Stability and Benefits of Suboptimal Utility Maximization

Tian Lan\(^1\), Xiaojun Lin\(^2\), Mung Chiang\(^1\), Ruby Lee\(^3\)

\(^1\)Department of Electrical Engineering, Princeton University, NJ 08544, USA
\(^2\)School of Electrical and Computer Engineering, Purdue University, IN 47907, USA

Abstract—Network utility maximization has been widely used to model resource allocation and network architectures. But in practice often it cannot be solved optimally due to complexity reasons. Thus motivated, we address the following two questions in this paper: can suboptimal utility maximization maintain queue stability? Can under-optimization of utility objective function in fact benefit other network design objectives? We quantify that a resource allocation that is suboptimal with respect to a utility maximization formulation maintains the maximum flow-level stability when the utility gap is sufficiently small and the information delay is bounded, and can still provide a guaranteed size of stability region otherwise. Utility-suboptimal rate allocation can also enhance other network performance metrics, e.g., it may reduce link saturation. Quantifying these intuitions, this paper provides a theoretical support for turning attention from optimal but complex solutions of network optimization to those that are simple even though suboptimal.

I. INTRODUCTION

The framework of Network Utility Maximization (NUM) has been very extensively studied over the last decade since [1]. Formulating many resource allocation problems as maximization of an increasing and concave utility function over a convex constraint set, a large number of publications have developed iterative, distributed algorithms that converge to the optimum.

Achieving optimality is clearly desirable for two reasons. Not only does this attain the benchmark of the highest value of network utility, it also guarantees flow-level stochastic stability. The number of flows varies over time as they are randomly generated by users and served by the network. This system can be viewed as a queuing system where the service rate depends on the resource allocation (e.g., rate control) policy employed by the network. For convex NUM and under the assumptions of Poisson arrivals, exponentially-distributed files sizes, and zero information delay (i.e. perfect queue-length information), it has been shown that for all rate allocation policies maximizing \(\alpha\)-fair utilities with \(\alpha > 0\), flow-level stochastic stability can be achieved if and only if the traffic intensity lies within the rate region, see, e.g., [2], [3], [5], [4]. In other words, rate region in the \(\alpha\)-fair utility maximization problem is also the maximum stability region under arrival and departure dynamics.

Utility-optimality and flow-level stability are strong benefits of optimizing NUM. However, in practice it is often prohibitive to solve NUM optimally, due to computational complexity and information delay. The impact of suboptimal solution is not well-studied in the existing literature. It is of practical importance to sharpen our understanding for two reasons:

- All rate-control algorithms require a non-negligible amount of execution time before they can reach the optimal rate allocation that maximizes the system utility. (In fact, most distributed rate-control algorithms cannot reach the optimal rate allocation in any finite number of iterations.) The time for the iterations to converge is often so long that the network states, e.g., the composition of the user population, change many times before convergence occurs. As a result, practical rate allocation algorithms are subject to a positive, and possibly random, time delay. Further, since we cannot afford to run the algorithms until it converges, each instance of the utility maximization problem can only be solved suboptimally.
- In wireless networks, the scheduling problem has an exponential computational complexity despite the fact that the rate region of the system is convex. For example, when the feasible rate region of a network is obtained by time-sharing among different subsets of users, a non-convex multi-user/link scheduling problem still needs to be solved in order to find the exact rate region achieved by time-sharing [6]. Such high computational complexity further increases the amount of time that is needed to reach the optimal rate allocation. In addition, many cross layer optimization algorithms that implement rate schedulers as an inner loop require the scheduling iterations to stop at some suboptimal point, for example, due to timescale separation assumption. There are many theoretical studies that investigate the use of low-complexity and even distributed scheduling algorithms. However, with these lower-complexity scheduling algorithms the rate-allocation algorithm will either take longer convergence time [7], or will not converge to the optimal rate-allocation any more [6]. In either case, if we are limited to a finite number of iterations, suboptimality in the rate-allocation becomes the only realistic outcome.

The gap between elegant theory and useful practice thus leads us to the following question: between optimality and simplicity, which one should we pick in solving NUM? Driven by the practical need for simple yet suboptimal solutions, we focus on suboptimal utility maximization, and then quantify the effects of information delay and utility-gap on flow-level stability, and on other important network performance metrics.

This work has been in part supported by the National Science Foundation through awards CCF-0448012, CNS-0430487, CNS-0519880, CCF-0635202, CNS-0720570, CNS-0721484, and ONR YIP award N00014-07-1-0864. An earlier version of this paper has appeared in IEEE Infocom 2008.
such as link saturation.

In [8], the authors show that for a class of rate allocation algorithms based on so-called dual solutions, the optimal stability region can be achieved even if the algorithm does not converge to the optimal rate allocation at any time. Similar observations have also been made in switching [9] and scheduling [10] problems. In this paper, we take a different approach. We characterize the capability of a resource allocation algorithm by two features: (1) the gap between its utility and the optimal utility; and (2) the time delay of the queue-length information. We study stability as a function of both the utility gap and the information delay. Our results apply to a class of general NUM formulations, in which the flow-level queueing models are not first-order Markov, thus making our proof technique of independent interest to general flow-level queueing model. Intuitively, one would think that the maximum stability region may be retained if the utility gap is small and the time delay is bounded, while only a reduced stability region can be achieved when the utility gap becomes large. This is indeed true. In Section III, we show that when information delay is uniformly bounded by a constant and the ratio of the utility gap (caused by a suboptimal rate allocation policy) to the maximum utility approaches zero as queue length tends to infinity, the maximum stability region can be retained. However, when the utility gap is proportional to the maximum utility, only a reduced stability region can be achieved. In this case, we can still provide a lower bound for the achievable stability region under rate allocation policies satisfying the information delay and the utility gap conditions. These results characterize the stability of a broad class of suboptimal rate allocation policies.

When information delay is bounded, since suboptimal rate allocations with a small enough utility gap are capable of achieving the maximum stability, we investigate the potential benefits of allowing such a utility gap, i.e., the upside of under-optimizing utility objectives. It is clear that by deliberately under-optimizing a utility, we can improve network performance in other metrics. What remains unclear is precisely how much improvement we can possibly achieve by under-optimizing the utility with a given allowable gap. We formulate the potential performance improvement as a function of utility gap, and derive a first-order approximation for the tradeoff curve based on local sensitivity (shadow price) analysis. This formulation generalizes that in [11], which focuses on how network performance can be affected by the choice of α-fair utility and assumes that optimality always holds. Our result not only illustrates the potential benefits of under-optimizing a utility, but also quantitatively characterizes the tradeoff between sacrificing utility value and improving other network performance metrics, e.g. link saturation. Our analysis can be easily extended beyond the class of α-fair utilities.

The results in this paper explore a new perspective to look at suboptimal solutions of the utility maximization problem. We show that suboptimal rate-allocation policies may not always be inferior in performance. More precisely, by under-optimizing a utility and allowing a certain optimization gap, we can still retain the maximum flow-level stability and obtain network performance improvements in other metrics. The rest of the paper is organized as follows: In Section II, we introduce the class of utility functions considered in this paper and define the utility gap for suboptimal rate allocations. Two stability results are stated next. In Section III.A, a sufficient condition on utility gap and information delay for achieving the maximum flow-level stability is provided. In Section III.B, when the utility gap is proportional to the maximum utility, we show that the achievable stability region can be strictly smaller, and we further obtain a lower bound for all achievable stability regions. In Section IV, we analyze the tradeoff between the utility gap and link saturation. Results based on sensitivity analysis are derived to measure the benefits of under-optimizing α-fair utility. Simulation results are provided at the end of section III and IV respectively. For smoother flow of the main results, we collect all the proofs in the appendices.

Throughout this paper, we use the following notations: Vectors are denoted in small letters, e.g., $x$, with their $i$th component denoted by $x_i$. Matrices are denoted by capitalized letters, e.g., $A$, with $A_{ij}$ denoting the $(i,j)$th component. Vector inequalities denoted by $x \geq y$ are considered component-wise. The superscript $(\cdot)^T$ denotes the matrix transpose. $\mathbb{P}(M)$ is the probability of an event $M$. We use $\mathcal{R}$ to denote a set of vectors and $\mathcal{R}$ for its interior.

II. Utility Maximization and Gap

Consider a communication network shared by a set of data flows, which belong to $N$ distinct flow classes. We refer to the vector $x = [x_1, \ldots, x_N]$ as the network state, where $x_i$ denotes the number of flows of class $i$ that remain in the system. The problem of rate allocation is to determine the total rate allocated to class-$i$ flows at state $x$, denoted by $\phi_i(x)$. Rate $\phi_i(x)$ is equally shared by all class-$i$ flows, each assigned rate $\phi_i(x)/x_i$. We refer to the vector $\phi(x) = [\phi_1(x), \ldots, \phi_N(x)]$ as the rate allocation at state $x$. Let $\mathcal{R} \subset \mathbb{R}^N$ be a set of all possible rate allocation vectors. Rate allocation $\phi(x)$ is restricted by $\phi(x) \in \mathcal{R}$, which means that the network can support the rate vector $\phi(x)$. In this paper, we only require the set $\mathcal{R}$ to be convex, compact, and coordinate-convex\footnote{A set $\mathcal{R}$ is coordinate convex when the following is true: If $x \in \mathcal{R}$, then $y \in \mathcal{R}$ for all $y$ component-wisely less than $x$.}, which holds in many settings, e.g., [3], [6].

