Курс: Байесовские методы машинного обучения, 2011

Методы Монте Карло по схеме марковских цепей (Markov Chain Monte Carlo, MCMC)

Дата: 19 ноября 2011

Методы Монте Карло в байесовском подходе

Рассмотрим вероятностное распределение $p(T)$. Методы Монте Карло (методы статистических испытаний) предполагают генерацию выборки из этого распределения:

$$T_1, \ldots, T_N \sim p(T).$$

Данная выборка может быть использована для оценки вероятностных интегралов вида

$$\mathbb{E}_T f(T) = \int f(T)p(T)dT \simeq \frac{1}{N}\sum_{n=1}^{N} f(T_n). \tag{1}$$

К интегралам такого вида сводятся многие шаги при осуществлении байесовского вывода в вероятностных моделях. Например, такими интегралами являются функция обоснованности $p(\boldsymbol{t}|X, \boldsymbol{\alpha})$ и прогнозное распределение $p(t_{new}|\boldsymbol{x}_{new}, \boldsymbol{t}, X, \boldsymbol{\alpha})$ в модели RVM, а также функционал на M-шаге EM-алгоритма $\mathbb{E}_{q(T)} \log p(X, T|\Theta)$. Здесь стоит отметить, что плотность $p(T)$ в подобных вероятностных интегралах часто известна с точностью до нормировочной константы

$$p(T) = \frac{1}{Z}\tilde{p}(T).$$

Выборка $T_1, \ldots, T_N$ может быть также использована для оценки моды распределения $p(T)$:

$$\max_T p(T) \simeq \max_n p(T_n),$$

т.к. появление точек выборки наиболее вероятно в областях больших значений плотности.

Основной вопрос, раскрываемый в дальнейшем, состоит в том, как эффективно сгенерировать выборку $T_1, \ldots, T_N$ из вероятностного распределения, заданного своей плотностью $p(T)$ или своей ненормированной плотностью $\tilde{p}(T)$.

Простейшие методы генерации

Рассмотрим одномерную случайную величину X и ее функцию распределения

$$f(x) = \mathbb{P}\{X < x\}.$$

Рассмотрим случайную величину $f(X)$. Легко показать, что она имеет равномерное распределение в интервале $[0, 1]$. Отсюда получаем простейший способ генерации

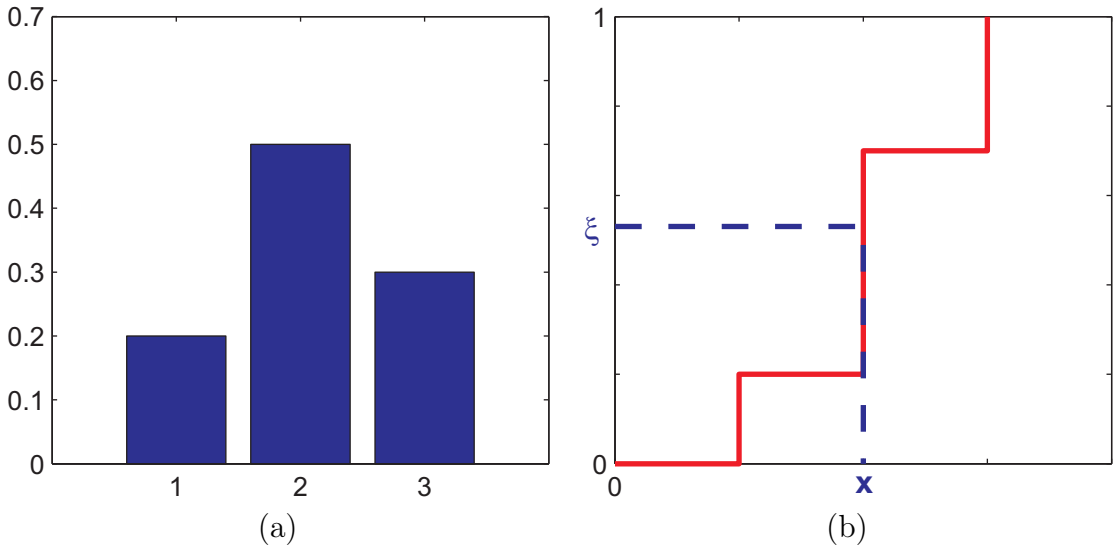

Рис. 1: Иллюстрация генерации выборки из конечного дискретного распределения. (a) – плотность распределения, (b) – соответствующая ей функция распределения.

