Delay Estimation and Fast Iterative Scheduling Policies for LTE Uplink

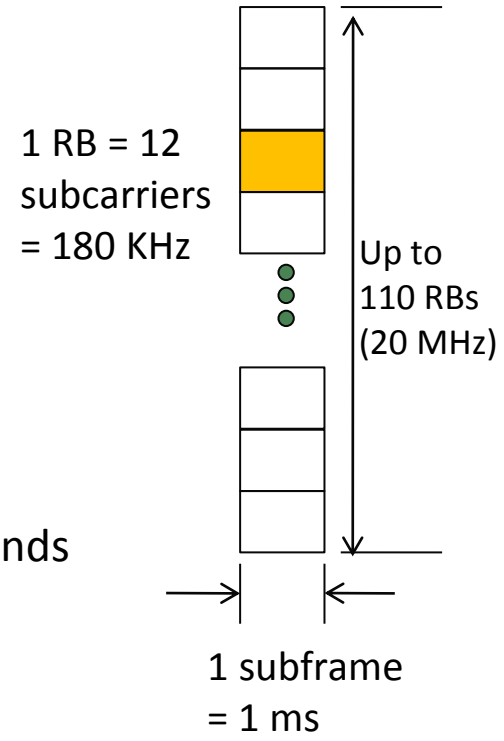
Akash Baid
WINLAB, Rutgers University

Ritesh Madan
Accelera Mobile Broadband

Ashwin Sampath
Qualcomm

LTE Resource Allocation Problem

- Each subframe, decide which RB to assign to which UE at what power:
  - 100s of UE per cell, 1000s of connections
  - 110 RBs in 20 MHz bandwidth
  - At most 1.5ms to make scheduling decisions



- Assign RBs based on:
  - Instantaneous channel gains on different sub-bands
  - Buffered queue length and packet delays
  - Average rate in the past

- Needs to account for re-transmissions, inter-cell interference, HARQ re-transmission target, adapt rate to channel variation

⇒LTE resource scheduling requires fast algorithms

Key Approaches in Scheduling Theory

Max weight Schedulers:

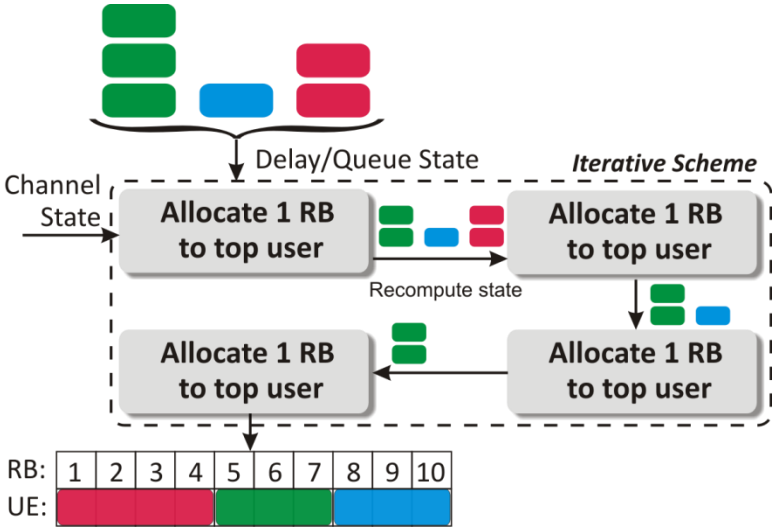
Given instantaneous spectral efficiency $s_i(t)$, at each time step t, select the user with maximum $K_i(t) * s_i(t)$ and assign all resources to that user

- $K_i(t)$ could be function of:
  - rate for elastic flows
  - delay or queue length for delay-sensitive flows

- [Stoylar'04, Whiting'04]: this maximizes the sum-utility of long term average rates
- [Tassiulas'93]: maximizing queue length times rate leads to stable queues

Key Approaches in Scheduling Theory

Iterative Scheduling: Make assignments 1 RB at a time; assign user with maximum recompute assign the next RB in the same way to (potentially) different user



- In large bandwidth systems, iterative algorithms allow more users to be multiplexed, lowers delay [Bodas'10, Lin'11]
- But past results based on restrictive models: No power and interference constraints; on-off rate model

