УДК 681.3.06

ОПТИМИЗАЦИЯ ОДНОРОДНЫХ НЕМАРКОВСКИХ СЕТЕЙ МАССОВОГО ОБСЛУЖИВАНИЯ

В.Н. Задорожный

Предложен новый эффективный аналитико-имитационный метод оптимизации немарковских сетей массового обслуживания. Экспериментально оценена скорость сходимости и точность метода. Даны практические рекомендации по его применению.

Ключевые слова: сеть массового обслуживания, оптимизация, аналитико-имитационное моделирование.

ВВЕДЕНИЕ

Производительность организационно-технических систем, предназначенных для обработки или обслуживания дискретных потоков каких-либо однотипных единиц (заявок), часто оценивается по времени прохождения заявок через эти системы. Унифицированным формализованным представлением подобных систем выступает сеть массового обслуживания (СеМО) со статистически однородными заявками — однородная сеть [1].

В виде СеМО традиционно представляются, например, информационно-вычислительные системы [1—5]. Заявки рассматриваются как передаваемые сообщения или как пользовательские запросы, обрабатываемые ресурсами системы. Время прохождения заявки через СеМО будем называть временем ответа. Среднее время ответа E зависит от распределения имеющихся ресурсов сети между ее узлами.

В последние годы успешно развиваются методы оптимизации марковских [1, 2] СеМО — сетей с экспоненциальными распределениями времени обслуживания заявок и (если сеть не замкнута) с пуассоновскими входными потоками. Задача оптимального распределения ресурса замкнутой марковской сети сводится к системе нелинейных уравнений, эффективно решаемой численными методами [1]. Решается задача оптимизации замкнутых марковских сетей с несколькими классами сообщений [6]. В работах [1, 7] для немарковских сетей рассматриваются методы приближенного расчета, основанные на аппроксимации произвольных распределений распределениями, допускающими рациональное преобразование Лапласа. Однако этот путь сопряжен со значительными вычислительными трудностями даже для одновариантного расчета сетей с небольшим числом узлов (особенно если приходится аппроксимировать распределения случайных величин, ограниченных узкими диапазонами возможных значений [8]). В работе [9] предпринята попытка решения задач оптимизации немарковских СеМО путем аппроксимации их узлов аналитическими выражениями, учитывающими первые два момента времени обслуживания и интервалов поступления заявок в узлы, с учетом межузловых взаимодействий. Однако до решения оптимизационных задач эта работа не доводится.

Таким образом, в общем случае для оптимизации немарковских сетей приходится использовать имитационное моделирование (ИМ). При этом если задача оптимизации содержит более двух-трех независимых переменных, то ее решение становится практически невозможным без привлечения градиентных методов. Но расчет градиентов в ИМ существенно затрудняется стохастической погрешностью вычисляемых оценок $\hat{E}$ отклика E [7, 10]. Известные методы [7, 11—13] решения этой проблемы либо пригодны только при ИМ изолированных систем массового обслуживания [7, 12, 13] и не распространяются на СеМО, либо их теоретически возможное применение к СеМО на практике приводит к значительным трудностям. Например, применению методов, опирающихся на имитацию большого числа периодов регенерации [11—14] препятствует то обстоятельство, что в СеМО, как правило, эти периоды бывают практически бесконечными. Успешные применения подобных методов ограничиваются классами СеМО, учитывающими специфику конкретных сетевых объектов при конкретных диапазонах их параметров [3].

Для задачи оптимального распределения ресурса по узлам однородной немарковской СеМО в статье предлагается метод, позволяющий эффективно решать проблему градиентов путем приме-

нения простой сепарабельной аппроксимации целевой функции.

Далее формулируется решаемая задача.

1. ЗАДАЧА ОПТИМИЗАЦИИ ОДНОРОДНОЙ НЕМАРКОВСКОЙ СЕТИ

Рассмотрим сначала *открытую* сеть, в которую поступает рекуррентный поток заявок с интенсивностью Λ. Интервалы поступления заявок — независимые случайные величины (сл. в.) с функцией распределения вероятностей (ф. р. в.) $A(t)$. Заявка из входного потока сети с вероятностью p_{0i} попадает в i-й узел, $i = \overline{1, n}$. В любом из K_i каналов i-го узла время обслуживания заявки (независимая сл. в.) имеет ф. р. в. $B_i(t)$. После обслуживания в i-м узле заявка случайно и независимо, в соответствии с заданными переходными вероятностями p_{ij}, выбирает один из узлов j для продолжения своего маршрута или, с вероятностью p_{i0}, уходит из сети. Вероятности p_{ij} ($i, j = \overline{0, n}$) задаются неразложимой стохастической матрицей $\mathbf{P} = \|p_{ij}\|$.

Стационарное среднее время E прохождения заявки через сеть (среднее время ответа) можно представить в виде:

$$E = \sum_{i=1}^{n} \alpha_i u_i = \sum_{i=1}^{n} \alpha_i (w_i + b_i) = \sum_{i=1}^{n} \alpha_i \left(w_i + \frac{1}{\mu_i} \right), \quad (1)$$

где α_i — среднее число посещений i-го узла заявкой за время ее прохождения через сеть, u_i — среднее время пребывания заявки в i-м узле, w_i — среднее время ожидания заявки в очереди i-го узла, b_i — среднее время обслуживания заявки в i-м узле, $\mu_i = b_i^{-1}$ — интенсивность обслуживания заявки каналом i-го узла.

Коэффициенты α_i однозначно определяются из системы уравнений баланса:

$$\alpha_i = \sum_{j=0}^{n} \alpha_j p_{ji}, \quad i = \overline{0, n}, \quad \alpha_0 \equiv 1.$$

Через коэффициенты α_i последовательно определяются интенсивности $\lambda_i = \Lambda \cdot \alpha_i$ входных потоков узлов, их коэффициенты загрузки $\rho_i = \lambda_i/(\mu_i K_i)$, и проверяются условия стационарности $\rho_i \leqslant 1$ (или $\mu_i \geqslant \lambda_i/K_i$), $i = \overline{1, n}$. Значения w_i в формуле (1) определяются посредством ИМ.