Various network rate control policies can be derived as solving some utility maximization problem with different utility functions:

$$\phi_{\text{opt}}(x) = \arg \max_{\phi \in \mathcal{R}} \sum_{i,x_i \geq 1} x_i U_i(\phi_i(x)/x_i)$$

where $U_i(\cdot)$ is a utility function for flow class $i$. In this paper, we assume that the function $U_i(\cdot)$ is continuous and twice differentiable on $(0, +\infty)$. In addition, we assume the utility functions satisfy the following conditions:

(a) $U_i(z) \geq 0 \forall z$ and $U_i(0) = 0$, or $U_i(z) \leq 0 \forall z, i$.

(b) $U_i(z)$ is concave and monotonically increasing.

(c) $\lim_{z \to 0} U_i'(z) = \infty, \forall i$.

(d) There exists $s$, s.t. $\frac{z U_i'(z)}{U_i(z)} \geq -s, \forall z, i$.

Assumptions (a) and (b) are commonly used in the literature [4]. Assumption (c) can be interpreted as one that prevents
starvation, since it implies that slope of the utility function increases to infinity as the rate of the flow class approaches zero. Condition (d) requires that the utility function does not have sharp changes. One example of such utility functions satisfying assumptions (a-d) is a class of the well-used α-fair utility functions [12], defined by

\[ U_i(z) = \begin{cases} \frac{z^{1-\alpha}}{1-\alpha}, & \alpha > 0 \text{ and } \alpha \neq 1 \\ \log z, & \alpha = 1 \end{cases} \quad (2) \]

where \( \alpha \) is a positive constant. It is easy to verify that the assumptions (a-d) are satisfied with \( s = \alpha \). Parameter \( \alpha \geq 0 \) models the level of fairness, which includes several special cases such as proportional fairness and max-min fairness. For example, maximizing the total utility corresponds to minimizing weighted throughput as \( \alpha \to 0 \), weighted proportional fairness as \( \alpha = 1 \), minimum potential delay as \( \alpha = 2 \) and max-min fairness as \( \alpha \to \infty \) [3].

It may be impractical to solve the rate allocation problem (1) optimally for all network states. In this paper, we consider a more general scenario where rate allocations are not optimal and thus could possibly reduce network performance. This work is motivated by the following two issues in practical networks: First, all practical rate allocation policies are subject to a positive delay due to the time requirement for gathering network information and for algorithm convergence. In other words, the practical rate allocation vector \( \phi(\hat{x}) \) can at best correspond to the optimal rate allocation \( \phi_{\text{opt}}(\hat{x}) \) for some vector \( \hat{x} \), where \( \hat{x} \neq x \) is the network state observed by a practical rate allocation policy. Second, due to computational overhead or requirement for distributed computation, even given the network state \( \hat{x} \), a practical rate allocation policy may still not be able to solve the NUM problem (1) optimally. In other words, there may exist a utility gap due to the suboptimality of the rate allocation policy. We quantify the suboptimality of a practical rate allocation \( \phi(\hat{x}) \) with respect to state \( \hat{x} \) by a utility gap as follows

\[ \Delta(\hat{x}) = \sum_{i: \hat{x}_i \geq 1} \hat{x}_i U_i \left( \frac{\phi_{\text{opt},i}(\hat{x}_i)}{\hat{x}_i} \right) - \hat{x}_i U_i \left( \frac{\phi_i(\hat{x})}{\hat{x}_i} \right). \quad (3) \]

The gap \( \Delta(\hat{x}) \) measures the difference between suboptimal rate allocations and the optimal allocation, caused only by the imperfect computation of the rate allocation algorithm. Given certain conditions on the utility gap \( \Delta(\hat{x}) \) and a model for the observed network state \( \hat{x} \), in section III we will characterize the stability region of networks with an arbitrary suboptimal rate allocation policy. In section IV, we will formulate and analyze the tradeoff between utility gap and two network performance metrics.

### III. Information Delay and Stability

Consider a network where class-\( i \) flows arrive as a Poisson process of intensity \( \lambda_i \geq 0 \) and have i.i.d. exponential file sizes of mean \( 1/\mu_i \). A flow is considered to have left the network when its file transfer is completed. Let \( \rho_i = \lambda_i/\mu_i \) be the traffic intensity of class-\( i \) flows. We formulate a stochastic process of the network state, denoted by \( x(t) \). To model the observed network state \( \hat{x}(t) \) at time \( t \), we introduce an information delay process \( \tau(t) \), such that the observed state of class-\( i \) flows is given by \( \hat{x}_i(t) = x_i(t - \tau_i(t)) \), for an information delay \( \tau_i(t) \). Since network information may arrive out of order, information delay \( \tau_i(t) \) is not necessarily increasing over time \( t \). Therefore, the rate allocation at time \( t \) is given by

\[ \phi(\hat{x}(t)) = \phi(x_i(t - \tau_i(t))). \quad (4) \]

We say that flow-level stability holds under rate allocations \( \phi(\hat{x}(t)) \) if there exists a positive non-decaying function \( f(z) \) with \( \lim_{z \to \infty} f(z) = \infty \), such that the resulting queue-length process \( x(t) \) satisfies

\[ \limsup_{T \to \infty} \frac{1}{T} \int_0^T E \left[ \sum_{i=1}^N f(x_i(t)) \right] \, dt < \infty. \quad (5) \]

The stability condition in (5) is usually refer to as stability in the mean [10], [15]. If we further assume that the rate allocation policy and the information delay make the queuing system an aperiodic Markov chain with a single communicating class (which is the case where there is no information delay and utility gap), then the stability in the mean property in (5) further implies that the Markov chain is positive recurrent [15].

**Remark 1:** For ease of exposition, in the above model we have assumed that the actual rate allocation is a given function \( \phi(\hat{x}) \) of the observed network states \( \hat{x} \). The result of this paper could also be extended to the case when this function \( \phi(\cdot) \) is replaced by a function \( \phi_i(\cdot) \) that also varies with time. In that case, \( \Delta(\hat{x}) \) in (3) can be defined as an upper bound on the utility gap caused by the functions \( \phi_i(\cdot) \) over all time \( t \).

When there is no utility gap or information delay, flow-level stability has been studied in [2], [3], [4], [5] using first-order Markov models. If the feasible rate region \( \mathcal{R} \) is compact and convex, and the optimal rate allocation \( \phi_{\text{opt}}(x(t)) \) that maximizes problem (1) with an \( \alpha \)-fair utility is implemented at each time \( t \), it is proven that such rate allocation achieves the maximum stability region (i.e. the interior of the feasible rate region \( \mathcal{R} \)). In other words, a sufficient and necessary condition for stability is that the traffic intensity vector must belong to the interior of the feasible rate region: \( \rho \in \mathcal{R} \).

However, due to the information delay, the rate allocation \( \phi(\hat{x}(t)) \) depends on previous network states at time \( t - \tau_i(t) \), for \( i = 1, \ldots, N \). Thus, the usual method of flow-level stability analysis in [2], [3], [4], [5], which requires a first-order Markov model of the queue-length process, is inadequate. In this paper, we consider non-optimal rate allocation policy and prove stability by evaluating an expected Lyapunov drift. Let \( h > 0 \) be a small time interval. The evolution of the \( i \)-th queue is described by the following equation:

\[ x_i(t + h) = x_i(t) + a_i(t, h) - d_i(t, h), \quad (6) \]

where \( a_i(t, h) \) is the number of flows arriving to flow class \( i \) during time \( t \) to \( t + h \) and \( d_i(t, h) \) is the number of departing flows. Due to utility gap, the departure rate now depends on the sub-optimal rate allocation \( \phi(\hat{x}(t)) \), given an observed state \( \hat{x}(t) = x_i(t - \tau_i(t)) \). In this section, we will derive a sufficient condition for achieving maximum stability. When the condition is not satisfied, we prove that the achievable stability region may be strictly smaller than the feasible rate.
A. A Sufficient Condition for Maximum Stability

**Theorem 1:** For an arbitrary suboptimal rate allocation \( \phi(\tilde{x}) \), if the information delay \( \tau(t) \) is uniformly bounded by a constant \( \Omega > 0 \) and the order of the utility gap (which is non-negative due to the definition in (3)) caused by the imperfectness of rate allocation algorithm is less than the order of the optimal utility when the number of active flows grows large, i.e.,

\[
\limsup_{\max \hat{x}_i \to \infty, i \to \infty} \frac{\Delta(\hat{x})}{x_i} = 0,
\]

(7)

then the network is stable if the traffic condition \( \rho \in \mathcal{R} \) is satisfied, i.e., the maximum stability region can be obtained.

There are two key difficulties in the proof. First, to account for the utility gap, we need to derive a relationship between the traffic intensity and the sub-optimal rate allocation \( \phi(\tilde{x}(t)) \), which is of a stronger form than those used in [3], [4] (see Lemma 4 in Appendix B). Second, the information delay leads to a further gap between \( \phi(\tilde{x}(t)) \) and \( \phi(x(t)) \). We need to carefully bound the effect of this gap, especially when the difference between \( x(t) \) and \( \tilde{x}(t) \) is large (see the detailed comment after Lemma 5 in Appendix B). We refer the readers to Appendix B for the detailed proof of Theorem 1.