Our Contributions

- Generalize delay based iterative schedulers to the LTE uplink system model via a novel objective function
- Obtain low complexity iterative algorithms
  - Sub-gradient analysis for frequency flat fading
  - Specialized interior point method for frequency selective fading
- Design a novel mechanism for inferring packet delays approximately from buffer status reports
- Demonstrate performance via detailed LTE simulations

LTE Uplink System Model

- Single Carrier FDMA with 180 kHz x 1 ms RB
- We consider *M* sub-bands of equal bandwidth *B*, with *B* < coherence bandwidth of each user
- **Constraints :**
  - Fractional power control: limits inter-cell interference

$$p_{ij} \leq \gamma_{ij} b_{ij}, \forall i, j$$

  - Peak power constraint: from regulatory requirements

$$\sum_{j=1}^{M} p_{ij} \leq P$$

  - Total bandwidth: based on available bandwidth

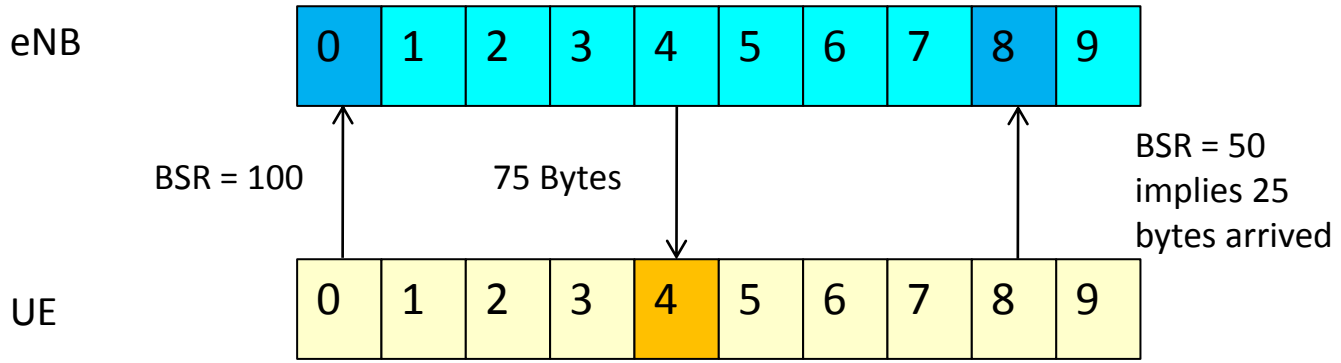
$$\sum_{i=1}^{N} b_{ij} = B, \ \forall j$$

  - Achievable rate:

$$r_{ij} = b_{ij} \psi \left(G_{ij} p_{ij} / (b_{ij} I_j) \right)$$

Estimating Delay from BSR

- UEs send periodic Buffer Status Report (BSR) with info about no. of total bytes in buffer
- We use the BSR reports along with knowledge of number of bytes scheduled in each subframe to estimate the packet delays:



- Main complexity is due to re-transmissions which can lead to BSR report arriving out of order