случайной величины, заданной своей функцией распределения: сначала генерируем $\xi \sim R[0,1]$, а затем вычисляем $x = f^{-1}(\xi)$. Данный метод генерации получил название метода обратной функции. С его помощью можно сгенерировать выборку из произвольного дискретного распределения с конечным носителем (см. рис. 1).

Рассмотрим пример применения метода обратной функции для непрерывного распределения Коши:

$$p(x) = \frac{1}{\pi(1+x^2)}, \quad f(x) = \frac{1}{\pi}\arctan x + \frac{1}{2} = \xi \;\Rightarrow\; x = \tan(\pi(\xi - \frac{1}{2})).$$

Очевидно, что метод обратной функции применим только в ограниченном числе случаев, т.к. требует аналитического вычисления обратной функции к функции распределения. В частности, метод обратной функции не применим для нормального распределения.

Для генерации выборки из нормального распределения можно воспользоваться центральной предельной теоремой. Рассмотрим набор независимых равномерно-распределенных на интервале $[0,1]$ случайных величин $\xi_1, \ldots, \xi_N$ и их среднее значение $S = \frac{1}{N}\sum_{n=1}^{N}\xi_n$. Тогда величина

$$\frac{S - \mathbb{E}S}{\sqrt{\mathbb{D}S}} = \frac{\frac{1}{N}\sum_{n=1}^{N}\xi_n - \frac{1}{2}}{\sqrt{\frac{1}{N^2}\frac{N}{12}}} = \sqrt{\frac{12}{N}}\left(\sum_{n=1}^{N}\xi_n - \frac{N}{2}\right) = \{N=12\} = \sum_{n=1}^{N}\xi_n - 6$$

имеет приближенное нормальное распределение.

Одним из общих методов генерации, который может быть применен практически для любой одномерной непрерывной случайной величины, является метод Rejection sampling. Пусть необходимо сгенерировать выборку из распределения $\tilde{p}(x)$, известного с точностью до нормировочной константы. Возьмем некоторое предложное распределение $q(x)$, из которого мы можем генерировать выборку и удовлетворяющего

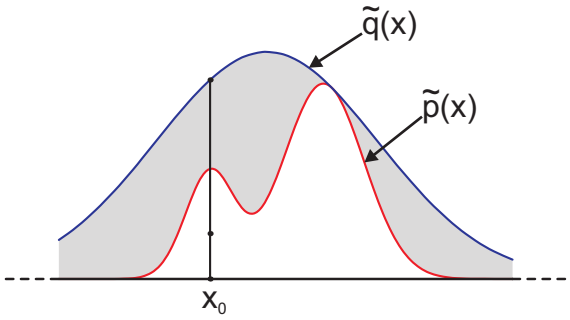

Рис. 2: Иллюстрация метода rejection sampling

свойству $\tilde{p}(x) \leq \tilde{q}(x)$ (см. рис. 2). Здесь $\tilde{q}(x)$ – ненормированная плотность $q(x)$. Сгенерируем сначала точку x_0 из распределения $q(x)$, а затем сгенерируем точку u_0 из равномерного распределения на отрезке $[0, \tilde{q}(x_0)]$. В результате набор пар (u_0, x_0) будет распределен равномерно в области под кривой $\tilde{q}(x)$. Теперь отбросим все точки (u_0, x_0) такие, что $u_0 > \tilde{p}(x_0)$. Оставшиеся пары (u_0, x_0) будут распределены равномерно в области под кривой $\tilde{p}(x)$. Теперь отбросим компоненты u_0 и получим выборку из распределения $p(x)$.