**Remark 2:** Theorem 1 establishes a sufficient condition for achieving maximum stability: It shows that it is not necessary to solve the optimal solution to the utility optimization problem (1) and to require perfect information on the network states \( x(t) \). The condition (7) could be used to construct stopping rules for designing sub-optimal rate-control policies that can achieve maximum stability. We note that such a stopping rule could be designed without knowledge of the optimal rate-allocation \( \phi_{\text{opt}} \). For example, consider a rate controller based on a dual algorithm for solving the NUM problem (1). Note that the optimal utility is upper bounded by the objective value of the dual problem, which can be easily calculated from the current dual variables. Similarly, the current primal variables can be used to generate a feasible rate-allocation and to compute the achievable utility values. Therefore, if the controller stops whenever the difference between the dual objective value and the achievable utility values is smaller than a threshold \( \Gamma \), then we can show that the resulting utility gap (between the optimal utility and the achievable utility value) is also bounded by \( \Delta(\hat{x}(t)) \leq \Gamma \).

Condition (7) is then satisfied, since:

\[
\limsup_{\max \hat{x}_i \to \infty, i \to \infty} \frac{\Delta(\hat{x})}{x_i} \leq 0,
\]

(8)

Hence, the maximum stability region can be achieved by this stopping rule according to Theorem 1. Although the above example assumes a centralized controller to check the stopping rule, we envision that distributed versions of the stopping rule are also possible, which we leave for future work. The condition in (7) is useful because it provides the guideline for designing such stopping rules for general networks.

B. A Lower Bound on the Achievable Stability Region

When the condition (7) in Theorem 1 is not satisfied and the utility gap is on the same order as that of the optimal utility, the achievable stability region could be smaller than the feasible rate region, even if the delay \( \tau(t) \) is zero.

**Proposition 1:** There exists a suboptimal rate allocation \( \phi(\tilde{x}) \) such that the utility gap is on the same order as the order of the optimal utility and the information delay is zero, i.e., for some constant \( \eta \in (0, 1) \),

\[
\Delta(\hat{x}) \leq \eta,
\]

(9)

but the achievable stability region is strictly smaller than \( \mathcal{R} \), even if the rate vector \( \phi(\tilde{x}) \) is Pareto-optimal (i.e. \( \phi(\tilde{x}) \) lies on the boundary of the feasible rate region).

Proposition 1 implies that if the utility gap is large, there exists a suboptimal rate allocation policy whose achievable stability region is strictly smaller than the interior of the feasible rate region \( \mathcal{R} \), regardless of the information delay. Raised from this example, a challenge is to answer the question: what is the minimum stability region that a suboptimal rate allocation policy can achieve given that condition (8) is satisfied?

In the next theorem, we show that (1 - \( \eta \))^\frac{1}{1-s} \mathcal{R} is a lower bound of all achievable stability regions, if the ratio of the utility gap and the optimal utility is asymptotically bounded by a constant \( \eta < 1 \) as the number of active flows grows large. This lower bound is tight in the sense that there exists a suboptimal rate allocation policy whose stability region is exactly (1 - \( \eta \))^\frac{1}{1-s} \mathcal{R}.

**Theorem 2:** For an arbitrary suboptimal rate allocation \( \phi(\tilde{x}) \), if the information delay is uniformly bounded by a constant and the order of the utility gap is the same as that of the optimal utility, i.e.,

\[
\limsup_{\max \hat{x}_i \to \infty, i \to \infty} \frac{\Delta(\hat{x})}{x_i} \leq \eta,
\]

(9)

then the achievable stability region is lower bounded by (1 - \( \eta \))^\frac{1}{1-s} \mathcal{R}, where \( s \) is the parameter defined in Assumption (d) in Section II. There also exists a suboptimal rate allocation policy satisfying (9) whose stability region is exactly (1 - \( \eta \))^\frac{1}{1-s} \mathcal{R}, i.e., the lower bound is tight.

According to Lemma 1, if the utility function satisfies Assumptions (a-d) and is positive, then we have \( U(a) \geq (\frac{a}{b})^{1-s} U(b) \), i.e. the utility function \( U(\cdot) \) shows a polynomial growth rate with exponent \( 1-s \). This implies that when
the order of the utility gap is the same as that of the optimal utility, the dependence of the performance bound on $s$ is due to the fact that utility function has a polynomial growth rate with the exponent $|1 - s|$. Hence, when the ratio between the utility gap and the optimal utility is $\eta$, the difference in the rate allocation can be on the order of $(1 - \eta)^{1/\eta}$.

Remark 3: Theorem 2 provides a lower bound for achievable stability regions. Of course, under condition (9), there might still exist certain suboptimal rate allocation policies that are capable of achieving the maximum stability. However, the lower bound in Theorem 2 is tight in the sense that there exist a suboptimal rate allocation policy with zero information delay and its stability region is exactly $(1 - \eta)^{1/\eta} R$. Proposition 1 and Theorem 2 together characterize the stability of a broad class of suboptimal rate allocation policies.

C. Numerical Examples

Fig. 1. A ring network with ten users and ten flow classes.

Consider a ring network with $N = 10$ flow classes and $L = 10$ unit-capacity links as shown in Fig.1. Flow class $i$ is initiated by user $i$ and contains $x_i(t) \geq 0$ active flows at time $t$. Let $\mathcal{R}$ be the shortest-distance routing matrix for this ring network. For an $\alpha$-fair utility function with $\alpha = 1$ (i.e. a logarithmic utility), we compute the optimal rate allocation $\phi_{\text{opt}}(x(t))$ for each time $t$ and then perturb it randomly to construct a set of suboptimal rate allocations, such that the resulting information delay and utility gap are constants:

$$\tau_0 = \tau_i(t) \text{ and } \Delta_0 = \Delta(\hat{x}(t)), \forall t.$$ (10)

According to Remark 1, since both utility gap and information delay are constants, the suboptimal rate allocation policy $\phi(\cdot)$ constructed above will achieve the maximum stability region that is equal to the interior of the feasible rate region $\mathcal{R} = \{\phi \in \mathbb{R}_+^N : R\phi \leq \mathbf{1}, \phi \geq \mathbf{0}\}$, where $\mathcal{R}$ is the $L \times N$ routing matrix: $R_{li} = 1$ if class-$i$ flows use link $l$, and $R_{li} = 0$ otherwise. Therefore, this rate region is a polytope determined by a set of linear link-capacity constraints.

Figure 2 illustrates flow-level stability of the network under different suboptimal rate allocation policies for $(\Delta_0, \tau_0) = \{(0, 5), (5, 0), (0, 2), (5, 2)\}$ respectively, by plotting the average total queue length vs. traffic load. In this simulation, we assume that the flow arrival rates for all flow classes are equal, i.e. $\rho_i = \rho_0$ for $i = 1, \ldots, 10$. For $\rho_0 \in [0, \frac{1}{2}]$, we have $\rho = \rho_0 \cdot \mathbf{1} \in \mathcal{R}$, which implies that the expected queue-length should remain finite. Figure 2 also shows that the average queue-length of a suboptimal policy $\phi(\hat{x}(t))$ approaches that of the optimal rate allocation policy, when utility gap and information delay decrease.

IV. Utility Gap and Network Performance

Section III showed that a suboptimal rate allocation policy that under-optimizes a certain utility still achieves the maximum stability region. Since each utility function is designated to capture one particular network objective, allowing a nonzero utility gap (or, equivalently, under-optimizing the utility) gives us freedom to potentially improve other network performance objectives, such as the maximum link saturation discussed in this section. More precisely, there exists a tradeoff between the utility gap and the maximum network performance improvement we can potentially achieve. In this section we first provide a formulation of this tradeoff. Then we develop an approximation of the tradeoff curve based on local sensitivity analysis. To obtain closed-form solutions, we focus on $\alpha$-fair utility functions in this section, although our result can be extended to all concave utility functions. Our approach is different from [11], which is restricted to a throughput-fairness tradeoff for optimal solutions only. In contrast, in this section, we addresses the following question pertaining to suboptimality: by under-optimizing an utility with gap $\Delta$, what is the maximum performance improvement we can possibly achieve?

A. Model and Analysis

We focus on the following model for wireline networks, which is an important special case of the model described in section III. Consider a network of $L$ links, indexed by $l$, each with a finite link capacity $c_l$. Then the feasible rate regions are defined by $\mathcal{R} = \{\phi : R\phi \leq c, \phi \geq 0\}$, where $c$ is the vector of link capacities and $R$ is the $L \times N$ routing matrix: $R_{li} = 1$ if class-$i$ flows use link $l$, and $R_{li} = 0$ otherwise. At
each state $x$, the optimal rate allocation is obtained by solving problem (1) with $\alpha$-fair utility, i.e.,

$$
\max_{\phi} \sum_{i=1}^{N} x_i^\alpha \phi^{1-\alpha} \frac{1}{1-\alpha} \tag{11}
$$

subject to $R \phi \leq c$, $\phi \geq 0$

Let $\phi_{opt}$ be the optimal rate allocation that solves the maximization problem (11). We say a rate allocation $\phi$ under-optimizes the $\alpha$-fair utility by a gap $\Delta$ if

$$
\Delta = U_{opt} - \sum_{i=1}^{N} x_i^\alpha \phi_{opt}^{1-\alpha} \frac{1}{1-\alpha} \tag{12}
$$

where $U_{opt} = \sum_{i=1}^{N} x_i^\alpha \phi_{opt}^{1-\alpha} / (1 - \alpha)$ is the optimal utility achieved by rate allocation $\phi_{opt}$. Since the $\alpha$-fair utility is designated for achieving fairness, under-optimizing the $\alpha$-fair utility with a gap $\Delta$ relaxes the maximization problem (11). Thus, it gives freedom to system designers to potentially improve other network performance objectives, such as maximum link saturation. However, it is unclear how much performance improvement we can achieve by under-optimizing the $\alpha$-fair utility with a given allowable gap. For example, if we prepare to sacrifice 5% of the utility, how much is the reduction of link saturation we could expect in return? We formulate this type of tradeoff functions and provide a local sensitivity analysis based on examining the Karush-Kuhn-Tucker (KKT) conditions at the optimum allocation.