Если сеть *замкнута*, то входной поток заявок отсутствует (параметр Λ не задается), и в сети циркулирует заданное постоянное число заявок R. Завершением цикла обслуживания заявки считается ее переход по некоторой «терминальной» дуге с «выхода» на «вход» сети. Среднее время ответа E замкнутой сети (среднее время цикла обслуживания) также определяется формулой (1) и может вычисляться посредством ИМ.

Теперь рассмотрим следующую обобщенную версию сформулированной в работе [1] задачи оптимизации однородной замкнутой экспоненциальной (марковской) сети.

Стоимость (ресурс) M однородной сети как функция вектора $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ интенсивностей обслуживания в узлах $i = \overline{1, n}$ задается в виде $M(\vec{\mu}) = \sum_{i=1}^{n} c_i \mu_i^{\beta_i}$, где c_i — стоимостные коэффициенты, $\beta_i > 0$ — коэффициенты нелинейности. Требуется найти вектор $\vec{\mu} = \vec{\mu}_{\text{opt}}$, доставляющий минимум функции $E = E(\vec{\mu})$:

$$E(\vec{\mu}) = \sum_{i=1}^{n} \alpha_i \left(w_i(\vec{\mu}) + \frac{1}{\mu_i} \right) \to \min_{\vec{\mu}}, \quad (2)$$

и принадлежащий следующей области допустимых решений (ОДР):

$$M(\vec{\mu}) = \sum_{i=1}^{n} c_i \mu_i^{\beta_i} = M^*, \quad \mu_i \geqslant \mu_{i\min}, \quad i = \overline{1, n}, \quad (3)$$

где для открытой сети $\mu_{i\min} = \lambda_i/K_i$ (граница области стационарности), а для замкнутой $\mu_{i\min} = 0$. Для ресурса M^* в выражении (3) должно выполняться условие $M^* \geqslant M_{\min}$, где для открытой сети $M_{\min} = \sum_{i=1}^{n} c_i \mu_{i\min}^{\beta_i} = \sum_{i=1}^{n} c_i (\lambda_i/K_i)^{\beta i}$, а для замкнутой $M_{\min} = 0$. В задаче (2), (3) имеется в виду, что изменение любой интенсивности μ_i приводит к изменению среднего $b_i = \mu_i^{-1}$ и к соответствующему масштабному изменению ф. р. в. $B_i(t)$. Вид ф. р. в. $B_i(t)$ не изменяется, поскольку ассоциируется со случайной трудоемкостью заявок, тогда как варьируемый параметр μ_i определяется производительностью каналов i-го узла.

2. ОБЩАЯ СТРУКТУРА И ОПОРНЫЕ ЭЛЕМЕНТЫ МЕТОДА «НАПРАВЛЯЮЩИХ ГИПЕРБОЛ»

Предлагаемый метод решения задачи (2), (3) состоит из двух этапов.

Этап I: ускоренный градиентный поиск точки $\vec{\mu}_{\text{opt}}$ *методом «направляющих гипербол»* (НГ), использующим ИМ сети и сепарабельную аппроксимацию целевой функции (название метода отражает роль и вид функций одного переменного — слагаемых этой аппроксимации).

Этап II (не обязательный): уточнение найденного решения методом циклического покоординатного спуска [15] (модифицированным), не использующим аппроксимаций.

Предлагаемый в настоящей статье метод решения задачи (2), (3) использует уточненную версию предложенной в работе [16] аппроксимации функции $E(\vec{\mu})$. Аппроксимация $E^{ap}(\vec{\mu})$ среднего времени ответа $E(\vec{\mu})$ на каждой итерации $k \geqslant 2$ поиска оптимального решения $\vec{\mu}_{\text{opt}}$ формируется по результатам ИМ сети в точках $\vec{\mu} = \vec{\mu}^{k-1}$ и $\vec{\mu} = \vec{\mu}^{k}$ и применяется для определения следующей точки $\vec{\mu} = \vec{\mu}^{k+1}$.

Определим опорные элементы метода НГ.

Центр $\vec{\mu}_c$ ОДР (3) при ресурсе $M^* > M_{\min}$ определяется условием равной загрузки узлов: $\rho_i = \lambda_i/(\mu_i K_i) = \alpha_i\lambda_0/(\mu_i K_i) = \rho_c$, $i = \overline{1, n}$, где λ_0 — интенсивность на терминальной дуге (для открытой сети $\lambda_0 = \Lambda$). Отсюда $\mu_i = (\alpha_i\lambda_0)/(K_i\rho_c)$, $\mu_i/\mu_1 = (\alpha_i/K_i)(K_1/\alpha_1)$ и $\mu_i = (\alpha_i/K_i)(K_1/\alpha_1)\mu_1$. Подставляя последнее выражение μ_i в формулу (3), имеем:

$$\sum_{i=1}^{n} c_i\Big(\frac{\alpha_i}{K_i}\cdot\frac{K_1}{\alpha_1}\cdot\mu_1\Big)^{\beta i} = M^*, \quad \mu_i = (\alpha_i/K_i)(K_1/\alpha_1)\mu_1, \quad i = \overline{2, n}. \tag{4}$$

Отсюда численными методами легко находится единственный положительный корень μ_1, определяющий все остальные координаты μ_i центра $\vec{\mu}_c$ ОДР. Если сеть открытая, то в центре $\vec{\mu}_c$ сразу определяются все $\rho_i = \rho_c \leqslant 1$, $i = \overline{1, n}$.

Если все $\beta_i = 1$, то из формулы (4) координаты центра ОДР можно выразить явно:

$$\mu_i = M^*(\alpha_i/K_i)\Big(\sum_{j=1}^{n} c_j\alpha_j/K_j\Big)^{-1}, \quad i = \overline{1, n}.$$

Открытая сеть имеет в центре $\vec{\mu}_c$ максимум пропускной способности $V(\vec{\mu})$, определяемой для нее как $V(\vec{\mu}) = \max\{\Lambda\colon \rho_i = \alpha_i\Lambda/(\mu_i K_i) \leqslant 1,\ i = \overline{1, n}\}$.