Очевидно, что метод Rejection sampling будет эффективным только в том случае, если функция $\tilde{q}(x)$ является достаточно точной оценкой сверху для $\tilde{p}(x)$ (серая область на рис. 2 имеет малую площадь). Понятно, что в одномерном случае плотность распределения всегда можно ограничить сверху кусочно-постоянной функцией (гистограммой), выборку из которой можно легко сгенерировать. На этом принципе построен один из наиболее эффективных алгоритмов генерации выборки из нормального распределения Ziggurat.

В заключение этого раздела рассмотрим вопрос генерации выборки из многомерного нормального распределения $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$. Если матрица ковариации Σ является диагональной, то все компоненты нормального распределения могут быть сгенерированы независимо из соответствующих одномерных нормальных распределений. Пусть Σ является произвольной положительно-определенной матрицей. Рассмотрим для нее разложение Холецкого $\Sigma = RR^T$, где R является верхнетреугольной матрицей. Сгенерируем точку $\boldsymbol{x}$ из стандартного многомерного нормального распределения $\mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, I)$. Тогда величина $\boldsymbol{y} = R\boldsymbol{x}$ будет распределена по закону $\mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, \Sigma)$. Действительно, линейная комбинация компонент нормального распределения распределена нормально, а мат.ожидание и матрица ковариации равны соответственно $\mathbb{E}\boldsymbol{y} = R\mathbb{E}\boldsymbol{x} = \boldsymbol{0}$ и $\mathrm{Cov}(\boldsymbol{y}) = \mathbb{E}\boldsymbol{y}\boldsymbol{y}^T = \mathbb{E}R\boldsymbol{x}\boldsymbol{x}^T R^T = R\mathbb{E}\boldsymbol{x}\boldsymbol{x}^T R^T = RIR^T = \Sigma$.

Идея MCMC

Рассмотрим теперь вопрос генерации выборки из распределения $p(T)$ в многомерном пространстве с помощью методов Монте Карло по схеме марковской цепи (MCMC). В этих методах вводится некоторая марковская цепь с априорным распределением $p_0(T)$ и вероятностями перехода в момент времени n $q_n(T_{n+1}|T_n)$, а генерация выборки

происходит следующим образом:

$$\begin{aligned} &T_1 \sim p_0(T), \\ &T_2 \sim q_1(T_2|T_1), \\ &\vdots \\ &T_N \sim q_{N-1}(T_N|T_{N-1}). \end{aligned} \tag{2}$$

Заметим, что при таком подходе генерируемая выборка не является набором независимых случайных величин. Однако, она подходит для оценки вероятностных интегралов вида (1) или оценки моды распределения. В том случае, если необходимо получить набор независимых величин, достаточно проредить полученный набор $T_1, \ldots, T_N$, взяв каждый m-ый отсчет, где m достаточно велико.

В дальнейшем рассматривается вопрос о том, как выбрать вероятности перехода $q_n(T_{n+1}|T_n)$ таким образом, чтобы выборка, генерируемая по схеме (2), была бы выборкой из интересующего нас распределения $p(T)$.

Теоретические свойства марковских цепей

Марковская цепь называется однородной, если вероятность перехода $q_n(T_{n+1}|T_n)$ не зависит от момента времени n, т.е. $q_n(T_{n+1}|T_n) = q(T_{n+1}|T_n)$. В дальнейшем будем рассматривать только однородные марковские цепи. Рассмотрим маргинальное распределение точек выборки в момент времени $n-1$, генерируемой с помощью однородной марковской цепи, и обозначим его через $p_{n-1}(T_{n-1})$. Тогда маргинальное распределение точек выборки в момент времени n можно вычислить следующим образом:

$$p_n(T_n) = \int q(T_n|T_{n-1}) p_{n-1}(T_{n-1}) dT_{n-1}.$$

Распределение $\pi(T)$ называется инвариантным относительно марковской цепи с вероятностью перехода q, если