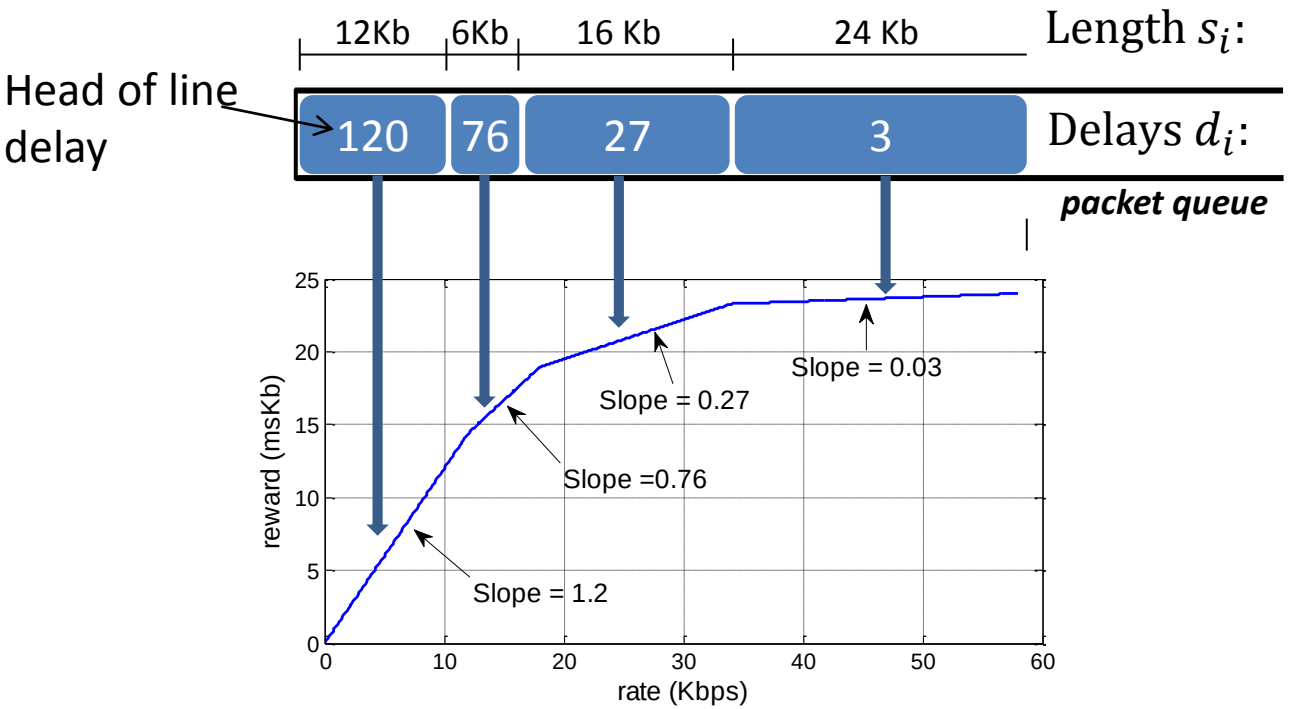
Reward Functions

Scheduler aims to maximize $\sum_{i=1}^{N} f_i(r_i)$, defined as:

- ***Best Effort:***
  - $f_i(r_i) = \frac{1}{\alpha_i} U_i((1-\alpha_i)x_i(t) + \alpha_i r_i)$
  - $U_i : \mathbb{R}_+ \mapsto \mathbb{R}$, strictly concave increasing function

Reward Functions

- ***Delay QoS Traffic:***
  - Piecewise linear continuous concave function:



$$f_i(r_i) = \sum_{j=1}^{n_i^{\text{serv}}(r_i)} s_i(j)d_i(j) + \left(r_i\Delta - \sum_{j=1}^{n_i^{\text{serv}}(r_i)} s_i(j) \right) d_i(n_i^{\text{serv}}(r_i)+1)$$

Frequency Flat Fading

- Given a total of *B* RBs, we want to solve:

$$\max. \quad \sum_{i=1}^{N} f_i \left(b_i \psi \left(\frac{G_i \min(\gamma_i b_i, P)}{I b_i} \right) \right)$$

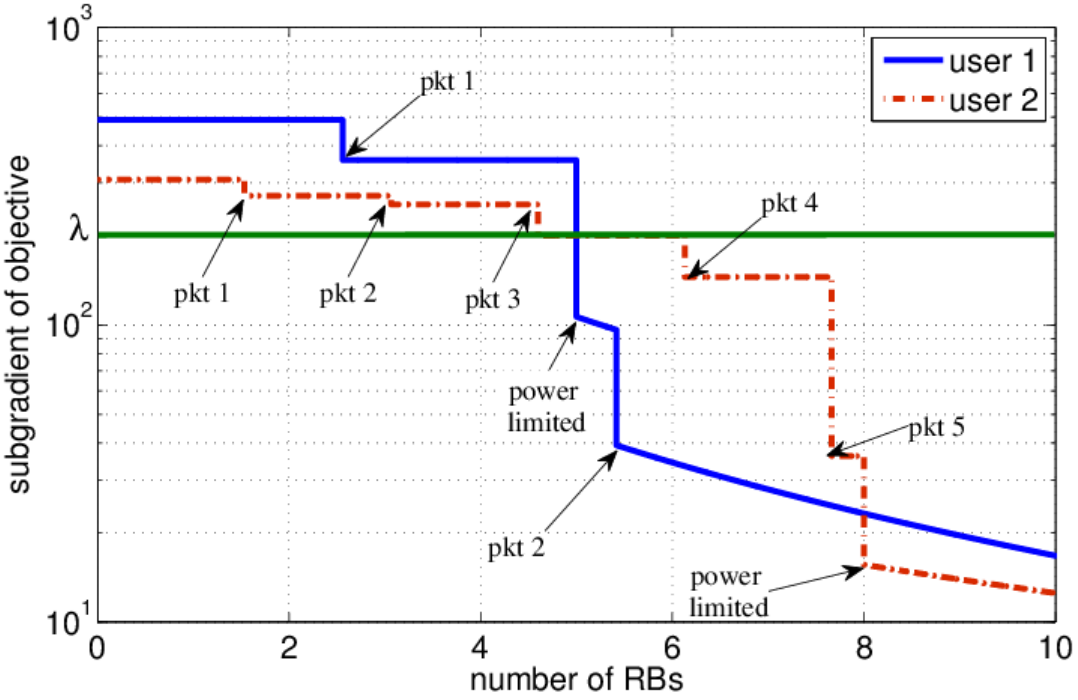
$$\text{s.t.} \quad 0 \le b_i \le b_i^{\max}, \ \forall i, \quad \sum_{i=1}^{N} b_i \le B$$

