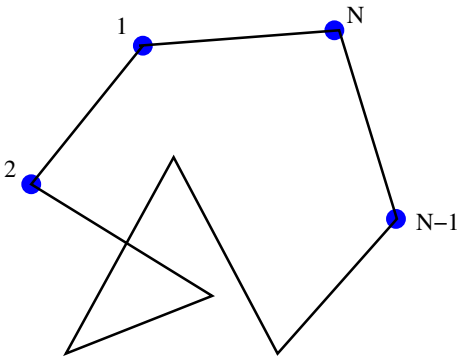
Lecture 9 Simulated Annealing and QMC *

Tiejun Li

1 Simulated Annealing

We already have very efficient algorithms for traditional convex programming. But how about the non-convex programming problems, such as the following combinatorial optimization problem?

Example 1. *(Traveling Salesman Problem) Suppose there are N cities and there exists one path ($l_{ij} = l_{ji}$) for each two. Try to find a minimal path passing all the cities such that each city is passed and only passed one time.*



$$\min_{x\in X} H(x) = \sum_{i=1}^{N} l_{x_i x_{i+1}}, \quad x_{N+1} := x_1.$$

$X = \{(x_1, \ldots, x_N),\ x_1, \ldots, x_N$ is a permutation of $1, 2, \ldots, N\}$

Figure 1: Traveling Salesman Problem

The number of all the possible paths is $\frac{N!}{2}$. It is a typical combinatorial explosion problem(NP problem). This number increases exponentially with N, and there isn't any rules for the function $H(x)$. The traditional algorithms are inapplicable here.

Example 2. *(Image smoothing problem) Suppose there are J pixels for an image, and there are 256 colors for each pixel. Any image can be represented as one element in*

$$X = \{(x_1, \ldots, x_J) \ : \ x_i \in \{0, 1, \ldots, 255\}\}.$$

The smoothness of an image is defined as

$$H(x) = \alpha \sum_{<s,t>} (x_s - x_t)^2, \qquad \alpha > 0,$$

*School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

where $< s, t >$ means the neighboring pixels in the lattice among $x = (x_1, \ldots, x_J)$. Then define the comparison function for images x and y where y is the reference image

$$H(x|y) = \alpha \sum_{<s,t>} (x_s - x_t)^2 + \frac{1}{2\sigma^2} \sum_s (x_s - y_s)^2.$$

An image recovering problem for polluted y may be proposed as minimizing the following function:

$$\min_{x \in X} H(x|y).$$

The number of all the possible states is 256^J! Traditional algorithms are still inapplicable here!

Simulated annealing algorithm is one of the framework to handle this kind of non-convex global optimization problem from stochastics recent years [3, 5]. While the effectivity is still under discussion.

1.1 Basic framework

For optimization problem

$$\min_{x \in X} H(x),$$

Define the global minimizers of $H(x)$

$$M = \{x_0 \ : \ H(x_0) = \min_{x \in X} H(x)\},$$

and introduce the parameter $\beta > 0$, define

$$\Pi^\beta(x) = \frac{1}{Z_\beta} e^{-\beta H(x)}, \qquad Z_\beta = \sum_{x \in X} \exp(-\beta H(x)),$$

then $\Pi^\beta(x)$ is a probability distribution on X.

Theorem 1. *$\Pi^\beta(x)$ has the property*

$$\lim_{\beta \to +\infty} \Pi^\beta(x) = \begin{cases} \frac{1}{|M|} & \text{if } x \in M, \\ 0 & \text{else.} \end{cases}$$

and if β is sufficiently large, then $\Pi^\beta(x)$ is monotonely increasing as a function of β for any $x \in M$, and $\Pi^\beta(x)$ is monotonely decreasing as a function of β for any $x \notin M$.

Proof. Rewrite

$$\begin{aligned} \Pi^\beta(x) &= \frac{e^{-\beta(H(x)-m)}}{\sum_{z:H(z)=m} e^{-\beta(H(z)-m)} + \sum_{z:H(z)>m} e^{-\beta(H(z)-m)}} \\ &\overset{\beta \to +\infty}{\longrightarrow} \begin{cases} \frac{1}{|M|}, & x \in M, \\ 0, & x \notin M, \end{cases} \end{aligned}$$

where $m = \min_x H(x)$.

If $x \in M$, we have

$$\Pi^\beta(x) = \frac{1}{|M| + \sum_{z:H(z)>m} e^{-\beta(H(z)-m)}},$$

then $\Pi^\beta(x)$ monotonely increases with β increasing.

