

# Метод отжига

Лопатин Андрей \*

## 1 Введение

Будем рассматривать реализацию и свойства различных вариантов метода отжига. Метод отжига (синонимы: метод обжига, метод симуляции отжига, метод модельной закалки, *simulated annealing*) - это техника оптимизации, использующая упорядоченный случайный поиск на основе аналогии с процессом образования веществом кристаллической структуры с минимальной энергией при охлаждении. В настоящее время метод отжига применяется для решения многих оптимизационных задач - финансовых, компьютерной графики, комбинаторных, в телекоммуникационных сетях, и многих других. Зачастую метод отжига используют для обучения нейронных сетей. Несмотря на такую широкую область применения, скорость сходимости метода отжига всё ещё мало изучена. История метода отжига начинается с 1953 года. В этом году Н. Метрополисом был разработан алгоритм симуляции установления равновесия в системе с множеством степеней свободы при заданной температуре. В начале 80-х у С. Киркпатрика впервые появилась идея использовать этот алгоритм не только для моделирования физических систем, но и для решения некоторых задач оптимизации. Огромным преимуществом метода отжига является свойство избежать "ловушки" в локальных минимумах оптимизируемой функции, и продолжить поиск глобального минимума. Это достигается за счет принятия не только изменений параметров, приводящих к уменьшению значения функции, но и некоторых изменений, увеличивающих ее значение, в зависимости от т.н. температуры  $T$  характеристики моделируемого процесса. Чем выше температура, тем большие "ухудшающие" изменения (аналогичные случайным флуктуациям в нагретом веществе) допустимы, и больше их вероятность. Еще одним преимуществом является то, что даже в условиях нехватки вычислительных ресурсов для нахождения глобального минимума, метод отжига, как правило, выдает весьма неплохое решение (один из локальных минимумов). Л. Ингбером показано, что метод отжига и его модификации являются одним из наиболее эффективных методов случайного поиска оптимального решения для большого класса задач. Л. Ингбером проведен сравнительный анализ адаптивного метода отжига (*Adaptive Simulated Annealing - ASA*) и генетических алгоритмов, из которого следует, что на большинстве задач метод отжига не проигрывает генетическим алгоритмам, а на многих и

---

\*Электронный конспект Крысталь Борис

выигрывает. К настоящему времени разработано множество различных вариантов метода отжига, как общих, так и их специализаций для решения конкретных задач. В данной работе исследованы известные на настоящий момент общие схемы отжига, а также изучено их поведение на различных задачах оптимизации. Предложен способ решения задачи нахождения какого-либо корня функции с помощью метода отжига. Также предложен новый подход к выбору параметров метода отжига.

## 2 Алгоритм метода отжига

Метод отжига служит для поиска глобального минимума некоторой функции  $f(x)$ , заданной для из некоторого пространства  $S$ , дискретного или непрерывного. Элементы множества  $S$  представляют собой состояния воображаемой физической системы ("энергетические уровни"), а значение функции  $f$  в этих точках используется как энергия системы  $E = f(x)$ . В каждый момент предполагается заданной температура системы, как правило, уменьшающаяся с течением времени. Находясь в состоянии при температуре, следующее состояние системы выбирается в соответствии с заданным порождающим семейством вероятностных распределений  $G(x, T)$ , которое при фиксированных и задаёт случайный элемент  $G(x, T)$  со значениями в пространстве  $S$ . После генерации нового состояния  $x^l = G(x, T)$ , система с вероятностью  $h(\Delta E, T)$  переходит к следующему шагу в состояние  $x^l$ , в противном случае процесс генерации  $x^l$  повторяется. Здесь  $\Delta E$  обозначает приращение функции энергии  $f(x^l) - f(x)$ . Величина  $h(\Delta E, T)$  называется вероятностью принятия нового состояния. Как правило, в качестве функции  $h(\Delta E, T)$  выбирается либо точное значение соответствующей физической величины

$$h(\Delta E, T) = \frac{1}{1 + \exp(\Delta E/T)}$$

либо приближенное значение

$$h(\Delta E, T) = (-\Delta E/T)$$

Вторая формула используется наиболее часто. При её использовании  $h(\Delta E, T)$  оказывается больше единицы в случае  $\Delta < 0$ , и тогда соответствующая вероятность считается равной 1. Таким образом, если новое состояние даёт лучшее значение оптимизируемой функции, то переход в это состояние произойдет в любом случае. Итак, конкретная схема метода отжига задается следующими параметрами:

