Joint Congestion Control and Routing Optimization: An Efficient Second-Order Distributed Approach

Jia Liu[†] Ness B. Shroff[†] Cathy H. Xia[*] Hanif D. Sherali[‡]
[†]Department of Electrical and Computer Engineering, The Ohio State University
[*] Department of Integrated Systems Engineering, The Ohio State University
[‡]Grado Department of Industrial and Systems Engineering, Virginia Tech

Abstract

Optimization-based algorithms for joint congestion control and routing have received a significant amount of attention recently. To date, however, most of the existing schemes follow a key idea called the back-pressure algorithm. Despite having many salient features, the first-order subgradient nature of the back-pressure based congestion control and routing necessitates small step-sizes, hence slowing down convergence and resulting in poor delay performance. To overcome these limitations, in this paper, we make a first attempt at developing a second-order joint congestion control and routing optimization framework that offers rate-optimality, queuing stability, fast convergence, and low delays. Our contributions in this paper are three-fold: i) we propose a new second-order joint congestion control and routing framework based on a primal-dual interior-point approach; ii) we establish rate-optimality and queuing stability of the proposed second-order method; and iii) we show how to implement the proposed second-order method in a distributed fashion.

1 Introduction

With the rapid integration of new applications and technologies, recent years have witnessed a growing challenge in making communication networks work more efficiently. To date, while there exists a large body of work on joint congestion control and routing for both wireline and wireless networks (see, e.g., [1–4] and many other follow-ups and extensions), most of these schemes follow a key idea called the "back-pressure" algorithm, which traces its roots to the celebrated paper in [5] published more than two decades ago. The enduring popularity of the back-pressure algorithm is primarily due to: i) a provable throughput optimality, ii) elegant cross-layer extensions, and iii) a distributed queue-length differential based routing that stabilizes all queues in the network. Researchers have also uncovered a fundamental connection between the back-pressure based congestion control and

the Lagrangian dual decomposition framework plus the subgradient method in classical nonlinear optimization theory [1, 3], where (scaled) queue-lengths play the role of Lagrangian dual variables and the queue-length updates correspond to subgradient directions. This enlightening insight has unified techniques that originated independently from control and optimization theory.

However, despite all the salient features, the subgradient nature of the back-pressure based congestion control and routing schemes turns out to be a factor that plagues their performance in practice. Being a first-order method (subgradients can be viewed as a first-order support of the dual function), back-pressure based joint congestion control and routing schemes neglect the curvature of the objective function contour, which is characterized by the eigenvalue condition number of the Hessian matrix that usually becomes increasingly ill-conditioned as the iterates approach an optimal solution [6]. As a result, it necessitates a small update in each iteration [1–4, 7], which subsequently slows down convergence and undermines the performance of optimization. This limitation motivates us to pursue a *second-order* design approach for joint congestion control and routing. The fundamental rationale behind our approach is that, as in classical nonlinear optimization theory [6], by considering the second-order Hessian information in congestion control and routing, we can expect to alleviate the inherent ill-conditioned behavior of first-order methods, thus leading to much faster convergence and hence better performance in practice.

However, due to a number of technical difficulties, developing a second-order congestion control and routing optimization theory is highly challenging and, to our knowledge, results in this area remain scarce. First, unlike the relatively obvious queue-length based connection between the back-pressure and the subgradient algorithms, it remains unclear how one can utilize the insights drawn from existing static second-order network optimization algorithms [8–11] to guide the design of joint congestion control and routing in practice. The main challenge here is that most of the algorithms in [8–11] operate with long-term rates rather than evolve with actual time instants. Also, their connection to observable network state information (e.g., queue-lengths, etc.) is still missing. Second, after constructing a second-order scheme, it remains a difficult task to prove its rate-optimality and queuing stability. This is because the incorporation of the second-order Hessian information significantly complicates the computational schemes and necessitates new theoretical approaches in performance analysis. Lastly, how to implement the developed second-order scheme in a distributed fashion (comparable to first-order methods) is still an open question. Similar to the (static) second-order optimization algorithms in [8–11], one would have to face the challenges arising from decentralizing the Hessian and Laplacian matrix inverse computations.

The key contribution of this paper is that, for the first time, we successfully develop a second-order joint congestion control and routing framework to address the aforementioned technical difficulties and establish an analytical foundation that offers fast convergence and high performance. The main

results and technical contributions of this paper are as follows:

- We propose a second-order joint congestion control and routing framework based on a *primal-dual* interior-point approach, in which we modify the step-size control strategies such that the resultant scheme is well-suited for implementation in practical networks. Our primal-dual approach exposes a deep connection between observable network state information and the primal-dual interior-point optimization theory, which itself is an active research field in operations research today (see, e.g., [12] for a survey).

- We establish the rate-optimality and the queuing stability of the proposed second-order framework. Our theoretical analysis unveils the fundamental reason behind the fast convergence in the proposed second-order framework. Interestingly, our analytical results naturally lead to a rate-optimality and queue-length trade-off relationship governed by the *barrier parameter* of the interior-point method. We compare this trade-off relationship to those in first-order methods and contrast their similarities and differences, thus further advancing our understanding of both first- and second-order methods in network optimization theory.

- We suggest several approaches to implement the proposed second-order method in a *distributed* fashion. In particular, for the distributed dual Newton direction computation (the most challenging part in our second-order method), we propose a new Sherman-Morrison-Woodbury (SMW) based iterative approach. We show that, on a L-link network, the SMW-based approach obtains the *precise* solution in $2L$ iterations, rather than asymptotically as in [8–10].

Collectively, our results in this paper contribute to an exciting development of a cross-layer network control and optimization theory with second-order techniques. The remainder of this paper is organized as follows. In Section 2, we review related works. Section 3 introduces the network model and problem formulation. Section 4 presents the algorithm and performance analysis of our second-order scheme. Section 5 develops the principal components of the distributed computations. Section 6 presents some numerical results, and Section 7 concludes the paper.

2 Related Work

In this section, we review the state-of-the-art of both first- and second-order methods that are closely related to this paper. As mentioned earlier, there is a large body of work on first-order back-pressure based joint congestion control and routing (e.g., [1–4, 7, 13]). Among these works, the scheme in [3] is the most related and can be directly compared to our work since it is also a primal-dual based controller, where the primal and dual variables are updated jointly (hence relatively more convenient to implement in practice). Thanks to the second-order structure, our approach requires a much less

conservative step-size selection, while achieving a steeper negative Lyapunov drift rate and inducing a much faster (three orders of magnitude numerically) convergence than in [3]. On the other hand, the schemes in [1,2,4] can be categorized as dual-based controllers, where an inner subproblem defined in terms of primal variables needs to be solved for each fixed set of dual variables. Thus, a counterpart of primal-dual step-size selection does not exist. However, similar Lyapunov drift rate analysis and numerical results also indicate a slow convergence performance due to their first-order nature.

In the second-order domain, recent (centralized and distributed) interior-point based methods for network optimization can be found in [8–11, 14–17]. In particular, significant efforts have been made to decentralize the second-order computations, including a Gaussian belief propagation technique in [15–17] and a matrix-splitting approach in [8] for flow control (with fixed routing); and a consensus-based local averaging scheme for minimum cost routing (with fixed source rates) in [11]. Finally, in our previous work [9,10], we developed distributed second-order methods for cross-layer optimization (joint flow control, routing, and scheduling) in both wireline and wireless networks. However, all these second-order methods operate with *long-term rates* and do *not* consider queuing stability. Moreover, they were all based on the classical barrier interior-point approach and none of them adopted the latest advances in primal-dual interior-point theory [12]. Therefore, the development of our primal-dual second-order method in this paper is novel.

3 Network Model and Problem Formulation

We first introduce the notation style in this paper. We use boldface to denote matrices and vectors. We let $\mathbf{A}^T$ denote the transpose of $\mathbf{A}$. Diag $\{\mathbf{A}_1, \ldots, \mathbf{A}_N\}$ represents the block diagonal matrix with $\mathbf{A}_1, \ldots, \mathbf{A}_N$ on its main diagonal. We let $(\mathbf{A})_{ij}$ represent the entry in the i-th row and j-th column of $\mathbf{A}$ and let $(\mathbf{v})_m$ represent the m-th entry of $\mathbf{v}$. We let $\mathbf{I}_K$ denote the K-dimensional identity matrix, and let $\mathbf{1}_K$ and $\mathbf{0}_K$ denote the K-dimensional vectors whose elements are all ones and zeros ("K" may be omitted for brevity if the dimension is clear from the context). We let $\lambda_{\min}\{\mathbf{A}\}$ and $\lambda_{\max}\{\mathbf{A}\}$ denote the smallest and largest eigenvalues of $\mathbf{A}$, respectively.

Network model: We consider a time-slotted communication network system with time slot units being indexed by $t = 0, 1, 2, \ldots$. As shown in Fig. 1, we represent the communication network by a directed graph $\mathcal{G} = \{\mathcal{N}, \mathcal{L}\}$, where $\mathcal{N}$ and $\mathcal{L}$ are the sets of nodes and links, with $|\mathcal{N}| = N$ and $|\mathcal{L}| = L$, respectively. We assume that $\mathcal{G}$ is connected. There are F end-to-end sessions in the network, indexed by $f = 1, \ldots, F$. Each session f has a source node and a destination node, represented by $\mathrm{Src}(f), \mathrm{Dst}(f) \in \mathcal{N}$, respectively. To avoid triviality, we assume that $\mathrm{Src}(f) \neq \mathrm{Dst}(f)$ for all f. The data of session f travel from $\mathrm{Src}(f)$ to $\mathrm{Dst}(f)$ through the network, possibly via multi-hop and multi-path routing.

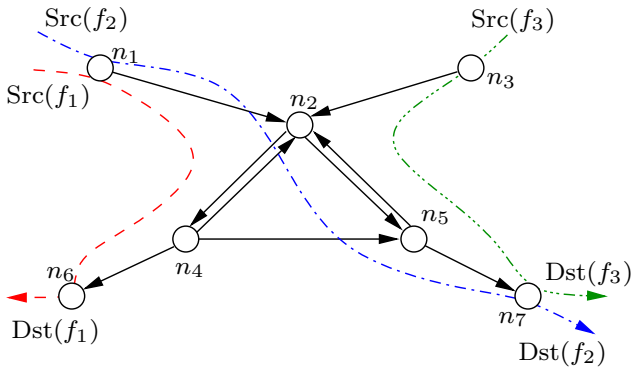

Figure 1: An illustrative example of the network model.

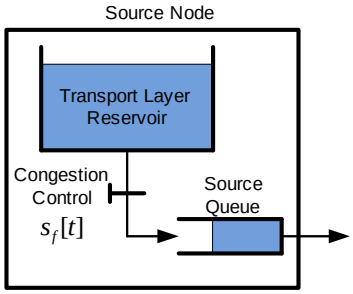

Figure 2: An illustrative example of source node congestion control.

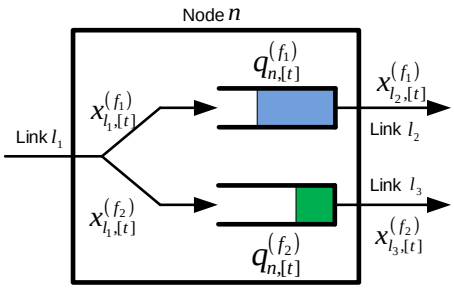

Figure 3: An illustrative example of routing at an intermediate node.