$$\pi(T) = \int q(T|S)\pi(S)dS, \tag{3}$$

Очевидно, что для генерации выборки из распределения $p(T)$ по схеме марковской цепи необходимо потребовать, чтобы распределение $p(T)$ было инвариантным относительно этой марковской цепи. Достаточным условием инвариантности распределения $\pi(T)$ является выполнимость уравнения детального баланса:

$$\pi(S)q(T|S) = \pi(T)q(S|T).$$

Действительно,

$$\int q(T|S)\pi(S)dS = \{\text{ур-е детального баланса}\} = \int q(S|T)\pi(T)dS = \\ = \pi(T) \underbrace{\int q(S|T)dS}_{=1} = \pi(T).$$

Марковская цепь может иметь более одного инвариантного распределения. Пусть $\pi(T)$ – ее инвариантное распределение. Тогда марковская цепь называется эргодичной, если

$$\forall\ p_0(T) \xrightarrow[n\to+\infty]{} \pi(T).$$

Здесь $p_0(T)$ – начальное (априорное) распределение. Очевидно, что эргодичная марковская цепь имеет только одно инвариантное распределение. Достаточным условием эргодичности однородной марковской цепи является следующее свойство:

$$\forall S,\ \forall T : \pi(T) \neq 0 :\ \ q(T|S) > 0.$$

Теперь для генерации выборки из интересующего нас распределения $p(T)$ по схеме (2) достаточно потребовать, чтобы наша марковская цепь был однородной и эргодичной, а распределение $p(T)$ было инвариантным относительно нашей марковской цепи. Тогда, вне зависимости от начального распределения $p_0(T)$, начиная с некоторого момента времени n выборка, генерируемая по схеме (2), будет выборкой из распределения $p(T)$.

Схема Метрополиса-Хастингса

Пусть необходимо сгенерировать выборку из распределения $p(T)$, известного с точностью до нормировочной константы:

$$p(T) = \frac{1}{Z}\tilde{p}(T).$$

Рассмотрим шаг генерации по схеме Метрополиса-Хастингса. Пусть на шаге n сгенерирована конфигурация T_n. Тогда на шаге $n+1$ сначала генерируется конфигурация T_* из некоторого предложного распределения $r(T|T_n)$. Затем вычисляется величина

$$A(T_*, T_n) = \min\left(1, \frac{\tilde{p}(T_*)r(T_n|T_*)}{\tilde{p}(T_n)r(T_*|T_n)}\right)$$

и точка T_* принимается в качестве следующей точки T_{n+1} с вероятностью $A(T_*, T_n)$. В противном случае, $T_{n+1} = T_n$. Таким образом, мы ввели марковскую цепь с вероятностью перехода

$$q(T_{n+1}|T_n) = \begin{cases} r(T_{n+1}|T_n)A(T_{n+1}, T_n), & \text{если } T_{n+1} \neq T_n, \\ 1 - r(T_{n+1}|T_n)A(T_{n+1}, T_n), & \text{если } T_{n+1} = T_n. \end{cases}$$

Покажем, что распределение $p(T)$ является инвариантным относительно введенной марковской цепочки. Если $T_{n+1} = T_n$, то инвариантность сохраняется, т.к. значение T_n не изменяется. Для случая $T_{n+1} \neq T_n$ проверим выполнимость уравнения детального баланса:

$$p(T_n)q(T|T_n) = \min(p(T_n)r(T|T_n), p(T)r(T_n|T)) = \min(p(T)r(T_n|T), p(T_n)r(T|T_n)) = p(T)q(T_n|T).$$

Для эргодичности введенной марковской цепи достаточно потребовать выполнение $r(T|S) > 0,\ \forall\ T, S.$

В том случае, если предложное распределение является симметричным, т.е. $r(T|S) = r(S|T)$, $\forall\ S, T$, то схема Метрополиса-Хастингса переходит в классическую схему Метрополиса. Согласно этой схеме, если значение плотности в новой точке T_* оказалось выше, чем значение плотности в предыдущей точке T_n, то эта точка гарантированно принимается в качестве следующей точки выборки. Если плотность в новой точке оказалась меньше, то такая точка тоже может быть принята, но с вероятностью, пропорциональной величине уменьшения плотности.