If $x \notin M$, we have

$$\begin{aligned}\frac{\partial \Pi^\beta(x)}{\partial \beta} &= \frac{1}{\tilde{Z}_\beta^2}\Big(e^{-\beta(H(x)-m)}(m-H(x))\tilde{Z}_\beta - e^{-\beta(H(x)-m)}\sum_{z\in X} e^{-\beta(H(z)-m)}(m-H(z))\Big)\\ &= \frac{1}{\tilde{Z}_\beta^2}\Big(e^{-\beta(H(x)-m)}\big[(m-H(x))\tilde{Z}_\beta - \sum_{z\in X} e^{-\beta(H(z)-m)}(m-H(z))\big]\Big),\end{aligned}$$

where

$$\tilde{Z}_\beta \triangleq \sum_{z\in X}\exp(-\beta(H(z)-m)).$$

Pay attention that

$$\lim_{\beta\to+\infty}\big[(m-H(x))\tilde{Z}_\beta - \sum_{z\in X} e^{-\beta(H(z)-m)}(m-H(z))\big] = |M|(m-H(x)) < 0,$$

The proof is completed.

Remark 1. *The construction of $\Pi^\beta(x)$ opens a new way to optimize $H(x)$ via stochastics. Theorem 1 shows that if we can generate the random sequence with distribution $\Pi^\beta(x)$, then the random numbers will finally jump among the minimizers when $\beta = +\infty$. This procedure is called* **annealing**. *It corresponds to the physical crystallization. In physics, β corresponds to $1/T$, where T is temperature. Global energy minimization means a perfect crystal without defects. The observed crystals with defects in nature can be understood as the local minimum state. In order to obtain a perfect crystal, one may image the following process: The crystals will take the form of liquids in the high temperature, then one decreases the temperature very slowly until the perfect crystal forms at the zero temperature. This is the basic idea of simulated annealing.*

The random number generation with distribution $\Pi^\beta(x)$ can be created by Metropolis algorithm.

1.2 Theoretical results

Assuming the Metropolis sampler for simulated annealing is

$$P^\beta(\sigma,\sigma') = \begin{cases} G(x,y)\frac{\pi^\beta(y)}{\pi^\beta(x)}, & \pi^\beta(y) < \pi^\beta(x) \text{ and } x \neq y,\\ G(x,y), & \pi^\beta(y) \geq \pi^\beta(x) \text{ and } x \neq y,\\ 1-\sum_{z\neq x} P^\beta(x,z) & x = y.\end{cases}$$

where $G(x,y)$ is the proposal matrix. It is symmetric as before.

In order to state the fundamental theorem of simulated annealing, we define the follows.

Definition 1. *(Neighborhood system) The neighborhood system of x is defined as $N(x) = \{y \in X | x \neq y, G(x,y) > 0\}$.*

Definition 2. *Given x and y, if there exists sequence $x = u_0, u_1, \ldots, u_{\sigma(x,y)} = y$ such that $u_{j+1} \in N(u_j)$ for any $j = 0, 1, \ldots, \sigma(x,y)-1$, then we say that the states x and y communicate, where $\sigma(x,y)$ is the length of the shortest path along which x and y communicate.*

Definition 3. *The maximal local increase of energy is defined as*

$$\Delta = \max\{H(y) - H(x) : x \in X, y \in N(x)\}.$$

Theorem 2. *(Fundamental theorem of simulated annealing) Suppose that X is a finite set, $H(x)$ is a nonconstant function, $G(x, y)$ is a symmetric irreducible proposal matrix,*

$$\tau = \max\{\sigma(x, y) : x, y \in X\}.$$

If the annealing procedure is chosen such that $\beta(n) \leq \frac{1}{\tau\Delta} \ln n$, then for any initial distribution ν, we have

$$\lim_{n \to +\infty} \|\nu P^{\beta(1)} \cdots P^{\beta(n)} - \Pi^{\infty}\| = 0.$$

The proof of the theorem may be refereed to [5].