Congestion control: As in [2, 3], we assume that the source node Src(f) has a continuously-backlogged transport layer reservoir that contains session f's data, as illustrated in Fig. 2. Similar to a valve, in each time-slot t, a transport layer congestion controller determines the amount of data $s_f[t]$ to be released from this reservoir into a network layer source queue, where the data await to be routed to node Dst(f) through the network. In other words, $\{s_f[t]\}$ acts as the arrival process to the source queue. To control the burstiness, we let $s_f[t] \leq s_f^{\max}$, $\forall t$. We let $\bar{s}_f \geq 0$ denote the time-average rate at which data of session f is injected at Src(f) under congestion control, i.e., $\bar{s}_f = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} s_f[t]$. Each session is associated with a utility function $U_f(\bar{s}_f)$, which represents the utility gained by session f when data is injected at rate $\bar{s}_f$. We assume that $U_f(\cdot)$ is strictly concave, monotonically increasing, and twice continuously differentiable.

Routing: We let $x_{l,[t]}^{(f)} \geq 0$ denote the rate offered to route session f's data in time-slot t at link l, as shown in Fig. 3. We let $\bar{x}_l^{(f)} \triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} x_{l,[t]}^{(f)}$ represent the time-average routing rate of session f at link l. We use $\bar{\mathbf{s}} \triangleq [\bar{s}_1, \ldots, \bar{s}_F]^T$ and $\bar{\mathbf{x}}^{(f)} \triangleq [\bar{x}_1^{(f)}, \ldots, \bar{x}_L^{(f)}]^T$ to group all congestion control and session f's routing rates. We denote the capacity of link l as C_l and assume that it is fixed, which is an appropriate model for wireline networks. We note that the theoretical results and

algorithms in this paper can be readily extended to wireless networks by replacing C_l with the convex hull of the wireless link capacity region, similar to [1–4].

As in [1, 3, 18], we define the *network capacity region* as the largest set of congestion control rates $\bar{\mathbf{s}}$ such that there exists a routing policy for which the time-average routing rates $\{\bar{\mathbf{x}}^{(f)}, \forall f\}$ satisfy the following constraints:

$$\sum_{l\in\mathcal{O}(n)} \bar{x}_l^{(f)} \geq \sum_{l\in\mathcal{I}(n)} \bar{x}_l^{(f)} + \bar{s}_f \mathbb{1}_f(n),\ \forall f, \forall n \neq \mathrm{Dst}(f), \tag{1}$$

$$\sum_{f=1}^{F} x_{l,[t]}^{(f)} \leq C_l, \quad \forall l, t, \tag{2}$$

where $\mathcal{O}(n)$ and $\mathcal{I}(n)$ represent the sets of outgoing and incoming links at node n, respectively; $\mathbb{1}_f(n)$ is an indicator function that takes the value 1 if $n = \mathrm{Src}(f)$ and 0 otherwise.

For convenience, we use a *node-arc incidence matrix* (NAIM) [19] $\mathbf{A}^{(f)} \in \mathbb{R}^{(N-1)\times L}$ and a *source vector* $\mathbf{b}^{(f)} \in \mathbb{R}^{N-1}$ to represent the network topology. Let $\mathrm{Tx}(l)$ and $\mathrm{Rx}(l)$ denote the transmitting and receiving nodes of link l, respectively. The entries $(\mathbf{A}^{(f)})_{nl}$ and $(\mathbf{b}^{(f)})_n$, $n \neq \mathrm{Dst}(f)$, are defined as follows:

$$(\mathbf{A}^{(f)})_{nl} = \begin{cases} 1 & \text{if } n = \mathrm{Tx}(l), \\ -1 & \text{if } n = \mathrm{Rx}(l), \\ 0 & \text{otherwise,} \end{cases} \quad (\mathbf{b}^{(f)})_n = \begin{cases} 1 & \text{if } n = \mathrm{Src}(f), \\ 0 & \text{otherwise.} \end{cases}$$

Then, the constraint in (1) can be compactly written as: $\mathbf{A}^{(f)}\bar{\mathbf{x}}^{(f)} - \bar{s}_f \mathbf{b}^{(f)} \geq \mathbf{0}, \quad \forall f = 1, 2, \ldots, F.$

Queuing stability: We assume that each node maintains a separate queue for each session f, as shown in Fig. 3. We let $q_{n,[t]}^{(f)} \geq 0$ represent the amount of data in session f's queue at node n at time t. Since data leave the network upon reaching destinations, we have $q_{\mathrm{Dst}(f),[t]}^{(f)} = 0$, $\forall t$. The evolution of $q_{n,[t]}^{(f)}$, $n \neq \mathrm{Dst}(f)$, is given by:

$$q_{n,[t+1]}^{(f)} = \left(q_{n,[t]}^{(f)} - \sum_{l\in\mathcal{O}(n)} x_{l,[t]}^{(f)} \right)^+ + \sum_{l\in\mathcal{I}(n)} \widehat{x}_{l,[t]}^{(f)} + s_{f,[t]} \mathbb{1}_f(n), \tag{3}$$

where $(\cdot)^+ \triangleq \max\{0, \cdot\}$ and $\widehat{x}_{l,[t]}^{(f)}$ is the *actual* routing rate. Note that $\widehat{x}_{l,[t]}^{(f)} \leq x_{l,[t]}^{(f)}$ since $\mathrm{Tx}(l)$ may have less than $x_{l,[t]}^{(f)}$ amount of data to transmit. Let $\mathbf{q}_{[t]} \triangleq [q_{n,[t]}^{(f)}, \forall f, \forall n \neq \mathrm{Dst}(f)]^T$ group all queue lengths at time t. In this paper, we adopt the same notion of queuing stability as in [3]: Under a congestion control and routing scheme, we say that the network is *stable* if the steady-state total queue length remains finite, i.e., $\limsup_{t\to\infty} \sum_{n=1}^{N} \sum_{f=1}^{F} q_{n,[t]}^{(f)} \leq \infty$.

Problem formulation: In this paper, our goal is to develop an optimal joint congestion control and routing scheme to maximize the total utility $\sum_{f=1}^{F} U_f(\bar{s}_f)$, subject to the network capacity region

constraints and that the network is stable. Putting together the models presented earlier yields the following joint congestion control and routing (JCCR) optimization problem:

JCCR:

$$\begin{aligned}
&\text{Maximize} && \sum_{f=1}^{F} U_f(\bar{s}_f) && \\
&\text{subject to} && 1)\ \mathbf{A}^{(f)}\bar{\mathbf{x}}^{(f)} - \bar{s}_f\mathbf{b}^{(f)} \geq \mathbf{0}, && \forall f, \\
& && 2)\ \sum_{f=1}^{F} x_{l,[t]}^{(f)} \leq C_l, && \forall l, t \\
& && 3) \text{ Stability of all network queues.} &&
\end{aligned}$$

As mentioned earlier, several first-order schemes based on the back-pressure idea [5] have been proposed (e.g., [1–4]) to solve Problem JCCR. However, the convergence behavior of these first-order schemes is slow, which could lead to poor performance in practice. In what follows, we will investigate a new second-order joint congestion control and routing framework.

4 A Second-Order Congestion Control and Routing Optimization Framework

In Section 4.1, we first present our second-order joint congestion control and routing algorithm along with the main results on rate-optimality and queuing stability. Then, in Section 4.2, we explain the design rationale of our second-order approach. Section 4.3 focuses on performance analysis and provides the proofs for the main theorems in Section 4.1. In Section 4.4, we discuss the key insights and intuition related to the results in Section 4.1.

4.1 The Algorithm and Main Theoretical Results

We start with some necessary notation that will be used throughout the paper. First, we use $\mathbf{y}_{[t]}$ to denote all instantaneous *joint congestion control and routing decisions* at time t, which are arranged in the following link-based order: $\mathbf{y}_{[t]} \triangleq \left[s_{1,[t]} \cdots s_{F,[t]}, x_{1,[t]}^{(1)} \cdots x_{1,[t]}^{(F)}, \cdots, x_{L,[t]}^{(1)} \cdots x_{L,[t]}^{(F)}\right]^T$. We let $\mathbf{M} \triangleq \left[\begin{array}{cccc} \mathbf{B} & \mathbf{A}_1 & \cdots & \mathbf{A}_L \end{array}\right]$, where $\mathbf{B}$ and $\mathbf{A}_l$ are defined as $\mathbf{B} \triangleq \text{Diag}\{\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(F)}\}$, and $\mathbf{A}_l \triangleq \text{Diag}\{-\mathbf{a}_l^{(1)}, \ldots, -\mathbf{a}_l^{(F)}\}$, and where in the definition of $\mathbf{A}_l$, the vector $\mathbf{a}_l^{(f)}$ is the l-th column of the matrix $\mathbf{A}^{(f)}$ in Problem JCCR (i.e., $\mathbf{A}^{(f)} = \left[\mathbf{a}_1^{(f)}, \mathbf{a}_2^{(f)}, \ldots, \mathbf{a}_L^{(f)}\right]$). Also, we let $\mathbf{N} \triangleq \text{Diag}\left\{\mathbf{0}_F^T, \mathbf{1}_F^T, \ldots, \mathbf{1}_F^T\right\} \in \mathbb{R}^{(L+1)\times(L+1)F}$ and $\mathbf{c} \triangleq [0, C_1, \ldots, C_L]^T \in \mathbb{R}^{L+1}$. Then, it can be verified that the first two constraints in Problem JCCR can be compactly written as $\mathbf{M}\mathbf{y}_{[t]} \leq \mathbf{0}$ (in

each time slot rather than on average) and $\mathbf{N}\mathbf{y}_{[t]} \leq \mathbf{c}$. Next, we define the following μ-scaled barrier augmented objective function:

$$f_\mu(\mathbf{y}_{[t]}) \triangleq -\mu \sum_{f=1}^{F} U_f(s_{f,[t]}) - \sum_{l=1}^{L} \log\left(C_l - \sum_{f=1}^{F} x_{l,[t]}^{(f)}\right) - \sum_{f=1}^{F} \log(s_{f,[t]}) - \sum_{l=1}^{L}\sum_{f=1}^{F} \log(x_{l,[t]}^{(f)}), \tag{4}$$

where $\mu > 0$ is called the barrier parameter (its meaning will be clear soon in Section 4.2). We let $\mathbf{g}_{[t]} \triangleq \nabla f_\mu(\mathbf{y}_{[t]})$ and $\mathbf{H}_{[t]} \triangleq \nabla^2 f_\mu(\mathbf{y}_{[t]})$ denote the gradient vector and Hessian matrix of $f_\mu(\cdot)$ evaluated at $\mathbf{y}_{[t]}$, respectively.