We consider the maximum link saturation as a network performance metric, defined by

$$
Z = \max_{t \in L} \frac{\sum_{i=1}^{N} R_{it} \phi_i}{c_t} \tag{13}
$$

By under-optimizing the utility, it is possible to reduce the maximum link saturation and then balance the network traffic over all links, an important goal in the operation of large networks by the Internet Service Providers. Moreover, reducing $Z$ could potentially minimize the occurrence of ‘bottleneck’ links in the network, reduce packet delay, and also make the network more robust to link capacity fluctuation and traffic bursts.

To characterize the optimal tradeoff between the utility gap and the maximum link saturation, we compute the minimum $Z$ that can be achieved by under-optimizing the $\alpha$-fair utility with a designated gap. This tradeoff function $Z(\Delta)$ can be formulated as follows

$$
Z(\Delta) = \min_{\phi} \max_{t \in L} \frac{\sum_{i=1}^{N} R_{it} \phi_i}{c_t} \tag{14}
$$

subject to $R \phi \leq c$, $\phi \geq 0$

$$
\sum_{i=1}^{N} x_i^\alpha \phi_i^{1-\alpha} \frac{1}{1-\alpha} \geq U_{opt} - \Delta
$$

Remark 4: For the tradeoff function defined by (14), it is easy to see that increasing utility gap relaxes the constraint set of the optimization problem, and leads to a smaller optimal objective value. Thus, the maximum link saturation $Z$ is a monotonically decreasing function of the utility gap $\Delta$. Furthermore, it is easy to verify that the optimization problem (14) is convex.

Now we conduct a local sensitivity analysis for the saturation-utility tradeoff defined by optimization problem (14). Again, we make the assumption that the active constraint set in the problem (14) is unchanged when the utility gap $\Delta$ is perturbed locally. The main result is summarized in the next theorem. We denote $Z_0$ as the link saturation for $\Delta = 0$.

**Theorem 3:** When the utility gap is small, the saturation-utility tradeoff function can be approximated using its first order expansion:

$$
Z - Z_0 = \left[ \frac{dZ}{d\Delta} \bigg|_{\Delta=0} \right] \Delta + o(\Delta). \tag{15}
$$

The first order derivative (shadow price) of the saturation-utility tradeoff function is given by

$$
\frac{dZ}{d\Delta} \bigg|_{\Delta=0} = -\frac{1}{c^2 (RD^{-1}R^T)^{-1}} D, \tag{16}
$$

where $D = \alpha \cdot \text{diag}\{x_1^\alpha \phi_{opt,1}^{1-\alpha}, \ldots, x_N^\alpha \phi_{opt,N}^{1-\alpha}\}$ is a diagonal matrix.

**B. Numerical Examples**

In this subsection, we plot the tradeoff curve and its first-order approximations obtained in Theorem 3 for the ring network described in Section III.C. Since all links have unit capacity, the feasible rate region is given by $R = \{\phi : R \phi \leq 1, \phi \geq 0\}$, where $R$ is the routing matrix for the ring network. Let $x_i$ denote the number of active flows for source $i$. We can solve the convex optimization problem (14) for varying utility gap $\Delta$ to obtain the exact tradeoff curve $Z(\Delta)$, which is plotted in Figure 3 using solid lines. We assume that the number of active flows is $x_1 = 10, \forall i$. A proportional fairness utility function corresponding to $\alpha = 1$ is considered.

When the utility gap $\Delta$ is small, the saturation-utility tradeoff curves can be approximated by its first order expansions given by (15). Using the closed-form solution in Theorem 3, we compute the first order gradient $\frac{dZ}{d\Delta} \bigg|_{\Delta=0} = -0.010$. Thus the tradeoff curve can be approximated by

$$
T(\Delta) \approx Z(0) - 0.01\Delta \tag{17}
$$

Figure 3 shows that the saturation-utility tradeoff defined in (14) can be well approximated by its first order expansion, given by the closed-form expression in Theorem 3. This tradeoff curve allows us to predict how much performance improvement we can possibly achieve by under-optimizing the utility with a designated small utility gap. For example, if we under-optimize the utility by 1%, i.e., $\Delta = 1\% |U_{opt}| = 0.312$, it is clear from equation (17) that a link saturation reduction of 0.31% ($Z - Z_0 = -0.0031$) could be expected in return. This result not only illustrates the potential benefits of under-optimizing an $\alpha$-fair utility, but also quantitatively characterizes the tradeoff between sacrificing utility value and achieving network performance improvement. Whether this particular tradeoff is worth making or not depends on operator’s preference, but it is important to provide the choices of tradeoff through results like those in this section.
V. CONCLUDING REMARKS

Suboptimal resource allocation with a utility gap is simply an inevitable phenomenon in real networking. Fortunately, it may still be able to maintain stability region and even enhance other network performance metrics. Intuition on stability and utility-versus-saturation tradeoff are quantified with closed-form expressions in this paper. There are open questions to be addressed in studying the impact of suboptimal solutions (a-d), then we have (i) 

\[
\frac{U(x)}{U'(x)} \geq -\frac{1}{\nu}
\]

Choose \(a \geq b > 0\) and integrate both side of the inequality from \(b\) to \(a\). We obtain \(U'(a) \geq (\frac{a}{b})^{s-a} U'(b)\). By switching \(a\) and \(b\), we can prove part (b). To show part (c), if the utility function is negative (i.e. case 2 in Assumption (a)), we fix \(b\) in the inequality in part (a) and integrate it from \(a = b\) to \(a = +\infty\), i.e.

\[
b^a U'(b) \int_b^{+\infty} \frac{1}{y^s} dy \leq U(\infty) - U(b) \leq -U(b) \tag{18}
\]

where \(U(\infty)\) exists because the utility function is monotonically increasing and upper bounded by zero as in Assumption (a). This implies that the integration on the left hand side also exists and thus \(s > 1\). We can derive \(-\frac{U'(b)}{U(b)} \leq \frac{s-1}{b} \cdot \frac{s}{b}\). Integrating it again from \(a\) to \(b \geq a\), we obtain \(U(a) \leq \left(\frac{a}{b}\right)^{1-s} U(b)\), which is the desired result. Similarly, when the utility function is positive, we consider the integral of \(U'(a) \geq \left(\frac{a}{b}\right)^{1-s} U'(b)\) from \(b = 0\) to \(b = a\) and derive the result in Lemma 1.

Lemma 2: For \(C \geq 0\), there exist a constant \(K_c\) such that \((1 + \frac{C}{z})^s \leq 1 + \frac{K_c}{z} x\) holds for all \(z \geq 1\), and \((1 - \frac{C}{z})^s \geq 1 - \frac{K_c}{z} x\) holds for all \(z \geq C\).

Proof: The proof is straightforward by comparing the first order derivatives with respect to \(z\).

B. Proof of Theorem 1.

Proof: We first sketch the main steps of the proof. To prove stability, we define a Lyapunov function \(V(x(t))\) and analyze its expectation \(W(t) = \mathbb{E}[V(x(t))]\) as a function of time\(^2\). We first derive an expression for \(\dot{W}(t)\), the drift of the expected value of the Lyapunov function. Here we need to use the Dominated Convergence Theorem in order to exchange the order of a limit and an expectation (Lemma 3). Then, using Lemmas 4-6, we upper bound the drift \(\dot{W}(t)\) by a negative function of the network state \(x(t)\), plus some positive constants. Finally, integrating the drift \(\dot{W}(t)\) and its upper bound from time \(t = 0\) to \(t = T\) establishes the stability condition in (5) and completes the proof. For ease of presentation, in this section we present the main flow of the proof with statements of Lemmas 3-6. The detailed proofs are summarized in Appendices C-F. We remind the readers that we will use the Lemmas in Appendix A.

Consider the following Lyapunov function

\[
V(x(t)) = \sum_{i=1}^{N} V_i(x_i(t)) = \sum_{i=1}^{N} \sum_{n=1}^{\infty} \frac{1}{\mu_i} U_i' \left( \frac{c \mu_i}{n} \right), \tag{19}
\]

where \(c > 0\) is a constant defined later in the proof. Let \(F_i = \sigma \{ x(u), u \leq t \} \) denote the \(\sigma\)-field generated by the history up to time \(t\). For \(h > 0\), we derive

\[
\lim_{h \to 0} \frac{W(t+h) - W(t)}{h} = \lim_{h \to 0} \frac{\mathbb{E}[\mathbb{E}[V(x(t+h))|F_t] - V(x(t))]}{h} \tag{20}
\]

\[
= \lim_{h \to 0} \mathbb{E} \left[ \sum_{i=1}^{N} \sum_{n=-X_i(t)}^{\infty} \frac{\mathbb{P}(x_i(t+h) - x_i(t) = n|F_t)}{h \mu_i} \cdot (V_i(x_i(t) + n) - V_i(x_i(t))) \right].
\]

\(^2\) A similar problem with feedback delay is treated in [10] although the model there does not involve any flow-level dynamics.
In order to move the limit (as \( h \to 0 \)) inside the expectation and the summation on the right hand side of (20), we will use the Dominated Convergence Theorem and the following Lemma.

