Open- and Closed-Loop Traffic Regulation Schemes for ATM Networks *

Mohamed Abdelaziz† Ioannis Stavrakakis‡
Department of Electrical Engineering and Computer Science
University of Vermont
Burlington, VT 05405

Abstract *In this paper, some new traffic regulation schemes are defined in terms of a relief-spacing (or spacing of the allowance for cell delivery to the network) function. This function may depend only on some user-state, thus defining the class of open-loop relief-spacing traffic regulators; or it may depend on both some user- and network-state, thus defining the class of closed-loop relief-spacing traffic regulators. Meaningful criteria for optimizing the performance of both classes are defined. A queueing model is presented to analyze their performance and Markov decision theory approaches are used to identify the optimal elements in the proposed classes.*

1 Introduction

Preventive control is considered to be a promising approach for traffic congestion management in the emerging high-speed Asynchronous Transfer Mode (ATM) networks. An extended survey of mechanisms - called bandwidth enforcement mechanisms or traffic regulators (TR) - developed for the implementation of such control can be found in [1]. The object function of a TR is to control the flow of the user traffic to the network in a way that unacceptable network congestion be avoided.

The prevailing approach for the design of efficient TR's has been the following: At first, certain key measures of basic traffic characteristics - which have a significant impact on the network performance - are identified. For instance, such measures can be the cell rate and burst length of the user-traffic. Then, the user and the network agree on some limits on the values of these measures (contract), so that the network be able to plan based on the maximum expected amount of stress (potential for congestion) coming from the particular user. Network congestion may then be controlled through an effective TR, whose basic function is to provide for the smoothing of the user traffic. A TR may also be seen as a mechanism which enforces compliance with the contract agreements, thus protecting network resources [2]. The insightful description of the basic functions that need to be implemented by a TR, presented in [3], as well as numerical results, have pointed to the ineffectiveness of the leaky bucket TR and the improved performance delivered by Spacer-Controller type TR's [3,4,5,6]. By defining the relief rate to be the rate of *allowance* for cell delivery to the network, the class of $\sigma - Relief$ TR's was introduced in [3], as a practical implementation of a fixed (and equal to σ) relief function in the slotted ATM environment. In this paper, this class is extended to include TR's implementing non-fixed, state-dependent relief functions. The open-loop relief-spacing regulation scheme considered, is defined to be one whose relief function depends only on user-state information. The optimal elements in the resulting class of the TR's are identified by defining some meaningful optimality criteria and following a Markov decision approach. Finally, the class of closed-loop relief-spacing TR's is also defined and studied. The associated relief function depends on both user and network state information. The closed-loop approach is considered to establish the magnitude of performance improvement that can be achieved through the availability of network-state information and to also gain insight toward the development of more efficient open-loop schemes or hybrid schemes.

2 The open-loop traffic regulators

With the exception of [7,8], the open-loop TR's proposed in the past can be classified as user-state-independent. In this section, the class of open-loop, user-state-dependent, relief-spacing TR's is introduced and its performance at both the user and network premises is evaluated.

To avoid practical problems associated with the implementation of relief (service) rates in a time-slotted environment, the considered TR's are defined in terms of a relief-spacing function. This function represents the time (spacing in slots) between potential consecutive server visits to the user-queue, T(). For an open-loop relief-spacing TR, T() is a function of both the source state (s) and the user-queue-occupancy (q_u), $T(s, q_u)$. The class of such TR's is described in terms of the class of all possible mappings:

$$T : S_s \times S_u \rightarrow Z_B^+ \equiv \{1, 2, ..., B\}, B < \infty \qquad (1)$$

*Research supported in part by the Advanced Research Project Agency (ARPA) under Grant F4963-1-0564.
†email : mohamed@emba.uvm.edu
‡email : stavraka@emba.uvm.edu