Next, we introduce dual variables $p_{n,[t]}^{(f)} > 0$, $\forall f$, $\forall n \neq \text{Dst}(f)$ to be associated with the constraint in (1) in each time slot t (i.e,. replacing $\bar{x}_l^{(f)}$ and $\bar{s}_f$ by $x_{l,[t]}^{(f)}$ and $s_f^{[t]}$, respectively). These dual variables play the role of *prices* charged to session f for using node n. We let $\mathbf{p}_{[t]} = [p_{n,[t]}^{(f)}, \forall f = 1, \forall n \neq \text{Dst}(f)]^T$ group all dual variables. We further introduce two diagonal matrices: $\mathbf{P}_{[t]} \triangleq \text{Diag}\left\{\mathbf{p}_{[t]}\right\}$ and $\mathbf{Q}_{[t]} \triangleq \text{Diag}\left\{\mathbf{M}\mathbf{y}_{[t]}\right\}$. We note that $\mathbf{Q}_{[t]}$ is intrinsically related to *queue-length evolutions*, since each diagonal entry of $\mathbf{Q}_{[t]}$ is of the form: $q_{n,[t]}^{(f)} - \sum_{l\in\mathcal{O}(n)} x_{l,[t]}^{(f)} + \sum_{l\in\mathcal{I}(n)} x_{l,[t]}^{(f)} + s_{f,[t]}\mathbb{1}_f(n)$ (cf. (3)). With these notation and assuming that the initial primal and dual variables are strictly feasible at time-slot $t = 0$ (i.e., $\mathbf{M}\mathbf{y}_{[0]} < \mathbf{0}$, $\mathbf{N}\mathbf{y}_{[0]} < \mathbf{c}$, and $\mathbf{p}_{[0]} > \mathbf{0}$), our proposed second-order algorithm is illustrated in Algorithm 1.

Remark 1. *Several remarks on the properties of Algorithm 1 are in order: i) Algorithm 1 is a primal-dual scheme in which the primal and dual variables are updated jointly as in (7) (more convenient for implementation in practice). This is unlike dual-based controllers (e.g., [1, 2, 4]), where a coupled subproblem defined in terms of primal variables is solved in each dual iteration. ii) Algorithm 1 operates on a "time-slot by time-slot" basis and captures the queue-length evolution $\mathbf{Q}_{[t]}$, thus exposing an observable network state information for practical implementations. In contrast, all existing second-order methods in [8–11, 14–17] only optimize "long-term rates" and fail to provide such a key connection. iii) For ease of performance analysis in this section, the primal-dual Newton directions in (5) and (6) are expressed in matrix form for now. Their explicit and distributed computational schemes will be derived later in Section 5.*

The following theorem says that the average rate obtained under Algorithm 1 can be made *arbitrarily close* to the optimal solution by increasing the barrier parameter μ.

Theorem 1 (Rate-optimality)**.** *Let $\bar{\mathbf{y}}^*$ represent the optimal average rate solution to Problem JCCR. Under Algorithm 1 and for some given μ, if the step-size π scales as $O(\frac{1}{\mu})$, then there exists some constant $B \in (0, \infty)$ independent of μ such that $\limsup_{T\to\infty} \left|\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right| \leq \frac{B}{\sqrt{\mu}}$.*

Algorithm 1 A second-order joint congestion control and routing optimization algorithm (for a given μ).

1. In time-slot t, determine the *second-order* primal (joint congestion control and routing) and dual (pricing) Newton directions $\Delta\mathbf{y}_{[t]}$ and $\Delta\mathbf{p}_{[t]}$ as follows:

$$\Delta\mathbf{y}_{[t]} = -(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M})^{-1}(\mathbf{g}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}), \tag{5}$$

$$\Delta\mathbf{p}_{[t]} = -(\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]})^{-1}\times [\mathbf{M}\mathbf{H}_{[t]}^{-1}(\mathbf{g}_{[t]} + \mathbf{M}^T\mathbf{p}_{[t]}) - (\mathbf{Q}_{[t]} + \mathbf{P}_{[t]}^{-1})\mathbf{1}]. \tag{6}$$

2. Update primal and dual variables *jointly* as:

$$\begin{bmatrix} \mathbf{y}_{[t+1]} \\ \mathbf{p}_{[t+1]} \end{bmatrix} = \left(\begin{bmatrix} \mathbf{y}_{[t]} \\ \mathbf{p}_{[t]} \end{bmatrix} + \pi \begin{bmatrix} \Delta\mathbf{y}_{[t]} \\ \Delta\mathbf{p}_{[t]} \end{bmatrix} \right)_{\mathcal{S}_\epsilon^M}, \tag{7}$$

where $0 < \pi \leq 1$ is an appropriate step-size, and $(\cdot)_{\mathcal{S}_\epsilon^M}$ represents the projection onto the set $\mathcal{S}_\epsilon^M$ defined as:

$$\mathcal{S}_\epsilon^M \triangleq \left\{ (\mathbf{y}, \mathbf{p}) \;\middle|\; \begin{array}{ll} \epsilon\mathbf{1} \leq \mathbf{y} \leq M\mathbf{1}, & \mathbf{M}\mathbf{y} \leq -\epsilon\mathbf{1}, \\ \mathbf{N}\mathbf{y} \leq \mathbf{c} - \epsilon\mathbf{1}, & \mathbf{p} \geq \epsilon\mathbf{1}. \end{array} \right\}, \tag{8}$$

where the constant $\epsilon > 0$ can be made arbitrarily close to zero and the constant $M > 0$ is used for burstiness reduction. Let $t \leftarrow t+1$ and go to Step 1.

For a time-varying matrix $\mathbf{A}_{[t]}$, we let $\lambda_{\min}\{\mathbf{A}\} \triangleq \inf_t \left\{\lambda_{\min}\left\{\mathbf{A}_{[t]}\right\}\right\}$. The following proposition explains why Algorithm 1 enjoys a *fast* convergence performance.

Proposition 2 (Lyapunov drift rate)**.** *If $\bar{\mathbf{y}}$ is outside of $[\bar{\mathbf{y}}^* - \frac{B}{\sqrt{\mu}},\ \bar{\mathbf{y}}^* + \frac{B}{\sqrt{\mu}}]$, where $\bar{\mathbf{y}}^*$ and B are as defined in Theorem 1, then there is a negative Lyapunov drift that drives $\bar{\mathbf{y}}$ toward this interval, and the drift rate R can be lower bounded by $R \geq \frac{\lambda_{\min}\{\mathbf{H}\}}{\lambda_{\min}\{\mathbf{H} - \mathbf{M}^T\mathbf{Q}^{-1}\mathbf{P}\mathbf{M}\}}$. Particularly, $R \geq 1$ as $\mu \to \infty$.*

Proposition 2 highlights that the second-order scaling term $(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M})^{-1}$ in (5) is crucial to the average rate convergence of Algorithm 1. Without this term (replacing it by an identity matrix $\mathbf{I}$), we essentially "rediscover" a first-order back-pressure based method (with $\mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}$ being the "pressure differential"). Proposition 2 indicates that, thanks to the appearance of this term in the denominator of R, the "pulling force" of the negative Lyapunov drift is strong, allowing our scheme to approach the desired region *at least as fast as* at a constant rate R that is *insensitive* to the objective function contour. In contrast, the Lyapunov drift rate in first-order methods can be characterized by $\inf_t\{\lambda_{\min}\{\text{Diag}\{-U_f''(s_{f,[t]}), \forall f\}\}\}$ (see, e.g., [3, Eq.(32)] and the discussion thereafter), which is clearly sensitive to the objective function contour and could be very small (i.e., induce stalling).

The next theorem states that, under Algorithm 1, the queue lengths are asymptotically bounded,

and hence induce queuing stability for the network.

Theorem 3 (Queuing stability)**.** *Under Algorithm 1 and for some given μ, letting $\epsilon = O(1/\mu)$, there exists a constant $K < \infty$ that scales as $O(\mu)$ such that* $\limsup_{t\to\infty} \left\|\mathbf{q}_{[t]}\right\| \leq K$.

The proofs of Theorems 1, 3 and Proposition 2 will be given in Section 4.3. In what follows, we first explain the design rationale behind Algorithm 1.

4.2 The Rationale behind the Algorithmic Design

The design of Algorithm 1 is inspired by, and mirrors, a primal-dual interior-point method for directly solving (static) Problem JCCR in terms of long-term average rates. In what follows, we outline the main steps in our algorithmic design.

Step 1) A perturbed KKT system: We start with reformulating Problem JCCR using the standard interior-point approach as follows: We first apply a logarithmic barrier function to the link capacity constraints and non-negativity constraints and then accommodate them in the objective function. As a result, the augmented objective function (to be minimized) can be written as follows:

$$\begin{aligned}\hat{f}_{\mu}^{(0)}(\bar{\mathbf{y}}) = &-\sum_{f=1}^{F} U_f(\bar{s}_f) - \frac{1}{\mu}\sum_{l=1}^{L}\log\left(C_l - \sum_{f=1}^{F}\bar{x}_l^{(f)}\right) - \frac{1}{\mu}\sum_{f=1}^{F}\log(\bar{s}_f) - \frac{1}{\mu}\sum_{l=1}^{L}\sum_{f=1}^{F}\log(\bar{x}_l^{(f)}) \\ &- \frac{1}{\mu}\sum_{f=1}^{F}\sum_{n\neq \mathrm{Dst}(f)}\log\left(\sum_{l\in\mathcal{O}(n)}\bar{x}_l^{(f)} - \sum_{l\in\mathcal{I}(n)}\bar{x}_l^{(f)} - \bar{s}_f\mathbb{1}_f(n)\right), \end{aligned} \tag{9}$$

where $\mu > 0$ is the same as in Section 4.1. Then, we can reformulate Problem JCCR as the following *unconstrained* optimization problem:

$$\textbf{R-JCCR: Minimize} \quad \hat{f}_{\mu}^{(0)}(\bar{\mathbf{y}}), \tag{10}$$

where, as $\mu \to \infty$, the original objective function of Problem JCCR dominates the barrier functions, and hence the solution of Problem R-JCCR approaches that of Problem JCCR asymptotically [12,20]. Next, we take the first derivatives of $\hat{f}_{\mu}^{(0)}(\bar{\mathbf{y}})$ and set them equal to zero (i.e., by way of the first-order (KKT) condition) to obtain:

$$\frac{\partial \hat{f}_{\mu}^{(0)}(\bar{\mathbf{y}})}{\partial \bar{s}_{(f)}} = -U'(\bar{s}_f) - \frac{1}{\mu\bar{s}_f} - \frac{1}{\mu(\sum_{l\in\mathcal{O}(\mathrm{Src}(f))}\bar{x}_l^{(f)} - \sum_{l\in\mathcal{I}(\mathrm{Src}(f))}\bar{x}_l^{(f)} - \bar{s}_f)} = 0, \tag{11}$$

$$\begin{aligned}\frac{\partial \hat{f}_{\mu}^{(0)}(\bar{\mathbf{y}})}{\partial \bar{x}_l^{(f)}} = &\frac{1}{\mu(C_l - \sum_{f'=1}^{F}\bar{x}_l^{(f')})} - \frac{1}{\mu x_l^{(f)}} - \frac{1}{\mu(\sum_{l\in\mathcal{O}(\mathrm{Tx}(l))}\bar{x}_l^{(f)} - \sum_{l\in\mathcal{I}(\mathrm{Tx}(l))}\bar{x}_l^{(f)} - \bar{s}_f\mathbb{1}_f(\mathrm{Tx}(l)))} \\ &+ \frac{1}{\mu(\sum_{l\in\mathcal{O}(\mathrm{Rx}(l))}\bar{x}_l^{(f)} - \sum_{l\in\mathcal{I}(\mathrm{Rx}(l))}\bar{x}_l^{(f)} - \bar{s}_f\mathbb{1}_f(\mathrm{Rx}(l)))} = 0 \end{aligned} \tag{12}$$