- Выбором закона изменения температуры  $T(k)$ , где  $k$  - номер шага.
- Выбором порождающего семейства распределений  $\zeta(x, T)$ .
- Выбором функции вероятности принятия  $h(\delta E, T)$

Алгоритм:

1. Случайным образом выбирается начальная точка  $x = x_0, x_0 \in S$ . Текущее значение энергии  $E$  устанавливается в значение  $f(x_0)$ .
2.  $k$ -я итерация основного цикла состоит из следующих шагов:
  - (a) Сравнить энергию системы  $E$  в состоянии  $x$  с найденным на текущий момент глобальным минимумом. Если  $E = f(x)$  меньше, то изменить значение глобального минимума.
  - (b) Сгенерировать новую точку  $x^l = G(x, T(k))$ .
  - (c) Вычислить значение функции в ней  $E^l = f(x^l)$ .
  - (d) Сгенерировать случайное число  $\alpha$  из интервала  $[0; 1]$
  - (e) Если  $\alpha < h(E^l - E, T(k))$ , то установить  $x \leftarrow x^l, E \leftarrow E^l$  и перейти к следующей итерации. Иначе повторить шаг (b), пока не будет найдена подходящая точка  $x^l$ .

Известны следующие модификации этого алгоритма:

- Модификация А. На шаге 2.(e) переход к следующей итерации происходит и в том случае, если точка  $x^l$  не являлась подходящей. При этом следующая итерация начинается с точки  $x$ , но уже с новым значением температуры.
- Модификация Б. В качестве оценки точки минимума возвращается последнее значение  $E$ . Это может незначительно ускорить алгоритм в случае большой размерности  $S$ , но с небольшой вероятностью может привести к тому, что будет получено худшее решение (особенно, если температура к моменту завершения алгоритма остаётся значительно больше нуля).
- Модификация В. На шаге 2.(b)  $x^l$  вычисляется рекуррентно с использованием формулы  $x^l = G(x^l, T(k))$ . Изначально на шаге 1 устанавливается  $x^l \leftarrow x_0$ . Это позволяет избежать "застревания" алгоритма, однако такая реализация теряет множество преимуществ метода отжига, т.к. не очень сильно отличается от обычного случайного поиска (особенно, если это комбинируется с вариантом 1 - в этом случае проверку  $h(\Delta E, T)$  вообще можно исключить). В следующем разделе будут рассмотрены основные схемы выбора параметров метода отжига.

## 3 Общие схемы метода отжига

### 3.1 Больцмановский отжиг

Исторически первой схемой метода отжига является т.н. *схема Больцмановского отжига*. Именно эта схема использовалась Н. Метрополисом для вычисления многомерных интегралов пути в задачах статистической физики, а

также С. Киркпатриком для решения задачи нахождения оптимальной разводки микросхем. В Больцмановском отжиге изменение температуры задаётся формулой

$$T(k) = \frac{T_0}{\ln(1+k)}, k > 0.$$

Семейство распределений  $Q(x, T)$  выбирается как семейство нормальных распределений с математическим ожиданием и дисперсией, т.е. задаётся плотностью

$$g(x, x^l, T) = (2\pi T)^{-D/2} \cdot \exp(-|x^l - x|^2/(2T)),$$

где  $D$  - размерность пространства состояний. Пространство состояний предполагается метрическим. Для Больцмановской схемы доказано, что при достаточно больших  $T_0$  и общем количестве шагов  $k$ , выбор такого семейства распределений гарантирует нахождение глобального минимума.