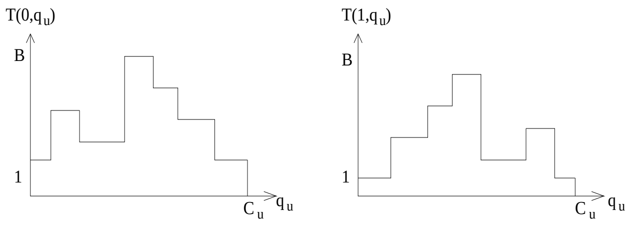

Figure 1: An example of the open-loop regulation function.

and is denoted by $\langle R - T(s, q_u)\rangle$, that is,

$$\langle R - T(s, q_u)\rangle \equiv \{T(s, q_u), (s, q_u)\epsilon S_s \times S_u : T(s, q_u)\epsilon Z_B^+\}$$

An example of a relief-spacing function $T(s, q_u)$ is shown in Fig. 1, where the state space of the Markov source representing the user-traffic is $S_s = \{0, 1\}$ and the state space of the user-queue occupancy process is $S_u = \{1, .., C_u\}$. A Spacer-Controller TR [3,4,5,6] can easily be viewed as a relief-spacing TR with constant relief-spacing function T. Such a TR does not utilize locally available user-state information and is expected to deliver sub-optimal performance.

2.1 Performance analysis for the open-loop relief-spacing TR

In this subsection, the cell loss probabilities induced at both the user premises and the network premises are calculated under an arbitrary relief-spacing function defined by the mapping (1).

The queueing model at user premises, shown in Fig. 2, will be considered for the performance study of the TR. It is assumed that the Markov source generates cells only when in state 1 (on) and generates no cells when in state 0 (off); the probabilities governing source transitions from state i to state j are denoted by p(i,j), the source cell rate by λ and the source burstiness coefficient is defined by $\gamma = p(1, 1) - p(0, 1)$. The probability that k cells are generated by the source while moving from state i to state j in m consecutive slots is denoted by $f^m(i, j, k)$ and can be easily computed [3]. Slots where cell departures from the user-queue occur are denoted as departure slots. The system state is defined at the beginning of departure slots in

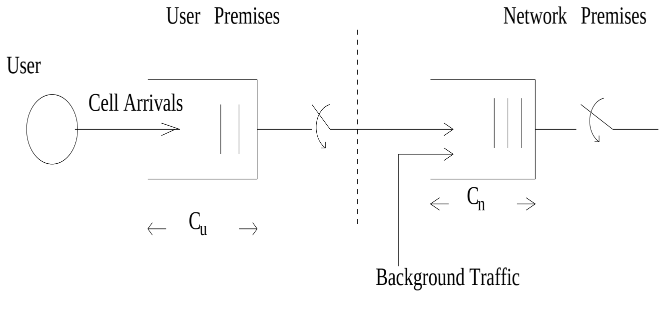

Figure 2: A queueing model for the relief-spacing TR.

terms of the quantities (s, q_u) where s and q_u describe the states of the source process $\{I_l\}_{l\geq 0}$, and the user-queue-occupancy process $(\{Q_l^u\}_{l\geq 0})$, at the $\underline{lth}$ departure slot, respectively. At each visit, the server remains at the queue for one time slot, provides one unit of service if a cell is available, then takes a vacation of length $T(s, q_u) - 1$. If no cells are available, the server will only switch away after providing service to the first arriving cell.

It is easy to establish that $\{I_l, Q_l^u\}_{l\geq 0}$ is a Markov chain with a state space $\{(i_1, n_1) : i_1\epsilon S_s, n_1\epsilon S_u\}$. The probability that the Markov chain $\{I_l, Q_l^u\}_{l\geq 0}$ moves from state (i_1, n_1) at the $\underline{lth}$ departure slot to state (i_2, n_2) at the $l + 1\underline{st}$ departure slot will be denoted by $a(i_1, n_1, i_2, n_2)$ and can be found in [12]; the corresponding transition matrix will be denoted by A. The steady state probabilities $\pi(i, n), i\epsilon S, n\epsilon S_u$ are obtained from the matrix equation $\overline{\Pi} = A\overline{\Pi}$, $\overline{\Pi} = [\pi(i, n)]$ with the normalizing condition $\overline{\Pi}\overline{e} = 1$ where $\overline{e}$ is a unity column vector. The cell loss probability L_u at the user premises can then be obtained from the expression:

$$L_u = \frac{\sum_{i\epsilon S_s}\sum_{n\epsilon S_u} L_u(i, n)\pi(i, n)}{\lambda\sum_{i\epsilon S_s}\sum_{n\epsilon S_u}\overline{ID}(in)\pi(i, n)}$$

where $ID(i, n)$ denotes the time in slots between two consecutive cell departures and can be easily related to the relief-spacing $T(i, n)$, [12]. $\overline{ID}(i, n)$ denotes the expected value of $ID(i, n)$. $L_u(i, n)$ denotes the expected number of cells lost over the interval $ID(i, n)$ and is given by:

$$L_u(i, n) = \sum_{j\epsilon S_u}\sum_{k=C_u-n+1}^{T(i,n)} (k - C_u + n - 1)f^{T(i,n)}(i, j, k)$$

To establish the performance of the TR's at the network premises, the network-queue of capacity $C_n < \infty$ is considered under the First-In-First-Out (FIFO) service policy (Fig. 2). In addition to the regulated user traffic, the network-queue is also fed by a background Bernoulli traffic, with rate λ_b. The cell loss probability at the network premises (L_n) is computed by defining the system state in terms of the three dimensional process $\{I_l, Q_l^u, Q_l^n\}_{l\geq 0}$, where $\{Q_l^n\}_{l\geq 0}$ describes the network-queue-occupancy process with state space $S_n = \{0, 1, .., C_n\}$. Proceeding as above, the transition probabilities, the stationary probabilities of the three dimensional system process and the loss probability are derived. In contrast to the approximate analyses in [7,8,9], the adopted analysis is exact, and involves no approximations of the output regulated traffic. To avoid the numerical complexity associated with state space of the above process, the recursive equations technique used in [3] is employed.

3 Optimal relief-spacing traffic regulation schemes

The objective in regulating the user-traffic in an ATM environment is to increase the network utilization while guaranteeing the Quality of Service (QoS) of all supported users. While the network performance (QoS) can be improved through increased regulation of the user traffic streams, excessive regulation may result in unacceptable performance (QoS) at the user premises. An optimal TR can thus be defined as one achieving an optimal balance between the delivered QoS at the user and network premises.

Let the Quality of Service Margin ($QoSM$) be defined as the maximum amount of "disturbance" of the user traffic which can be tolerated before the necessary QoS is not provided. The $QoSM$ is basically described in terms of probabilistic measures on cell delay and/or loss (as for the QoS). Let $QoSMR_u$ and $QoSMR_n$ denote the value of the $QoSM$ Reduction at the user and network premises, respectively. A feasible optimization problem, leading to the derivation of some optimal TR, could be set up in terms of the following objective:

$$\min: \ QoSMR_n \ subject \ to: \ QoSMR_u \leq w \qquad (2)$$

In this paper, the cell loss probability is considered to be the measure of the provided QoS. Thus, $w = 0$ corresponds to an optimization formulation identifying the loss-free at the user premises TR, which minimizes the network induced losses.

