

# Оценка параметров модели случайного процесса нагрузки сервера, обрабатывающего заявки

Владимир Руцкий, 5057/12

Санкт-Петербургский государственный политехнический университет

14 июня 2011

# План презентации

- 1 Постановка задачи
- 2 Решение в случае бесконечного времени обработки заявки
  - Итеративный метод
  - Оценивание ЕМ-алгоритмом
- 3 Решение в случае конечного времени обработки заявки
  - Метод наименьших квадратов
- 4 Результаты работы

## Сервер

- Сервер обрабатывает поступающие заявки
- Для обработки заявки используются ресурсы сервера
  - Количество используемых ресурсов сервера — загрузка сервера — скалярная величина, например процессорное время

## Задача

Дан лог загрузки сервера. Необходимо:

- 1 идентифицировать моменты поступления заявок,
- 2 оценить:
  - интенсивность поступления заявок,
  - загрузку сервера в фоновом режиме,
  - использование ресурсов сервера для обработки одной заявки

# Математическая модель (1)

Загрузка сервера — случайный процесс  $X(t): \mathbb{R} \rightarrow \mathbb{R}$

Фоновая загрузка сервера

Сумма постоянной загрузки и винеровского процесса (шум):

$$B(t) = m + \sigma \mathcal{W}(t)$$

Загрузка сервера при обработке заявки

Заявка, поступившая в  $t = t_c$ , увеличивает загрузку сервера на:

$$K_{t_c}(t) = \mathcal{N}(m_c, \sigma_c^2) \cdot I(t - t_c) \cdot e^{-\lambda_c(t-t_c)}$$

$$I(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \quad \text{— Функция Хевисайда}$$

## Математическая модель (2)

$T_c = \{t_{c1}, \dots, t_{cR}\}$  — множество моментов времени, когда поступили заявки (всего  $R$  заявок за время наблюдения за сервером)

### Интенсивность поступления заявок

Интервал времени между поступлением двух последовательных заявок распределён экспоненциально:

$$(t_{c_i} - t_{c_{i-1}}) \sim \text{Exp}(\lambda)$$

### Общая загрузка сервера

$$X(t) = B(t) + \sum_{t_c \in T_c} K_{t_c}(t)$$

### Перейдём к дискретному случайному процессу

$$X(t_i), \quad t_i = t_0 + i \cdot \Delta t \quad (t_0 \text{ и } \Delta t \text{ даны})$$

## Лог загрузки сервера

Траектория  $X(t_i)$ :  $\{x_i \mid i = 1, \dots, N\}$

Считаем, что  $\Delta t$  достаточно мало и  $t_{c_j} = t_0 + i \cdot \Delta t$  — заявки поступили в некоторые наблюдаемые моменты времени  $t_i$ .

Рассматривается случай  $t_0 = 0$

## Задача

По траектории  $X(t_i)$

- ❶ идентифицировать моменты поступления заявок,
- ❷ оценить параметры модели:
  - $m, \sigma$  — параметры фоновой загрузки,
  - $\lambda$  — интенсивность поступления заявок,
  - $m_c, \sigma_c, \lambda_c$  — параметры загрузки сервера при обработке заявок

# Случай бесконечного времени обработки заявки

$$\lambda_c \approx 0$$

При поступлении заявки в момент времени  $t_c$  загрузка сервера увеличивается на  $\Delta x \sim \mathcal{N}(m_c, \sigma_c)$

$$K_{t_c}(t) = \mathcal{N}(m_c, \sigma_c^2) \cdot I(t - t_c)$$



# Разностный аналог производной

$$dX(t)$$

Рассмотрим ненормированный разностный аналог производной:

$$dX(t) = X(t) - X(t - \Delta t)$$

$dX(t)$  в момент времени поступления заявки  $t_c$

$$dX(t_c) = \mathcal{N}(m_c, \sigma^2 \Delta t + \sigma_c^2)$$

(при условии, что в момент времени  $(t_c - \Delta t)$  заявки не было)

$dX(t)$  в момент времени отсутствия заявок  $t$

$$dX(t) = \mathcal{N}(0, \sigma^2 \Delta t)$$

(при условии, что в момент времени  $(t - \Delta t)$  заявки не было)

# Итеративный метод идентификации заявок (1)

Предположим, что в отрезке времени  $[t_k, t_{k+n}]$  не поступило ни одной заявки.