In (11) and (12), with respect to the final terms, we define dual variables (also called "barrier multipliers", see [12, Section 3.1]) as follows:

$$\hat{p}_n^{(f)} = \frac{1}{\mu\left(\sum_{l\in\mathcal{O}(n)} \bar{x}_l^{(f)} - \sum_{l\in\mathcal{I}(n)} \bar{x}_l^{(f)} - \bar{s}_f \mathbb{1}_f(n)\right)}, \quad \forall f, \forall n \neq \mathrm{Dst}(f). \tag{13}$$

Clearly, if $\bar{\mathbf{y}}$ is strictly primal feasible, we have $\hat{p}_n^{(f)} > 0$. We use the vector $\widehat{\mathbf{p}} \triangleq [\hat{p}_n^{(f)}, \forall f, \forall n \neq \mathrm{Dst}(f)]^T$ to group all dual variables. Also, we let $\hat{f}_\mu(\bar{\mathbf{y}}) = -\sum_{f=1}^F U_f(\bar{s}_f) - \frac{1}{\mu}\sum_{l=1}^L \log\left(C_l - \sum_{f=1}^F \bar{x}_l^{(f)}\right) - \frac{1}{\mu}\sum_{f=1}^F \log(\bar{s}_f) - \frac{1}{\mu}\sum_{l=1}^L\sum_{f=1}^F \log(\bar{x}_l^{(f)})$. Substituting (13) in (11) and (12) and then using $\hat{f}_\mu(\bar{\mathbf{y}})$, $\widehat{\mathbf{p}}$ and the property of $\mathbf{M}$, we arrive at the following *perturbed* Karush-Kuhn-Tucker (KKT) system that contains stationarity (ST), primal feasibility (PF), dual feasibility (DF), and perturbed complementary slackness (CS) conditions:

$$\begin{aligned}
&\text{(ST): } \nabla \hat{f}_\mu(\bar{\mathbf{y}}) + \mathbf{M}^T \widehat{\mathbf{p}} = 0,\\
&\text{(PF): } \bar{\mathbf{y}} > \mathbf{0}, \quad \mathbf{M}\bar{\mathbf{y}} < \mathbf{0},\\
&\text{(DF): } \widehat{\mathbf{p}} > \mathbf{0},\\
&\text{(CS): } -\mathrm{Diag}\{\mathbf{M}\bar{\mathbf{y}}\}\,\widehat{\mathbf{p}} = (1/\mu)\mathbf{1}.
\end{aligned}$$

Compared to the classical KKT conditions [6], the only difference in this perturbed KKT system is that the right-hand side (RHS) of the CS condition is changed from $\mathbf{0}$ to $\frac{1}{\mu}\mathbf{1}$. As a result, as $\mu \to \infty$, the perturbed KKT point $(\bar{\mathbf{y}}, \widehat{\mathbf{p}})$ "almost" satisfies the classical KKT conditions, implying a *near-optimality*. Also, we point out that we have specially used the substitution (13) with respect to the $\mathbf{M}\bar{\mathbf{y}} < \mathbf{0}$ restrictions in order to handle them *explicitly* via the (PF) and (CS) conditions in the perturbed KKT system and enable the subsequent queuing design and analysis.

To simplify notation and algebraic derivations, we let $f_\mu(\bar{\mathbf{y}}) \triangleq \mu \hat{f}_\mu(\bar{\mathbf{y}})$. Accordingly, we let $\mathbf{p} = \mu\widehat{\mathbf{p}}$ absorb the μ-factor and work with the μ-scaled perturbed KKT system as follows:

$$\begin{aligned}
&(\mu\text{-ST}): \nabla f_\mu(\bar{\mathbf{y}}) + \mathbf{M}^T \mathbf{p} = 0, && (14)\\
&(\mu\text{-PF}): \bar{\mathbf{y}} > \mathbf{0}, \quad \mathbf{M}\bar{\mathbf{y}} < \mathbf{0}, && (15)\\
&(\mu\text{-DF}): \mathbf{p} > \mathbf{0}, && (16)\\
&(\mu\text{-CS}): -\mathrm{Diag}\{\mathbf{M}\bar{\mathbf{y}}\}\,\mathbf{p} = \mathbf{1}. && (17)
\end{aligned}$$

Step 2) Second-order Newton's method: We will now apply Newton's method to the perturbed KKT conditions (14)–(17), which is a *second-order* algorithm. We first work with the μ-ST and μ-CS conditions, while the μ-PF and μ-DF conditions will be handled later explicitly when determining the step-size to be taken along the Newton direction. Note that finding a primal-dual pair $(\bar{\mathbf{y}}, \mathbf{p})$ that satisfies the μ-ST and μ-CS conditions amounts to computing the roots of a *nonlinear*

equality system consisting of (14) and (17), which does not have analytic solutions in general and necessitates numerical methods. By using the Newton's method and given a feasible primal-dual pair $(\bar{\mathbf{y}}^k, \mathbf{p}^k)$, one can compute the Newton direction $[(\Delta\bar{\mathbf{y}}^k)^T, (\Delta\mathbf{p}^k)^T]^T$ as (see [6]):

$$\left[\begin{array}{cc} \mathbf{H}_k & \mathbf{M}^T \\ -\mathbf{P}_k\mathbf{M} & -\mathbf{Q}_k \end{array}\right]\left[\begin{array}{c} \Delta\bar{\mathbf{y}}^k \\ \Delta\mathbf{p}^k \end{array}\right] = -\left[\begin{array}{c} \mathbf{g}^k + \mathbf{M}^T\mathbf{p}^k \\ -(\mathbf{P}_k\mathbf{Q}_k + \mathbf{I})\mathbf{1} \end{array}\right], \tag{18}$$

where we let $\mathbf{g}^k \triangleq \nabla f_\mu(\bar{\mathbf{y}}^k)$, $\mathbf{H}_k \triangleq \nabla^2 f_\mu(\bar{\mathbf{y}}^k)$, $\mathbf{P}_k \triangleq \text{Diag}\left\{\mathbf{p}^k\right\}$, and $\mathbf{Q}_k \triangleq \text{Diag}\left\{\mathbf{M}\bar{\mathbf{y}}^k\right\}$. Note that, due to the perturbed KKT conditions, (18) is *different* from the Newton systems in existing second-order methods (cf. [8, Eq.(4)], [9, Eq.(8)], [10, Eq.(9)]). Also, directly solving (18) is undesirable due to its complex structure. A better way for solving (18) is to derive a reduced linear system by Gaussian elimination to obtain (assuming $\mathbf{p}^k > \mathbf{0}$ and hence $\mathbf{P}_k$ is non-singular, which can be ensured by the step-size control described next):

$$\Delta\bar{\mathbf{y}}^k = -\left(\mathbf{H}_k - \mathbf{M}^T\mathbf{Q}_k^{-1}\mathbf{P}_k\mathbf{M}\right)^{-1}(\mathbf{g}^k - \mathbf{M}^T\mathbf{Q}_k^{-1}\mathbf{1}), \tag{19}$$

$$\begin{aligned}\Delta\mathbf{p}^k = &-\left(\mathbf{M}\mathbf{H}_k^{-1}\mathbf{M}^T - \mathbf{P}_k^{-1}\mathbf{Q}_k\right)^{-1} \\ &\times [\mathbf{M}\mathbf{H}_k^{-1}(\mathbf{g}_k + \mathbf{M}^T\mathbf{p}^k) - (\mathbf{Q}_k + \mathbf{P}_k^{-1})\mathbf{1}].\end{aligned} \tag{20}$$

Now, it is not difficult to recognize the structural similarity between (5)–(6) and (19)–(20).

Next, we handle the μ-PF and μ-DF conditions by step-size control: In iteration k, we update the primal and dual variables as $\bar{\mathbf{y}}^{k+1} = \bar{\mathbf{y}}^k + \pi^k\Delta\bar{\mathbf{y}}^k$ and $\mathbf{p}^{k+1} = \mathbf{p}^k + \pi^k\Delta\mathbf{p}^k$. In standard primal-dual interior-point methods [12], the step-size control is based on *two* rules: The first one is to satisfy primal-dual feasibility by finding:

$$\max\left\{\pi \in [0,1] \;\middle|\; \begin{array}{l} \bar{\mathbf{y}}^k + \pi\Delta\bar{\mathbf{y}}^k \geq \epsilon\mathbf{1}, \\ \mathbf{M}(\bar{\mathbf{y}}^k + \pi\Delta\bar{\mathbf{y}}^k) \leq -\epsilon\mathbf{1}, \\ \mathbf{N}(\bar{\mathbf{y}}^k + \pi\Delta\bar{\mathbf{y}}^k) \leq \mathbf{c} - \epsilon\mathbf{1}, \\ \mathbf{p}^k + \pi\Delta\mathbf{p}^k \geq \epsilon\mathbf{1}, \end{array}\right\}, \tag{21}$$

where $\epsilon > 0$ is some arbitrarily small constant. Note that a full Newton step is taken if $\pi^k = 1$. The second step-size selection rule is to guarantee a *decreasing residual*. Specifically, let $\mathbf{r}_\mu(\bar{\mathbf{y}}^k, \mathbf{p}^k) \triangleq [(\mathbf{g}^k + \mathbf{M}^T\mathbf{p}^k)^T,\ (-\mathbf{P}_k\mathbf{Q}_k\mathbf{1} - \mathbf{1})^T]^T$ be the residual of μ-ST and μ-CS at $\bar{\mathbf{y}}^k$ (i.e., the right-hand side (RHS) of (18)). The second rule is to choose π^k to satisfy [12]:

$$\|\mathbf{r}_\mu(\bar{\mathbf{y}}^{k+1}, \mathbf{p}^{k+1})\| < \|\mathbf{r}_\mu(\bar{\mathbf{y}}^k, \mathbf{p}^k)\|. \tag{22}$$

Under the step-size rules in (21) and (22), the convergence and second-order convergence speed analysis follow from standard primal-dual interior-point methods (see [12]).

Step 3) Back to Algorithm 1: Now, we can see that Algorithm 1 indeed mimics the foregoing approach to adjust $\mathbf{y}_{[t]}$ in *every time-slot*, rather than the average rate $\bar{\mathbf{y}}^k$. Moreover, Algorithm 1

has a much simplified step-size selection rule: We do *not* require a delicate line search to determine π^k as in (21) and have the residuals (in the form of the RHS of (18)) decrease, both of which are expensive to check due to a large number of gradient and constraint evaluations in each time-slot. Rather, we use a fixed step-size $\pi \in (0, 1]$ and a projection to maintain primal-dual feasibility (a basic requirement in an interior-point method). Surprisingly, even with this much simplified and relaxed step-size rule, we are still able to show that the time-average of $\{\mathbf{y}_{[t]}, \mathbf{p}_{[t]}\}_{t=0}^{\infty}$ converges to a bounded region around the optimal solution as indicated in Theorem 1, which is *exactly the goal* of Problem JCCR.

4.3 Proofs of the Main Theorems

In this section, we provide sketched proofs for the theorems in Section 4.1 for better readability. The detailed proof derivations can be found in the appendices. First, we show a basic property of the dual sequence $\{\mathbf{p}_{[t]}\}_{t=0}^{\infty}$ that will be useful in proving Theorems 1 and 3.