Let L_u and L_n denote the cell loss probabilities at the user and network premises, respectively. The optimal element in the class of open-loop, user-state-dependent relief-spacing TR's, $\langle R - T(s, q_u)\rangle$, according to the criterion in (2) will be:

$$R_2 - T(s, q_u) = arg\{\min_{T:S_s \times S_u \rightarrow Z_B^+} \{L_n\} : L_u \leq w\} \qquad (3)$$

Since $B^{2\times C_u}$ elements are contained in $\langle R - T(s, q_u)\rangle$, it is apparent that exhaustive search for $R_2 - T(s, q_u)$, is unrealistic for most practical cases. A tractable formulation of the optimization problem can be obtained by interpreting the relief-spacing process $\{T_l\}_{l\geq 0}$ as a set of decisions. Each decision (T_l) is taken upon a visit to some state (s, q_u) of the Markov process $\{I_l, Q_l\}_{l\geq 0}$, and determines the relief-spacing applied at the $\underline{lth}$ departure slot. By associating some cost with each decision, a total cost associated with a given relief-spacing process $\{T_l\}_{l\geq 0}$ can be computed. Thus the optimal relief-spacing function (or $R - T(s, q_u)$ TR), minimizing the total cost, can be identified by following a semi-Markov decision approach. Notice that the cost associated with a decision must be completely determined in terms of the decision T_l taken and the state of $\{I_l, Q_l^u\}_{l\geq 0}$. The advantage of such a formulation of the optimization problem is that non-exhaustive, computationally appealing approaches exist for the derivation of the optimal policy, such as the policy and value iteration algorithms and the linear programming approach [10].

Unfortunately, since no network-state information is included in the state description of the open-loop control, the cost associated with the optimization problem described in (2), that is cell loss at the network premises, cannot be determined by the state of $\{I_l, Q_l^u\}_{l\geq 0}$ and $\{T_l\}_{l\geq 0}$. Consequently, a network performance measure (cost)- to be minimized- which can be determined by the state of $\{I_l, Q_l^u\}_{l\geq 0}$, and the decision taken, is considered. In this section, the variance of the cell interdeparture process is considered as such a measure of the smoothness of the regulated traffic and the potentially induced network congestion. Thus, the optimal $R - T(s, q_u)$ is defined by:

$$R_v - T(s, q_u) = arg\{\min_{T:S_s \times S_u \rightarrow Z_B^+} \{v(T)\} : L_u \leq w\} \qquad (4)$$

where $v(T)$ denotes the variance of the cell interdeparture process given by:

$$v(T) = \frac{\sum_{(i,n)\epsilon S_s \times S_u} c(i, n, T(i, n))\pi(i, n)}{\sum_{(i,n)\epsilon S_s \times S_u} \tau(i, n, T(i, n))\pi(i, n)}$$

where: $c(i, n, T(i, n)) = (\overline{ID}(i, n) - \frac{1}{\lambda})^2$ is the cost associated with the decision $T(i, n)$; $\tau(i, n, T(i, n)) = \overline{ID}(i, n)$. The linear programming approach [10] is used to identify the optimal element $R_v - T(s, q_u)$. To illustrate the performance improvement obtained by utilizing the user-state-dependent relief-spacing functions, the following TR subclasses are defined.

$$\langle R - T\rangle \equiv \{T(s, q_u) = c, (s, q_u)\epsilon S_s \times S_u : c\epsilon Z_B^+\}$$

$$\langle R - T(q_u)\rangle \equiv \{T(s, q_u) = T(q_u), (s, q_u)\epsilon S_s \times S_u : T(q_u)\epsilon Z_B^+\}$$

Note that the relief-spacing function is user-state-independent in $\langle R - T\rangle$ and is source-state-independent in $\langle R - T(q_u)\rangle$. Let $R_v - T$ and $R_v - T(q_u)$ denote the optimal elements in these classes in the sense that they minimize the variance of the cell interdeparture process. The improvement in performance obtained through the adoption of the user-state-dependent relief-spacing TR's is illustrated in Fig. 3, where the cell loss probability at the network premises, L_n, is plotted for the loss-free optimal policies $R_v - T, R_v - T(q_u)$ and $R_v - T(s, q_u)$;