  - Convex problem but non-differentiable objective

- Solution through Lagrange dual problem
  - There exists a $\lambda^\star$ s.t. the following are equalized $\forall i$:
  - For B.E. users: **marginal utility** x **incremental rate**
  - For Delay QoS users: **HoL delay** x **incremental rate**
  - $O(N \log L)$ for N users and L points of discontinuity

Computation of optimal solution

- The optimal solution can be found through a bisection search on the sub-gradient; however:
  - Discontinuity due to max power could result in convergence to unstable points
  - Discontinuity due to design of delay utility function could result in non-convergence of bisection search



Two user example:

Packet Delays

U1	450	330	…
U2	170	150	

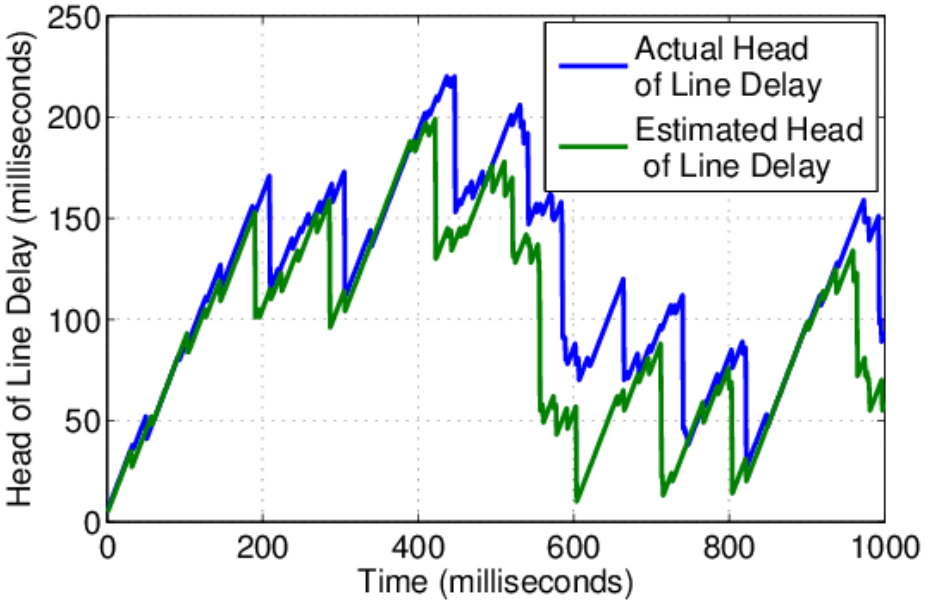
$b_i^{\max}$ **threshold**

U1	5
U2	8

Simulation Framework & Topologies

- Detailed system simulator:
  - PHY layer performance (channel, power, rate)
  - MAC layer signaling
  - 19 cell simulation with wrap around to compute IoT
  - Single cell simulation to calculate resulting rate, delay
- Mix of two types of traffic:
  - Live Video: On-Off Markov model, 300 kbps when On
  - Streaming Video: packet inter-arrival time, packet length both drawn from truncated Pareto distr., adaptive rate streaming with mean rate at 80% of achievable rate at full buffer