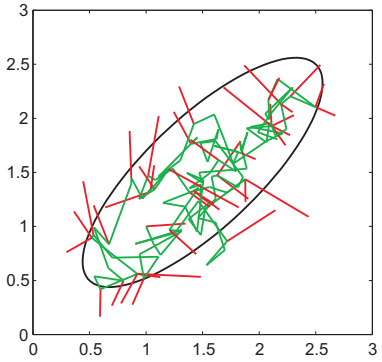

Рис. 3: Иллюстрация генерации выборки из двухмерного нормального распределения по схеме Метрополиса. Красные — отвергаемые шаги, зеленые — принимаемые шаги.

Рассмотрим модельный пример применения схемы Метрополиса (см. рис. 3). Пусть нам необходимо сгенерировать выборку из двухмерного нормального распределения с недиагональной матрицей ковариации. Возьмем в качестве предложного распределения двухмерное нормальное распределение с матрицей ковариации, пропорциональной единичной: $r(T|S) = \mathcal{N}(T|S, \sigma I)$. Это распределение, очевидно, является симметричным.

Значение параметра σ в значительной степени определяет эффективность процесса генерации выборки. Если значение σ слишком велико, то большинство новых точек будет отвергаться. Если значение σ слишком мало, то шаги в пространстве будут также маленькими, и понадобится очень много времени, чтобы покрыть область больших значений плотности распределения.

Схема Гиббса

Пусть необходимо сгенерировать выборку из многомерного распределения $p(T)$, где $T = \{t_1, \ldots, t_P\}$. Рассмотрим шаг генерации по схеме Гиббса. Пусть на шаге n сгенерирована конфигурация $T^n = \{t_1^n, \ldots, t_P^n\}$. Тогда генерация следующей точки выборки T^{n+1} происходит следующим образом:

$$
\begin{aligned}
t_1^{n+1} &\sim p(t_1|t_2^n, t_3^n, \ldots, t_P^n), \\
t_2^{n+1} &\sim p(t_2|t_1^{n+1}, t_3^n, t_4^n, \ldots, t_P^n), \\
t_3^{n+1} &\sim p(t_3|t_1^{n+1}, t_2^{n+1}, t_4^n, \ldots, t_P^n), \\
&\ldots \\
t_P^{n+1} &\sim p(t_P|t_1^{n+1}, t_2^{n+1}, \ldots, t_{P-1}^{n+1}).
\end{aligned}
\tag{4}
$$

Здесь через $p(t_i|T_{\setminus i})$ обозначено маргинальное одномерное распределение значений i-ой компоненты при условии всех остальных. Таким образом, согласно схеме Гиббса генерация выборки из многомерного распределения заменяется на итерационную генерацию точек из одномерных распределений. По аналогии с методами одномерной оптимизации генерация выборки из одномерного распределения является существенно более простой задачей, чем генерация выборки из многомерного распределения.

Докажем, что распределение $p(T)$ является инвариантным относительно введенной марковской цепи. Рассмотрим один шаг генерации очередной компоненты $t_p \sim p(t_p|T_{\setminus p})$. По предположению индукции $T_{\setminus p} \sim p(T_{\setminus p})$. Тогда совместная конфигурация $(t_p, T_{\setminus p}) \sim p(t_p|T_{\setminus p})p(T_{\setminus p}) = p(T)$. Отсюда, совместное распределение является инвариантным относительно одного шага процесса генерации (4). Следовательно, оно является инвариантным и относительно всего процесса (4).