### 3.2 Отжиг Коши (быстрый отжиг)

Основным недостатком Больцмановского отжига является очень медленное убывание температуры. Например, для того, чтобы понизить исходную температуру в 40 раз, требуется  $e^{40} \approx 2.35 \cdot 10^{17}$  итераций, что уже вряд ли приемлемо при решении каких-либо задач. Ввиду этого Цу и Хартли предложили алгоритм, который позволяет использовать для изменения температуры схему

$$\frac{T(k)}{k} = T_0$$

без потери гарантии нахождения глобального минимума. Это достигается за счет использования в качестве  $Q$  распределений Коши с плотностью

$$g(x^l; x, T) = \frac{T}{(|x^l - x|^2 + T^2)^{(D+1)/2}}$$

соответствующим образом нормированных. Например, в случае  $D = 1$  приходим к плотности

$$g(x^l; x, T) = \frac{1}{\pi} \frac{T}{(|x^l - x|^2 + T^2)}$$

К сожалению, это распределение не очень удобно моделировать в пространстве размерности больше 1. Этого можно избежать, например, с помощью перемножения  $D$  одномерных распределений Коши:

$$g(x^l; x, T) = \frac{1}{\pi^D} \prod_{i=1}^D \frac{T}{(|x_i^l - x_i|^2 + T^2)}$$

но в этом случае нахождение глобального минимума гарантируется только при законе изменения температуры не быстрее чем:

$$T(k) = \frac{T}{k^{1/D}}$$

что гораздо медленнее первой схемы.

### 3.3 Сверхбыстрый отжиг

Недостатки двух предыдущих методов привели к тому, что в 1989 году американским исследователем Л. Ингбером был разработан метод сверхбыстрого отжига (Very Fast Annealing). В нём пространство  $S$  считается состоящим из  $D$ -мерных векторов  $(x_1, \dots, x_D)$  где  $x_i \in [A_i, B_i]$ . Кроме этого, температура по каждой из координат может различаться, таким образом,  $T$  также является вектором размерности  $D$ . Семейство распределений строится следующим образом. Вводится функция

$$g_T(y) = \prod_{i=1}^D \frac{1}{2(|y| + T_i) \ln(1 + 1/T_i)} \equiv \prod_{i=1}^D g_{(i;T)}(y_i), y_i \in [-1, 1]$$

В качестве  $y$  для получения плотности распределений  $\zeta$  используется  $\frac{\Delta x}{B_i - A_i}$ , таким образом, новое значение  $l_i$  вычисляется по формуле  $x_i^l = x_i + z_i(B_i - A_i)$ , где  $z_i$  - случайная величина с плотностью  $g_{(i;T)}$  на  $[-1, 1]$ . При этом выходящие за границы интервала значения параметра генерируются заново (пока не произойдет попадание в интервал) или приравниваются соответствующим границам. Такую случайную величину легко промоделировать:

$$z_i = \operatorname{sgn}\left(\alpha_i - \frac{1}{2}\right) T_i \left( (1 + l/T_i)^{(2\alpha_i - 1)} - 1 \right)$$

где  $\alpha_i$  - независимые случайные величины, распределенные равномерно на  $[0, 1]$ . Доказано, что закон изменения температуры

$$T_i(k) = T_{(i;0)} \exp(-c_i k^{\frac{1}{D}}), c_i > 0$$

даёт статистическую гарантию нахождения глобального минимума. Для вероятности принятия также используется отдельная шкала температуры, изменяющаяся по такому же закону. Как правило, при реализации этого метода  $c_i$ , управляется двумя параметрами:

$$c_i = m_i \exp(-n_i/D).$$

Преимущества такого метода очевидны. Во-первых, экспоненциальное убывание температуры гораздо быстрее достижимого в предыдущих методах. Во-вторых, разделение размерностей может дать большой

выигрыш, как и благодаря отдельным температурам (т.к. специфика конкретной задачи может сильно различать параметры), так и благодаря ускорению процесса, в случае, если не нужно менять все координаты одновременно. Кроме того, в отличие от отжига Коши, сверхбыстрый отжиг, как было показано, допускает очень быстрое моделирование распределения  $\zeta$  независимо от размерности  $S$ . Среди недостатков этого метода можно назвать то, что ввиду большого количества параметров иногда требуется по несколько месяцев, чтобы хорошо настроить его для решения конкретной задачи.