HoL Delay Estimation



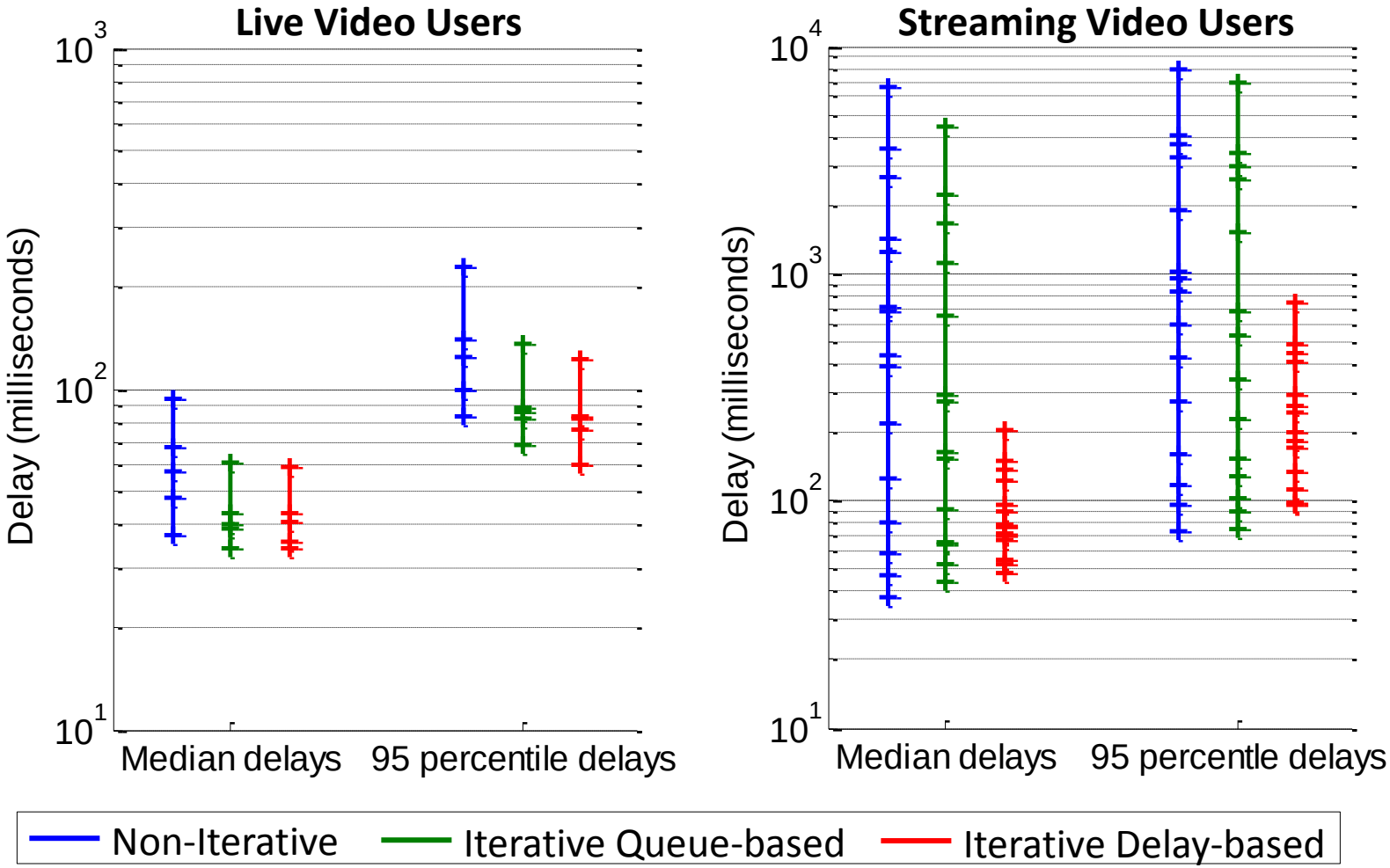
- Delay estimation performance over a 1 second run for a particular UE in a 20 UE simulation.
- Main source of error is the difference in the time at which UE and BS updates its delay counters

Delay performance

- Comparing three scheduling algorithms:
  - **Non-iterative maximum weight:** a UE with the highest queue length times spectral efficiency for first RB is allocated bandwidth until the queue is drained or the UE becomes power limited before allocation to the next UE.
  - **Iterative Queue:** minimizes sum-of-squares of queue lengths [8]
  - **Iterative Delay***:* maximizes the reward function we defined above