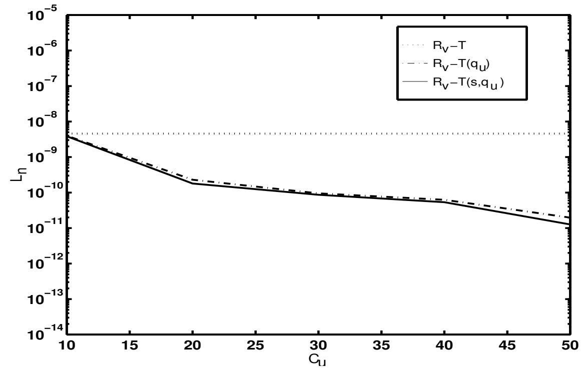

Figure 3: Cell loss probability L_n versus C_u, induced by $R_v - T$, $R_v - T(q_u)$ and $R_v - T(s, q_u)$.

$C_n = 50$, $10 \leq C_u \leq 50$, $\lambda = 0.09$, $\gamma = 0.8$ and $\lambda_B = 0.8$. It can be observed that lower cell loss probabilities L_n can be achieved by $R_v - T(q_u)$ or $R_v - T(s, q_u)$ as C_u increases.

4 The closed-loop traffic regulators

Motivated by the potential for improved performance when network state information can be available at the user premises, the class of closed-loop relief-spacing traffic regulation schemes is introduced in this section. Unlike the class of open-loop TR's presented in section 2, the relief-spacing function associated with the closed-loop TR's is assumed to depend on the network-state as well as the user-state. As shown earlier, the network premises has been defined to consist of the network-queue receiving directly the regulated user-traffic together with some Bernoulli background traffic. With this definition, the network-state is sufficiently described by the network-queue occupancy process $\{Q_l^n\}_{l \geq 0}$. Accordingly, if the user and the network queues are not far apart, it may not be unrealistic to assume that such information can be timely available, and therefore the effects of propagation delay can be assumed negligible. Similar network-feedback-based traffic regulation schemes have been recently proposed, in an attempt to improve on the performance of entirely preventive congestion control approaches [11].

4.1 Formulation of the optimization problem

A closed-loop, user-state-dependent relief-spacing TR is defined in terms of the associated relief-spacing function $T(s, q_u, q_n)$, $(s, q_u, q_n) \epsilon S_s \times S_u \times S_n$ and is denoted by $R - T(s, q_u, q_n)$. The class of such TR's is defined in terms of all possible mappings:

$$T : S_s \times S_u \times S_n \rightarrow Z_B^+ = \{1, 2, ..., B\},\ B < \infty$$

and is denoted by:

$$\langle R - T(s, q_u, q_n)\rangle \equiv \{T(s, q_u, q_n), (s, q_u, q_n) \epsilon S_s \times S_u \times S_n : T(s, q_u, q_n) \epsilon Z_B^+\}$$

As in section 2, it is easy to establish that a semi Markov decision approach can be followed for the derivation of optimal elements in $\langle R - T(s, q_u, q_n)\rangle$. The semi-Markov process $\{(I_l, Q_l^u, Q_l^n)\}_{l \geq 0}$ defined at departure slots is used to completely describe the evolution of the system under a given relief-spacing function $T(s, q_u, q_n)$ and determine the induced performance measure (cost) involved in the optimization problem formulation. The optimal TR's in $\langle R - T(s, q_u, q_n)\rangle$ - $\langle < R_t - T(s, q_u, q_n)\rangle$ and $\langle R_w - T(s, q_u, q_n)\rangle$ - are derived in the following two senses:

$$R_t - T(s, q_u, q_n) = arg\{\min_{T : S_s \times S_u \times S_n \rightarrow Z_B^+}\{L_u + L_n\}\} \tag{5}$$

$$R_w - T(s, q_u, q_n) = arg\{\min_{T : S_s \times S_u \times S_n \rightarrow Z_B^+}\{L_n\} : L_u \leq w\} \tag{6}$$

