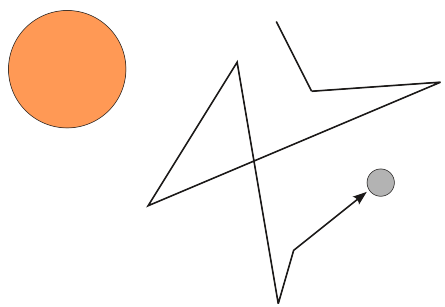
Статистические характеристики случайных процессов в присутствии механизма эпизодического перезапуска

Эпизодическое прерывание процесса случайного блуждания (например движения частицы, помещенной в жидкость, в которой поддерживается течение со случайной компонентой скорости) в окрестности поглощающей области с последующем возобновлением новой статистически независимой реализации того же процесса может существенно повлиять на среднее время, требуемое для достижения заданной области. При этом оказывается возможным решить задачу в общем случае - то есть определить среднее время первого достижения для произвольного процесса и при произвольной частоте событий перезапуска, а также вывести некоторые утверждения справедливые не только для периодического, но произвольного протокола перезапуска. Будут рассмотрены известные результаты касательно задачи оптимизации средней длительности случайных процессов методом перезапуска, а также оптимизации некоторых других статистических характеристик.



Пример 2: временные затраты на выполнение алгоритма

Время выполнения алгоритма определяется числом элементарных действий необходимым для получения конечного результата.

Время выполнения может существенно зависеть от реализации данных, подаваемых на вход алгоритма. Средняя сложность определена как математическое ожидание $\langle T \rangle$ времени выполнения алгоритма, вычисленное в предположении, что все элементы множества входных данных равновероятны.

Эта величина определяет эффективность работы алгоритма: если на вход алгоритма подается равнораспределенный поток данных, то за большое время t алгоритм выполнит $\approx t/\langle T \rangle$ задач (предполагаем, что $\langle T \rangle$ конечно).

На практике поток данных, поступающих на вход алгоритма, может характеризоваться неравномерным распределением вероятности на множестве возможных входов. Для алгоритмов с сильно различающимся временем обработки среднего и худшего случаев это может приводить к существенному снижению эффективности в ситуации, когда типичные реализации входа близки к худшему сценарию.

Введение элемента случайности в алгоритм позволяет обойти вышеуказанную проблему, делая среднее время выполнения нечувствительным к статистическим особенностям входных данных.

Простой пример: алгоритм линейного поиска в неупорядоченном массиве.

N - длина одномерного массива (списка) неупорядоченных элементов; a - элемент, который требуется найти (вход); n_a - позиция запрашиваемого элемента a в списке (выход);

Алгоритм: начиная с первого элемента в списке, движемся от ячейки к ячейке, до тех пор, пока либо не найдём то, что ищем.

$T = n_a$ - время выполнения алгоритма (число сравнений, требуемых чтобы найти элемент списка равный запрашиваемому).

Среднее время поиска в предположении, что поток входных данных принадлежит равнораспределенному статистическому ансамблю, т. е. $p_0(n_a) = \frac{1}{N}\sum_{n=1}^{N}\delta(n_a - n)$, равно $\langle T\rangle = \frac{N+1}{2}$. Между тем, в худшем случае $p_{worst}(n_a) = \delta(n_a - N)$ имеем $\langle T\rangle = N$.

Рандомизируем алгоритм. А именно, каждый раз перед тем как начать новый поиск будем подбрасывать симметричную монетку, инициализируя затем поцедуру поиска, начиная либо с первого элемента списка, либо с последнего, в зависимости от того орел выпал или решка. Случайное время выполнения рандомизированного алгоритма можно записать как

$$T = n_a I(q = 1) + (N - n_a + 1)I[q = 0], \tag{1}$$

где q - бинарная случайная величина равная 0 или 1 с равными вероятностями, $I(\ldots)$ - индикаторная переменная. Усредним по статистике случайных величин n_a и q

$$\langle T\rangle = \frac{1}{2}\langle n_a\rangle + \frac{1}{2}(N - \langle n_a\rangle + 1) = \frac{N+1}{2}. \tag{2}$$

Данный ответ справедлив для любой статистики входных данных.