Диаметр D ОДР определим как длину максимального из диапазонов варьирования переменных μ_i: $D = \max\{l_i\}$, где $l_i = \mu_{i\max} - \mu_{i\min}$ и, согласно формуле (3), $\mu_{i\max} = \Big[\Big(M^* - \sum\limits_{j\neq i} c_j\mu_{j\min}^{\beta_j}\Big)c_i^{-1}\Big]^{1/\beta_i}$, $i = \overline{1, n}$.

Малый шаг размером, например, $D\cdot 10^{-4}$, будем применять при построении и сканировании траекторий на поверхности ограничений, определяемой уравнением (3). Шаг выбирается с учетом требований к точности оптимизации.

Аппроксимация $E^{ap}(\vec{\mu})$ целевой функции $E(\vec{\mu})$, используемая для приближенного вычисления градиента, представляет собой сепарабельную функцию варьируемых переменных μ_i:

$$E^{ap}(\vec{\mu}) = \sum_{i=1}^{n}\alpha_i\Big(W_i(\mu_i) + \frac{1}{\mu_i}\Big),$$

$$\text{где } W_i(\mu_i) = \begin{cases} \dfrac{R_i}{\mu_i - S_i}, \text{ если } \hat{w}_i^k \neq \hat{w}_i^{k-1}, \\ \hat{w}_i^k, \text{ если } \hat{w}_i^k = \hat{w}_i^{k-1}, \end{cases} \tag{5}$$

и на каждом шаге k оптимизации заново настраивается (посредством коэффициентов R_i и S_i) по оценкам $\hat{w}_i^{k-1}$ и $\hat{w}_i^k$ среднего времени ожидания, найденным для узлов $i = \overline{1, n}$ с помощью ИМ сети в точках $\vec{\mu} = \vec{\mu}^{k-1}$ и $\vec{\mu} = \vec{\mu}^{k}$. При $\hat{w}_i^k \neq \hat{w}_i^{k-1}$ выражение $R_i/(\mu_i - S_i)$ в формуле (5) аппроксимирует соответствующую функцию $w_i(\vec{\mu})$в выражении (2) так, что его значение в точках $\vec{\mu} = \vec{\mu}^{k-1}$ и $\vec{\mu} = \vec{\mu}^{k}$ совпадает с оценками $\hat{w}_i^{k-1} \approx w_i(\vec{\mu}^{k-1})$ и $\hat{w}_i^k \approx w_i(\vec{\mu}^k)$. Таким образом, имеем $R_i/(\mu_i^{k-1} - S_i) = \hat{w}_i^{k-1}$ и $R_i/(\mu_i^k - S_i) = \hat{w}_i^k$, откуда

$$S_i = \frac{\hat{w}_i^k\mu_i^k - \hat{w}_i^{k-1}\mu_i^{k-1}}{\hat{w}_i^k - \hat{w}_i^{k-1}}, \quad R_i = \hat{w}_i^{k-1}(\mu_i^{k-1} - S_i), \quad i = \overline{1, n}, \tag{6}$$

(верхний индекс здесь везде соответствует шагу оптимизации).

При $\hat{w}_i^k = \hat{w}_i^{k-1}$ расчет значений R_i и S_i в формуле (5) не нужен, но для определения *валидной части* $[L]$ (см. далее) полагаем $S_i \to \infty$. Такая настройка функции $E^{ap}(\vec{\mu})$ обеспечивает ее совпадение с целевой функцией $E(\vec{\mu})$ в точках $\vec{\mu}^{k-1}$ и $\vec{\mu}^k$ (с точностью до стохастической погрешности оценок ИМ). В других точках $\vec{\mu}$ точность аппроксимации $E^{ap}(\vec{\mu})$ тем хуже, чем они дальше от точек $\vec{\mu}^{k-1}$ и $\vec{\mu}^k$, и чем «менее сепарабельна» $E(\vec{\mu})$, т. е. чем сильнее изменение интенсивностей μ_i в одних узлах влияет на среднее время w_j в других ($i \neq j$).

Градиент $\nabla E^{ap}(\vec{\mu})$ в точке $\vec{\mu} = \vec{\mu}^k$ есть аппроксимация градиента $\nabla E(\vec{\mu})$ в этой точке, и вычисляется с помощью выражения, полученного дифференцированием выражения (5):

$$\nabla E^{ap}(\vec{\mu}^k) = \Big(\alpha_1\frac{\partial W_1}{\partial\mu_1} - \frac{\alpha_1}{(\mu_1)^2}, \ldots, \alpha_n\frac{\partial W_n}{\partial\mu_n} - \frac{\alpha_n}{(\mu_n)^2}\Big), \tag{7}$$

где $\dfrac{\partial W_i}{\partial \mu_i} = \begin{cases} \dfrac{-R_i}{(\mu_i^k - S_i)^2}, \hat{w}_i^k \neq \hat{w}_i^{k-1}, \\ 0, \hat{w}_i^k = \hat{w}_i^{k-1}. \end{cases}$ $i = \overline{1, n}$.

Валидная часть $[L]$ исходящей из точки $\vec{\mu}^k$ траектории L поиска точки $\vec{\mu}^{k+1}$ лежит между $\vec{\mu}^k$ и первой на L точкой $\vec{\mu}$, у которой какая-либо координата μ_i достигает границы ОДР $\mu_i = \mu_{i\min}$ или полюса $\mu_i = S_i$ аппроксимации (5).

Приведем пошаговое описание метода НГ с расчетными формулами.