Lemma 4. *For a given μ and under Algorithm 1, if $\|\mathbf{p}_{[0]}\| < \infty$, then $\|\mathbf{p}_{[t]}\| < \infty$ for all t.*

We prove Lemma 4 by induction. Suppose that $\|\mathbf{p}_{[t]}\| < \infty$. We let $\widetilde{\mathbf{p}}_{[t+1]}$ be obtained by taking a full Newton step (i.e., $\pi = 1$). Note that once we show $\|\widetilde{\mathbf{p}}_{[t+1]}\| < \infty$, the result stated in Lemma 4 immediately follows from the fact that $\mathbf{p}_{[t+1]}$ is a convex combination of $\mathbf{p}_{[t]}$ and $\widetilde{\mathbf{p}}_{[t+1]}$, and hence its norm must also be upper bounded. We relegate the proof details to Appendix A.

Sketch of the proof of Theorem 1. The main idea and key steps for proving Theorem 1 are as follows. First, we consider the one-slot drift of the following particular choice of quadratic Lyapunov function:

$$V\left(\mathbf{y}_{[t]}, \mathbf{p}_{[t]}\right) \triangleq \frac{1}{2\pi}\left\|\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right\|^2 + \frac{1}{2\mu^3\pi}\left\|\mathbf{p}_{[t]} - \mathbf{p}^*\right\|^2,$$

which can be interpreted as measuring the (unscaled) distance between a primal-dual iterate $(\mathbf{y}_{[t]}, \mathbf{p}_{[t]})$ and a perturbed KKT point $(\bar{\mathbf{y}}^*, \mathbf{p}^*)$ satisfying (14)–(17). For simplicity, we let $\mathbf{F}_{[t]}$ and $\mathbf{G}_{[t]}$ be defined as follows: $\mathbf{F}_{[t]} \triangleq \mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M}$ and $\mathbf{G}_{[t]} \triangleq \mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}$. Then, after some algebraic derivations and upper-bounding (see Appendices B.1 and B.2 for derivation details), we obtain the following relationship:

$$\Delta V\left(\mathbf{y}_{[t]}, \mathbf{p}_{[t]}\right) \triangleq V\left(\mathbf{y}_{[t+1]}, \mathbf{p}_{[t+1]}\right) - V\left(\mathbf{y}_{[t]}, \mathbf{p}_{[t]}\right) \leq -R\|\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\|^2 + \pi B_1 + \frac{1}{\mu}B_2 + \frac{1}{\mu}B_3, \quad (23)$$

where $R \triangleq \frac{\lambda_{\min}\{\mathbf{H}\}}{\lambda_{\min}\{\mathbf{F}\}} > 0$ and is independent of μ; and B_1, B_2, and B_3 are some positive constants

that do *not* scale with μ, and are defined as follows:

$$B_1 \triangleq \frac{1}{2\lambda_{\min}^2\{\mathbf{F}\}}\sup_t\left\{\left[\left\|\mathbf{g}_{[t]}-\mathbf{g}^*\right\|^2+\left\|\mathbf{M}^T\left(\mathbf{Q}_{[t]}^{-1}-\mathbf{Q}_*^{-1}\right)\right\|^2\right]\right\},$$

$$B_2 \triangleq \frac{\|\mathbf{M1}\|}{\mu^2\lambda_{\min}\{\mathbf{G}\}}\sup_t\left\{\left\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right\|\left\|\mathbf{p}_{[t]}-\mathbf{p}^*\right\|\right\},$$

$$B_3 \triangleq \frac{1}{2\mu^2\lambda_{\min}^2\{\mathbf{G}\}}\sup_t\left\{\left[\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\|\|\mathbf{M1}\|\right.\right.$$
$$\left.\left.+\|\mathbf{M}(\mathbf{H}_{[t]}^{-1}\mathbf{g}^*-\bar{\mathbf{y}}^*)\|\|\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T\mathbf{p}_{[t]}\|+\|\mathbf{P}_{[t]}^{-1}\mathbf{1}\|\right]\right\}.$$

It can be seen from (23) that if $\pi = O(1/\mu)$, we have $V(\mathbf{y}_{[t+1]},\mathbf{p}_{[t+1]}) - V(\mathbf{y}_{[t]},\mathbf{p}_{[t]}) \leq -R\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\|^2 + \frac{1}{\mu}\widehat{B}$, where $\widehat{B} \triangleq \alpha B_1 + B_2 + B_3$ for some $\alpha > 0$. Telescoping T via one-slot drift expressions for $t = 0,\ldots,T-1$ yields:

$$V\left(\mathbf{y}_{[T]},\mathbf{p}_{[T]}\right) - V\left(\mathbf{y}_{[0]},\mathbf{p}_{[0]}\right) \leq -R\sum_{t=0}^{T-1}\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\|^2 + \frac{T}{\mu}\widehat{B}.$$

Next, dividing both sides by TR, rearranging terms, and taking T to infinity, we have $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\|^2 \leq \frac{B^2}{\mu}$, where we let $B^2 \triangleq \widehat{B}/R$. Then, the proof is complete because when T is large, we have

$$\left|\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right)\right| \overset{(a)}{\leq} \left(\frac{1}{T}\sum_{t=0}^{T-1}\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\|^2\right)^{\frac{1}{2}} \leq \frac{B}{\sqrt{\mu}},$$

where (a) follows from the triangular inequality and the basic relationship between l_1- and l_2-norms. We note that the most challenging step in the proof lies in the one-slot drift analysis, where we repeatedly exploit the key relationships in the perturbed KKT system in (14)–(17). We relegate the derivation details to Appendix B. □

Proof of Proposition 2. The results in Proposition 2 follow immediately from (23) and noting the fact that $\frac{\lambda_{\min}\{\mathbf{H}\}}{\lambda_{\min}\{\mathbf{F}\}} \to 1$ as μ gets large. □

Proof of Theorem 3. The basic idea to prove Theorem 3 is based on analyzing the one-slot drift of the following quadratic Lyapunov function: $\hat{V}\left(\mathbf{q}_{[t]}\right) \triangleq \frac{1}{2}\left\|\mathbf{q}_{[t]}\right\|^2$. For convenience, we let $\widehat{\mathbf{y}}_{[t]} \triangleq \left[s_{1,[t]}\cdots s_{F,[t]}, \widehat{x}_{1,[t]}^{(1)}\cdots\widehat{x}_{1,[t]}^{(F)},\cdots,\widehat{x}_{L,[t]}^{(1)}\cdots\widehat{x}_{L,[t]}^{(F)}\right]^T$ group all source and *actual* routing rates. Note that $\widehat{\mathbf{y}}_{[t]} \leq \mathbf{y}_{[t]}$ since $\widehat{x}_{l,[t]}^{(f)} \leq x_{l,[t]}^{(f)}$. Then, the queuing dynamic can be written as $\mathbf{q}_{[t+1]} = \mathbf{q}_{[t]} + \mathbf{M}\widehat{\mathbf{y}}_{[t]}$ and the one-slot drift $\Delta\hat{V}$ can be bounded as:

$$\begin{aligned}\Delta\hat{V} &= \frac{1}{2}\|\mathbf{q}_{[t+1]}\|^2 - \frac{1}{2}\|\mathbf{q}_{[t]}\|^2 = \mathbf{q}_{[t]}^T\mathbf{M}\widehat{\mathbf{y}}_{[t]} + \frac{1}{2}\widehat{\mathbf{y}}_{[t]}^T(\mathbf{M}^T\mathbf{M})\widehat{\mathbf{y}}_{[t]}\\ &\leq \mathbf{q}_{[t]}^T\mathbf{M}\widehat{\mathbf{y}}_{[t]} + \frac{1}{2}\mathbf{y}_{[t]}^T(\mathbf{M}^T\mathbf{M})\mathbf{y}_{[t]}\\ &\overset{(a)}{\leq} \mathbf{q}_{[t]}^T\mathbf{M}\mathbf{y}_{[t]} + NL\max_{\forall l}\{C_l\} + \frac{1}{2}\mathbf{y}_{[t]}^T(\mathbf{M}^T\mathbf{M})\mathbf{y}_{[t]},\end{aligned} \tag{24}$$

where (a) is due to [3, Lemma 1]. Now, we let $B_4 \triangleq NL\max_{\forall l}\{C_l\} + \frac{1}{2}\lambda_{\max}\{\mathbf{M}^T\mathbf{M}\}\sup_t\{\|\mathbf{y}_{[t]}\|^2\}$. Note that $NL\max_{\forall l}\{C_l\}$ and $\lambda_{\max}\{\mathbf{M}^T\mathbf{M}\}$ are determined by the network topology and $\sup_t\{\|\mathbf{y}_{[t]}\|^2\} \leq (\max\{M, \max_{\forall l} C_l\})^2$. As a result, B_4 depends only on the network and is independent of μ. On the other hand, according to our step-size control in (8) and that $\epsilon = O(\frac{1}{\mu})$, we have $\mathbf{M}\mathbf{y}_{[t]} \leq -\frac{\beta}{\mu}\mathbf{1}$ for some $\beta > 0$. Therefore, we have

$$\begin{aligned}\Delta\hat{V} &\leq \mathbf{q}_{[t]}^T\mathbf{M}\mathbf{y}_{[t]} + B_4 \leq -\frac{\beta}{\mu}\mathbf{q}_{[t]}^T\mathbf{1} + B_4\\ &= -\frac{\beta}{\mu}\sum_{f=1}^{F}\sum_{n\neq \mathrm{Dst}(f)} q_n^{(f)}[t] + B_4. \end{aligned} \tag{25}$$

So it follows that when $\sum_{f=1}^{F}\sum_{n\neq\mathrm{Dst}(f)} q_n^{(f)}[t] \geq \frac{\mu}{\beta}(B_4+\epsilon_1)$, where $\epsilon_1 > 0$ is some constant, we have $\Delta\hat{V}(\mathbf{q}_{[t]}) \leq -\epsilon_1$, i.e., the first term in (25) dominates B_4 and results in a negative drift when the total queue length is large.