При реализации схемы Гиббса на практике часто допускается следующая ошибка: вместо шага

$$t_p^{n+1} \sim p(t_p|t_1^{n+1}, \ldots, t_{p-1}^{n+1}, t_{p+1}^n, \ldots, t_P^n)$$

делается шаг

$$t_p^{n+1} \sim p(t_p|t_1^n, \ldots, t_{p-1}^n, t_{p+1}^n, \ldots, t_P^n),$$

т.е. в условие подставляются значения компонент только с предыдущей итерации. При таком подходе вероятность перехода в марковской цепи определяется как

$$q(T|T^n) = \prod_{p=1}^{P} p(t_p|T_{\setminus p}^n). \tag{5}$$

Распределение $p(T)$ не является инвариантным относительно данной марковской цепи! Эту ситуацию легко исправить, если взять схему Метрополиса-Хастингса, где в качестве предложного распределения фигурирует распределение (5). Заметим, что в отличие от схемы Гиббса, схема Метрополиса-Хастингса с предложным распределением (5) легко распараллеливается и на практике в некоторых ситуациях может работать быстрее, чем схема Гиббса.

Применение схемы Гиббса для дискретной марковской сети

Рассмотрим марковскую сеть с графом-решеткой с K-значными переменными. Распределение вероятности для конфигурации T этой марковской сети может быть записано как

$$p(T) = \frac{1}{Z} \exp\left[-\sum_{p=1}^{P} h_p(t_p) - \sum_{(i,j)\in\mathcal{E}} f_{ij}(t_i, t_j)\right], \quad t_p \in \{1, \ldots, K\}.$$

Здесь Z — нормировочная константа распределения, а h_p и f_{ij} – некоторые функции дискретного аргумента. Для применения схемы Гиббса необходимо уметь генерировать выборку из всех одномерных маргинальных распределений вида $p(t_p|T_{\setminus p}^n)$. В данном случае это распределение легко найти по следующей формуле:

$$p(t_p|T_{\setminus p}^n) \propto \exp\left[-h_p(t_p) - \sum_{i:(p,i)\in\mathcal{E}} f_{pi}(t_p, t_i^n)\right].$$

При этом константа данного распределения легко считается путем суммирования K величин. Это распределение является дискретным, и, следовательно, выборку из него легко получить путем генерации равномерно распределенной случайной величины.

Алгоритм 1: Оценка отношения нормировочных констант двух распределений с помощью схемы Гиббса

Вход: Ненормированные распределения $\tilde{p}_A(\boldsymbol{t})$ и $\tilde{p}_B(\boldsymbol{t})$.
Выход: Оценка отношения нормировочных констант Z_B/Z_A.

1: Построить последовательность распределений $\tilde{p}_k(\boldsymbol{t}) = [\tilde{p}_A(\boldsymbol{t})]^{1-\alpha_k}[\tilde{p}_B(\boldsymbol{t})]^{\alpha_k}$ для набора значений $0 = \alpha_1 < \alpha_2 < \cdots < \alpha_{K-1} < \alpha_K = 1$.
2: для $m = 1, \ldots, M$
3: Сгенерировать $\boldsymbol{t}^1$ из распределения $p_1(\boldsymbol{t})$;
4: Сделать шаг по схеме Гиббса $\boldsymbol{t}^2 \sim T_2(\boldsymbol{t}, \boldsymbol{t}^1)$ генерации из распределения $p_2(\boldsymbol{t})$;
5: ...
6: Сделать шаг по схеме Гиббса $\boldsymbol{t}^{K-1} \sim T_{K-1}(\boldsymbol{t}, \boldsymbol{t}^{K-2})$ генерации из распределения $p_{K-1}(\boldsymbol{t})$;
7: $w_m = \prod_{k=1}^{K-1}(\tilde{p}_{k+1}(\boldsymbol{t}^k)/\tilde{p}_k(\boldsymbol{t}^k))$;
8: $Z_B/Z_A \simeq \frac{1}{M}\sum_{m=1}^{M} w_m$.