3. МЕТОД «НАПРАВЛЯЮЩИХ ГИПЕРБОЛ»

Начальная фаза оптимизации. Задаем число итераций $N > 2$ (выбирается с учетом допустимых затрат машинного времени) и две точки $\vec{\mu}^1 = \vec{\mu}_c$ и $\vec{\mu}^2 \neq \vec{\mu}^1$, принадлежащие ОДР (3). С помощью ИМ вычисляем в этих точках оценки среднего времени ответа $\hat{E}^1$ и $\hat{E}^2$, и, соответственно, оценки среднего времени ожидания $(\hat{w}_1^1, ..., \hat{w}_n^1)$, $(\hat{w}_1^2, ..., \hat{w}_n^2)$ в узлах 1, ..., n. Полагаем $k = 2$.

Основной цикл. Известны точки $\vec{\mu}^{k-1}$ и $\vec{\mu}^k$ с оценками $(\hat{w}_1^{k-1}, ..., \hat{w}_n^{k-1})$, $\hat{E}^{k-1}$ и $(\hat{w}_1^k, ..., \hat{w}_n^k)$, $\hat{E}^k$ откликов $(w_1^{k-1}, ..., w_n^{k-1})$, E^{k-1} и $(w_1^k, ..., w_n^k)$, E^k.

1. Используя оценки $(\hat{w}_1^{k-1}, ..., \hat{w}_n^{k-1})$ и $(\hat{w}_1^k, ..., \hat{w}_n^k)$, находим по формулам (6) коэффициенты R_i и S_i, $i = \overline{1, n}$, аппроксимации $E^{ap}(\vec{\mu})$.

Вычисляем градиент (7) функции E^{ap}. Направление $-\nabla E^{ap}(\vec{\mu}^k)$ наискорейшего убывания функции $E^{ap}(\vec{\mu})$ проецируем на поверхность ограничений (3). Если $\beta_i = 1$ для всех $i = \overline{1, n}$, то поверхность (3) является гиперплоскостью, и проекция L на нее направления вектора $-\nabla E^{ap}(\vec{\mu}^k)$ есть направление вектора $-\nabla E_{pr}^{ap}(\vec{\mu}^k) = -\nabla E^{ap}(\vec{\mu}^k) + \vec{n}(\vec{n}\nabla E^{ap}(\vec{\mu}^k)) = \vec{e}$, где $\vec{n} = \vec{c}/|\vec{c}|$ — нормаль к гиперплоскости ограничений, $\vec{c} = (c_1, ..., c_n)$ — вектор стоимостных коэффициентов, $|\vec{x}|$ — длина вектора $\vec{x}$. Валидная часть $[L]$ проекции L ограничена точками $\vec{\mu}^k$ и $\vec{\mu} = \vec{\mu}^k + h\vec{e}$, где $h = \min\{h_1, h_2\}$,

$$h_1 = \min\{h_{1i}: h_{1i} > 0; i = \overline{1, n}\},$$

$$h_2 = \min\{h_{2i}: h_{2i} > 0; i = \overline{1, n}\},$$

$$h_{1i} = -(\mu_i^k - \mu_{i\min})/e_i, \quad h_{2i} = -(\mu_i^k - S_i)/e_i, \quad i = \overline{1, n}.$$

Если не все β_i равны 1, то проекцию L направления антиградиента строим пошагово, как исходящую из точки $\vec{\mu}^k$ ломаную, узлы $\vec{\mu}$ которой суть проекции на поверхность (3) равноотстоящих с малым шагом точек направления $-\nabla E^{ap}(\vec{\mu}^k)$. Для каждого очередного узла $\vec{\mu}$ проверяются условия его валидности $\mu_i > \mu_{i\min}$ и $(\mu_i - S_i)(\mu_i^k - S_i) > 0$, $i = \overline{1, n}$. Равенство знаков разностей $(\mu_i - S_i)$ и $(\mu_i^k - S_i)$ означает, что координата μ_i текущего узла $\vec{\mu}$ траектории L и координата μ_i^k начальной ее точки $\vec{\mu}^k$ находятся по одну сторону от полюса S_i аппроксимации. Построение $[L]$ завершается получением и отбрасыванием первого невалидного узла либо выяснением, что очередная точка в направлении антиградиента $-\nabla E^{ap}(\vec{\mu}^k)$ уже не имеет проекции на поверхность ограничений.

2. В качестве следующей точки $\vec{\mu}^{k+1}$ выбираем решение (получаемое методом сканирования) задачи одномерной оптимизации $E^{ap}(\vec{\mu}) \to \min$, $\vec{\mu} \in [L]$.

3. Полагаем $k = k + 1$. С помощью ИМ вычисляем оценки $(\hat{w}_1^k, ..., \hat{w}_n^k)$ и $\hat{E}^k$. Если $k < N$, то переходим к шагу 1, иначе — к шагу 4.

4. Точку $\vec{\mu}^* \in \{\vec{\mu}^1, ..., \vec{\mu}^N\}$ с оценкой $\hat{E}(\vec{\mu}^*) = \min\{\hat{E}^1, ..., \hat{E}^N\}$ принимаем в качестве приближенного решения задачи. *Конец алгоритма.*

4. УСКОРЕННЫЙ ПОКООРДИНАТНЫЙ СПУСК (ВТОРОЙ ЭТАП ОПТИМИЗАЦИИ)

Погрешность первого этапа (метода НГ) включает в себя две составляющие: *стохастическую* и *детерминированную*. Стохастическая составляющая контролируется расчетом доверительных интервалов для оценок $(\hat{w}_1^k, ..., \hat{w}_n^k)$ и $\hat{E}^k$, и ее можно снижать путем удлинения прогонов модели. Детерминированная составляющая обусловлена применением сепарабельной аппроксимации $E^{ap}(\vec{\mu})$ для несепарабельной (в общем случае) функции $E(\vec{\mu})$. Это приводит к тому, что решение $\vec{\mu}^*$, найденное методом НГ, отличается от искомого $\vec{\mu}_{\text{opt}}$ даже при условии полного устранения стохастических погрешностей оценок ИМ. Поэтому решение $\vec{\mu}^*$ в общем случае целесообразно уточнять (или проверять) методом, не использующим аппроксимацию $E^{ap}(\vec{\mu})$.