Тогда  $dx_k, \dots, dx_{k+n}$  — наблюдения  $\mathcal{N}(0, \sigma^2 \Delta t) = dX(t)$ .

Оценим  $\sigma^2$  по  $[x_k, x_{k+n}]$ :

$$\hat{\sigma}^2 \Delta t = \frac{1}{n-1} \sum_{i=1}^n ((x_{k+i} - x_{k+i-1}) - 0)^2$$

## Гипотеза $H_0$

В отрезке времени  $[t_{k+n}, t_{k+n} + \Delta t]$  не поступило ни одной заявки

## Критерий принятия $H_0$ с уровнем значимости $\alpha$

Разность значений наблюдений  $(x_{k+n+1} - x_{k+n})$  лежит в  $(1 - \alpha)$  квантиле нормального распределения  $\mathcal{N}(0, \hat{\sigma}^2 \Delta t)$ :

$$H_0 \text{ принимается} \iff (x_{k+n+1} - x_{k+n}) < \mathcal{N}_{1-\alpha}$$

## Итеративный метод идентификации заявок (2)

### Алгоритм итеративной идентификации первой заявки

FIND-REQUEST-INDEX ( $D = \{dx_{i+\text{offset}}\}, n, \alpha$ )

```
1  for  $k \leftarrow n$  to  $\text{length}[D] - 1$ 
2      do  $H_0\text{-criterion} \leftarrow \text{BUILD-H0-CRITERION}(D[1..k], \alpha)$ 
3      if not  $H_0\text{-criterion}(D[k+1])$ 
4          then  $\triangleright (k+1)\text{-й выброс отвергает } H_0 \implies \text{заявка}$ 
5              return  $k+1$ 
6  return 0  $\triangleright$  Заявка не обнаружена
```

# Итеративный метод идентификации заявок (3)

## Алгоритм итеративной идентификации всех заявок

FIND-REQUESTS-ITERATIVE ( $D = \{dx_i\}, n, \alpha$ )

```
1   $\widehat{T}_c \leftarrow \emptyset$ 
2   $idx \leftarrow 0$   $\triangleright$  Индекс последней идентифицированной заявки
3  while TRUE
4      do  $\triangleright$  Находим индекс следующей заявки
5           $idx \leftarrow \text{FIND-REQUEST-INDEX} \left( D[idx+1..length[D]], n, \alpha \right)$ 
6          if  $idx \neq 0$ 
7              then  $\widehat{T}_c \leftarrow \widehat{T}_c \cup \{t_{idx}\}$ 
8              else return  $\widehat{T}_c$ 
```

## Интенсивность поступления заявок

$$\hat{\lambda} = \frac{1}{|\widehat{T_c}|} \sum_{i=1}^{|\widehat{T_c}|} (t_{c_{i+1}} - t_{c_i})$$

## Параметры фоновой загрузки

$$\hat{m} = \frac{1}{K-1} \sum_{i=1}^{K-1} x_i, \quad \hat{\sigma}^2 = \frac{1}{K-2} \sum_{i=1}^{K-1} (x_i - \hat{m})^2, \quad K = t_{c_1}/\Delta t$$

## Параметры загрузки при обработке заявок

$$\widehat{m_c} = \frac{1}{|\mathcal{T}_c|} \sum_{t_c \in \mathcal{T}_c} dx_{t_c/\Delta t}, \quad \widehat{\sigma}^2 \Delta t + \widehat{\sigma_c}^2 = \frac{1}{|\mathcal{T}_c| - 1} \sum_{t_c \in \mathcal{T}_c} \left( dx_{t_c/\Delta t} - \widehat{m_c} \right)^2$$

$Z_i$

Введём скрытые случайные величины  $Z_i$ ,  $i = 1, \dots, N$

- $Z_i = 1$ , если в момент времени  $t_i$  поступила заявка,
- $Z_i = 2$ , если в момент времени  $t_i$  не поступила заявка

Количество поступивших за время наблюдения заявок — случайная величина, распределённая по закону Пуассона  $\mathcal{P}(\lambda)$ .