Метод перезапуска

Да, рандомизация алгоритма позволяет избавиться от чувствительности его средней производительности к статистике ансамбля входных данный. Между тем, понятно, что всегда присутствует некоторая ненулевая вероятность встретить неудачное сочетание в паре "конкретная реализация вероятностного алгоритма + конкретные входные данные" , для которого время выполнения будет близко к худшему сценарию.

Повысить среднюю производительность вероятностного алгоритма, характеризующегося относительно большими флуктуациями случайного времени выполнения, помогает перезапуск, то есть эпизодическое прерывание алгоритма с последующим запуском новой статистически независимой реализации.

Идея проста: если текущая реализация алгоритма слишком затянулась, то можно ожидать, что реализовалось нежелательное сочетание входных данных и параметров алгоритма, и лучше попробовать перезапустить счет с новыми случайными параметрами, чем ждать окончания текущей попытки.

Модель: стохастический процесс с перезапуском

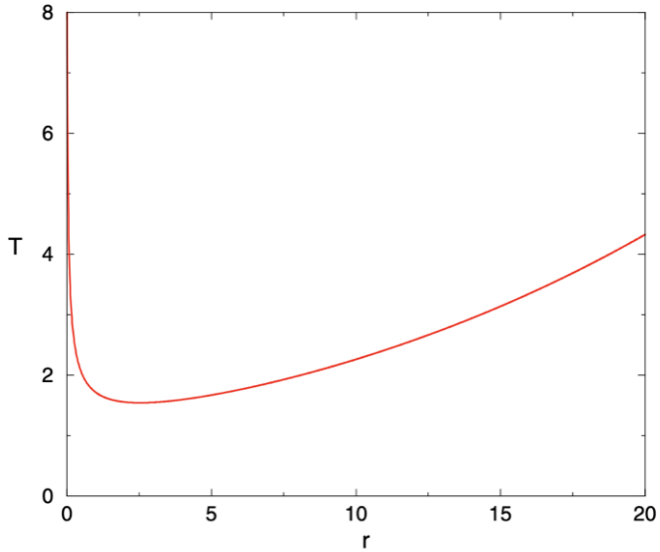
T - случайная длительность стохастического процесса;
$P(T)$ - плотность распределения случайной величины T;
$\mathcal{R} = \tau_1, \tau_2, \ldots$ - протокол перезапуска, заданный набором интервалов времени между последовательными событиями перезапуска.

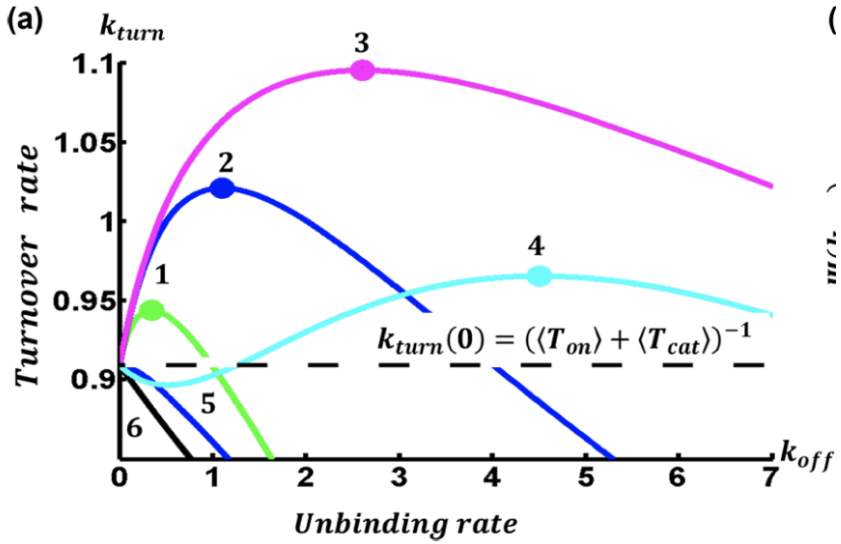
Если процесс успел завершиться не позднее времени τ_1, то на этом история заканчивается. В противном случае, текущая попытка прерывается и инициализируется новая статистически независимая реализация процесса. Далее, если эта реализация в свою очередь не завершается за время τ_2, то она также прерывается и запускается третья попытка. И так далее, пока процесс наконец не завершится.

Будем обозначать через $T_{\mathcal{R}}$ случайную длительность процесса в присутствии протолока перезапуска $\mathcal{R}$. Протокол считается эффективным, если $\langle T_{\mathcal{R}} \rangle < \langle T \rangle$.