Для этого на втором этапе оптимизации в окрестности решения $\vec{\mu}^*$ определяется $2(n-1)$ пробных точек, каждая из которых отличается от $\vec{\mu}^*$ лишь одной из $(n-1)$ «свободных» координат μ_i

(например, одной из координат $\mu_1, ..., \mu_{n-1}$) на величину $\pm\Delta\mu$. «Связанная» координата пробной точки (например, μ_n) определяется через известные «свободные» ее координаты путем решения уравнения (3), чтобы обеспечить принадлежность всех пробных точек ОДР. Отклонение $\Delta\mu$ выбирается с учетом допустимой погрешности решения и с учетом возможности надежного сравнения откликов $E(\vec{\mu})$ в соответствующих точках по их оценкам $\hat{E}(\vec{\mu})$. Далее решение $\vec{\mu}^*$ уточняется путем ИМ сети в пробных точках и замены $\vec{\mu}^*$ тем пробным решением $\vec{\mu}$, для которого $E(\vec{\mu}) < E(\vec{\mu}^*)$. Если такое не находится, процесс завершается, иначе для нового решения цикл уточнения повторяется. Для исключения зацикливания возврат к пройденным решениям запрещается. Для ускоренного сравнения откликов $E(\vec{\mu})$ и $E(\vec{\mu}^*)$ с заданной надежностью по их оценкам $\hat{E}(\vec{\mu})$ и $\hat{E}(\vec{\mu}^*)$ применяется метод «общих случайных чисел» [10].

5. ИСПЫТАНИЯ МЕТОДА «НАПРАВЛЯЮЩИХ ГИПЕРБОЛ» НА ОТКРЫТЫХ СЕТЯХ

Подробно охарактеризовать возможности метода НГ можно на примере оптимизации открытой версии тестовой СеМО-1 (рис. 1). Суммарный ресурс $M = 30$ распределяется здесь при $\vec{c} = (c_1, ..., c_n) = (K_1, ..., K_9) = (1, 2, 1, 1, 1, 1, 1, 3, 1)$, т. е. для каждой системы массового обслуживания стоимостный коэффициент равен числу ее каналов. Типы распределений $B_i(t)$ для узлов $i = 1, ..., 9$ определены как R, R, R, M, M, E^2, E^2, E^2, R соответственно, где M — экспоненциальное распределение, R — равномерное (на отрезке от 0 до двух средних), E^m — эрланговское распределение m-го порядка. Входной поток СеМО-1 пуассоновский и имеет интенсивность $\Lambda = 1$. Переходные вероятности указаны на рис. 1.

Тестовая сеть СеМО-1 применялась для сравнения метода НГ с базовым алгоритмом оптимизации, в котором градиент рассчитывался методом малых приращений. В каждом прогоне модели СеМО-1 имитировалось прохождение через сеть около 1 млн. заявок. На рис. 2 приведена типичная траектория изменений целевой функции в процессе оптимизации, выполняемой методом НГ. Приближение к точке оптимума происходит за 7—11 шагов, т. е. число эффективных итераций близко к размерности n факторного пространства. Вблизи точки оптимума растут стохастические ошибки аппроксимации, перенастраиваемой по двум сближающимся точкам факторного пространства, и происходит «отбрасывание» очередного приближения $\vec{\mu}^{k+1}$ от искомой точки оптимума, вследствие чего значение целевой функции резко возрастает. Затем процесс вновь возвращается к точке оптимума. На рис. 3 показана типичная траектория значений целевой функции, получаемая при использовании базового метода. Заметно выигрывая у этого метода в глубине оптимизации, метод НГ превосходит его и в быстродействии. Это достигается как благодаря резкому снижению числа итераций, так и потому, что для вычисления градиента на каждой итерации в методе НГ используется лишь один прогон имитационной модели, а не $(n + 1)$ прогонов, как в базовом методе. Решение, определяемое для СеМО-1 методом НГ за 7—11 итераций, обеспечивает среднее время ответа $E \approx 7{,}49...7{,}51$. Базовый метод за 100...140 итераций (т. е. за 1000...1400 прогонов модели) дает более слабый результат $E \approx 9{,}3...9{,}5$.

Тестовая задача решалась также с помощью оптимизатора OptQuest [17]. Программой OptQuest за

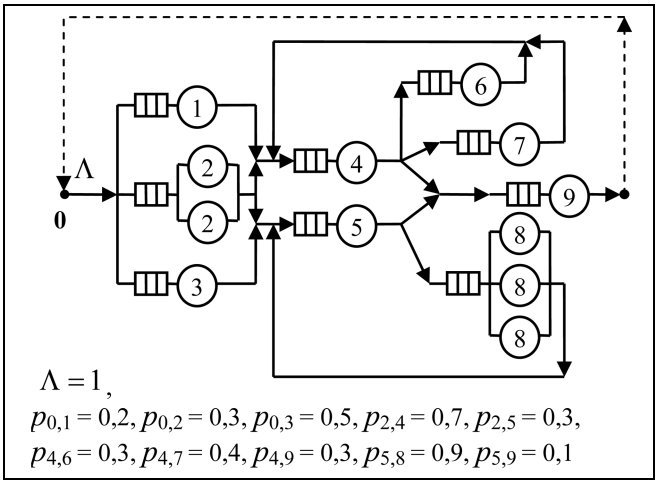

Рис. 1. Тестовый пример СеМО-1. Штриховая дуга соответствует замкнутой версии сети

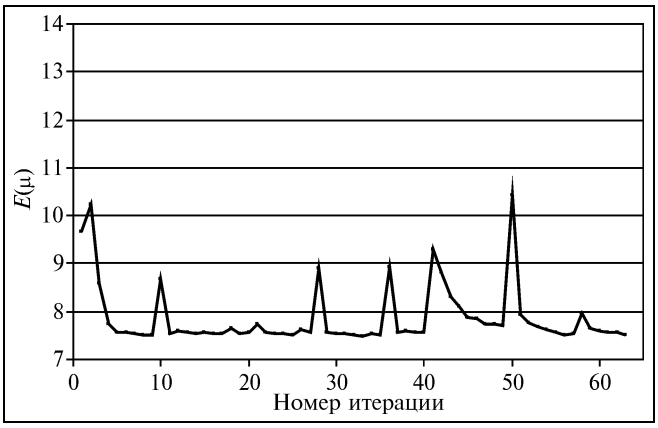

Рис. 2. Изменения отклика E в ходе поиска оптимума методом «направляющих гипербол»

750...1000 прогонов модели, т. е. при стократной «форе» по времени, достигается среднее время ответа $E \approx 7{,}91$, также проигрывающее результату метода НГ.