$\mathcal{P}(\lambda)$  можно аппроксимировать Биномиальным законом  $\mathcal{B}(N, \tau)$ .

$\implies Z_i$  можно считать распределённой по закону Бернулли.

$\mathbf{P}(Z_i = 1) = \tau_1$ ,  $\mathbf{P}(Z_i = 2) = \tau_2 = 1 - \tau_1$

$$\begin{aligned} dX(t_i)|(Z_i = 1) &\sim \mathcal{N}(\mu_1, \sigma_1^2) = \mathcal{N}(m_c, \sigma^2 \Delta t + \sigma_c^2), & (t_i \in T_c), \\ dX(t_i)|(Z_i = 2) &\sim \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(0, \sigma^2 \Delta t), & (t_i \notin T_c). \end{aligned}$$

Введём обозначения:  $\theta = (\tau_1, \tau_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ ,  $d\mathbf{x} = (dx_1, \dots, dx_N)$ ,  
 $\mathbf{z} = (z_1, \dots, z_N)$

## Функция правдоподобия

$$L(\theta; \mathbf{dx}, \mathbf{z}) = \mathbf{P}(\mathbf{dx}, \mathbf{z} | \theta) = \prod_{i=1}^N \sum_{j=1}^2 \mathbb{I}(z_i = j) \tau_j f(\mathbf{dx}_i, \mu_j, \sigma_j^2),$$

$$\mathbb{I}(\text{expr}) = \begin{cases} 0, & \text{expr} = \text{False} \\ 1, & \text{expr} = \text{True} \end{cases} \quad \text{— функция индикатор}$$

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{— плотность нормального распределения}$$

В экспоненциальной форме:

$$L(\theta; \mathbf{dx}, \mathbf{z}) = \exp \left\{ \sum_{i=1}^N \sum_{j=1}^2 \mathbb{I}(z_i = j) \left[ \log \tau_j - \frac{1}{2} \log(2\pi) - \log(\sigma_j) - \frac{(\mathbf{dx}_i - \mu_j)^2}{2\sigma_j^2} \right] \right\}$$

## ЕМ-алгоритм

EM-ALGORITHM()

```
1   $\theta \leftarrow \text{ESTIMATE-INITIAL-PARAMETERS}()$ 
2  while TRUE
3      do  $\theta_{\text{next}} \leftarrow \text{M-STEP}(\text{E-STEP}())$ 
4          if  $|\theta - \theta_{\text{next}}| < \varepsilon$ 
5              then return  $\theta$ 
6          else  $\theta \leftarrow \theta_{\text{next}}$ 
```



$\tau_j$

$$\tau_j^{(0)} = 0.5, \quad j = 1, 2$$

$\mu_j$

Построим полигон частот  $dx_i$ . В качестве  $\mu_1^{(0)}$  возьмём последний локальный максимум частот, а в качестве  $\mu_2^{(0)}$  — первый

$\sigma_j$

$$\sigma_j^{(0)} = \frac{1}{3} \left( \mu_1^{(0)} - \mu_2^{(0)} \right), \quad j = 1, 2$$

## Условная вероятность принадлежности $i$ -го наблюдения

По Т. Байеса:

$$\begin{aligned} T_{j,i}^{(k)} &= \mathbf{P} \left( Z_i = j \mid d\mathbf{X}(t_i) = d\mathbf{x}_i; \theta^{(k)} \right) = \\ &= \frac{\tau_j^{(k)} \cdot f \left( d\mathbf{x}_i; \mu_j^{(k)}, \sigma_j^{(k)} \right)}{\tau_1^{(k)} \cdot f \left( d\mathbf{x}_i; \mu_1^{(k)}, \sigma_1^{(k)} \right) + \tau_2^{(k)} \cdot f \left( d\mathbf{x}_i; \mu_2^{(k)}, \sigma_2^{(k)} \right)} \end{aligned}$$

