ИСПОЛЬЗОВАНИЕ АДАПТИВНЫХ ПРИБЛИЖЕНИЙ В АЛГОРИТМАХ ПАРАМЕТРИЧЕСКОЙ ОПТИМИЗАЦИИ СЕТЕЙ С ОЧЕРЕДЯМИ

В. Н. Задорожный, Е. С. Ершов, О. Н. Канева (Омск)

Известно, что наличие статистической погрешности в численных результатах имитационного моделирования (ИМ) обусловливает значительные потери машинного времени при использовании пошаговых методов параметрической оптимизации (в частности, градиентных методов [1]). Эта же проблема возникает при численном решении уравнений или неравенств, задающих в факторном пространстве (ФП) $R_M \subset R^n$ модели $\vec{y}(x_1,...,x_n)$ границы искомых областей. Одним из возможных способов решения проблемы применения градиентов и в целом численных методов в ИМ является использование аналитической информации о свойствах зависимости $\vec{y}(x_1,...,x_n)$ в окрестностях «текущей» точки ФП, в частности, использование адаптивных аппроксимаций. Под адаптивной аппроксимацией мы понимаем упрощенное аналитическое выражение зависимости $\vec{y}(x_1,...,x_n)$, легко настраиваемое на приближенное описание $\vec{y}$ в нужном участке ФП. Пригодность такой аппроксимации зависит от согласованности ее свойств с общими свойствами модели $\vec{y}(x_1,...,x_n)$. Ее выбор может быть основан на эвристических соображениях, опыте или интуиции исследователя.

Постановка задачи

Эффективность применения в ИМ адаптивных аппроксимаций можно продемонстрировать решением следующей задачи оптимального распределения ресурсов по узлам «канонической» сети с очередями.

При наличии некоторого суммарного ресурса производительности

$$M = c_1\mu_1 + \ldots + c_n\mu_n = \text{const}, \qquad (1)$$

представленного, например, в стоимостном выражении, задача его оптимального распределения по узлам состоит в нахождении для узлов таких интенсивностей $\vec{\mu} = (\mu_1, \ldots, \mu_n)$, которые при заданных удельных расходах $\vec{c} = (c_1,\ldots,c_n)$ ресурса *M* на единицу производительности удовлетворяют условию (1) и минимизируют время ответа (среднее) *E*. Под временем ответа понимается время прохождения заявки через сеть.

«Канонической» будем называть сеть, обладающую следующими свойствами.

Входной поток, имеющий среднюю интенсивность поступления заявок Λ, рекуррентный. Время обслуживания заявок любым из K_i каналов *i*-го узла – это независимая случайная величина с одной и той же функцией распределения вероятностей $B_i(t)$. Поэтому интенсивности обслуживания μ_i у этих каналов также одинаковы.

После обслуживания в *i*-м узле заявка случайно и независимо в соответствии с заданными переходными вероятностями $p_{ij} > 0$ выбирает один из возможных узлов *j* для продолжения своего маршрута (вероятность p_{i0} соответствует выходу из сети, p_{0i} –попаданию из входного потока сети в *i*-й узел).

В канонической сети время ответа можно представить в виде суммы

$$E = \sum_{i=1}^{n} U_i = \sum_{i=1}^{n} \alpha_i w_i + \sum_{i=1}^{n} \alpha_i b_i = \sum_{i=1}^{n} \alpha_i w_i + \sum_{i=1}^{n} \alpha_i \cdot \frac{1}{\mu_i},$$

где U_i –вклад *i*-го узла в среднее время ответа, *c*; α_i – среднее число (частота) посещений *i*-го узла одной заявкой; w_i – среднее время ожидания заявки в очереди *i*-го узла, *c*; b_i – среднее время обслуживания заявок в *i*-м узле, *c*; $\mu_i = b_i^{-1}$ – интенсивность обслуживания заявки каналом *i*-го узла, c^{-1}; *n* – число узлов сети.

Частоты α_i посещения заявкой узлов однозначно определяются переходными вероятностями и могут быть найдены из системы уравнений баланса:

$$\alpha_i = \sum_{j=0}^{n} \alpha_j p_{ji}, \quad i = \overline{0,n}, \tag{2}$$

при $\alpha_0 \equiv 1$. Применяя (2) к тестовой сети, изображенной на рис. 1, находим:

$$\vec{\alpha} = (\alpha_1,...,\alpha_9) = (0.2,\ 0.3,\ 0.5,\ 1.3667,\ 5.9,\ 0.41,\ 0.54667,\ 5.31,\ 1).$$

Для всякого *i*-го узла интенсивность λ_i его входного потока заявок определяется равенством $\lambda_i = \Lambda \cdot \alpha_i$.