В более полный набор тестов включены открытые сети СеМО-2, ..., СеМО-5.

Сеть СеМО-2 отличается от СеМО-1 лишь равномерным на отрезке [0,5; 1,5] распределением интервалов поступления заявок в сеть. Сеть СеМО-3 состоит из 30-ти узлов и представляет собой каскадное (последовательное) соединение трех сетей, подобных СеМО-1, усложненное возвратами заявок с выходов всех каскадов на их входы и на входы предшествующих каскадов. Сеть СеМО-4 содержит 20 узлов, ее маршрутная матрица сформирована с помощью датчика случайных чисел при соблюдении требования связности сети. Число каналов в каждом узле также разыгрывалось как случайное и не превышало шести. Сеть СеМО-5 содержит 100 узлов и имеет структуру переходов, представленную на рис. 4. Соответствующие переходные вероятности, число каналов в узлах и вид законов распределения $B_i(t)$ определялись датчиком случайных чисел. Кроме распределений, использованных в СеМО-2, в СеМО-5 применялись

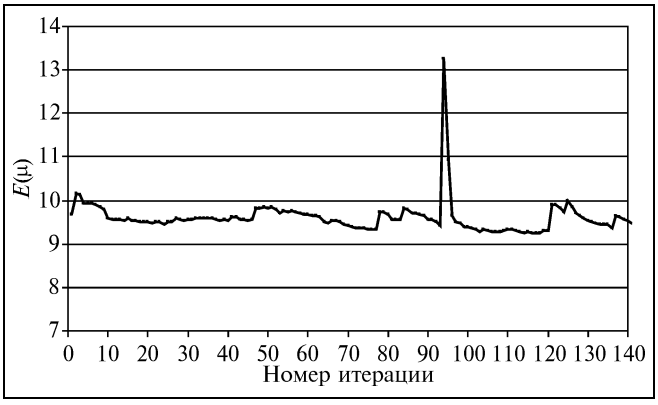

Рис. 3. Изменения отклика *E* при применении базового метода малых приращений

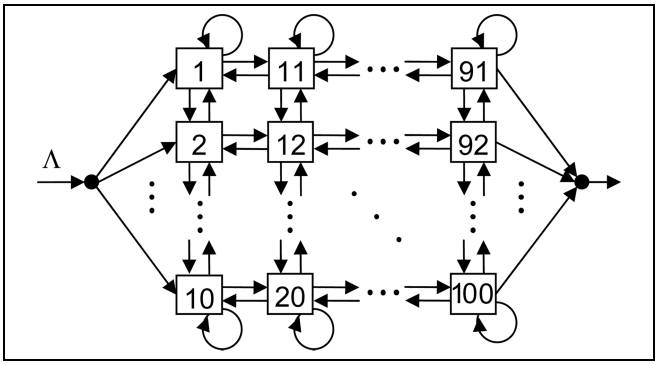

Рис. 4. Структура переходов в СеМО-5

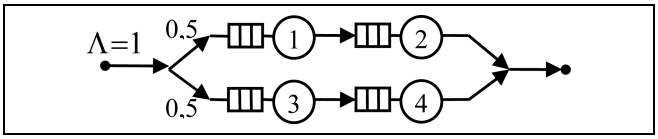

Рис. 5. Тестовая двухлинейная система

следующие распределения: эрланговское 3-го порядка; треугольное; гамма-распределение с параметром формы $\alpha = 2{,}5$; гиперэкспоненциальные распределения с коэффициентами вариации (к. в.) $C = 2$ и $C = 3$ и распределения Вейбулла с к. в. $C = 2$, $C = 3$ и $C = \sqrt{5}$. Последнему распределению принадлежат и интервалы поступления заявок в СеМО-4 и СеМО-5.

В табл. 1 приведены результаты оптимизации этих четырех тестовых сетей с помощью метода НГ. Каждая сеть оптимизировалась дважды: при распределяемом ресурсе $M \approx 1{,}5M_{\min}$ и при $M \approx 2M_{\min}$. Время T_M итерации (практически совпадающее со временем прогона модели) указано в минутах для двухядерной ПЭВМ с тактовой частотой процессора 2 ГГц и объемом оперативной памяти 0,97 ГБ.

Как видно из табл. 1, при числе узлов и каналов, достигающем сотен (СеМО-5), когда высокие затраты времени на имитацию существенно ограничивают возможность увеличения числа итераций, метод НГ позволяет достигать значительного сокращения времени E при числе итераций N, меньшем размерности n факторного пространства.

6. СПЕЦИАЛЬНЫЕ ИСПЫТАНИЯ МЕТОДА «НАПРАВЛЯЮЩИХ ГИПЕРБОЛ»

Для оценки качества приближенного решения $\vec{\mu}^*$ введем три показателя: эффект оптимизации $Q = (E_c - E^*)/E_c$, фазовую погрешность $\delta = |\vec{\mu}_{\text{opt}} - \vec{\mu}^*|/|\vec{\mu}_{\text{opt}}|$ и упущенный эффект $\varepsilon = (E^* - E_{\text{opt}})/E_{\text{opt}}$, где $E_c = E(\vec{\mu}_c)$, $E^* = E(\vec{\mu}^*)$, $E_{\text{opt}} = E(\vec{\mu}_{\text{opt}})$.