Примеры

- $P(T) = \frac{L}{2\sqrt{\pi D} T^{3/2}} \exp\left(-\frac{L^2}{4DT}\right)$ (задача о времени достижения заданной точки одномерным диффузионным процессом).
- Сумма двух экспонент $P(T) = p\gamma_1 e^{-\gamma_1 T} + (1-p)\gamma_2 e^{-\gamma_2 T}$





Результаты

- *Общее решение задачи*
  Среднее время выполнения для произвольного расписания бесконечной последовательности $\mathcal{R} = \tau_1, \tau_2, \ldots$ перезапусков:

$$\langle T_{\mathcal{R}} \rangle = \sum_{k=1}^{\infty} \left(\frac{\int_0^{\tau_k} dT P(T) T}{\int_{\tau_k}^{\infty} dT P(T)} + \tau_k \right) \prod_{i=1}^{k} \int_{\tau_i}^{\infty} dT P(T) \tag{3}$$

  Частный случай строго периодического перезапуска: $\langle T_\tau \rangle = \frac{\int_0^\tau P(T) T dT + \tau \int_\tau^\infty P(T) dT}{\int_0^\tau P(T) dT}$

- *Для каких процессов перезапуск эффективен?*
  Простые достаточные условия существования эффективного протокола перезапуска:
  – степенной хвост плотности распределения времени завершения $P(T)$;
  – $\frac{\sigma(T)}{\langle T \rangle} > 1$, где $\sigma(T)$ - среднеквадратическое отклонение случайной длительности процесса в отсутствии перезапуска;
  – ...

- *Какой протокол перезапуска наиболее эффективен?*
  Оптимальным всегда является строго периодический перезапуск: если найден период перезапуска $\tau_* \geq 0$ (возможно $\tau_* = +\infty$) такой что для любого $\tau \geq 0$ справедливо $\langle T_{\tau_*} \rangle \leq \langle T_\tau \rangle$, то $\langle T_{\tau_*} \rangle \leq \langle T_{\mathcal{R}} \rangle$ для любого протокола $\mathcal{R}$.

- *Общие свойства оптимально перезапускаемых процессов*
  Для любого стохастического процесса справедливо неравенство

$$\frac{\sigma(T_{\tau_*})}{\langle T_{\tau_*} \rangle} \leq 1, \tag{4}$$

  где τ_* - оптимальный период строго периодического перезапуска.

Усреднение рекурсивного уравнения

Справедливо следующее рекурсивное уравнение

$$\boxed{T_\tau = T \cdot I(T < \tau) + (\tau + T'_\tau) \cdot I(T \geq \tau)} \tag{5}$$

где T'_τ - статистически независимая копия T_τ, а $I(\ldots)$ - индикаторная переменная. Усредняя, находим

$$\langle T_\tau \rangle = \langle T \cdot I(T < \tau) \rangle + \tau \langle I(T \geq \tau) \rangle + \langle T'_\tau \cdot I(T \geq \tau) \rangle, \tag{6}$$

Далее, так как $\langle T'_\tau \cdot I(T \geq \tau) \rangle = \langle T'_\tau \rangle \cdot \langle I(T \geq \tau) \rangle$ и $\langle T'_\tau \rangle = \langle T_\tau \rangle$, то

$$\langle T_\tau \rangle = \langle T \cdot I(T < \tau) \rangle + \tau \langle I(T \geq \tau) \rangle + \langle T_\tau \rangle \cdot \langle I(T \geq \tau) \rangle, \tag{7}$$

и, следовательно,

$$\langle T_\tau \rangle = \frac{\langle T \cdot I(T < \tau) \rangle + \tau \langle I(T \geq \tau) \rangle}{1 - \langle I(T \geq \tau) \rangle} = \frac{\int_0^\tau P(T) T dT + \tau \int_\tau^\infty P(T) dT}{\int_0^\tau P(T) dT}. \tag{8}$$

В последней строчке мы использовали соотношения $\langle T I(T < \tau) \rangle = \int_0^\tau P(T) T dT$, $\langle I(T \geq \tau) \rangle = \int_\tau^\infty P(T) dT$, $1 - \langle I(T \geq \tau) \rangle = \langle I(T < \tau) \rangle = \int_0^\tau P(T) dT$.

