p - x|)$ models the effect of the distance between x_p and x on $f_p(x_p|x)$. Notice this does not contain any information of energy density function. It only cares about the distance between x_p and x.
- $Q(x_p)$ models the effect of the energy density function $P(x_p)$ on $f_p(x_p|x)$. It is not $P(x_p)$ but has close relationships with $P(x_p)$.
- Notice that $f_p(x_p|x)$ is a pdf. Thus we need a normalizaing constant K(x) where $K(x) = \int_{-\infty}^{\infty} G(|y - x|)Q(y)\,dy = G(|x|) * Q(x)$, where * is the convolution operator.

Separable Model

- After defining $f_p(x_p|x)$ (conditional pdf of proposed energy), we can express $f(x_p|x)$ (conditional pdf of the energy at the next time step).
- $$f(x_p|x) = \begin{cases} f_p(x_p|x)e^{-s(x_p-x)} & x_p > x \\ f_p(x_p|x) + \delta(x_p - x)\int_x^\infty f_p(y|x)(1 - e^{-s(y-x)})dy & x_p \leq x \end{cases}$$
- $f(x_p|x) = \frac{1}{K(x)}G(|x_p - x|)Q(x_p) + \delta(x_p - x)\int_x^\infty [G(|y - x|) - A_s(y - x)]Q(y)dy$
- where
$$A_s(x_p - x) = \begin{cases} G(|x_p - x|)e^{-s|x_p-x|} & x_p > x \\ G(|x_p - x|) & x_p \leq x \end{cases}$$

Moments of Energy Increment

- Define $\xi(s)$ to be the energy increment. Then we have the n-th moment of the energy increment $\mathbf{E}(\xi^n(s)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-x)^n f(y,x)dydx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-x)^n f(y|x)P_s(x)dydx$ by property of conditional density.
- After substituting z = y-x, we have $\rho_n(s) = \mathbf{E}(\xi^n(s)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} z^n \frac{1}{K(y-z)} A_s(z)Q(y)\frac{P(y-z)e^{-s(y-z)}}{Z(s)}dydz$
- Recall that K and Z are normalizing terms, where K is the normalizing term for $f_p(x_p|x)$ and x_p and Z is the normalizing term for $P_s(x)$ (stationary pdf at temperature s).

HOLD ON!

- So what is our goal?
- Our goal is to maximize ρ_2! Now we have some solid knowledge on that.
- Then what should we do next?
- In the expression on previous pages, we need to fill in some gaps. That is, we need to specify choices for G, P and Q.
- After we specify G, P and Q, what should we do?
- Ideally we just maximize ρ_2 by setting the derivative to 0 and check the second-order derivative.

Exponential Locality Model and Energy Density Model

- $G(|z|) = e^{-\beta|z|}$, $\beta > 0$ and $z = |x_p - x|$. We consider this for two reasons: 1. it is easy enough and include an exponential term; 2. intuition is that the closer the proposed energy is to the current energy, the higher the probability is proposed.
- By some derivations, Lam showed that P and Q are closely related and we can actually express (approximate) Q using P. Therefore the rest of our job is to specify only one of them.
- Lam chose Gamma distribution for Q because of simplicity.

HOLD ON (Again)

- Recall where we begin: $f_p(x_p|x) \propto G(|x_p - x|)Q(x_p)$. Such a Gamma-exponential conjugate pair (usually seen in Bayesian context) indeed makes it possible to analyze.
- My comment on the rest of the derivations: it seems a little hard for me to trust. However we can possibly start from here to continue the derivations (maybe not now but after I finished my current things in the plate). Also we may try normal-normal pair since it seems to be even easier to analyze. Remember the final goal is to (possibly) make it work in the parallel case.

Recap: the Whole Structure

- In the first step we define quasi-stationarity. Then we discuss about the autoregressive modeling on the process and define the most important term $\rho_2 = \mathbf{E}([\underline{X}(s_n) - \underline{X}(s_{n-1})]^2)$.
- Our goal is to maximize ρ_2. However this could not be analyzed directly. Also to analyze this, Lam proposes separable model and energy density model.
- Today's talk will be focused on how (intuitively) we can relate ρ_0 (acceptance ratio) to ρ_2. Specifically, use ρ_0 to express (approximately) ρ_2.
- Notation recap: G models the effect of the distance between x_p and x on $f_p(x_p|x)$, $Q(x_p)$ models the effect of the energy density function $P(x_p)$ on $f_p(x_p|x)$. $f_p(x_p|x) \propto G(|x_p - x|)Q(x_p)$.