**Lemma 3:** There exists an integrable function \( g(x_i(t)) \) such that, for all \( 0 < h < 1 \) and \( m \geq 0 \),
\[
\left| \sum_{n=0}^{+\infty} \frac{\mathbb{P}(x_i(t+h) - x_i(t) = n \mathcal{F}_t)}{h \mu_i} \right| \leq g(x_i(t))
\]
(21)

This lemma provides the bound needed for the Dominated Convergence Theorem to hold. Therefore, we can move the limit inside the expectation on the right hand side of (20). Due to the orderliness property of Poisson process (which implies that arrivals do not occur simultaneously), we can take the limit (as \( h \to \infty \)) and narrow down the conditional probability terms in (20) to get
\[
\hat{W}(t) = \sum_{i=1}^{N} \mathbb{E}\left[ \lambda_i(V_i(x_i(t) + 1) - V_i(x_i(t))) \right] - \sum_{i=1}^{N} \mathbb{E}\left[ \phi_i(\hat{t}(t))\mu_i(V_i(x_i(t)) - V_i(x_i(t) - 1)) \right].
\]

Using \( \rho_i = \lambda_i/\mu_i \) and plugging in the Lyapunov in (19) function into above, we obtain
\[
\hat{W}(t) = \sum_{i=1}^{N} \mathbb{E}\left[ \rho_i U_i' \left( \frac{c \rho_i}{x_i(t) + 1} \right) \right] - \sum_{i=1}^{N} \mathbb{E}\left[ \phi_i(\hat{t}(t))U_i' \left( \frac{c \rho_i}{x_i(t)} \right) 1_{\{x_i(t) \geq 1\}} \right].
\]
(22)

**Derive a bound for (23).** In order to bound \( \hat{W}(t) \) by a negative function of \( x(t) \), we first derive a bound for the second summation (23) and replace the rate allocation \( \phi(\hat{x}(t)) \) by a function of traffic intensity \( \rho_i \). The resulting bound will then have a form similar to (22) so that we can compare the difference. We make use of the following lemma:

**Lemma 4:** For any traffic intensity \( \rho \in \mathcal{R} \) and constant \( C > 0 \), if the suboptimal rate allocation \( \phi(\hat{x}(t)) \) satisfies the utility gap condition in (7), there exist positive constants \( \gamma > 0 \) and \( \epsilon > 0 \) such that, for all \( |\epsilon| < C \) and for any network state satisfying \( \max_i \hat{x}_i(t) > \gamma \), the following inequality holds:
\[
\sum_{i:x_i(t) \geq 1} \rho_i(1 + \epsilon)^3 U_i' \left( \frac{c \rho_i}{x_i(t) + 1} \right) - \phi_i(\hat{t}(t))U_i' \left( \frac{c \rho_i}{x_i(t) + r} \right) \leq 0,
\]
where \( c = (1 + \epsilon)^4 \) is a constant.

We choose \( \tau = x(t) - \hat{x}(t) \) in (24). Note that in order to use Lemma 4 to bound (23), we need the network state to satisfy \( \max_i \hat{x}_i(t) > \gamma \) and the quantity \( |x(t) - \hat{x}(t)| \) to be bounded by \( C \). Toward this end, we define the event \( \mathcal{E}_t = \{|x(t-u) - x(t-\Omega)| > C/2, \forall u \in [0, \Omega] \} \), i.e., it is the event that the maximum change of network state within time \( t - \Omega \) to \( t \) is bounded by \( C/2 \). Since the information delay is bounded by \( \tau(t) < \Omega \), event \( \mathcal{E}_t \) implies the following:
\[
||x(t) - \hat{x}(t)||_1 \leq ||x(t) - x(t - \Omega)||_1 + ||\hat{x}(t) - x(t - \Omega)||_1 \leq C
\]

Therefore, we can bound (23) by
\[
- \sum_{i=1}^{N} \mathbb{E}\left[ \phi_i(\hat{t}(t))U_i' \left( \frac{c \rho_i}{x_i(t)} \right) 1_{\{x_i(t) \geq 1\}} \right] \leq A_1 - \sum_{i:x_i(t) \geq 1} \mathbb{E}\left[ \rho_i U_i' \left( \frac{c \rho_i}{x_i(t)} \right) 1_{\{x_i(t) \geq 1\}} \right]
\]
(25)

Note that \( \hat{x}_i(t) \leq C \) if \( x_i(t) = 0 \) under event \( \mathcal{E}_t \). To change the summation from \( \{i:x_i(t) \geq 1\} \) to \( \{i: \hat{x}_i(t) \geq 1\} \) in the first step above, we have added
\[
A_1 = \sum_{i=1}^{N} \psi U_i' \left( \frac{c \rho_i}{C} \right).
\]
(26)
where we choose \( \phi_i(\hat{t}(t)) \leq \psi \) since the feasible rate region \( \mathcal{R} \) is compact. The last step of (25) follows from Lemma 4. In order to compare the difference with (22), we still need to replace the \( \hat{x}(t) \) on the right hand side by \( x(t) \). Let \( \mathcal{G} = \{x(t) : \max_i x_i(t) \leq \xi \} \) be a bounded region with \( \xi = \gamma + 2C \).

We can prove the following result:

**Lemma 5:** There exists \( A_2, A_3 > 0 \) such that the right hand side of (25) can be further bounded by
\[
-(1 + \epsilon)^2 \sum_{i: \hat{x}_i(t) \geq 1} \mathbb{E}\left[ \rho_i U_i' \left( \frac{c \rho_i}{x_i(t)} \right) 1_{\{x_i(t) \geq 1\}} \right] \leq A_2 + 3A_3 - \sum_{i:x_i(t) \geq 1} \mathbb{E}\left[ \rho_i U_i' \left( \frac{c \rho_i}{x_i(t)} \right) 1_{\{x_i(t) \geq 1\}} \right].
\]

The intuition behind Lemma 5 is that: when \( \mathcal{E}_t \) occurs, the absolute difference between \( x(t) \) and \( \hat{x}(t) \) is bounded by \( C \). If in addition the network state is large, then the relative difference between \( x(t) \) and \( \hat{x}(t) \) will be small, and hence the corresponding values of \( U_i' (\cdot) \) will be close to each other. Proving Lemma 4 turns out to be non-trivial. The challenge is that \( \hat{t}(t) \) and \( \mathcal{E}_t \) in (27) are dependent. To handle this difficulty, we consider the \( \sigma \)-field \( \mathcal{F}_t \cap \Omega = \sigma \{x(u), u \leq t-\Omega\} \). If we introduce the network state \( x(\Omega) = x(t-\Omega) \) as an auxiliary variable, then \( \hat{t}(t) \) and \( \mathcal{E}_t \) can be bounded with respect to \( x(\Omega) \), separately. Details are provided in Appendix D.

**Derive a bound for (22).** We next provide a corresponding bound for (22) to compare with (27). Note that region \( \mathcal{G} \) is compact. As part of the proof of Lemma 5, \( A_2 \) is chosen in (49) such that
\[
\sum_{i=1}^{N} \rho_i U_i' \left( \frac{c \rho_i}{x_i(t) + 1} \right) 1_{\{x_i(t) \in \mathcal{G}\}} \leq A_2 < \infty.
\]
(28)

Applying \( A_2 \) to (22), we have
\[
\sum_{i=1}^{N} \mathbb{E}\left[ \rho_i U_i' \left( \frac{c \rho_i}{x_i(t) + 1} \right) 1_{\{x_i(t) \in \mathcal{G}\}} \right] \leq A_2.
\]
(29)
where \(A_1\) is added to change the summation to \(i : x_i(t) \geq 1\). To further bound (29), we make use of Lemma 1 (b), with \(a = c \rho_i / x(t)\) and \(b = c \rho_i / (x(t) + 1)\):

\[
\begin{aligned}
&\sum_{i : x_i(t) \geq 1} \mathbb{E} \left[ \rho_i U_i' \left( \frac{c \rho_i}{x_i(t) + 1} \right) 1_{\{x_i(t) \in G^c\}} \right] \\
\leq &\sum_{i : x_i(t) \geq 1} \mathbb{E} \left[ \rho_i \left( \frac{1 + 1}{x_i(t)} \right) U_i \left( \frac{c \rho_i}{x_i(t)} \right) 1_{\{x_i(t) \in G^c\}} \right] \\
\leq &\sum_{i : x_i(t) \geq 1} \mathbb{E} \left[ \rho_i \left( \frac{1 + K_1}{x_i(t)} \right) U_i \left( \frac{c \rho_i}{x_i(t)} \right) 1_{\{x_i(t) \in G^c\}} \right]
\end{aligned}
\]

where \(K_1\) in the last step is the constant defined in Lemma 2. Finally, we use the following result:

**Lemma 6:** For any \(K_1, C, \epsilon > 0\) there exists \(\xi\) such that, for any \(\max \tilde{x}_i(t) > \xi\), we have

\[
\sum_{i : \tilde{x}_i(t) \geq 1} \rho_i \left( \frac{K_1 C}{\tilde{x}_i(t)} - \frac{\epsilon}{2} \right) U_i' \left( \frac{c \rho_i}{\tilde{x}_i(t)} \right) \leq 0.
\]

Combining Lemma 6 with (29) and (30), we derive the following upper bound for (22):

\[
\begin{aligned}
\sum_{i = 1}^{N} \mathbb{E} \left[ \rho_i U_i' \left( \frac{c \rho_i}{x_i(t) + 1} \right) \right] \\
\leq A_1 + A_2 + \sum_{i : x_i(t) \geq 1} \mathbb{E} \left[ \rho_i \left( 1 + \frac{\epsilon}{2} \right) U_i \left( \frac{c \rho_i}{x_i(t)} \right) 1_{\{x(t) \in G^c\}} \right]
\end{aligned}
\]