Remark 2. *Theorem 2 shows that the annealing rate must be slow enough such that $\beta(n) \leq \frac{1}{\tau\Delta} \ln n$. It is a very very slow rate because $n \geq \exp(\tau\Delta\beta(n))$, we need $n \sim \exp(N_0)$ if $\beta(n) = N_0 \gg 1$. This means high accuracy needs exponential computing time, which is impossible for realistic computation. We should take more rapid annealing rates such as $\beta(n) \sim p^{-n}$ $(p \lesssim 1)$ or others. Of course, it has no theoretical foundations. The implementation details may be refereed to [2].*

2 Quasi-Monte Carlo Method

The standard MC is of $O(\frac{\sigma}{\sqrt{N}})$. In order to improve the accuracy, one has two choices

- Take N very large — Huge computational effort;
- Variance reduction techniques.

In the follows we will introduce the QMC to replace the pseudo-random sequence with quasi-random sequence. It improves the convergence rate to $O((\ln N)^k N^{-1})$, where k depends on the space dimension. Finally we will find that QMC is essentially a deterministic method which is very similar with MC. The main contents may be refereed to [1].

2.1 Discrepancy

The concept of discrepancy is an estimate of the uniformity of the points. For N points $\{x_n\}_{n=1}^N$ belonging to the unit d-cube $I^d = [0, 1]^d$, define

$$R_N(J) = \frac{1}{N} \#\{x_n \in J\} - m(J) \tag{1}$$

for any set $J \subset I^d$, where $\#\{x_n \in J\}$ means the number of the points in set J, and $m(J)$ is the measure of J. Intuitively $R_N(J)$ is the difference between the exact volume and the random sampling estimate of the volume.

Definition 4. *Define the whole set of rectangles in I^d as*

$$E = \{J(x, y) \ : \ (0, 0, \ldots, 0) \leq x \leq y \leq (1, 1, \ldots, 1)\},$$

where $x \leq y$ means $x_i \leq y_i$,$i = 1, \ldots, d$, $J(x, y)$ means the set of rectangles with the lower left node x and the upper right node y. Define

$$E^* = \{J(0, y) \ : \ (0, 0, \ldots, 0) \leq y \leq (1, 1, \ldots, 1)\}.$$

Definition 5. *The L^∞-discrepancy of a sequence $\{x_n\}_{n=1}^N$ is defined as*

$$D_N = \sup_{J \in E} |R_N(J)|;$$

and the L^2-discrepancy

$$T_N = \Big(\int_{(x,y) \in I^{2d}, x \leq y} R_N(J(x, y))^2 dxdy \Big)^{\frac{1}{2}}.$$

The L^p-discrepancy can be defined similarly. Specially we define the discrepancy

$$D_N^* = \sup_{J \in E^*} |R_N(J)|,$$
$$T_N^* = \left(\int_{I^d} R_N(J(0, x))^2 dx \right)^{\frac{1}{2}}.$$

2.2 Total variation

In 1D case, the total variation of a function is defined as the sum of the jumps:

$$V[f] = \sup_{\tau} \sum_i |f(x_{i+1}) - f(x_i)|,$$

where τ is taken to all the possible partitions of the domain. If f is differentiable, then

$$V[f] = \int_0^1 |df| = \int_0^1 |f'(x)| dx.$$

The total variation of function f in unit d-cube $[0, 1]^d$ is defined as

$$V[f] = \int_{I^d} \Big| \frac{\partial^d f}{\partial x_1 \cdots \partial x_d} \Big| dx_1 \cdots dx_d + \sum_{i=1}^d V[f_1^{(i)}],$$

where $f_1^{(i)}$ is the restriction of f on the boundary $x_i = 1$. It is a recursive definition of total variation.

Theorem 3. *(Koksma-Hlawka) For any sequence $\{x_n\}_{n=1}^N \subset I^d$ and the function f with bounded variation in I^d, the integration error $\mathcal{E}$ obeys the following inequality*

$$\mathcal{E}[f] \leq V[f] D_N^*,$$

where $\mathcal{E}[f] \triangleq |I[f] - I_N[f]| = | \int_{I^d} f(x) dx - \frac{1}{N} \sum_{i=1}^N f(x_i)|$.

Proof. We only present the intuitive proof here.