Сравнивая выражения (2) и (5), можно видеть, что детерминированная составляющая погрешности, обусловленная использованием сепарабельной аппроксимации (5) для несепарабельной функции (2), должна быть более заметной при оптимизации тех сетей, в которых изменение интенсивностей обслуживания μ_i в одних узлах сильнее влияет на среднее время ожидания w_j в других. Но большое влияние μ_i на w_j возникает тогда, когда заявки переходят из узла i непосредственно в узел j, и при этом к. в. распределения $B_j(t)$ значительно отличается от к. в. распределения $B_i(t)$ и (или) от к. в. интервалов поступления заявок в i-й узел [1, 9].

С учетом этих факторов для специальных испытаний точности метода НГ сконструирована

«неудобная» для него двухлинейная система (ДС) (рис. 5). Интервалы поступления заявок в ДС распределены равномерно на отрезке [0, 2]. Время обслуживания в узлах *1* и *4* детерминированное (к. в. $C_1 = C_4 = 0$). В узлах *2* и *3* время обслуживания разыгрывается как функция x стандартной сл. в. η: $x = d[(\eta + \varepsilon)^{-1} - (1 + \varepsilon)^{-1}]$. Здесь сл. в. η распределена равномерно на отрезке [0, 1], масштабный коэффициент d определяет требуемое среднее $M(x) = b$, а параметр $\varepsilon > 0$ позволяет формировать требуемый к. в. $C_x = C$ ($C \to \infty$ при $\varepsilon \to 0$). Если для узлов *2* и *3* взять $\varepsilon = 0{,}216$, $d = 1{,}106$, то получим в них $b = 1$ и $C = 1$ (вариант ДС1 — *ординарный* тест). При $\varepsilon = 0{,}019$, $d = 0{,}332$ имеем $b = 1$, $C = 2$ (вариант ДС2 — *усложненный* тест), а при $\varepsilon = 0{,}001$, $d = 0{,}169$ получаем $b = 1$, $C = 5{,}22$ (вариант ДС3 — *жесткий* тест). Все коэффициенты c_i и β_i взяты равными единице.

В табл. 2 представлены решения, полученные методом НГ (этап I) и последующим покоординатным спуском (этап II) для тестов ДС1, ДС2 и ДС3. На этапе II результат, полученный методом НГ, предварительно округлялся до сотых долей (с учетом равенства $\mu_1 + \mu_2 + \mu_3 + \mu_4 = M$) и затем уточнялся ускоренным методом покоординатного спуска при отклонениях свободных координат $\Delta\mu_i = \pm 0{,}01$ (N' — число получившихся шагов). Средние значения времени ответа E_{I} и E_{II}, полученные на этапах I и II, приведены со всеми точно установленными значащими цифрами. Известные интервалы между сравниваемыми пробными точками позволили оценить фазовую погрешность δ сверху. Как видно из табл. 2, погрешность δ метода НГ выходит за пределы 1 % лишь при весьма больших перепадах к. в. C_i. При $C_2 = C_3 \approx 5$ ($C_1 = C_4 = 0$) она может достигать 3 %. При этом упущенный эффект оптимизации ε не превышает половины процента. Значение $\varepsilon < 0$ в табл. 2 получено для случая, когда решение $\vec{\mu}^*$ на этапе II ухудшилось из-за округления.

Можно предположить, что в сетях с ветвящимися и сливающимися маршрутами точность метода НГ будет не хуже, поскольку рассеивание и смешивание потоков обусловливает тенденцию их

Таблица 1

Результаты испытаний алгоритма на тестовых сетях CeMO-2, ..., CeMO-5

Число узлов n, общее число каналов K	CeMO-2 $n = 9$, $K = 12$		CeMO-3 $n = 30$, $K = 134$		CeMO-4 $n = 20$, $K = 52$		CeMO-5 $n = 100$, $K = 260$	
Число итераций N × время итерации T_M	20 × 2		60 × 10		40 × 5		50 × 94	75 × 94
Распределяемый ресурс M	24	32	96	128	39	52	437	582
Отклик $\hat{E}_c$ в центре ОДР	13,98	8,043	49,92	27,88	78,08	38,19	278,1	163,3
Отклик $\hat{E}^* = \hat{E}(\vec{\mu}^*)$	10,83	6,225	41,53	22,97	55,12	29,94	249,6	143,8
Эффект оптимизации $(\hat{E}_c - \hat{E}^*)/\hat{E}_c \cdot 100$ %	22,5	22,6	17	18	29	22	10	12

Таблица 2

Результаты оптимизации тестовых ДС

Вариант теста	M	ρ_c	Этап	N; N'	Найденное приближение к точке оптимума				E_{I}; E_{II}	$\delta \cdot 100$ %	$\varepsilon \cdot 100$ %
					μ_1	μ_2	μ_3	μ_4			
Ординарный $C_2 = C_3 = 1$	4	0,5	I	48	1,105	0,946	0,890	1,059	2,925	0,6	0,00
			II	0	1,11	0,95	0,88	1,06			
Усложненный $C_2 = C_3 = 2$			I	62	1,198	0,913	0,772	1,117	4,320	0,8	0,02
			II	0	1,20	0,90	0,78	1,12	4,319		
Жесткий $C_2 = C_3 = 5{,}22$			I	122	1,282	0,837	0,646	1,235	12,921	3,0	0,4
			II	5	1,33	0,84	0,63	1,20	12,871		
Ординарный $C_2 = C_3 = 1$	2,5	0,8	I	4	0,654	0,603	0,599	0,644	9,790	1,0	0,2
			II	0	0,66	0,60	0,59	0,65	9,766		
Усложненный $C_2 = C_3 = 2$			I	29	0,692	0,596	0,557	0,655	19,972	0,6	−0,1
			II	0	0,69	0,60	0,56	0,65	19,987		
Жесткий $C_2 = C_3 = 5{,}22$			I	8	0,687	0,604	0,533	0,676	90,16	1,2	0,4
			II	1	0,70	0,60	0,53	0,67	89,76		

Таблица 3

Оценка точности метода НГ для СеМО-1

Этап	N; N'	Найденное приближение к точке оптимума									E_{I}; E_{II}	$\delta \cdot 100$ %	$\varepsilon \cdot 100$ %
		μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9			
I	33	0,818	0,597	1,405	3,025	9,490	1,272	1,682	2,965	2,219	7,514	0,4	0,01
II	5	0,79	0,60	1,43	3,02	9,49	1,28	1,69	2,96	2,22	7,513		

приближения к экспоненциальным, и, соответственно, форма $E(\vec{\mu})$должна в таких открытых сетях приближаться к сепарабельной. Это подтверждается данными о точности метода НГ в табл. 3, полученной при двухэтапной оптимизации СеМО-1.