### 3.4 Алгоритм Ксин Яо

Алгоритм Ксин Яо был получен повторным применением идеи предыдущего алгоритма. В качестве  $g_T(y)$  выбирается

$$g_T(y) = \prod_{i=1}^D g_{(i;T)}(y_i) = \prod_{i=1}^D \frac{1}{2(|y_i| + \frac{1}{\ln(1/T_i)}) \ln(1 + \ln(1/T_i))}$$

Утверждается, что при изменении температуры по закону

$$T_i(k) = T_{(i;0)} \exp(-\exp(b_i k^{1/D}))$$

достигается статистическая гарантия нахождения глобального минимума. Однако, показано, что увеличение скорости убывания температуры вовсе не означает ускорения в решении задачи. Более того, "размазанность" распределения приводит к тому, что метод генерирует огромное число "длинных" переходов, которые отвергаются в силу низкой вероятности их принятия. Таким образом, несмотря на то, что этот процесс можно итерировать до бесконечности, получая законы изменения температуры вроде

$$T_i(k) = T_{(i;0)} \exp(-\exp(\exp(\dots \exp(b_i k^{1/D}) \dots))),$$

ценность таких "улучшений" представляется сомнительной. Более того, легко видеть, что в пределе это приводит к тривиальному методу случайного поиска, которым является метод отжига при  $T = 0$ . Это в небольшой степени применимо и к методу сверхбыстрого отжига, так что вопрос о скорости сходимости этих методов, а также о других методах, обеспечивающих не такое быстрое убывание температуры, но большую скорость сходимости, остаётся открытым. Вполне возможны задачи, на которых вторая итерация вышеописанного процесса может давать неплохие результаты.

### 3.5 Методы "тушения"

Далеко не всегда хватает вычислительных ресурсов на поиск глобального минимума. Кроме того, зачастую достаточно не глобально оптимального

решения задачи, а достаточно близкого к нему. Методы "тушения" (simulated quenching) не гарантируют нахождения глобального минимума, но, как правило, быстро находят близкое решение, а на практике зачастую и сам оптимум. Основная идея этих методов заключается в том, чтобы скомбинировать семейство распределений  $\zeta$  одного из предыдущих четырёх методов с более быстрым законом убывания температуры. Например, можно рассматривать нормальные распределения  $\zeta$  из Больцмановского отжига, но при этом уменьшать температуру по закону

$$T_{k+1} = cT_k$$

Как правило, в этом случае  $c$  выбирается между 0.7 и 0.99. Такой метод очень быстро сходится, и для конкретных задач может давать весьма неплохое решение, близкое к оптимальному, в условиях реального времени. Зачастую методы тушения основаны либо на нормальном распределении, либо на распределении для сверхбыстрого отжига. Кроме того, встречаются специальные распределения, подобранные опытным путём для решения конкретных задач.

### 3.6 Масштабирование в ходе отжига

Зачастую при реализации сверхбыстрого отжига в задачах с большой размерностью используется технология масштабирования отжига (re-annealing), иногда также применяемая и к другим вариантам отжига. При использовании этой технологии периодически во время отжига производится следующая операция. Обозначим за  $s_i$  значение некоторой оценки производной целевой функции по  $i$ -й координате в точке текущего минимума:

$$s_i \geq \left| \frac{\partial f}{\partial x_i}(x_i^{(min)}) \right|.$$

Кроме того, пусть  $s_{max} = \max_{1 \leq i \leq D} s_i$ . После этого номер шага (называемый в этом алгоритме временем, отжига (annealing-time)) и температура для каждой размерности изменяются следующим образом:

$$T_i^l = T_i(k_i)(s_{max}/s_i)$$

$$k_i = (\ln(T_{(i;0)}/T_i^l)/c_i)^D.$$

Таким образом, значение  $T_{(i;0)}$  сохраняется, а время отжига (которое теперь может принимать не только целые значения) масштабируется согласно температуре. Температура, используемая для расчета вероятности принятия, также масштабируется. Способ масштабирования зависит от задачи, но, как правило, коэффициент масштабирования определяется значением целевой функции в точке текущего минимума (чем меньше

значение функции, тем меньше температура). Время отжига для пересчета температуры принятия также изменяется аналогичным образом.

Кроме приведенного, возможны другие способы масштабирования, зависящие от конкретной задачи. Как правило, они сводятся к другому выбору коэффициентов чувствительности  $s_i$ . Приведенный выше способ выбора  $s_i$  используется для нелинейных физических задач большой размерности.