Оценка нормировочной константы распределения с помощью схемы Гиббса

Предположим, что у нас имеется вероятностное распределение, известное с точностью до нормировочной константы $p(\boldsymbol{t}) = \tilde{p}(\boldsymbol{t})/Z$. Как было отмечено выше, схема Гиббса позволяет сгенерировать выборку из этого распределения $\boldsymbol{t}^1, \ldots, \boldsymbol{t}^N$, которая затем может быть использована для оценки статистики распределения $f(\boldsymbol{t})$ по формуле (1).

Рассмотрим задачу оценки нормировочной константы распределения Z. Эта нормировочная константа играет роль обоснованности модели и может быть использована для сравнения различных вероятностных моделей между собой, а также для оценки вероятности $p(\boldsymbol{t})$ для тестовых объектов. Нормировочная константа является «нулевой статистикой» распределения и поэтому не может быть оценена с помощью формулы (1).

Предположим, что у нас имеется два распределения $p_A(\boldsymbol{t}) = \tilde{p}_A(\boldsymbol{t})/Z_A$ и $p_B(\boldsymbol{t}) = \tilde{p}_B(\boldsymbol{t})/Z_B$. Тогда отношение двух нормировочных констант можно оценить по следующей схеме:

$$\frac{Z_B}{Z_A} = \frac{\int \tilde{p}_B(\boldsymbol{t})d\boldsymbol{t}}{Z_A} = \int \frac{\tilde{p}_B(\boldsymbol{t})}{Z_A}d\boldsymbol{t} = \int \frac{\tilde{p}_B(\boldsymbol{t})}{\tilde{p}_A(\boldsymbol{t})}p_A(\boldsymbol{t})d\boldsymbol{t} \simeq \frac{1}{M}\sum_{m=1}^{M}\frac{\tilde{p}_B(\boldsymbol{t}^m)}{\tilde{p}_A(\boldsymbol{t}^m)}. \tag{6}$$

Здесь $\boldsymbol{t}^1, \ldots, \boldsymbol{t}^M$ – выборка из распределения $p_A(\boldsymbol{t})$, которую можно сгенерировать, например, по схеме Гиббса. Если у распределения $p_A(\boldsymbol{t})$ нормировочная константа известна, то тогда мы можем оценить абсолютное значение Z_B.

К сожалению, схема (6) применима только в случае, когда распределение $p_A(\boldsymbol{t})$ является хорошим приближением для $p_B(\boldsymbol{t})$. На практике поиск хорошего аналитического приближения для интересующего нас распределения $p_B(\boldsymbol{t})$ может оказаться очень трудной задачей. В этом случае можно построить серию промежуточных распределений $p_A = p_1, p_2, \ldots, p_{K-1}, p_K = p_B$ и оценить нормировочную константу Z_B из

соотношения:

$$\frac{Z_B}{Z_A} = \frac{Z_K}{Z_1} = \frac{Z_2}{Z_1}\frac{Z_3}{Z_2}\cdots\frac{Z_{K-1}}{Z_{K-2}}\frac{Z_K}{Z_{K-1}}.$$

Здесь каждое отношение Z_{k+1}/Z_k оценивается по схеме (6) путем генерации выборки из распределения $p_k(\boldsymbol{t})$, а серия промежуточных распределений строится как $\tilde{p}_k(\boldsymbol{t}) = [\tilde{p}_A(\boldsymbol{t})]^{1-\alpha_k}[\tilde{p}_B(\boldsymbol{t})]^{\alpha_k}$ для некоторого набора значений $0 = \alpha_1 < \alpha_2 < \cdots < \alpha_{K-1} < \alpha_K = 1$.

Предположим, что для каждого промежуточного распределения $p_k(\boldsymbol{t}) = \tilde{p}_k(\boldsymbol{t})/Z_k$, известного с точностью до нормировочной константы, мы можем применить схему Гиббса генерации выборки из этого распределения. Обозначим один шаг такой схемы Гиббса через $\boldsymbol{t}_{next} \sim T_k(\boldsymbol{t}, \boldsymbol{t}_{pred})$. Тогда итоговую схему оценки нормировочной константы Z_B можно представить как Алгоритм 1.