Next, we claim that the following relationship is true:

$$\limsup_{t\to\infty}\hat{V}(\mathbf{q}_{[t]}) \leq \frac{\mu^2}{2\beta^2}(B_4+\epsilon_1)^2 + B_4. \tag{26}$$

This claim can be shown by the following argument: First, suppose that $\hat{V}(\mathbf{q}_{[t]}) \leq \frac{\mu^2}{2\beta^2}(B_4+\epsilon_1)^2$. From (25), we know that $\mathbf{q}_{[t]}^T\mathbf{M}\mathbf{y}_{[t]} \leq 0$, which further implies that $\Delta\hat{V}(\mathbf{q}_{[t]}) < B_4$. As a result, we have

$$\hat{V}(\mathbf{q}_{[t+1]}) = \hat{V}(\mathbf{q}_{[t]}) + \Delta\hat{V}(\mathbf{q}_{[t]}) \leq \frac{\mu^2}{2\beta^2}(B_4+\epsilon_1)^2 + B_4,$$

i.e., (26) is true. On the other hand, suppose that $\hat{V}(\mathbf{q}_{[t]}) > \frac{\mu^2}{2\beta^2}(B_4+\epsilon_1)^2$. From the basic relationship between l_1- and l_2-norms, we have $(2\hat{V}(\mathbf{q}_{[t]}))^{\frac{1}{2}} \leq \sum_{f=1}^{F}\sum_{n\neq\mathrm{Dst}(f)} q_n^{(f)}[t]$. This implies that if $\hat{V}(\mathbf{q}_{[t]}) > \frac{\mu^2}{2\beta^2}(B_4+\epsilon_1)^2$, we have $\Delta\hat{V}(\mathbf{q}_{[t]}) \leq -\epsilon_1$. This means that $\hat{V}(\mathbf{q}_{[t+1]}) < \hat{V}(\mathbf{q}_{[t]})$ and that the sequence $\{\hat{V}(\mathbf{q}_{[t]})\}$ will monotonically decrease at a rate at least ϵ_1. Therefore, there exists a time t' such that $\hat{V}(\mathbf{q}_{[t']}) \leq \frac{\mu^2}{2\beta^2}(B_4+\epsilon_1)^2$, and then the rest follows from the earlier discussions in the case where $\hat{V}(\mathbf{q}_{[t]}) \leq \frac{\mu^2}{2\beta^2}(B_4+\epsilon_1)^2$. Finally, we let $K^2 \triangleq 2\big[\frac{\mu^2}{2\beta^2}(B_4+\epsilon_1)^2 + B_4\big]$ and note that K^2 scales as $O(\mu^2)$. Then, the result stated in the theorem follows by multiplying both sides of (26) by two and taking the square root. This completes the proof. □

4.4 Key Insights for the Theoretical Results

It is insightful to compare our results with those of the first-order methods in [1–4]. First and foremost, as we mentioned earlier, Algorithm 1 reveals an important connection between our second-order method and an *observable* network state information: the potential queue-length changes $\mathbf{Q}_{[t]}$. As opposed to first-order methods where queue-length itself is *directly used* as a price, the $\mathbf{Q}_{[t]}$-terms

Table 1: Performance scaling-law comparisons.

	2nd-order	1st-order (Primal-dual: [3])	1st-order (Dual: [1, 2, 4])
Optimality gap	$O(\frac{1}{\sqrt{\mu}})$	$O(\frac{1}{\sqrt{V}})$	$O(\frac{1}{V})$
Queue-length	$O(\mu)$	$O(V)$	$O(V)$
Step-size	$O(\frac{1}{\mu})$	$O(\frac{1}{V^2})$	$O(\frac{1}{V})$

in (6) show that our pricing scheme is based on the *change of queue-length*, hence providing another perspective to interpret the name "second-order method." Also, Proposition 2 indicates that the negative Lyapunov drift rate tends to be *steeper* and *insensitive* to the objective function contour, thanks to the Hessian scaling factor. This avoids the potential ill-conditioned limitations in first-order methods and explains the fast convergence performance.

In addition to the aforementioned salient features, there are several key insights regarding the performance scaling laws in first- and second-order methods. We note that, like most first-order methods, Theorems 1 and 3 imply a *trade-off* relationship between optimality gap and queue-length (hence delay). Particularly, although having a fundamentally different algorithmic meaning, the barrier parameter μ in our second-order method does play a *similar* role in performance characterizations compared to the subgradient step-size scaling factor in first-order methods (e.g., "h" in [1, 4], "V" in [2], and "K" in [3]). Accordingly, we summarize the performance scaling laws of the first- and second-order methods in Table 1 (all parameters in first-order methods are standardized to "V").

First, we can see that all schemes have a similar queue-length scaling. The optimality gap scaling in [3] and our work are similar due to the common primal-dual nature. However, the $O(\frac{1}{\sqrt{V}})$-scaling in [3] is achieved at a slower convergence performance and under a more restrictive step-size scaling described next. For the dual-based controller in [2] (optimality gap scaling was not discussed in [1,4]), the optimality gap scales as $O(\frac{1}{V})$. Although this result appears to be better at first glance, a closer look reveals that such a direct comparison cannot be made. In [2], the gap is measured by (in our notation) $\sum_f U_f(s_f^*) - \sum_f U_f(s_f)$. In contrast, Theorem 1 measures the gap by $\|\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\|$. We point out that our metric is stronger since it measures the distance to the optimal solution in *every* coordinate, while the metric in [2] only addresses the objective value gap. Since the objective function is continuous, a coordinate-wise near-optimality implies a near-optimality in the objective value, but the reverse is *not* necessarily true.

For step-size scaling, we can see that, for the first-order primal-dual scheme in [3] to approach optimality, the step-size should scale as $O(\frac{1}{V^2})$, which is *much smaller* than our $O(\frac{1}{\mu})$-scaling. On the other hand, although there is no direct primal-dual step-size counterpart in dual-based controllers [1, 2,4], the dual step-size scaling therein can be understood as $O(\frac{1}{V})$, similar to our $O(\frac{1}{\mu})$. However, this

$O(\frac{1}{V})$-scaling is obtained under the dual-based architecture, which is more cumbersome to implement due to the coupled inner primal subproblem. Finally, we remark that our $O(\frac{1}{\mu})$ step-size scaling is *not* restrictive in practical implementations since it is just a sufficient condition to establish the result in Theorem 1. Given that the proof of Theorem 1 is a limiting argument where the bounding constants B_1, B_2, and B_3 are not tight, the choice of the constant in $O(\frac{1}{\mu})$-scaling does not have to be conservative. In practice, the more restrictive requirement is the primal and dual feasibility assurance, which plays a key role in offering queuing stability.

So far, we have designed a second-order joint congestion control and routing algorithm and established its optimality and queuing stability. However, given the more complex computational scheme in Algorithm 1, one question begs to be answered: *Can we design a distributed algorithm based on the proposed second-order method?* Moreover, although it is convenient to express (5) and (6) in matrix equations, they are cumbersome to use and more *explicit* scalar-based expressions are desired for implementations in practice. These issues constitute the main discussions in the next section.

5 Second-Order Distributed Algorithm Design

In this section, our main goal is to *decentralize* the proposed second-order method in Section 4. Note that the main computational complexity in (5) and (6) stems from the following two *dense* matrix inverse computations that require global network information:

$$\mathbf{F}_{[t]}^{-1} = \left(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M}\right)^{-1}, \tag{27}$$

$$\mathbf{G}_{[t]}^{-1} = \left(\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}\right)^{-1}. \tag{28}$$

Thus, our effort in this section is centered around tackling these two challenges. We first derive an alternative way for computing the primal and dual Newton directions in Section 5.1. Next, we develop distributed computational schemes for the primal and dual Newton directions in Sections 5.2 and 5.3, respectively.

5.1 An Alternative Approach for Computing the Newton Directions

Our first step toward designing a second-order distributed joint congestion control algorithm is to simplify the primal and dual Newton direction computational schemes in (5) and (6) in order to facilitate a distributed design. The rationale behind this simplification is based on the following observation: While (5) and (6) "cleanly" express $\mathbf{y}_{[t+1]}$ and $\mathbf{p}_{[t+1]}$ *only* in terms of $\mathbf{y}_{[t]}$ and $\mathbf{p}_{[t]}$ and enable all the subsequent optimality and queuing stability proofs, they also make the computational schemes unnecessarily more complex for practical implementations. Toward this end, we establish the following lemma that will be useful in Sections 5.2 and 5.3:

Lemma 5. *The primal and dual Newton directions in (5) and (6) can be alternatively computed as follows:*

$$\Delta \mathbf{y}_{[t]} = -\mathbf{H}_{[t]}^{-1} \left(\mathbf{g}_{[t]} - \mathbf{M}^T \widetilde{\mathbf{p}}_{[t+1]} \right), \tag{29}$$

$$\Delta \mathbf{p}_{[t]} = \widetilde{\mathbf{p}}_{[t+1]} - \mathbf{p}_{[t]}, \tag{30}$$

where $\widetilde{\mathbf{p}}_{[t+1]}$ *is obtained by starting from* $\mathbf{p}_{[t]}$ *and taking a unit step-size (i.e.,* $\pi[t] = 1$*), which can be computed as:*

$$\widetilde{\mathbf{p}}_{[t+1]} = \mathbf{G}^{-1} \left[\mathbf{M} \mathbf{H}_{[t]}^{-1} (-\mathbf{g}_{[t]}) + \mathbf{P}_{[t]}^{-1} \mathbf{1} \right]. \tag{31}$$

The key idea here is that, through the use of an auxiliary variable $\widetilde{\mathbf{p}}_{[t+1]}$, the expressions in (29) and (30) can be made much simpler. Clearly, (30) follows from the definition of $\widetilde{\mathbf{p}}_{[t+1]}$. Since the expressions in (29) and (31) are not obvious, we provide a proof in Appendix C. With Lemma 5, we are now in a position to derive a distributed scheme for computing primal and dual Newton directions.

5.2 Distributed Computation of the Primal Newton Direction

The first advantage of using the new scheme in (29) is that instead of having to deal with $\mathbf{F}$, which is the unstructured and dense matrix, we are now faced with $\mathbf{H}_{[t]}$, which has the following nice *block diagonal* structure:

$$\mathbf{H}_{[t]} = \text{Diag} \left\{ \mathbf{S}_{[t]}, \mathbf{X}_{1,[t]}, \ldots, \mathbf{X}_{L,[t]} \right\},$$

where $\mathbf{S}_{[t]}$ is a diagonal matrix defined as

$$\mathbf{S}_{[t]} \triangleq \text{Diag} \left\{ -\mu U_f''(s_f[t]) + \frac{1}{s_f^2[t]},\ f = 1, \ldots, F \right\} \in \mathbb{R}^{F \times F}; \tag{32}$$

and where $\mathbf{X}_l \in \mathbb{R}^{F \times F}$ is a symmetric matrix with entries defined as follows:

$$(\mathbf{X}_{l,[t]})_{f_1,f_2} = \begin{cases} \frac{1}{\delta_l^2[t]} + \frac{1}{\left(x_l^{(f_1)}[t]\right)^2} & \text{if } f_1 = f_2, \\ \frac{1}{\delta_l^2[t]} & \text{if } f_1 \neq f_2, \end{cases} \tag{33}$$

where $\delta_l[t] \triangleq C_l - \sum_{f=1}^{F} x_l^{(f)}[t]$ represents the *unused link capacity* of link l in time-slot t, which will occur frequently in the rest of the paper. It then follows from the block diagonal structure of $\mathbf{H}_{[t]}$ that

$$\mathbf{H}_{[t]}^{-1} = \text{Diag} \left\{ \mathbf{S}_{[t]}^{-1}, \mathbf{X}_{1,[t]}^{-1}, \ldots, \mathbf{X}_{L,[t]}^{-1} \right\}. \tag{34}$$

We note that this block diagonal structure of the Hessian is exactly the same as that in [9, Section V-C] (after replacing the long-term average rates by instantaneous rates in each time-slot t). Due to the same structure as their counterparts in [9], $\mathbf{S}_{[t]}^{-1}$ and $\mathbf{X}_{l,[t]}^{-1}$ can be computed in closed-form by using Lemma 4 and Theorem 5 in [9]. Further, by noting the similarity in structure to the primal Newton direction scheme in [9, Eq. (9)], we immediately have the following result:

Theorem 6. *Let $\widehat{\mathbf{x}}_l$ be defined as in [9, Theorem 6]. Given dual prices $\mathbf{p}_{[t]}$, the congestion control and routing directions $\Delta s_{f,[t]}$ and $\Delta x_{l,[t]}^{(f)}$ can be computed in closed-form using local information at each source node s and link l, respectively, as follows (omitting time-slot indexes "$[t]$" and "$[t+1]$" for simplicity):*

$$\Delta s_{f,[t]} = \frac{s_{f,[t]}\left(\mu s_{f,[t]} U_f'(s_{f,[t]}) + 1 - s_{f,[t]}\widetilde{p}^{(f)}_{\mathrm{Src}(f),[t]}\right)}{1 - \mu s_{f,[t]}^2 U_f''(s_{f,[t]})}, \qquad \forall f, \tag{35}$$

$$\Delta x_{l,[t]}^{(f)} = \left(x_{l,[t]}^{(f)}\right)^2 \left[\left(1 - \frac{(x_{l,[t]}^{(f)})^2}{\|\widehat{\mathbf{x}}_{l,[t]}\|^2}\right)\left(\frac{1}{x_{l,[t]}^{(f)}} - \frac{1}{\delta_{l,[t]}} + \widetilde{p}^{(f)}_{\mathrm{Tx}(l),[t]} - \widetilde{p}^{(f)}_{\mathrm{Rx}(l),[t]}\right) + \sum_{f'=1, f'\neq f}^{F} \frac{(x_{l,[t]}^{(f')})^2}{\|\widehat{\mathbf{x}}_{l,[t]}\|^2}\left(\frac{1}{x_{l,[t]}^{(f')}} - \frac{1}{\delta_{l,[t]}} + \widetilde{p}^{(f')}_{\mathrm{Tx}(l),[t]} - \widetilde{p}^{(f')}_{\mathrm{Rx}(l),[t]}\right)\right], \qquad \forall l, f. \tag{36}$$

,

The proof of Theorem 6 follows the same line as in [9]: (i) applying (34) as well as Lemma 4 and Theorem 5 of [9] in (29); and (ii) exploiting the second-order properties of $\mathbf{a}_l^{(f)}$ and $\mathbf{b}^{(f)}$ to simplify the result. Hence, we omit the proof of this theorem for brevity.

Remark 2. *Theorem 6 has two interesting networking interpretations. First, the dual price differential $(\widetilde{p}^{(f)}_{\mathrm{Tx}(l)} - \widetilde{p}^{(f)}_{\mathrm{Rx}(l)})$ in (36) plays a similar role of the queuing backlog differential in the back-pressure schemes. The main difference is that $\Delta x_l^{(f)}$ (i.e., to increase or decrease $x_{l,[t]}^{(f)}$) is based on not only the pressure differential of session f, but that of all sessions in link l. Moreover, unlike the "winner-take-all" policy in the back-pressure schemes (i.e., the session with the largest backlog differential uses up the link capacity), our second-order approach is more "democratic" in that every session gets a share of the link capacity as indicated in (36).*

5.3 Distributed Computation of the Dual Newton Direction

Recall that the dual Newton direction $\Delta\mathbf{p}_{[t]}$ can be computed indirectly by first solving for the auxiliary variable $\widetilde{\mathbf{p}}_{[t]}$ in (31). However, there remains one key technical challenge in this approach: The matrix $\mathbf{G} = \left(\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}\right)$ contains a weighted Laplacian matrix term $\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T$, which is dense and involves global information. As a result, it is generally intractable to derive a distributed and closed-form analytic expression for $\mathbf{G}^{-1}$ except for some simplistic network structures.

In what follows, we will first analyze the structure of $\mathbf{G}$ and then propose two strategies to compute the dual Newton direction in a distributed fashion. Recall that $\mathbf{M}$ can be written in a partitioned

matrix form as $\widetilde{\mathbf{M}} = \begin{bmatrix} \mathbf{B} & \mathbf{A}_1 & \cdots & \mathbf{A}_L \end{bmatrix}$. Hence, we can decompose $\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T$ as

$$\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T = \begin{bmatrix} \mathbf{B} & \mathbf{A}_1 & \cdots & \mathbf{A}_L \end{bmatrix} \begin{bmatrix} \mathbf{S}^{-1} & & & \\ & \mathbf{X}_1^{-1} & & \\ & & \ddots & \\ & & & \mathbf{X}_L^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{B}^T \\ \mathbf{A}_1^T \\ \vdots \\ \mathbf{A}_L^T \end{bmatrix} = \mathbf{B}\mathbf{S}^{-1}\mathbf{B}^T + \sum_{l=1}^{L} \mathbf{A}_l \mathbf{X}_l^{-1} \mathbf{A}_l^T. \tag{37}$$

Now, we consider each term in the decomposition in (37). For $\mathbf{B}\mathbf{S}^{-1}\mathbf{B}^T$, since $\mathbf{B}$ and $\mathbf{S}^{-1}$ are diagonal, we have

$$\mathbf{B}\mathbf{S}^{-1}\mathbf{B}^T = \mathrm{Diag}\left\{ \frac{1}{-\mu U_1''(s_1) + \frac{1}{(s_1)^2}} \mathbf{b}^{(1)}(\mathbf{b}^{(1)})^T, \ldots, \frac{1}{-\mu U_F''(s_F) + \frac{1}{(s_F)^2}} \mathbf{b}^{(F)}\mathbf{b}^{(F)} \right\}, \tag{38}$$

which is a block diagonal matrix. Moreover, from the definition of $\mathbf{b}^{(f)}$, each bock has the following structure:

$$\frac{1}{-\mu U_f''(s_f) + \frac{1}{(s_f)^2}} \mathrm{Diag}\left\{0 \ldots 1 \ldots 0\right\},$$

where the position of the only non-zero entry 1 corresponds to node $\mathrm{Src}(f)$. Next, consider the term $\sum_{l=1}^{L} \mathbf{A}_l \mathbf{X}_l^{-1} \mathbf{A}_l^T$, which is more involved. From [9, Theorem 5], we can decompose $\sum_{l=1}^{L} \mathbf{A}_l \mathbf{X}_l^{-1} \mathbf{A}_l^T$ as follows:

$$\begin{aligned} \sum_{l=1}^{L} \mathbf{A}_l \mathbf{X}_l^{-1} \mathbf{A}_l^T = \sum_{l=1}^{L} \Bigg\{ & \begin{bmatrix} -\mathbf{a}_l^{(1)} & & \\ & \ddots & \\ & & -\mathbf{a}_l^{(F)} \end{bmatrix} \Bigg(\begin{bmatrix} (x_l^{(1)})^2 & & \\ & \ddots & \\ & & (x_l^{(F)})^2 \end{bmatrix} \\ & - \frac{1}{\|\widehat{\mathbf{x}}_l\|^2} \begin{bmatrix} (x_l^{(1)})^2 \\ \vdots \\ (x_l^{(F)})^2 \end{bmatrix} \begin{bmatrix} (x_l^{(1)})^2 & \cdots & (x_l^{(F)})^2 \end{bmatrix} \Bigg) \begin{bmatrix} -\mathbf{a}_l^{(1)} & & \\ & \ddots & \\ & & -\mathbf{a}_l^{(F)} \end{bmatrix}^T \Bigg\}. \end{aligned} \tag{39}$$

Due to the block diagonal structure, the first term in (39) can be further written as

$$\sum_{l=1}^{L} \mathrm{Diag}\left\{ (x_l^{(1)})^2 \mathbf{a}_l^{(1)} (\mathbf{a}_l^{(1)})^T, \ldots, (x_l^{(F)})^2 \mathbf{a}_l^{(F)} (\mathbf{a}_l^{(F)})^T \right\},$$

which is also a block diagonal matrix. Moreover, we note that the term $-\mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}$ is a diagonal matrix, which can be written as:

$$-\mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]} = \mathrm{Diag}\left\{ \frac{1}{p_n^{(f)}[t]} \left[\sum_{l \in \mathcal{O}(n)} x_l^{(f)}[t] - s_f \mathbb{1}_f(n) - \sum_{l \in \mathcal{I}(n)} x_l^{(f)}[t] \right],\ f = 1, \ldots, F \right\},$$

where $\mathbb{1}_f(n)$ is an indicator function defined as:

$$\mathbb{1}_f(n) = \begin{cases} 1 & \text{if } n = \mathrm{Src}(f), \\ 0 & \text{otherwise.} \end{cases}$$

Thus, we can combine these two terms with $\mathbf{B}\mathbf{S}^{-1}\mathbf{B}^T$. For convenience, we let $\mathbf{D} \triangleq \mathbf{B}\mathbf{S}^{-1}\mathbf{B}^T + \sum_{l=1}^{L} \mathrm{Diag}\left\{(x_l^{(1)})^2 \mathbf{a}_l^{(1)}(\mathbf{a}_l^{(1)})^T, \ldots, (x_l^{(F)})^2 \mathbf{a}_l^{(F)}(\mathbf{a}_l^{(F)})^T\right\} - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}$. Clearly, $\mathbf{D}$ is also block diagonal, and can be written as $\mathbf{D} = \mathrm{Diag}\{\mathbf{D}_1, \ldots, \mathbf{D}_F\}$. Then, by using [9, Lemma 2], we obtain the following result, where the proof is relegated to Appendix D.

Lemma 7. *The matrix $\mathbf{D}$ is block diagonal and each block $\mathbf{D}_f$ on the main diagonal has the following structure:*

- *The diagonal entries $(\mathbf{D}_f)_{ii}$ are given by*

$$(\mathbf{D}_f)_{ii} = \begin{cases} \sum_{l\in\mathcal{O}(n)\cup\mathcal{I}(n)}(x_l^{(f)})^2 + \frac{1}{-\mu U_f''(s_f)+\frac{1}{(s_f)^2}} + \frac{1}{p_n^{(f)}}\left[\sum_{l\in\mathcal{O}(n)} x_l^{(f)} - s_f\mathbb{1}_f(n) - \sum_{l\in\mathcal{I}(n)} x_l^{(f)}\right] \\ \quad \textit{if row } i \textit{ corresponds to node } n \textit{ and } n = \mathrm{Src}(f), \\ \sum_{l\in\mathcal{O}(n)\cup\mathcal{I}(n)}(x_l^{(f)})^2 + \frac{1}{p_n^{(f)}}\left[\sum_{l\in\mathcal{O}(n)} x_l^{(f)} - s_f\mathbb{1}_f(n) - \sum_{l\in\mathcal{I}(n)} x_l^{(f)}\right] \\ \quad \textit{otherwise.} \end{cases}$$

- *The off-diagonal entries of $(\mathbf{D}_f)_{ij}$, $i \neq j$, are given by*

$$(\mathbf{D}_f)_{ij} = \begin{cases} -\sum_{l\in\Gamma(n_1,n_2)}(x_l^{(f)})^2 & \textit{if row } i \textit{ and column } j \textit{ correspond to two connected nodes } n_1 \textit{ and } n_2, \\ 0 & \textit{otherwise,} \end{cases}$$

  where $\Gamma(n_1, n_2) \triangleq \{l \in \mathcal{L} : \mathrm{Tx}(l) = n_1 \textit{ and } \mathrm{Rx}(l) = n_2, \textit{ or } \mathrm{Tx}(l) = n_2 \textit{ and } \mathrm{Rx}(l) = n_1\}$.