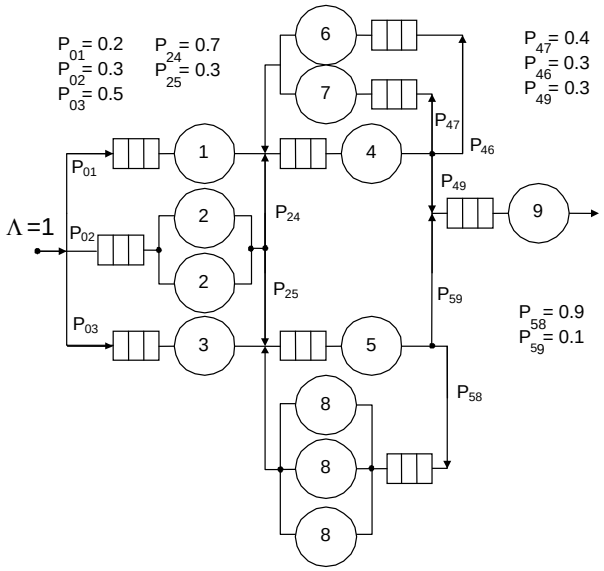

Рис. 1. Тестовый пример сети с очередями

С учетом сказанного задачу оптимального распределения ресурсов сети можно поставить в виде

$$E(\vec{\mu}) = \sum_{i=1}^{n} \alpha_i w_i(\vec{\mu}) + \sum_{i=1}^{n} \alpha_i \cdot \frac{1}{\mu_i} \to \min_{\vec{\mu}}; \tag{3}$$

$$\sum_{i=1}^{n} c_i \mu_i = M; \tag{4}$$

$$\mu_i K_i \geq \lambda_i, \ i = \overline{1,n}. \tag{5}$$

Здесь ограничения (5) задают для возможных распределений $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ область, в которой существует стационарный режим функционирования сети. Совокупность ограничений (3) и (4) задает область допустимых решений (ОДР) задачи (3)–(5).

Адаптивная аппроксимация поверхности отклика

Возможность использования градиентов при решении поставленной выше задачи можно обеспечить, применяя достаточно простую адаптивную аппроксимацию поверхности отклика $E(\mu_1, \ldots, \mu_n)$ в области текущих точек ФП, «просчитываемых» посредством ИМ. Предлагаемая ниже аппроксимация целевой функции $E(\vec{\mu})$ основана на следующих двух предположениях:

- время ожидания $w_i(\vec{\mu})$ в i-м узле при изменениях $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ определяется в основном интенсивностью обслуживания μ_i и слабо зависит от интенсивности обслуживания в других узлах;
- зависимость $w_i(\mu_i)$ на достаточно больших интервалах значений μ_i близка к гиперболической зависимости

$$w_i \approx \frac{R_i}{\mu_i - S_i} \tag{6}$$

с подобранными соответствующим образом константами R_i и S_i.

Исходя из вышесказанного, в качестве аппроксимации целевой функции $E(\vec{\mu})$ в задаче (3)–(5) можно использовать функцию:

$$E^{ap}(\vec{\mu}) = \sum_{i=1}^{n} \frac{\alpha_i R_i}{\mu_i - S_i} + \sum_{i=1}^{n} \frac{\alpha_i}{\mu_i}. \tag{7}$$

Настройка аппроксимации (7) осуществляется по результатам ИМ в двух точках $\vec{\mu}^{k-1} = (\mu_1^{k-1}, \ldots, \mu_n^{k-1})$ и $\vec{\mu}^{k} = (\mu_1^{k}, \ldots, \mu_n^{k})$, отличающихся всеми координатами. Верхним индексом буквы указывается порядковый номер очередной точки ФП, в которой на соответствующем шаге оптимизации выполняется ИМ. Получаемые с помощью ИМ значения времени ожидания w_i^{k-1} и w_i^{k} используются для определения коэффициентов R_i и S_i настраиваемой аппроксимации (6) i-го узла; $i = \overline{1,n}$. Нетрудно установить, что два коэффициента R_i и S_i гиперболы (6), проведенной через две точки (μ_i^{k-1}, w_i^{k-1}) и (μ_i^{k}, w_i^{k}), определяются следующим образом:

$$S_i = \frac{w_i^k \mu_i^k - w_i^{k-1} \mu_i^{k-1}}{w_i^k - w_i^{k-1}};$$

$$R_i = w_i^{k-1} \cdot (\mu_i^{k-1} - S_i), \; i = \overline{1,n}.$$

Подстановка всех найденных значений R_i и S_i в (7) дает «настроенную» аппроксимацию $E^{ap}(\vec{\mu})$ функции $E(\vec{\mu})$ (т. е. n-мерной поверхности отклика). На первом шаге оптимизации $E^{ap}(\vec{\mu})$ формируется по двум прогонам модели, а далее адаптируется к любой очередной точке всего лишь по одному прогону: на каждом новом шаге она формируется по данным ИМ в новой точке и по имеющимся данным в предыдущей точке.

Дифференцированием аппроксимации $E^{ap}(\vec{\mu})$ (7) нетрудно получить аппроксимацию градиента. С ее помощью на каждом шаге вычисляется (приближенно) градиент и его проекция на гиперплоскость (4). В уравнении (4) вектор $\vec{c} = (c_1, c_2, ..., c_n)$ является нормалью к этой гиперплоскости, и, следовательно, нормированный вектор нормали

$\vec{n}$, требуемый для построения проекций градиента, определяется в виде $\vec{n} = \vec{c} / |\vec{c}|$, где $|\vec{c}|$ – длина вектора $\vec{c}$.

Кроме того, аппроксимация $E^{ap}(\vec{\mu})$ позволяет находить оптимальную длину шага в направлении проекции антиградиента, определяя на этом направлении точку $\vec{\mu}$, в которой $E^{ap}(\vec{\mu})$ принимает минимальное значение. Такая точка $\vec{\mu}$ быстро устанавливается одномерным сканированием значений E^{ap} вдоль проекции антиградиента. Найденные координаты $\vec{\mu}$ передаются в имитационную модель, которая для этой точки $\vec{\mu}$ вычисляет отклик E и значения $\vec{w} = (w_1, \ldots, w_n)$ времени ожидания в узлах. Далее по этим значениям и соответствующим значениям из предыдущей точки ФП выполняется новая настройка (адаптация к последним двум точкам) аппроксимации (7), и градиентный поиск продолжается. Процесс завершается при достижении заданных условий останова. Адаптация аппроксимации к очередной точке ФП и определение очередного оптимального шага каждый раз выполняются на уровне аналитики, причем в тысячи раз быстрее, чем выполняется прогон имитационной модели в этой точке ФП.

Точное описание алгоритма оптимизации приводится в [2].

Результаты испытаний метода

Возможности метода продемонстрируем на примере оптимизации распределения ресурсов для тестовой сети, изображенной на рис. 1. Суммарный ресурс $M = 30$ будем распределять при векторе удельных расходов $\vec{c} = (c_1, \ldots, c_n) = (K_1, \ldots, K_9) = (1,2,1,1,1,1,1,3,1)$, т. е. в каждой СМО удельный расход ресурса равен числу ее каналов, и удельный расход ресурса на один канал везде одинаковый. Типы распределений $B_i(t)$ для $i = \overline{0,n}$, соответственно, следующие: (M,R,R,R,R,M,M,E^2,E^2,E^2,R), где M – экспоненциальное распределение, R – равномерное (на интервале от 0 до двух средних), E^2 – эрланговское распределение второго порядка, и $B_0(t)$ соответствует входному потоку.

Алгоритм оптимизации реализован следующим образом. Операции на уровне аналитики осуществляются внешней программой, которая автоматически запускает GPSS World с имитационной моделью сети и передает ей в виде текстового файла координаты очередной точки ФП. После ИМ программа принимает его результаты (также через текстовый файл) и вычисляет координаты следующей точки ФП для их передачи в GPSS (либо завершает поиск).

С помощью этой же внешней программы поиск оптимума проводился путем расчета градиентов стандартным (базовым) методом малых приращений, с целью получения сравнительной оценки эффективности для метода адаптивной аппроксимации. Длина прогона имитационной модели во всех точках ФП определялась прохождением через сеть 1 млн. заявок.