**Prove stability.** Let \(A_0 = 2A_1 + 3A_2 + 2A_3\). Replacing (22) and (23) by their upper bounds, we can bound the expected drift \(\bar{W}(t)\) by:

\[
\begin{aligned}
\bar{W}(t) &\leq A_0 - \frac{\epsilon}{2} \sum_{i \in \mathcal{X}_t} \mathbb{E} \left[ \rho_i U_i' \left( \frac{c \rho_i}{x_i(t)} \right) 1_{\{x(t) \in G^c\}} \right] \\
&\leq A_0 + A_2 - \frac{\epsilon}{2} \sum_{i \in \mathcal{X}_t} \mathbb{E} \left[ \rho_i U_i' \left( \frac{c \rho_i}{x_i(t)} \right) \right].
\end{aligned}
\]

where inequality (28) is used in the last step and a constant \(A_2\) is added. Rearranging the terms and integrating (33) from \(t = 0\) to \(t = T\), we obtain

\[
\limsup_{T \to \infty} \frac{1}{T} \int_{t=0}^{T} \mathbb{E} \left[ \sum_{i \in \mathcal{X}_t} \rho_i U_i' \left( \frac{c \rho_i}{x_i(t)} \right) \right] dt \\
= \limsup_{T \to \infty} \frac{2W(0) - 2W(T)}{Te} + 2A_0 + 2A_2 \\
\leq \frac{2A_0 + 2A_2}{\epsilon}
\]

where the last step uses the fact that \(W(T) \geq 0\) is positive. Since function \(U_i'()\) is a non-negative and non-decreasing, and \(\lim_{z \to \infty} U_i'(1/z) = \infty\), equation (34) implies the stability of the network, as claimed in the stability definition (5).

**C. Proof of Lemma 4**

**Proof:** According to the utility gap condition in (7), we can conclude that for any \(\delta \geq 0\), there exists a positive \(\gamma\) such that, for \(\hat{x}(t)\) satisfying \(\max \hat{x}_i(t) > \gamma\), we have

\[
\Delta(\hat{x}(t)) \leq \delta \sum_{i : \tilde{x}_i(t) \geq 1} \hat{x}_i(t) U_i' \left( \frac{\phi_{\text{opt}, i}(\hat{x}(t))}{\hat{x}_i(t)} \right) \left( \frac{\phi_{\text{opt}, i}(\hat{x}(t))}{\hat{x}_i(t)} \right)
\]

To remove the absolute value on the right hand side of (35), we first assume that the utility function is non-negative. Plugging the expression of utility gap \(\Delta(\hat{x}(t))\) into (35), we have

\[
0 \leq \sum_{i : \tilde{x}_i(t) \geq 1} \hat{x}_i(t) U_i' \left( \frac{\phi_{\text{opt}, i}(\hat{x}(t))}{\hat{x}_i(t)} \right) - (1 - \delta) \hat{x}_i(t) U_i \left( \frac{\phi_{\text{opt}, i}(\hat{x}(t))}{\hat{x}_i(t)} \right)
\]

Since \(\phi_{\text{opt}, i}(\hat{x}(t))\) is the optimal rate allocation for the NUM problem (1) at state \(\hat{x}_i(t)\), no rate vector \(u \in \mathcal{R}\) can achieve a higher utility value. Choose \(\epsilon > 0\) and \(\delta > 0\) such that

\[
u \equiv (1 + \epsilon)^4 (1 - \delta)^{-\frac{\mathbb{E}}{\mathbb{E}} - \rho / \mathcal{R} \in \mathcal{R}}
\]

Such \(\epsilon > 0\) and \(\delta > 0\) exist due to the traffic condition \(\rho \in \mathcal{R}\). Let \(\delta_0 = (1 - \delta)^{-\frac{\mathbb{E}}{\mathbb{E}} - \rho \in \mathcal{R}}\). We then have

\[
\begin{aligned}
0 &\leq \sum_{i : \tilde{x}_i(t) \geq 1} \hat{x}_i(t) U_i' \left( \frac{\phi_{\text{opt}, i}(\hat{x}(t))}{\hat{x}_i(t)} \right) - (1 - \delta) \hat{x}_i(t) U_i \left( \frac{\phi_{\text{opt}, i}(\hat{x}(t))}{\hat{x}_i(t)} \right) \\
&\leq \sum_{i : \tilde{x}_i(t) \geq 1} \hat{x}_i(t) U_i' \left( \frac{\phi_{\text{opt}, i}(\hat{x}(t))}{\hat{x}_i(t)} \right) - \hat{x}_i(t) U_i \left( \frac{\delta_0 u_i}{\hat{x}_i(t)} \right) \\
&\leq \sum_{i : \tilde{x}_i(t) \geq 1} \left[ \phi_{\text{opt}, i}(\hat{x}(t)) - (1 + \epsilon)^4 \rho_i \right] U_i' \left( \frac{1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right)
\end{aligned}
\]

where the second step uses Lemma 1 (iii) with \(a = u_i / \hat{x}_i(t)\) and \(b = \delta_0 u_i / \hat{x}_i(t)\), the third step uses the concavity of the utility function \(U_i()\), and the last step is from (37). Note that a similar expression for (38) can also be shown if the utility function is negative. Specifically, when the utility value in (35) is negative, using the same proof technique and choosing \(\delta_0 = (1 + \delta)^{-\frac{\mathbb{E}}{\mathbb{E}} - \rho / \mathcal{R} \in \mathcal{R}}\), we can show that the inequality (38) is also satisfied. Refer to [13] for the proof.

Note that (38) is almost the same as (24), except for a constant \(r\) in the denominator and an additional \((1 + \epsilon)\) factor. Let \(c = (1 + \epsilon)^4\). For \(|r| < C\), we have

\[
\sum_{i : \tilde{x}_i(t) \geq 1} -\phi_i(\hat{x}(t)) U_i' \left( \frac{c \rho_i}{\hat{x}_i(t)} \right) \left( \frac{c \rho_i}{\hat{x}_i(t)} + r \right)
\]

\[
\leq \sum_{i : \tilde{x}_i(t) \geq 1} -\phi_i(\hat{x}(t)) U_i' \left( \frac{c \rho_i}{\hat{x}_i(t)} \right) \left( \frac{c \rho_i}{\hat{x}_i(t)} - C \right)
\]

\[
\leq \sum_{i : \tilde{x}_i(t) \geq 1} \phi_i(\hat{x}(t)) \left( 1 - \frac{C}{\hat{x}_i(t)} \right)^s \left( \frac{c \rho_i}{\hat{x}_i(t)} \right)
\]
\[ \leq \sum_{i: \hat{x}(t) \geq C} (1 + \epsilon)^3 \rho_i - \frac{K_c C \phi_i(\hat{x}(t))}{\hat{x}(t)} U'_i \left( \frac{c\rho}{\hat{x}(t)} \right) \leq 0. \] (40)

Applying (40) to (39), we derive
\[ \sum_{i: \hat{x}(t) \geq 1} (1 + \epsilon)^3 \rho_i U'_i \left( \frac{c\rho}{\hat{x}(t)} \right) - \phi_i(\hat{x}(t))U'_i \left( \frac{c\rho}{\hat{x}(t) + \varphi} \right) \leq 0, \]
which completes the proof of Lemma 4.

\section*{D. Proof of Lemma 5}

\textit{Proof:} To prove (27), we use network state \( x_\Omega(t) = x(t - \Omega) \) at time \( t - \Omega \) as an auxiliary variable, and bound both sides of (27) with respect to \( x_\Omega(t) \), respectively.

\textbf{Bound the left hand side of (27).} Define the event \( \Lambda^\Omega_i = \{ \max_i x_\Omega^\Omega(t) > \xi - C \} \), i.e., it is the event that the maximum queue length at time \( t - \Omega \) is greater than \( \xi - C \). We start with
\[ \sum_{i: x_\Omega^\Omega(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c\rho}{x_\Omega^\Omega(t)} \right) \right] \leq A_3 + \sum_{i: \hat{x}(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c\rho}{x_\Omega^\Omega(t)} \right) \right] \]
where a constant \( A_3 \) is added to change the summation from \( \{ i: x_\Omega^\Omega(t) \geq 1 \} \) to \( \{ i: \hat{x} \geq 1 \} \). Such \( A_3 \) satisfies
\[ \sum_{i: x_\Omega^\Omega(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c\rho}{x_\Omega^\Omega(t)} \right) \right] \leq \sum_{i: \hat{x}(t) \geq 1} \mathbb{E} \left[ \rho_i \left( 1 + \epsilon \right)^3 \rho_i - \frac{K_c C \phi_i(\hat{x}(t))}{\hat{x}(t)} U'_i \left( \frac{c\rho}{\hat{x}(t)} \right) \right]\]
\[ \leq \sum_{i: \hat{x}(t) \geq 1} \mathbb{E} \left[ \rho_i (1 + \epsilon)^3 \rho_i - \frac{K_c C \phi_i(\hat{x}(t))}{\hat{x}(t)} U'_i \left( \frac{c\rho}{\hat{x}(t)} \right) \right] \leq A_3 < +\infty, \]
(42)

where \( A_3 \) in the last step exists because \( |\hat{x}(t) - x_\Omega^\Omega(t)| \) can be bounded by a Poisson random variable with mean \( \omega \Omega \). The second step uses Lemma 1 (ii). To proceed from (41), we have
\[ \sum_{i: x_\Omega^\Omega(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c\rho}{x_\Omega^\Omega(t)} \right) \right] \leq \sum_{i: \hat{x}(t) \geq 1} \mathbb{E} \left[ \rho_i \left( 1 + \epsilon \right)^3 \rho_i - \frac{K_c C \phi_i(\hat{x}(t))}{\hat{x}(t)} U'_i \left( \frac{c\rho}{\hat{x}(t)} \right) \right] \leq A_3 < +\infty, \]
(42)