7. ОПТИМИЗАЦИЯ ЗАМКНУТЫХ СЕТЕЙ

Оптимизация замкнутой сети методом НГ характеризуется более высокой детерминированной составляющей погрешности, так как функция $E(\vec{\mu})$ здесь «сильно» несепарабельна. Изменение интенсивности μ_i в любом узле замкнутой сети приводит к изменению интенсивностей λ_j входных потоков всех узлов j, $j = \overline{1, n}$, и, тем самым, непосредственно изменяет время w_j во всех узлах. Испытания метода НГ для замкнутой версии СеМО-1 показывают, что при линейном ограничении (3) значение δ может достигать 0,03 (в открытой версии СеМО-1 δ = 0,004). При нелинейном ограничении (3) (с разбросом коэффициентов β_i в пределах от 0,5 до 1,5) δ равно 0,06...0,08. Но упущенный эффект при этом остается небольшим ($\varepsilon < 0{,}01$). В этих испытаниях длина прогонов составляла около 1 млн. заявок, а число N итераций лежало в пределах нескольких десятков. Таким образом, и при оптимизации замкнутых сетей на практике часто можно ограничиваться применением только первого этапа оптимизации — метода НГ.

ЗАКЛЮЧЕНИЕ

Для оптимизации немарковских СеМО в общем случае приходится применять ИМ, которое характеризуется высокой сложностью вычисления градиентов, обусловленной стохастическим характером получаемых путем ИМ оценок. При решении задачи оптимального распределения ресурса проблема градиентов эффективно решается предлагаемым двухэтапным методом оптимизации, ядро которого составляет метод «направляющих гипербол». Он характеризуется относительно хорошей точностью и приемлемой вычислительной трудоемкостью, позволяющей рекомендовать его для практического применения при проектировании или модернизации сетей массового обслуживания, содержащих десятки и сотни узлов.

ЛИТЕРАТУРА

1. *Вишневский В.М.* Теоретические основы проектирования компьютерных сетей. — М.: Техносфера, 2003. — 512 с.
2. *Клейнрок Л.* Вычислительные системы с очередями. — М.: Мир, 1979. — 600 с.
3. *Вишневский В.М., Пороцкий С.М.* Моделирование ведомственных систем электронной почты // Автоматика и телемеханика. — 1996. — № 12. — С. 48—57.
4. *Жожикашвили В.А., Вишневский В.М.* Сети массового обслуживания. Теория и применение к сетям ЭВМ. — М.: Радио и связь, 1988. — 192 с.
5. *Феррари Д.* Оценка производительности вычислительных систем. — М.: Мир, 1981. — 576 с.
6. *Герасимов А.И.* Оптимизация замкнутых сетей массового обслуживания с несколькими классами сообщений // Проблемы передачи информации. — 1994. — Т. 30, № 1. — С. 85—96.
7. *Рыжиков Ю.И.* Имитационное моделирование. Теория и технологии. — СПб.: КОРОНА принт; М.: Альтекс-А, 2004. — 384 с.
8. *Клейнрок Л.* Теория массового обслуживания. — М.: Машиностроение, 1979. — 432 с.
9. *Gabriel R. Bitran, Reinaldo Morabito.* Open Queueing Networks: Optimization and Performance Evaluation Models for Discrete Manufacturing Systems / Сайт Массачусетского технологического института. — Режим доступа: http://dspace.mit.edu/bitstream/handle/1721.1/2537/SWP-3743-31904719.pdf?sequence=1.
10. *Клейнен Дж.* Статистические методы в имитационном моделировании. — М.: Статистика, 1978. — Вып. 1. — 221 с.
11. *Johnson M.E., Jackson J.* Infinitesimal Perturbation Analysis: a Tool for Simulation // J. of the Operational Res. Soc. — 1989. — Vol. 40, N 3. — P. 134—160.
12. *Rubinstein R.Y.* Sensitivity analysis of computer simulation models via the efficient score // Oper. Res. — 1989. — Vol. 37. — P. 72—81.
13. *Suri R, Zazanis M.* Perturbation Analysis Gives Strongly Consistent Sensitivity Estimates for the M|G|1 Queue // Mgmt Science. — 1988. — Vol. 34. — P. 39—64.
14. *Иглхарт Д.Л., Шедлер Д.С.* Регенеративное моделирование сетей массового обслуживания. — М.: Радио и связь, 1984. — 135 с.
15. *Базара М., Шетти К.* Нелинейное программирование. Теория и алгоритмы. — М.: Мир, 1982. — 583 с.
16. *Задорожный В.Н.* Методы двухуровневого моделирования систем с очередями // Тр. VII междунар. конф. «Идентификация систем и задачи управления» SICPRO'08. — Москва, 28—31 января 2008 / ИПУ РАН. — М., 2008. — С. 1484—1563.
17. *Карпов Ю.Г.* Имитационное моделирование систем. Введение в моделирование с AnyLogic-5. — СПб.: БХВ-Петербург, 2005. — 400 с.

Статья представлена к публикации членом редколлегии А.С. Манделем.

Задорожный Владимир Николаевич — канд. техн. наук, доцент, Омский государственный технический университет,
☎(3812) 65-20-84, ✉ zwn@yandex.ru.