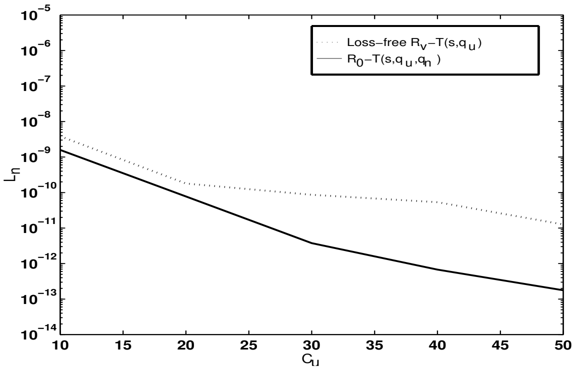

Figure 4: L_n versus C_u for the optimal open- and closed-loop TR's.

Notice that the calculation of L_n in terms of the semi-Markov process and the action taken, is now possible and, thus, it can be part of the quantity to be minimized. This is contrary to the open-loop case where only some traffic smoothness measure (determined by the user-state) was possible to consider. Also, note that no losses can occur at the user-queue when the relief spacing $T(s, q_u, q_n)$ is less than the remaining queue capacity $(C_u - q_u + 1)$. Thus, $R_0 - T(s, q_u, q_n)$ which is the optimal loss-free element at the user premises ($w = 0$ in $R_w - T(s, q_u, q_n)$) can be equivalently identified as a member of the sub-class of $\langle R_f - T(s, q_u, q_n)\rangle$, determined by,

$$\langle R_f - T(s, q_u, q_n)\rangle \equiv \{T(s, q_u, q_n),\ (s, q_u, q_n) \epsilon S_s \times S_u \times S_n : T(s, q_u, q_n) \epsilon Z_B^+,\ T(s, q_u, q_n) \leq C_u - q_u + 1\}. \tag{7}$$

and therefore, $R_0 - T(s, q_u, q_n)$ can be obtained by solving the equivalent unconstrained optimization problem,

$$R_0 - T(s, q_u, q_n) = arg\{\min_{T : R - T(s, q_u, q_n) \epsilon \langle R_f - T(s, q_u, q_n)\rangle}\{L_n\}\} \tag{8}$$

4.2 Numerical results

In Fig. 4, the performance of the optimal open- and closed-loop, user-state-dependent relief-spacing TR's inducing zero cell loss at the user premises is presented in terms of L_n versus C_u plots. As expected, the employment of additional (that is, network state) information in the closed-loop scheme, improves the performance achieved by the optimal open-loop scheme. It should be noted that the difference in performance increases as C_u increases since a more effective utilization of the network-queue information can be achieved increasing the advantage of the closed-loop scheme. In Fig. 5-6, some relief-spacing functions for $R_t - T(s, q_u, q_n)$ obtained under various background traffic rates λ_B are presented. These figures illustrate the adaptation of the optimal TR, $R_t - T(s, q_u, q_n)$, to the change in service requirements of both the regulated traffic and the background traffic. It can be seen that as the arrival rate of the background process (λ_B) decreases while the regulated traffic's parameters are held constant, the optimal relief- spacing TR provides lower relief-spacing for smaller values of q_n thus taking advantage of the reduction in service demand by the background process and offering this available capacity to the regulated process.

References

(1) J. Bae, T. Suda, "Survey of Traffic Control Schemes and Protocols in ATM Networks", Proc. IEEE, (79), 2, Feb. 1991.

(2) E. Rathgeb, "Modeling and Performance Comparisons of Policing Mechanisms for ATM Networks", IEEE J. Select. Areas Commun., (9), 3, April 1991.

(3) I. Stavrakakis, M. Abdelaziz and D. Hoag, "A User Relief Approach to Congestion Control in ATM Networks", Asynchronous Transfer Mode Networks, Plenum Press, New York, pp. 135-155, 1993.

(4) F. Guillemin, P. Boyer, A. Dupuis, L. Romoeuf, "Peak Rate Enforcement in ATM Networks", IEEE Infocom'92, pp. 753-758.