To bound (46), we define the event \( \mathcal{L}_i = \{ x_i(t) > x_\Omega^\Omega(t) \} \), and evaluate (46) over three events: \( \mathcal{E}_i^\gamma \cap \mathcal{L}_i \), \( \mathcal{E}_i^\gamma \cap \mathcal{L}_i^c \), and \( \mathcal{E}_i^c \), separately. For \( \mathcal{E}_i^\gamma \cap \mathcal{L}_i \), note that \( \mathcal{E}_i^\gamma \cap \mathcal{L}_i \) implies that the increase of \( x(t) \) in the interval \( t - \Omega \) to \( t \) is larger than \( C \). The number of arrivals in this interval must also be larger than \( C \). Since the arrival process is Poisson, when \( C \) is large, we can bound the probability of this event by a small number. Hence, the contribution to the expectation in (46) will be small. Specifically, let \( a_i(t) \) be the arrival within time \( t - \Omega \) to \( t \). We have
\[ \sum_{i: x_\Omega^\Omega(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c\rho}{x_\Omega^\Omega(t)} \right) \right] \leq \sum_{i: \hat{x}(t) \geq 1} \mathbb{E} \left[ \rho_i \left( 1 + \epsilon \right)^3 \rho_i - \frac{K_c C \phi_i(\hat{x}(t))}{\hat{x}(t)} U'_i \left( \frac{c\rho}{\hat{x}(t)} \right) \right] \leq A_3 < +\infty, \]
(42)
The first step uses the monotonicity of $U^i(\cdot)$ and $x_i(t) \leq x_i^{\Omega}(t)$ under $\mathcal{L}_i$. The fourth step is due to $\mathbb{E}[1_{\mathcal{E}_i} | \mathcal{F}_{t-\Omega}] \leq \mathcal{K}$ according to (48). The last step uses the inequality (49).

Finally, we evaluate (46) over $\mathcal{E}_i$. Note that when $\mathcal{E}_i$ occurs, the difference between $x(t)$ and $x_i^{\Omega}(t)$ is small. We will use this idea to replace the $x(t)$ in the right-hand side of (46) by $x_i^{\Omega}(t)$. Recognizing

$$\mathcal{E}_i \cap \mathcal{X}_i \subset \begin{cases} \|x_i^{\Omega}(t) - x_i(t)\|_1 < \frac{C}{2}, & \max_i x_i(t) > \xi \\ \max_i x_i^{\Omega}(t) > \xi - C \\ \end{cases} = \lambda_i^{\Omega},$$

we obtain

$$\sum_{i: x_i^{\Omega}(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c \rho_i}{x_i^{\Omega}(t)} \right) 1_{\mathcal{X}_i} 1_{\mathcal{E}_i} \right] \leq \sum_{i: x_i^{\Omega}(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c \rho_i}{x_i^{\Omega}(t) + c} \right) 1_{\mathcal{X}_i} 1_{\mathcal{E}_i} \right] \leq \sum_{i: x_i^{\Omega}(t) \geq 1} \mathbb{E} \left[ \rho_i \left( 1 + \frac{C}{x_i^{\Omega}(t)} \right)^s U'_i \left( \frac{c \rho_i}{x_i^{\Omega}(t)} \right) 1_{\mathcal{X}_i} 1_{\mathcal{E}_i} \right] \leq \sum_{i: x_i^{\Omega}(t) \geq 1} \mathbb{E} \left[ \rho_i \left( 1 + \frac{\epsilon}{2} \right) U'_i \left( \frac{c \rho_i}{x_i^{\Omega}(t)} \right) 1_{\mathcal{X}_i} 1_{\mathcal{E}_i} \right]$$

where the second step uses the monotonicity of $U^i(\cdot)$, and the third step uses Lemma 1 (ii), and the fourth step is from Lemma 2 and 6 (similar to the argument in (43)).

Note that $1_{\mathcal{E}_i} 1_{\mathcal{L}_i} + 1_{\mathcal{E}_i} 1_{\mathcal{L}_i} + 1_{\mathcal{E}_i} = 1$. Therefore, putting (49), (50), and (52) together, and plugging the result into (46), we derive

$$\sum_{i: x_i(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c \rho_i}{x_i(t)} \right) 1_{\mathcal{X}_i} \right] - A_3 - \frac{\epsilon}{2} A_2$$

We still need to replace the last constant $\epsilon/2$ by a function of $1_{\mathcal{E}_i}$. Toward this end, we can show the following the inequality, which is obtained in a similar way as the third and the last lines of (50):

$$\sum_{i: x_i^{\Omega}(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c \rho_i}{x_i^{\Omega}(t)} \right) 1_{\mathcal{X}_i} 1_{\mathcal{E}_i} \right] \leq \frac{\epsilon}{4} A_2 + \frac{\epsilon}{4} \sum_{i: x_i^{\Omega}(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c \rho_i}{x_i^{\Omega}(t)} \right) 1_{\mathcal{X}_i} 1_{\mathcal{E}_i} \right]$$

The first step uses the monotonicity of $U^i(\cdot)$ and $x_i(t) \leq x_i^{\Omega}(t)$ under $\mathcal{L}_i$. The fourth step is due to $\mathbb{E}[1_{\mathcal{E}_i} | \mathcal{F}_{t-\Omega}] \leq \mathcal{K}$ where the second step uses Lemma 1 (ii), and the third step uses the monotonicity of $U^i(\cdot)$ and $x_i(t) - x_i^{\Omega}(t) \leq a_i(t)$ (since $a_i(t)$ is the arrival time within time $t - \Omega$ to $t$), and the quantity $\mathcal{K}$ in the last step is chosen as:

$$\mathcal{K} = \mathbb{E} \left[ (1 + a_i(t))^s 1_{\mathcal{E}_i} | \mathcal{F}_{t-\Omega} \right].$$

The conditional expectation is taken over the $\sigma$-field $\mathcal{F}_{t-\Omega} = \sigma \{ x(u), u \leq t - \Omega \}$, generated by the history up to time $t - \Omega$. Given $\mathcal{F}_{t-\Omega}$, $a_i(t)$ is Poisson-distributed with mean $\lambda_i\Omega$. Therefore, using $\mathbb{E}[(1 + a_i(t))^s] < \infty$, we can use the Dominant Convergence Theorem to show

$$\lim_{C \to \infty} \mathcal{K} = \lim_{C \to \infty} \mathbb{E} \left[ (1 + a_i(t))^s (1 - 1_{\mathcal{E}_i}) | \mathcal{F}_{t-\Omega} \right] = \mathbb{E} \left[ (1 + a_i(t))^s (1 - \lim_{C \to \infty} 1_{\mathcal{E}_i}) | \mathcal{F}_{t-\Omega} \right] = 0.$$
(54) and rearrange the terms:

\[
\left(1 - \frac{\epsilon}{4}\right) \sum_{i : \hat{x}_i(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c p_i}{\hat{x}_i(t)} \right) X_i^t \right] \leq \sum_{i : \hat{x}_i(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c p_i}{\hat{x}_i(t)} \right) X_i^t \right] - A_3 - \epsilon A_2
\]

which establishes a lower bound for the right hand side of (27).

**Proof (27).** To show inequality (27), we combine the two bounds (44) and (56) to get

\[
\sum_{i : \hat{x}_i(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c p_i}{\hat{x}_i(t)} \right) 1_{\{x(t) \in \mathcal{G}'\}} \right] \leq \epsilon A_2 + (2 + \frac{3\epsilon}{2}) A_3 + (1 + \frac{\epsilon}{2})(1 + \frac{3\epsilon}{2}) \\
\cdot \sum_{i : \hat{x}_i(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c p_i}{\hat{x}_i(t)} \right) 1_{x_i(t) \geq (1 + \frac{3\epsilon}{2})} \sum_{i : \hat{x}_i(t) \geq 1} \mathbb{E} \left[ \rho_i U'_i \left( \frac{c p_i}{\hat{x}_i(t)} \right) X_i^t \right] - A_3 - \epsilon A_2
\]

Finally, note that \( \epsilon < 1, \frac{3\epsilon}{2} < 2, \) and \( (1 + \frac{\epsilon}{2})(1 + \frac{3\epsilon}{2}) < (1 + \epsilon)^2. \) Inequality (27) is immediate from (57).