Поиск границ эффективности перезапуска

Хотя оптимизация с помощью перезапуска широко используется в практике компьютерного программирования и представляет собой активную область академических исследований в статистической физике, пределы эффективности этого инструмента оптимизации до недавних пор оставались неизвестными.

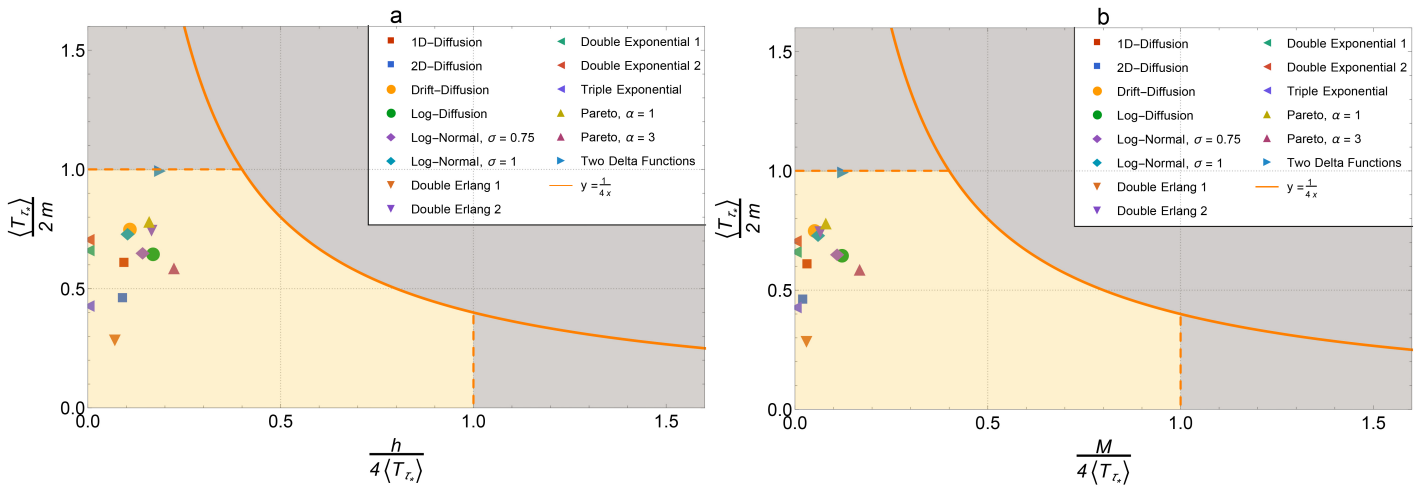
Пределы эффективности задаются вероятностными неравенствами вида

$$\langle T_{\mathcal{R}} \rangle \geq C_1 \cdot \mathcal{T}_1, \tag{9}$$

$$\langle T_{\tau_*} \rangle \leq C_2 \cdot \mathcal{T}_2, \tag{10}$$

где $C_1 > 0$ и $C_2 > 0$ - универсальные безразмерные константы, не зависящие от протокола перезапуска и статистики перезупускаемого процесса, $\mathcal{T}_1$ и $\mathcal{T}_2$ - некоторые масштабы времени, которые выражаются через статистические характеристики "невозмущенного" процесса (например, через математическое ожидание $\langle T \rangle$, дисперсию $\sigma(T)$, медианное значение m, вершину M и т.д.).

Результаты



Для любого стохастического процесса и любого протокола перезапуска справедливо:

$$\langle T_{\mathcal{R}} \rangle \geq \frac{1}{4} h, \tag{11}$$

где $h = (\int_0^\infty dT T^{-1} P(T))^{-1}$ - среднее гармоническое время завершения исходного процесса.
Если дополнительно предположить, что $P(T)$ это гладкая функция с единственным максимумом, то для любого протокола перезапуска будет справедливо

$$\langle T_{\mathcal{R}} \rangle \geq \frac{1}{4} M, \tag{12}$$

где $M = \text{argmax}_T P(T)$ - вершина плотности распределения $P(T)$, то есть значение времени завершения T, встречающееся наиболее часто.
Среднее время завершения произвольного стохастического процесса, который перезапускается с оптимально подобранным периодом τ_*, удовлетворяет неравенству

$$\langle T_{\tau_*} \rangle \leq 2m, \tag{13}$$

где m - медианное время завершения невозмущенного процесса ($\int_0^m P(T) dT = 1/2$).