(5) C. Ohta, H. Tode, M. Yamamoto, H. Okada and Y. Tezuke, "Peak Rate Regulation Scheme for ATM Networks and its performance", IEEE Infocom'93, pp. 680-689.

(6) F. M. Brochin, "A Cell Spacing Device for Congestion Control in ATM Networks", Performance Evaluation, pp. 107-127, 1992.

(7) F.Bernabei, L. Gratta, M. Listanti, M. Testa, "Analysis of Two Level Shaping for Multiplexing of ON-OFF ATM Sources", ICC'93 Conf.

(8) F.Bernabei, L. Gratta, M. Listanti, M. Testa, "Analysis of ON-OFF Source Shaping for ATM Multiplexing", IEEE Infocom'93, pp. 1330-1336.

(9) H. W. Lee and J. W. Mark, "ATM Network Traffic Characterization Using two types of On-Off Sources", IEEE Infocom'93, pp. 152-159.

(10) H. C. Tijms, Stochastic Modeling and Analysis: A Computational Approach. New York, Wiley, 1986.

(11) I. Rubin and K. D. Lin, "A Burst Level Adaptive Input-Rate Flow Control Scheme for ATM Networks", IEEE Infocom'93, pp. 386-394.

(12) M. Abdelaziz and I. Stavrakakis, "Some Optimal Traffic Regulation Schemes for ATM Networks: A Markov Decision Approach", IEEE/ACM Transactions on Networking, to appear.

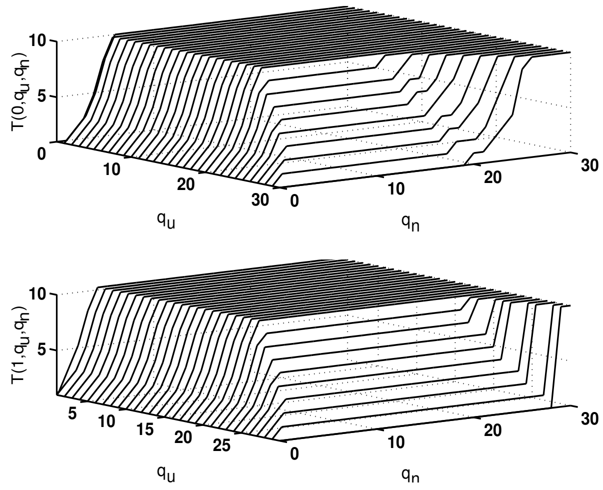

Figure 5: Relief-spacing function for $R_t - T(s, q_u, q_n)$ $\lambda_B = 0.9$, $\gamma = 0.8$ and $\lambda = 0.09$.

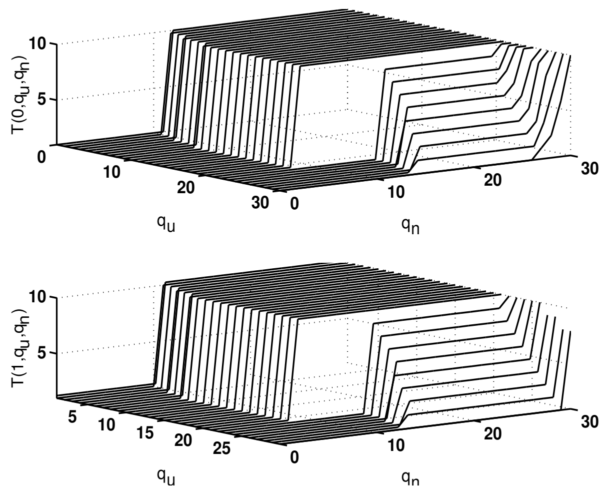

Figure 6: Relief-spacing function for $R_t - T(s, q_u, q_n)$ $\lambda_B = 0.8$, $\gamma = 0.8$ and $\lambda = 0.09$.