**E. Proof of Lemma 6.**

**Proof:** We first show that there exists a constant \( A_4, \) such that

\[
\sum_{i : \hat{x}_i(t) \geq 1} \rho_i \left( \frac{K_i C}{\hat{x}_i(t)} - \frac{\epsilon}{4} \right) U'_i \left( \frac{c p_i}{\hat{x}_i(t)} \right) \leq A_4.
\]

(58)

Since the left hand side of (58) is non-negative, if \( \hat{x}_i(t) \leq 4K_i C/\epsilon, \) we have

\[
\sum_{i : \hat{x}_i(t) \geq 1} \rho_i \left( \frac{K_i C}{\hat{x}_i(t)} - \frac{\epsilon}{4} \right) U'_i \left( \frac{c p_i}{\hat{x}_i(t)} \right) \leq \sum_{i : \hat{x}_i(t) \geq 1} \rho_i \left( \frac{K_i C}{\hat{x}_i(t)} - \frac{\epsilon}{4} \right) U'_i \left( \frac{c p_i}{\hat{x}_i(t)} \right) \leq A_4
\]

where negative terms are dropped in the first step, and the second step uses the monotonicity of \( U'(\cdot) \) and \( \hat{x}_i(t) \geq 1. \) Plugging this into the left hand side of (31), we have

\[
\sum_{i : \hat{x}_i(t) \geq 1} \rho_i \left( \frac{K_i C}{\hat{x}_i(t)} - \frac{\epsilon}{4} \right) U'_i \left( \frac{c p_i}{\hat{x}_i(t)} \right) \leq A_4 - \sum_{i : \hat{x}_i(t) \geq 1} \rho_i \frac{\epsilon}{4} U'_i \left( \frac{c p_i}{\hat{x}_i(t)} \right)
\]

According to Assumption (c) in Section II, for \( \xi \to \infty \) and \( \max_i \hat{x}_i(t) > \xi, \) we have (60) \( \to -\infty. \) Therefore, there exists large enough \( \xi, \) such that (60) \( \leq 0. \)

**F. Proof of Proposition 1.**

**Proof:** To prove that the network is unstable when the information delay is zero and the utility gap is on the same order as that of the optimal utility (8), we construct a counterexample, in which the expected total queue-length grows unbounded as time \( t \) increases. Consider a network with two classes of flows and a feasible rate region (which is convex and compact, e.g., an ellipse), depicted in Figure 4. For an \( \alpha \)-fair utility with \( \alpha = 1/2, \) let \( \phi_{opt}(x(t)) \) denote the optimal rate allocation for state \( x(t) \) at time \( t. \) We define a suboptimal rate allocation by

\[
\phi(\hat{x}(t)) = \begin{cases} \phi_{opt}(x(t)), & \text{if } \phi_{opt}(x(t)) \text{ does not lie on } \hat{AB} \\
\phi_A, & \text{otherwise, if } x_1(t) > x_2(t) \\
\phi_B, & \text{otherwise, if } x_1(t) \leq x_2(t) 
\end{cases}
\]

where \( \hat{AB} \) denotes the boundary of the rate region between points \( A \) and \( B, \) and \( \phi_A \) and \( \phi_B \) are the optimal rate vectors at points \( A \) and \( B \) respectively. To prove Proposition 1, we notice that \( \Delta(x(t)) = 0 \) for all \( \phi_{opt}(x(t)) \in \hat{AB}. \) It can be shown that the utility gap is on the same order as the optimal utility, i.e.,

\[
\limsup_{t \to \infty} \frac{\Delta(x(t))}{\sum_{i=1}^{N} x_i(t) \alpha_i(t)} = 1
\]

\[
= 1 - \liminf_{t \to \infty} \frac{\sqrt{x_1(t)\phi_{opt,1}(t)} + \sqrt{x_2(t)\phi_{opt,2}(t)}}{\sqrt{x_1(t)\phi_{opt,1}(t)} + \sqrt{x_2(t)\phi_{opt,2}(t)}}
\]

\[
\leq 1 - \liminf_{t \to \infty} \frac{\sqrt{x_1(t)\phi_{B,1}} + \sqrt{x_2(t)\phi_{A,2}}}{\sqrt{x_1(t)\phi_{A,1}} + \sqrt{x_2(t)\phi_{B,2}}}
\]

where the second step holds because \( \phi_{B,1} \leq \phi_{1}(t) \leq \phi_{A,1} \) and \( \phi_{A,2} \leq \phi_{2}(t) \leq \phi_{B,2} \) for any rate vector \( \phi(\hat{x}(t)) \) that lies on \( \hat{AB}. \) Hence, the rate allocation policy \( \phi(\hat{x}(t)) \) satisfies condition (8) as claimed.

Next, we choose a point \( C \) as the middle point of line \( \hat{AB} \) and show that for small enough \( \epsilon > 0, \) the network is unstable under traffic intensity \( \rho = (1 + \epsilon)\phi_C \in \hat{R}. \) Consider a Lyapunov function defined by the weighted sum queue-length

\[
V(x) = \mu_1^{-1} w_1 x_1 + \mu_2^{-1} w_2 x_2
\]

where \( w_1 = \phi_{B,2} - \phi_{A,2} \) and \( w_2 = \phi_{A,1} - \phi_{B,1} \) are two positive constants. Then, we can formulate the expected Lyapunov
function \( W(t) = \mathbb{E}[V(x(t))] \) and derive an expression for the drift \( \dot{W}(t) \) using the same manner as in (22) and (23). In the following, we show that the expected drift is strictly above zero for traffic intensity \( \rho = (1 + \epsilon)\phi C \in \mathbb{R} \) with a small enough \( \epsilon > 0 \), i.e.

\[
\dot{W}(t) = \mathbb{E}\left[ \sum_{i=1}^{2} \lambda_i w_i - w_i \phi_i(x(t)) \right] \\
= \epsilon(w_1 \phi_{C,1} + w_2 \phi_{C,2}) + \mathbb{E}[w_1(\phi_{C,1} - \phi_1(t))] + \mathbb{E}[w_2(\phi_{C,2} - \phi_2(t))] \\
\geq \epsilon(w_1 \phi_{C,1} + w_2 \phi_{C,2}) > 0
\]

where the third inequality holds since the suboptimal rate allocation \( \phi(x(t)) \) always lies below the straight line \( AB \), whose slope is \(-\frac{w_2}{w_1}\). Thus, for the choice of Lyapunov function (61) and the traffic intensity \( \rho = (1 + \epsilon)\phi C \in \mathbb{R} \), the expected drift \( \dot{W}(t) \) is strictly above zero by a constant \( \epsilon(w_1 \phi_{C,1} + w_2 \phi_{C,2}) \). This implies that the network is unstable, since \( \lim_{t \to \infty} W(t) = \infty \) as \( t \to \infty \).

\section*{G. Proof of Theorem 2.}

\textbf{Proof:} The proof is almost the same as that of Theorem 1, except that we need to prove a result similar to Lemma 3, but with traffic intensity \( \rho \in (1 - \eta)\frac{\partial}{\partial \eta} \mathcal{R} \). For the Lyapunov function \( V(x) = \sum_{i=1}^{N} \sum_{n=1}^{x_i} U_i'(\frac{c p_i}{x_i(t)}) / \mu_i \), we derive the same drift \( \dot{W}(t) \) as in (22) and (23). To bound \( W \) under the new utility gap condition (9), we prove the following Lemma for traffic intensity \( \rho \in (1 - \eta)\frac{\partial}{\partial \eta} \mathcal{R} \).

\textbf{Lemma 7:} Let \( \epsilon = (1 + \epsilon)^2 \). For any traffic intensity \( \rho \in (1 - \eta)\frac{\partial}{\partial \eta} \mathcal{R} \) and constant \( C > 0 \). If the suboptimal rate allocation \( \phi(x(t)) \) satisfies the utility gap condition in (9), there exist positive constants \( \gamma > 0 \) and \( \epsilon > 0 \) such that, for all \( \rho C_i < C \) and for any network state satisfying \( \max_i x_i(t) > \gamma \), the following inequality holds:

\[
\sum_{i: \tilde{x}_i(t) \geq 1} \rho_i (1 + \epsilon)^3 U_i''\left(\frac{c p_i}{x_i(t)}\right) \phi_i(x(t)) U_i'\left(\frac{c p_i}{x_i(t)}\right) \leq 0.
\]

\section*{H. Proof of Theorem 3.}

\textbf{Proof:} We first notice that optimization problem (14) can be rewritten as

\[
Z(\Delta) = \min_{\phi} Z
\text{ s.t. } \quad R\phi \leq Zc, \quad \phi \succeq 0
\]

\[
\sum_{i=1}^{N} x_i^{\alpha} \phi_i^{1-\alpha} \geq U_{\text{opt}} - \Delta
\]

Its Lagrangian is then given by

\[
L(\phi, Z, p, q) = Z + p^T(R\phi - Zc) + q(U_{\text{opt}} - V(\phi) - \Delta)
\]

At the optimal point of (67), the KKT conditions for optimality are given by

\[
R\phi = Zc, \quad V(\phi) = U_{\text{opt}} - \Delta
\]

\[
R^T p - q \nabla_V V(\phi) - \nabla_\phi(\sum_{i=1}^{N} \phi_i) = 0, \quad p^T c = 1
\]

From the implicit function theorem, variables \( \phi, Z, p \) and \( q \) can be viewed as implicit functions of \( \Delta \), which is uniquely defined by the KKT conditions (68) and (69). We define a
vector $y = [\phi; p; q; Z]$ and a residual

$$G(y, \Delta) = \begin{pmatrix} R^T p - q \nabla p h_1 V(\phi) \\ R\phi - Zc \\ U_{\text{opt}} - \Delta - V(\phi) \\ 1 - p^T c \end{pmatrix}$$ (70)

Then the KKT conditions (68) and (69) are equivalent to $G(y, \Delta) = 0$.

From the implicit function theorem, we have

$$\frac{dZ}{d\Delta} = -(\nabla_y G)^{-1} \nabla_\Delta G,$$ (71)

Plugging $\nabla_y G$ and $\nabla_\Delta G$ into above and performing some matrix manipulations, we can derive the result in Theorem 4.

**REFERENCES**