## Математическое ожидание логарифма функции правдоподобия

$$\begin{aligned} Q \left( \theta \mid \theta^{(k)} \right) &= \mathbf{E} \left[ \log L(\theta; d\mathbf{x}, \mathbf{z}) \right] = \\ &= \sum_{i=1}^N \sum_{j=1}^2 T_{j,i}^{(k)} \left[ \log \tau_j - \frac{1}{2} \log(2\pi) - \log(\sigma_j) - \frac{(d\mathbf{x}_i - \mu_j)^2}{2\sigma_j^2} \right] \end{aligned}$$

## Максимизация $Q(\theta|\theta^{(k)})$

$$\theta^{(k+1)} = \operatorname{argmax}_{\theta} Q(\theta|\theta^{(k)}).$$

$$\tau_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^N T_{j,i}^{(k)}, \quad \mu_j^{(k+1)} = \frac{\sum_{i=1}^N T_{j,i}^{(k)} dx_i}{\sum_{i=1}^N T_{j,i}^{(k)}},$$

$$\sigma_j^{(k+1)} = \frac{\sum_{i=1}^N T_{j,i}^{(k)} (dx_i - \mu_j^{(k+1)})^2}{\sum_{i=1}^N T_{j,i}^{(k)}}$$

Оценка  $\sigma^2$ ,  $m_c$ ,  $\sigma_c$

$$\widehat{m}_c = \mu_1, \quad \widehat{\sigma}^2 = \frac{\sigma_2^2}{\Delta t}, \quad \widehat{\sigma}_c^2 = \sigma_1^2 - \sigma_2^2$$

Оценка времени поступления заявок

$$\widehat{T}_c = \left\{ t_i \mid i = 1, \dots, N \mid T_{1,i}^{(k)} > T_{2,i}^{(k)} \right\}$$

Оценки  $m$  и  $\lambda$  получаем так же, как в итеративном методе

# Случай конечного времени обработки заявки

$$\lambda_c \gg 0$$

$$\begin{aligned} X(t) &= B(t) + \sum_{t_c \in T_c} K_{t_c}(t) = \\ &= m + \sigma W(t) + \sum_{t_c \in T_c} \mathcal{N}(m_c, \sigma_c^2) \cdot I(t - t_c) \cdot e^{-\lambda_c(t - t_c)} \end{aligned}$$

Считаем, что заявки поступают достаточно редко и обрабатываются достаточно быстро: для каждой заявки предыдущие заявки практически не влияют на уровень загрузки

# Разностный аналог производной

$dX(t)$  в момент времени поступления заявки  $t_c$

$$\begin{aligned}dX(t_c) &= X(t_c) - X(t_c - \Delta t) = \\&= \mathcal{N}(m_c, \sigma^2 \Delta t + \sigma_c^2).\end{aligned}$$

(при условии, что в момент времени  $(t_c - \Delta t)$  заявки не было)

$dX(t)$  в момент времени отсутствия заявок  $t$

$$\begin{aligned}dX(t) &= X(t) - X(t - \Delta t) = \\&= \mathcal{N}(0, \sigma^2 \Delta t) + \mathcal{N}(m_c, \sigma_c^2) \cdot \left( e^{-\lambda_c(t-t_c)} - e^{-\lambda_c(t-\Delta t-t_c)} \right).\end{aligned}$$

(при условии, что в момент времени  $(t - \Delta t)$  заявки не было)

- 1 Построим вариационный для  $dx_i$ :

$$dx_{(1)} \leq dx_{(2)} \leq \dots \leq dx_{(N)}.$$

Приходящие заявки вносят существенно большее изменение уровня загруженности ресурсов сервера, чем фоновая нагрузка  
 $\implies$  все наблюдения  $dX(t_c)$  находятся в правой части ряда.