For the function $f(x)$ which takes value 0 on the boundary of I^d, define

$$R(x) = R_N(J(0, x)),$$

then

$$dR(x) = \{\frac{1}{N}\sum_{i=1}^{N}\delta(x - x_i) - 1\}dx,$$

where $dR = \frac{\partial^d R}{\partial x_1 \cdots \partial x_d}$, $dx = dx_1 \cdots dx_d$. $\delta(x - x_i)$ is the Dirac-δ function centered at x_i, then we have

$$\begin{aligned}
\mathcal{E}[f] &= \Big|\int_{I^d} f(x)dx - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\Big| \\
&= \Big|\int_{I^d} \{1 - \frac{1}{N}\sum_{i=1}^{N}\delta(x - x_i)\}f(x)dx\Big| \\
&= \Big|\int_{I^d} R(x)df(x)\Big| \\
&\leq (\sup_x R(x))\int_{I^d} |df(x)| = D_N^* V[f].
\end{aligned}$$

Koksma-Hlawka theorem shows that the discretization error can be described by the total variation $V[f]$ and the discrepancy for the sample points. QMC gives some special quasi random sequences which have good discrepancy properties. It is a pure number theoretic result.

2.3 Quasi Monte Carlo integration

Definition 6. *A sequence $\{x_n\}_{n=1}^N \subset I^d$ is called quasi-random if*

$$D_N \leq C(\ln N)^k N^{-1},$$

in which c and k are constants that are independent of N, but may depend on the dimension d.

What follows are some typical quasi random sequences:

- Van der Corput sequence($d = 1$):

  The generation of sequence $\{x_i\}_{i=1}^N$ is composed of two steps:

  Step1. Write out n in base 2:

  $$n = \big(a_m a_{m-1} \cdots a_1 a_0\big)_2,$$

  where $(\cdot)_2$ means in base 2, $a_i \in \{0, 1\}$ is the i-th bit of n;

  Step2. Generate x_n in base 2

  $$x_n = \Big(0.\, a_0 a_1 \cdots a_m\Big)_2.$$

- Halton sequence($d > 1$):

  Denote $x_n = (x_n^1, x_n^2, \ldots, x_n^d)$, where the k-th component x_n^k is obtained by two steps.

Step1. Write out n in base p_k. (where p_k is the k-th prime number, e.g. $p_1 = 2, p_2 = 3$)

$$n = \left(a^k_{m_k} a^k_{m_k-1} \cdots a^k_1 a^k_0\right)_{p_k};$$

Step2. Generate x^k_n in base p_k:

$$x^k_n = \left(0.\, a^k_0 a^k_1 \cdots a^k_{m_k}\right)_{p_k}.$$

The number theorists has proved

$$D_N(\text{Halton}) \leq C_d (\ln N)^d N^{-1}.$$

Some other quasi random number sequences such as Sobol sequence, Faure sequence etc. may be refereed to [4].

2.4 Limitations of QMC

QMC has the following limitations:

- QMC are designed for integration and are not directly applicable to simulations. This is because of the correlations between the points of a quasi-random sequence.
- Because the theoretical basis of QMC is from Koksma-Hlawka theorem, and the generation style of quasi-random numbers is very special, it is commonly applied to the integral in rectangle with the form $\int_{I^d} f(x)dx$. For the powerful Metropolis algorithm in statistical physics, how to design the corresponding QMC version is an open problem.
- QMC is found to lose its effectiveness when the dimension of the integral becomes large. This can be anticipated from the bound $(\ln N)^d N^{-1}$ on discrepancy. For large dimension d, this bound is dominated by the $(\ln N)^d$ term unless $N > 2^d$;
- QMS is found to lose its effectiveness if the integrand f is not smooth. The factor $V[f]$ in the Koksma-Hlawka inequality is an indicator of this dependence.

All in all, QMC is suitable for the integration in which the space dimension is not so big, the integrand f is relatively smooth. Though it has better convergence rate than Monte Carlo method, its applicability is limited.

3 Homeworks

HW1. If we apply the argument in the simulated annealing to the continuous space case with smooth energy $V(x)$ and isolated minimizers, what can we say about the limit as $\beta \to \infty$?

References

[1] R.E. Caflisch, Monte Carlo and Quasi-Monte Carlo methods, Acta Numerica, Vol. 7, 1-49, 1998.

[2] Lishan Kang, Non-Numerical Parallel Algorithms: Simulated Annealing Algorithms, Science Press, Beijing, 1994.(In Chinese)

[3] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi, Optimization by Simulated Annealing, Science 220(1983), 671-680.

[4] W.H. Press et al., Numerical Recipes: the Art of Scientific Computing, Cambridge university press, Cambridge, 1986.

[5] G. Winkler, Image Analysis, Random Fields and Dynamic Monte Carlo Methods, Springer-Verlag, Berlin and New York, 1995.