Испытания метода проводились многократно при различном выборе двух стартовых точек и условий останова. На рис. 2 показаны соответствующие изменения целевой функции в зависимости от номера шага оптимизации. Первое приближение к оптимуму происходит за 8–10 итераций, т. е. число шагов примерно совпадает с размерностью ФП. Вблизи точки оптимума длина шага неизбежно становится малой, это приводит к большим стохастическим ошибкам аппроксимации и, рано или поздно, – к отбрасыванию на большое расстояние от точки оптимума. После этого за 8–10 итераций вновь происходит приближение к этой точке.

Типичная траектория значений целевой функции при использовании базового метода, не предусматривающего аппроксимации поверхности отклика, приведена на

рис. 2 справа. Здесь значения целевой функции стабильно отделены интервалом около 2 единиц от минимума $E \approx 7.49$, достигнутого предлагаемым методом, что характеризует меньшую точность базового метода. Предлагаемый градиентный метод имеет преимущество также и в скорости прохождения точек ФП, так как базовый метод приращений требует выполнения в каждой точке не одного, а $n+1$ прогонов имитационной модели. В рассматриваемом примере это увеличивает для него время одной итерации вдесятеро.

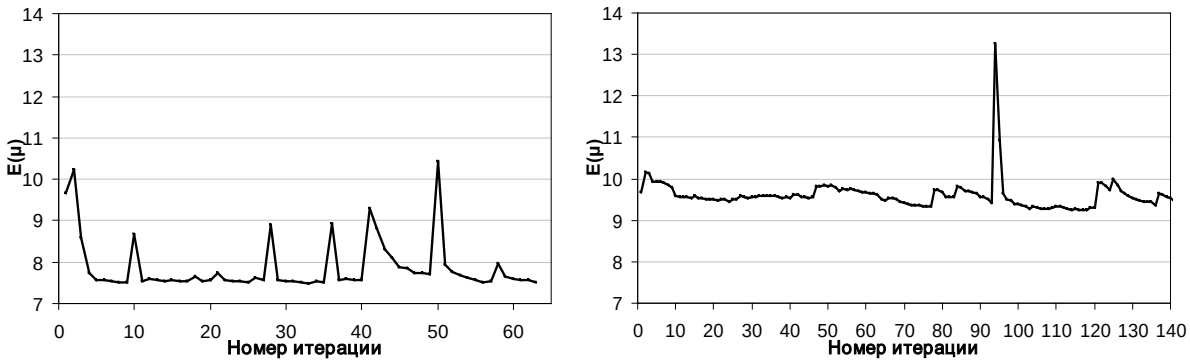

Рис. 2. Изменения *E* в ходе оптимизации: слева – при использовании адаптивной аппроксимации, справа – при использовании базового метода

Таким образом, выполненные испытания демонстрируют существенные преимущества предлагаемого метода градиентной оптимизации. Кроме этого «формального» результата испытаний отметим, что точка «справедливого» распределения ресурса, обеспечивающего равные коэффициенты загрузки узлов (она использовалась в качестве стартовой точки), не является оптимальной. В решенной задаче при оптимальном распределении ресурса коэффициенты загрузки разных узлов были разбросаны в диапазоне от 0,2 до 0,6.

Выводы

Предложенный градиентный метод оптимального распределения ресурсов сети, который основан на простой адаптивной аппроксимации поверхности отклика, обладает достаточно высокой точностью и быстродействием. При решении задачи оптимального распределения ресурсов число итераций, требующих одного обращения к имитационной модели, приблизительно равно размерности ФП, т. е. числу узлов сети, что близко к пределу эффективности методов нелинейной оптимизации [3].

Возможности этого метода позволяют применять его на практике для оптимизации сетей массового обслуживания, содержащих десятки и сотни узлов, например, для оптимизации компьютерных сетей, структур обслуживающих предприятий или транспортных сетей городов.

Литература

1. **Рыжиков Ю. И.** Имитационное моделирование. Теория и технологии. СПб.: КОРОНА принт. М.: Альтекс-А, 2004. 384 с.
2. **Задорожный В. Н., Ершов Е. С., Канева О. Н.** Двухуровневые градиентные методы для оптимизации сетей с очередями//Омский научный вестник. 2006. № 7(43). С. 123–131.
3. **Базара М., Шетти К.** Нелинейное программирование. Теория и алгоритмы/Пер. с англ. М.: Мир, 1982. 583 с.