- 2 Используя итеративный метод для идентификации заявок FIND-REQUEST-INDEX, найдём границу  $k$  наблюдений  $dX(t_c)$  в вариационном ряде:

$$dx_{(i)} \sim dX(t_c), \quad i \geq k.$$

- 3 По границе  $k$  выделим наблюдения из  $\widehat{T}_c$

# Метод наименьших квадратов (1)

В моменты времени между поступлениями заявок  $[t_k, t_{k+n}]$

$$\begin{aligned}dX(t) &= \mathcal{N}(0, \sigma^2 \Delta t) + \mathcal{N}(m_c, \sigma_c^2) \cdot \left( e^{-\lambda_c(t-t_c)} - e^{-\lambda_c(t-\Delta t-t_c)} \right) \\&= \left( \mathcal{N}(0, \sigma^2 \Delta t) + \mathcal{N}(0, \sigma_c^2) \cdot \left( e^{-\lambda_c(t-t_c)} - e^{-\lambda_c(t-\Delta t-t_c)} \right) \right) + \\&\quad m_c \cdot \left( e^{-\lambda_c(t-t_c)} - e^{-\lambda_c(t-\Delta t-t_c)} \right)\end{aligned}$$



# Метод наименьших квадратов (2)

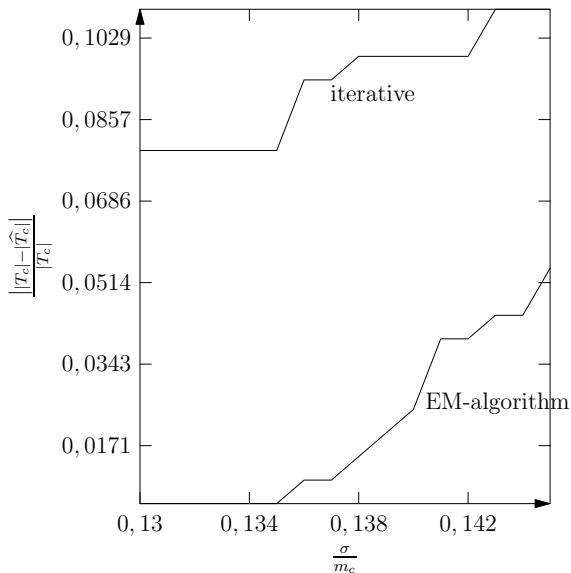
## Метод наименьших квадратов

$$\begin{aligned}(m_c, \lambda_c) &= \operatorname{argmin}_{m_c, \lambda_c} S(m_c, \lambda_c) = \\ &= \operatorname{argmin}_{m_c, \lambda_c} \sum_{i=1}^n \left( dx_i - m_c \cdot \left( e^{-\lambda_c(t_i - t_k)} - e^{-\lambda_c(t_i - \Delta t - t_k)} \right) \right)^2\end{aligned}$$

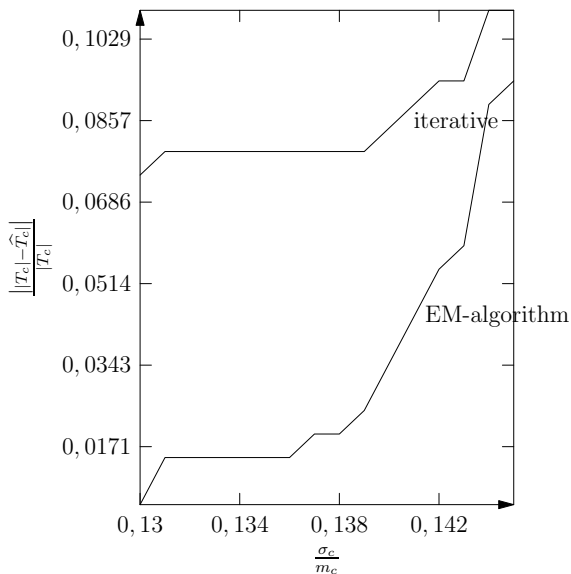
$S(m_c, \lambda_c)$  выпукла вниз на интересующей для анализа области — используем численный метод минимизации (метод покоординатного спуска)