## 4 Моделирование схем отжига

Для сравнения различных схем метода отжига одной из задач дипломной работы являлось их моделирование. Вопрос моделирования различных схем отжига разбивается на две части:

- Промоделировать соответствующее распределение.
- Промоделировать схему изменения температуры.

### 4.1 Больцмановский отжиг

Нормальное распределение моделировалось с помощью центральной предельной теоремы. Обозначим  $\alpha_i$  независимые реализации равномерного распределения на  $[0; 1]$ . Тогда  $E\alpha_i = 0$ ,  $D\alpha_i = 1/12$  и

$$\frac{\sum_{i=1}^n \alpha_i}{\sqrt{n/12}} \rightarrow_{n \rightarrow \infty} N(0, 1)$$

Таким образом, можно сложить  $n$  независимых реализаций  $\alpha_i$ , и разделив их сумму на  $\sqrt{n/12}$ , мы будем иметь достаточно хорошее приближение к нормальному распределению. Как показывает практика, достаточно взять  $n = 24$ . Для моделирования требуемого  $N(0; T)$  достаточно домножить получившееся число на  $\sqrt{T}$ . В случае размерности, большей единицы, по каждой из координат вводились независимые возмущения с таким распределением. Для моделирования убывания температуры использовалась непосредственно формула  $T_0/\ln k$ ,  $k = 2, 3, \dots$

### 4.2 Отжиг Коши

Распределение Коши моделировалось методом обратных функций. Пусть  $X$  - случайная величина, имеющая распределение Коши с плотностью

$$g(x) = \frac{1}{\pi} \frac{T}{x^2 + T^2}.$$

Вычислим  $P\{X < y\}$  :

$$P\{X < y\} = \int_{-\infty}^y \frac{1}{\pi} \frac{T}{x^2 + T^2} dx = \frac{1}{\pi} \left( \arctan\left(\frac{y}{T}\right) + \frac{\pi}{2} \right).$$

Следовательно, если  $\alpha$  - реализация равномерно распределенной на  $[0; 1]$  случайной величины, то величина

$$T \operatorname{tg}\left(\pi\alpha - \frac{\pi}{2}\right)$$

имеет требуемое распределение. В случае задач с размерностью  $D > 1$  использовалось произведение  $D$  одномерных распределений Коши, т.е. вносилось возмущение по каждой из координат, вычисленное по формуле выше для независимых  $\alpha_i$ . Для температуры в этом случае используется формула

$$T(k) = \frac{T_0}{k^{1/D}}$$

причем в случае  $D = 1$  она вычислялась как  $T_0/k$ , в случае  $D = 2$  - как  $T_0/\sqrt{k}$ , для увеличения скорости работы программ.

### 4.3 Сверхбыстрый отжиг

Если при  $D = 2$  отличие, аналогично предыдущему случаю, сводилось к вычислению  $k^{\frac{1}{2}}$  как  $\sqrt{k}$ , то при  $D = 1$  возможна более существенная оптимизация, т.к.

$$\frac{T(k+1)}{T(k)} = e^{-c_i} = \text{const}$$

Значение этой константы вычислялось заранее, и при каждом уменьшении температуры производилось умножение на нее вместо вычисления экспоненты и деления, что примерно в 200 раз быстрее. Для предотвращения деления на 0 в формулах температура была ограничена снизу значением  $10^{4000}$ .

### 4.4 Методы "тушения"

Кроме вышеописанных методов, были смоделированы четыре метода тушения:

- Больцмановский отжиг с уменьшением температуры как в методе Коши.
- Больцмановский отжиг с экспоненциальным уменьшением температуры ( $T_{k+1} = cT_k$ ,  $c = 0.99$ )

- Метод Коши с экспоненциальным уменьшением температуры. ( $T_{k+1} = cT_k, c = 0.99$ )
- Сверхбыстрое тушение.

Сверхбыстрое тушение получается из метода сверхбыстрого отжига с использованием следующих формул:

$$T_i(k) = T_{(i;0)} \exp(-c_i k^{Q/D}), c > 0$$
$$c_i = m_i \exp(-n_i/D),$$

где  $Q$  - так называемый множитель тушения. При тестировании множитель тушения брался равным 2.