Details of the Exponential Locality Model

- As we have discussed before, to further analyze on how to maximize ρ_2, we need to specify what G, Q and P are. Here we specify what G is.
- $G(|z|) = e^{-\beta|z|}$, $\beta > 0$ and $z = |x_p - x|$.
- From Lam (along with some details in my provided notes),

$$\rho_n(s) = \begin{cases} \frac{2\int_0^\infty z^n G(|z|)H_s(-z)dz}{\int_{-\infty}^\infty G(|z|)H_s(-z)dz} & n \text{ is even} \\ 0 & n \text{ is odd} \end{cases}$$

- where $H_s(z) = \int_{-\infty}^{\infty} Q(y)Q_s(y-z)dy = Q(z) * Q_s(-z)$ and $Q_s(x) = \frac{Q(x)e^{-sx}}{Z_Q(s)}$.
- Now plug in the G mentioned above.

Details of the Exponential Locality Model

- First decompose $G(|z|) = e^{-\beta|z|}$ into two parts. We have
- $$g_+(z) = \begin{cases} e^{-\beta z} & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases}$$
- $$g_-(z) = \begin{cases} e^{\beta z} & \text{if } z < 0 \\ 0 & \text{otherwise} \end{cases}$$
- Then consider the Fourier transform pair
- $F(g(x)) = \phi(\omega) = \int_{-\infty}^{\infty} g(x)e^{-j\omega x}dx$ and $F^{-1}(\phi(\omega)) = g(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi(w)e^{jwx}dx$. Here j is simply i in complex domain.

Rewrite ρ_n

- The denominator:

$$\int_{-\infty}^{\infty} G(|z|)H_s(-z)dz = \int_{-\infty}^{0} e^{\beta z}H_s(-z)dz + \int_{0}^{\infty} e^{-\beta z}H_s(-z)dz$$

- Note it equals to

$$\int_{-\infty}^{\infty} G(|z|)H_s(-z)dz = \int_{-\infty}^{+\infty} g_-(z)H_s(-z)dz + \int_{-\infty}^{+\infty} g_+(z)H_s(-z)dz$$

- Let $L_-^{(0)}(\beta, s)$ be the first component and $L_+^{(0)}(\beta, s)$ be the second component.

Rewrite ρ_n

- Similarly, we define $L_+^{(n)}(\beta, s) = \int_0^\infty z^n e^{-\beta z} H_s(-z) dz$
- Till now we can rewrite ρ_n in terms of $L_+^{(n)}$, $L_-^{(0)}$ and $L_+^{(0)}$.
- $$\rho_n(\beta, s) = \begin{cases} \frac{2L_+^{(n)}(\beta,s)}{L_-^{(0)}(\beta,s)+L_+^{(0)}(\beta,s)} & n \text{ is even} \\ 0 & \text{otherwise} \end{cases}$$
- Plancherel's theorem: if f and g are squared integrable, then $\langle f, g \rangle = \langle \hat{f}, \hat{g} \rangle$.
- Therefore we may write $L_-^{(0)}(\beta, s) = \int_{-\infty}^{+\infty} \frac{F(Q(\beta,x))F^*(Q_s(\beta,x))}{\beta+j\omega} d\omega$ where $\beta + j\omega$ comes from $F(g_-(z)) = \frac{1}{\beta+j\omega}$.
- We can also express $L_+^{(0)}$ and $L_+^{(n)}$ in similar way. We write in this way for further analysis.

A Brief Summary

- Till now we have an expression for ρ_2. The only constraint for that is β. Which means, the necessary condition for maximizing ρ_2 is $\frac{\partial \rho_2(\beta, s)}{\partial \beta} = 0$.
- However, this is still almost impossible to analyze directly. To further analyze this, we express this in terms of ρ_0, a.k.a. acceptance ratio.
- For the rest of the time, I plan not to provide the details of the derivations on how to relate those two terms. It is a very dense derivation with many approximations. Instead, I would like hold a brief discussion on how to understand ρ_0.

Discussion: ρ_0

- Recall:
- Define $\xi(s)$ to be the energy increment. **As n $\geq$ 1** we have the n-th moment of the energy increment $\mathbf{E}(\xi^n(s)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-x)^n f(y,x)dydx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-x)^n f(y|x)P_s(x)dydx$ by property of conditional density. Note $\mathbf{E}(\xi^0(s)) = 1$.
- After substituting z = y-x, we have $\rho_n(s) = \mathbf{E}(\xi^n(s)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} z^n \frac{1}{K(y-z)} A_s(z)Q(y)\frac{P(y-z)e^{-s(y-z)}}{Z(s)}dydz$ as $n > 0$.
- where

$$A_s(x_p - x) = \begin{cases} G(|x_p - x|)e^{-s|x_p - x|} & x_p > x \\ G(|x_p - x|) & x_p \leq x \end{cases}$$

- Recall that K and Z are normalizing terms, where K is the normalizing term for $f_p(x_p|x)$ and x_p and Z is the normalizing term for $P_s(x)$ (stationary pdf at temperature s).
- **But what about n = 0? Why it refers to the acceptance ratio (intuitively)?**