$m$  и  $\sigma$  оцениваются так же, как в итеративном методе

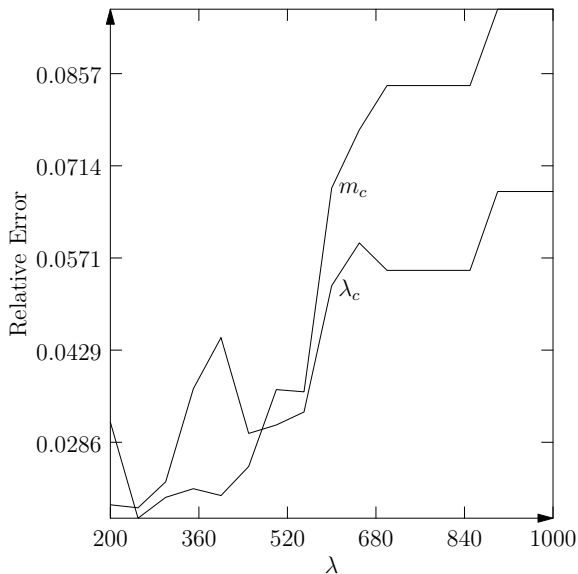
# Результаты работы (1)



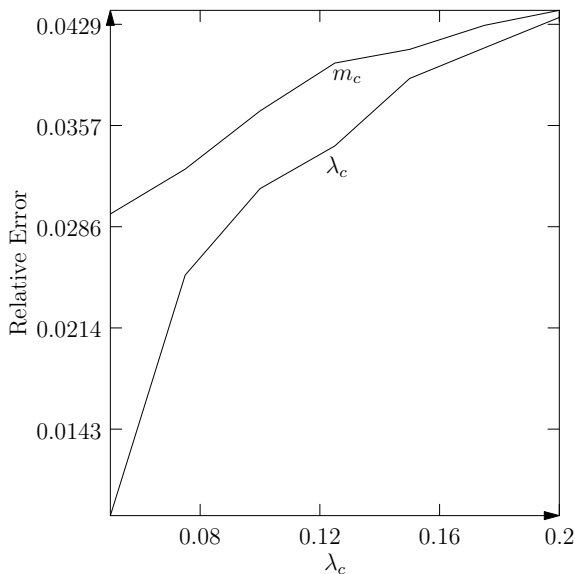
## Результаты работы (2)



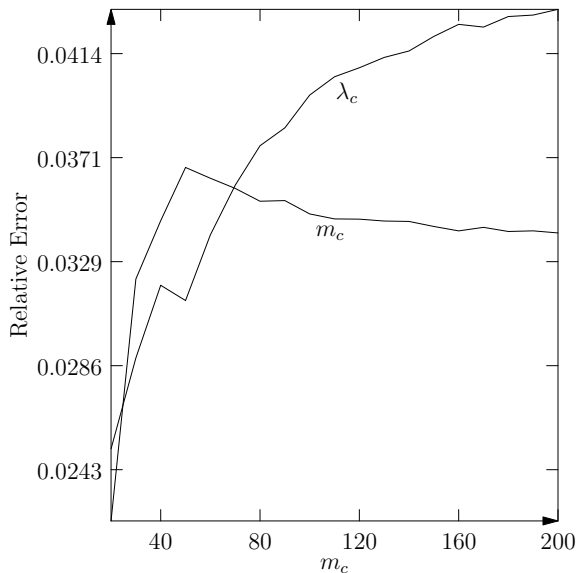
# Результаты работы (3)



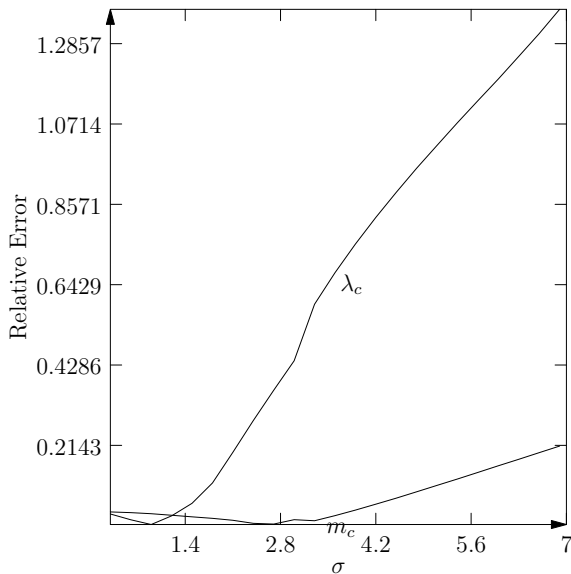
## Результаты работы (4)



## Результаты работы (5)



## Результаты работы (6)





## Результаты работы (7)