Побочный продукт: новое достаточное условие эффективности перезапуска

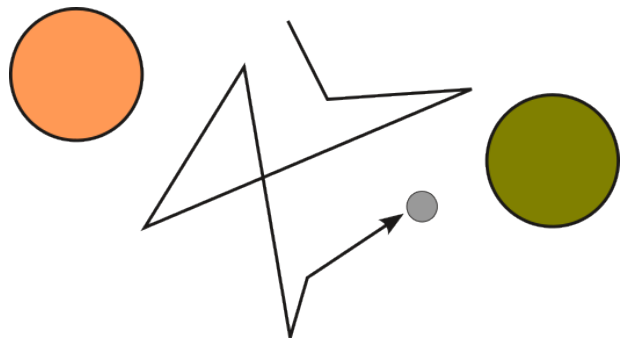
Опираясь на полученную границу эффективности оптимального протокола, можно сформулировать новое достаточное условие эффективности перезапуска. Так как $\langle T_{\tau_*} \rangle \leq 2m$, то неравенство

$$\langle T \rangle > 2m, \tag{14}$$

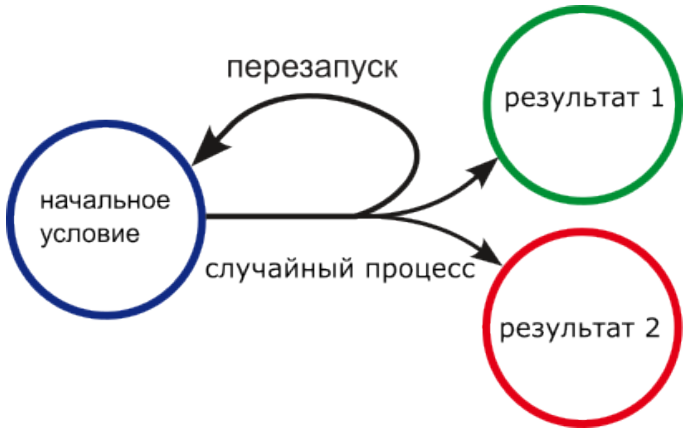
означает, что $\tau_* \neq \infty$ и, следовательно, найдется некоторый конечный период перезапуска, уменьшающий среднее время завершения процесса.

Процессы с несколькими исходами

Стохастический процесс может иметь несколько альтернативных сценариев завершения. Примеры:



Модель: эксперимент с двумя вариантами результата с перезапуском



Ингредиенты модели:

- Случайный процесс
  T - случайное время завершения
  $P(T) = P^s(T) + P^f(T)$ - плотность распределения
  $P^s(T)$ - вклад удачных попыток
  $P^f(T)$ - вклад неудачных попыток
  $p = \int_0^\infty P^s(T)dT$ - вероятность успеха
  $1 - p = \int_0^\infty P^f(T)dT$ - вероятность неудачи
- Протокол перезапуска
  $\mathcal{R} = \tau_1, \tau_2, \ldots$ - набор интервалов времени между последовательными событиями перезапуска

Будем обозначать через $p_\mathcal{R}$ вероятность успеха в присутствии протолока перезапуска $\mathcal{R}$. Протокол считается эффективным, если $p_\mathcal{R} > p$.

Результаты

- *Достаточные условия эффективности перезапуска*
  Простое достаточное условие существования эффективного протокола перезапуска:

$$\langle T^s \rangle < \langle T^f \rangle$$

  где $\langle T^s \rangle$ - среднее время завершение удачных реализаций, $\langle T^f \rangle$ - среднее время завершение неудачных реализаций

- *Наиболее эффективный протокол перезапуска*
  Оптимальным всегда является строго периодический перезапуск: если найден период перезапуска $\tau_* \geq 0$ (возможно $\tau_* = +\infty$) такой что для любого $\tau \geq 0$ справедливо $p_{\tau_*} \geq p_\tau$, то $p_{\tau_*} \geq p_{\mathcal{R}}$ для любого протокола $\mathcal{R}$

- *Общие свойства в статистике оптимально перезапускаемых процессов*
  Для любого стохастического процесса с двумя альтернативными исходами справедливо неравенство

$$\langle T^s_{\tau_*} \rangle \geq \langle T^f_{\tau_*} \rangle$$

  где τ_* - период строго периодического перезапуска, оптимизирующий вероятность успеха

- *Как выбрать эффективный период перезапуска*