Delay performance

- Macro-cell simulation: 20 UEs with pathloss between 100 dB and 135 dB



Frequency Selective Fading

- For frequency selective fading, we use Interior point method (with few Newton steps) to solve the combined optimization problem across bands
- General purpose interior points solutions have a complexity of $O(NM + NL)^3$ per iteration
- We exploit the structure to reduce it to $O(N(L^2 + M^2))$:

$$\left[\begin{array}{ccc|c} H_1 & & & A^T \\ & \ddots & & \\ & & H_N & \\ \hline & A & & 0 \end{array}\right] \left[\begin{array}{c} x_1 \\ \vdots \\ x_N \\ y \end{array}\right] = \left[\begin{array}{c} a \\ b \end{array}\right]$$

$$H_i \in \mathbb{R}^{(L+M)\times(L+M)}, \; A \in \mathbb{R}^{M\times N(L+M)}$$

Conclusions

- LTE resource allocation problems require low complexity algorithms
- A general iterative framework with different utility functions for different traffic types can be used to achieve respective goals
- Exploiting the structure of the problem can lead to design of fast scheduling algorithms:
  - per iteration for frequency flat fading
  - per iteration for frequency selective
- Simulation shows the benefit of iterative HoL delay based scheduling over non-iterative and iterative queue-based scheduling

Thanks !

Questions ?

Extras

Estimating Delay from BSR

- Each BSR is characterized by a three tuple:

$$\{B_i(1), \tau_i(1), \delta_i(1)\}$$

Reported Buffer Size — Time of Reception — Retransmission Delay

- Maintain a history of estimated queue length Q_i:
- Arrival time of packets can be estimated using Q_i:
  - Assume a BSR *b*:{1300, 24, 16} is received
  - Assume $Q_i(t)$ was 1000 at *t* = (24-16)
  - Then it can be deduced that 300 bytes arrived between *t* = *8* and the time of previous BSR arrival
- Main complexity is due to re-transmissions which can lead to BSR report arriving out of order

Estimating Delay from BSR

- Maintain a history of estimated queue length $Q_i(t)$
- $Q_i(t)$ entries are updated according to:

For every t, i

1) *Scheduled Bytes:* $Q_i(t) = Q_i(t-1) - C_i(t)$.
2) *Failed Bytes:* $Q_i(t) = Q_i(t) + F_i(t)$.
3) *BSR report:* If a BSR report is received at time t, i.e., there is n such that $\tau_i(n) = t$, then update queue state as follows: If the base-station has not received any BSR report created after time $t - \delta_i(n)$, then

$$Q_i(t - \delta_i(n) : t) = Q_i(t - \delta_i(n) : t) + A_i(t - \delta_i(n))$$

where arrival $A_i(t - \delta_i(n)) = B_i(t) - Q_i(t - \delta_i(n))$ otherwise for

$$\arg\min_{\{m:\ \tau_i(m) < t\}} [\tau_i(m) - \delta_i(m) - (\tau_i(n) - \delta_i(n))]$$

update

$$A_i(t - \delta_i(n)) = B_i(t) - Q_i(t - \delta_i(n))$$

$$A_i(\tau_i(m) - \delta_i(m)) = A_i(t - \delta_i(m)) - A_i(t - \delta_i(n))$$

$$Q_i(t - \delta_i(n) : \tau_i(m) - \delta_i(m) - 1) = Q_i(t - \delta_i(n) : \tau_i(m) - \delta_i(m) - 1) + A_i(t - \delta_i(n))$$

Bisection Search

Given starting value of $\underline{\lambda}$, $\overline{\lambda}$, and tolerance ϵ.
repeat
  Bisect: $\lambda = (\underline{\lambda} + \overline{\lambda})/2$.
  Allocate bandwidth for all i*:*
  if $\lambda > \max \partial f_i(0) \max \partial h_i(0)$ **then**
    set $b_i = 0$.
  else
    b_i is such that

$$\lambda \in [\min \partial f_i(r_i) \times \min \partial h_i(b_i), \max \partial f_i(r_i) \times \max \partial h_i(b_i)]$$

    where

$$r_i = \left(b_i \psi \left(\frac{G_i(t) \min(\gamma_i b_i, P)}{I b_i} \right) \right)$$

  end
  Update: if $\sum_{i=1}^{N} b_i - B > 0$, $\underline{\lambda} = \lambda$, else $\overline{\lambda} = \lambda$
until $\left| \sum_{i=1}^{N} b_i - B \right| < \epsilon$

$O(\log L)$ complexity if $f_i(r_i)$ is non-differentiable at at most L points.

Overall complexity: $O(N \log L)$ per iteration (with ~ 10 iterations) for N users and L points of non-differentiability in the objective function.