In Lemma 7, we have omitted the time-slot index "$[t]$" for notational simplicity. For the same reason, in the rest of the paper, the associated time-slot index "$[t]$" will be dropped whenever such an omission does not cause confusion.

Next, we study the second term in (39), denoted as $\mathbf{W}$, which is symmetric and has the following partitioned structure:

$$\mathbf{W} \triangleq \sum_{l=1}^{L}\left(\frac{1}{\|\widehat{\mathbf{x}}_l\|^2}\begin{bmatrix} -\mathbf{a}_l^{(1)} & & \\ & \ddots & \\ & & -\mathbf{a}_l^{(F)} \end{bmatrix}\begin{bmatrix} (x_l^{(1)})^4 & \cdots & (x_l^{(1)}x_l^{(F)})^2 \\ \vdots & \ddots & \vdots \\ (x_l^{(F)}x_l^{(1)})^2 & \cdots & (x_l^{(F)})^4 \end{bmatrix}\begin{bmatrix} -\mathbf{a}_l^{(1)} & & \\ & \ddots & \\ & & -\mathbf{a}_l^{(F)} \end{bmatrix}^T\right)$$

$$= \begin{bmatrix} \widehat{\mathbf{D}}_1 & \mathbf{J}_{12} & \cdots & \mathbf{J}_{1F} \\ \mathbf{J}_{21} & \widehat{\mathbf{D}}_2 & \cdots & \mathbf{J}_{2F} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{J}_{F1} & \mathbf{J}_{F2} & \cdots & \widehat{\mathbf{D}}_F \end{bmatrix},$$

where

$$\widehat{\mathbf{D}}_f = \sum_{l=1}^{L} \frac{(x_l^{(f)})^4}{\|\widehat{\mathbf{x}}_l\|^2} \mathbf{a}_l^{(f)} (\mathbf{a}_l^{(f)})^T, \qquad f = 1, \ldots, F,$$

$$\mathbf{J}_{f_1 f_2} = \sum_{l=1}^{L} \frac{(x_l^{(f_1)} x_l^{(f_2)})^2}{\|\widehat{\mathbf{x}}_l\|^2} \mathbf{a}_l^{(f_1)} (\mathbf{a}_l^{(f_2)})^T, \qquad f_1, f_2 = 1, \ldots, F, f_1 \neq f_2.$$

Noting the similarity between $\widehat{\mathbf{D}}_f$ and $\mathbf{D}_f$, and by using [9, Lemma 2] and following a similar derivation to that for Lemma 7, we obtain the following result for characterizing $\widehat{\mathbf{D}}_f$, where we omit the proof to avoid repetition.

Lemma 8. *The matrix $\widehat{\mathbf{D}}_f$ has the following structure:*

- *The diagonal entries $(\widehat{\mathbf{D}}_f)_{ii}$ are given by*

$$(\widehat{\mathbf{D}}_f)_{ii} = \sum_{l \in \mathcal{O}(n) \cup \mathcal{I}(n)} \frac{(x_l^{(f)})^4}{\|\widehat{\mathbf{x}}_l\|^2}.$$

- *The off-diagonal entries of $(\widehat{\mathbf{D}}_f)_{ij}$, $i \neq j$, are given by*

$$(\widehat{\mathbf{D}}_f)_{ij} = \begin{cases} -\sum_{l \in \Gamma(n_1, n_2)} \frac{(x_l^{(f)})^4}{\|\widehat{\mathbf{x}}_l\|^2} & \text{if row } i \text{ and column } j \text{ correspond to two connected nodes } n_1, n_2, \\ 0 & \text{otherwise.} \end{cases}$$

Using [9, Lemma 3], we can also characterize the structure of $\mathbf{G}_{f_1 f_2}$ as stated in Lemma 9 below, where the proof follows that of Lemma 7 and is therefore omitted for the sake of brevity.

Lemma 9. *The matrix $\mathbf{G}_{f_1 f_2}$ has the following structure:*

$$(\mathbf{G}_{f_1 f_2})_{ij} = \begin{cases} \sum_{l \in \mathcal{O}(n) \cup \mathcal{I}(n)} \frac{(x_l^{(f_1)} x_l^{(f_2)})^2}{\|\widehat{\mathbf{x}}_l\|^2} & \text{if row } i \text{ and column } j \text{ correspond to the same node } n, \\ -\sum_{l \in \Gamma(n_1, n_2)} \frac{(x_l^{(f_1)} x_l^{(f_2)})^2}{\|\widehat{\mathbf{x}}_l\|^2} & \text{if row } i \text{ and column } j \text{ correspond to two connected nodes } n_1, n_2, \\ 0 & \text{otherwise.} \end{cases}$$

So far, we have characterized the structures of $\widehat{\mathbf{D}}_f$ and $\mathbf{G}_{f_1 f_2}$. Hence, the structure of $\mathbf{W}$ is also known. Finally, recall that $\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T = \mathbf{D} - \mathbf{W}$. Therefore, combining the previous derivations, we have the following result for the structural property of $\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T$.

Theorem 10. *The matrix $\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T$ can be written as the following partitioned matrix:*

$$\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T = \begin{bmatrix} \mathbf{D}_1 - \widehat{\mathbf{D}}_1 & -\mathbf{W}_{12} & \cdots & -\mathbf{W}_{1F} \\ -\mathbf{W}_{21} & \mathbf{D}_2 - \widehat{\mathbf{D}}_2 & \cdots & -\mathbf{W}_{2F} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{W}_{F1} & -\mathbf{W}_{F2} & \cdots & \mathbf{D}_F - \widehat{\mathbf{D}}_F \end{bmatrix},$$

where the structural properties of the matrices $\mathbf{D}_f$, $\widehat{\mathbf{D}}_f$, *and* $\mathbf{W}_{f_1 f_2}$ *are specified in Lemmas 7, 8, and 9, respectively.*

With Theorem 10, we are now in a position to design a distributed iterative scheme to compute the dual Newton directions. We propose two approaches: i) matrix-splitting based approach and ii) Sherman-Morrison-Woodbury (SMW) based matrix inversion. The most appealing feature of the matrix-splitting based approach is that it only requires one-hop local information exchange. However, the major limitation of the matrix-splitting approach is that the obtained solution is only an approximation and would only converge to the true dual Newton direction asymptotically, regardless of the size of the network. In contrast, the SMW-based approach can yield the *exact* value of $\mathbf{G}^{-1}$ (and hence $\Delta\mathbf{p}_{[t]}$) within a *finite* number of steps equal to twice the number of links in the network. However, the efficiency gain is achieved at the expense of more than one hop of information exchange.

5.3.1 Matrix-Splitting Based Approach

In the literature, the idea of matrix splitting is a generic framework for solving linear equation systems in an iterative fashion [21]. Consider a consistent linear equation system $\mathbf{Kz} = \mathbf{d}$, where $\mathbf{K} \in \mathbb{R}^{n\times n}$ is a nonsingular matrix and $\mathbf{z}, \mathbf{d} \in \mathbb{R}^n$. Now, suppose that $\mathbf{K}$ is split into a nonsingular matrix $\mathbf{K}_1$ and another matrix $\mathbf{K}_2$ according to $\mathbf{K} = \mathbf{K}_1 - \mathbf{K}_2$. Also, let $\mathbf{z}^0$ be an arbitrary starting vector. Then, a sequence of approximate solutions can be generated by using the following iterative scheme:

$$\mathbf{z}^{k+1} = (\mathbf{K}_1^{-1}\mathbf{K}_2)\mathbf{z}^k + \mathbf{K}_1^{-1}\mathbf{d}, \quad k \geq 0. \tag{40}$$

Generally, $\mathbf{K}_1$ should be an easily invertible matrix (e.g., diagonal, etc). It can be shown that this iterative method is convergent to the unique solution $\mathbf{z} = \mathbf{K}^{-1}\mathbf{b}$ if and only if the spectral radius of the matrix $\mathbf{K}_1^{-1}\mathbf{K}_2$ is less than one, i.e., $\rho(\mathbf{K}_1^{-1}\mathbf{K}_2) < 1$, where $\rho(\cdot)$ represents the spectral radius of a matrix. The following result provides a sufficient condition for $\rho(\mathbf{K}_1^{-1}\mathbf{K}_2) < 1$ (see [8, 21] for more details):

Lemma 11. *Suppose that* $\mathbf{K}$ *is a real symmetric matrix. If both matrices* $\mathbf{K}_1 + \mathbf{K}_2$ *and* $\mathbf{K}_1 - \mathbf{K}_2$ *are positive definite, then* $\rho(\mathbf{K}_1^{-1}\mathbf{K}_2) < 1$.

Lemma 11 suggests that the convergence property of a given matrix splitting scheme can be verified by checking for the positive definiteness of the identified matrix. The following lemma states a sufficient condition for checking positive definiteness based on diagonal dominance [22, Corollary 7.2.3]:

Lemma 12. *If a symmetric matrix* $\mathbf{Q}$ *is strictly diagonally dominant, i.e.,* $|(\mathbf{Q})_{ii}| > \sum_{j\neq i} |(\mathbf{Q})_{ij}|$, *and if* $(\mathbf{Q})_{ii} > 0$ *for all* i, *then* $\mathbf{Q}$ *is positive definite.*

We are now ready to use the matrix splitting scheme in (40) to compute $\mathbf{p}_{[t]}$. First, we let $\mathbf{\Lambda}$ be the diagonal matrix having the same main diagonal of $\mathbf{G}$, i.e.,

$$\mathbf{\Lambda}_{[t]} = \text{Diag}\left\{\mathbf{G}\right\} = \text{Diag}\left\{\text{diag}\left\{\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}\right\}\right\}. \tag{41}$$

We let $\mathbf{\Omega}$ denote the matrix containing the remaining entries after subtracting $\mathbf{\Lambda}_{[t]}$ from $\mathbf{G}$, i.e.,

$$\mathbf{\Omega}_{[t]} = \mathbf{G} - \mathbf{\Lambda}_{[t]}. \tag{42}$$

Further, we define a diagonal matrix $\overline{\mathbf{\Omega}}_{[t]}$ where the diagonal entries are defined by

$$(\overline{\mathbf{\Omega}}_{[t]})_{ii} = \sum_j |(\mathbf{\Omega}_{[t]})_{ij}|. \tag{43}$$

Then, we can split $\mathbf{G}$ as $(\mathbf{\Lambda}_{[t]} + \alpha\overline{\mathbf{\Omega}}_{[t]}) - (\alpha\overline{\mathbf{\Omega}}_{[t]} - \mathbf{\Omega}_{[t]})$, where $\alpha > \frac{1}{2}$ is a parameter that serves the purpose of tuning convergence performance. Based on this splitting scheme, we have the following result:

Proposition 13. *Consider the matrix splitting scheme* $\mathbf{G}$ *as* $\mathbf{G} = (\mathbf{\Lambda}_{[t]}+\alpha\overline{\mathbf{\Omega}}_{[t]})-(\alpha\overline{\mathbf{\Omega}}_{[t]}-\mathbf{\Omega}_{[t]})$*, where* $\mathbf{\Lambda}_{[t]}$*,* $\mathbf{\Omega}_{[t]}$*, and* $\overline{\mathbf{\Omega}}_{[t]}$ *are defined in (41), (42), and (43), respectively. Then, the following sequence* $\{\mathbf{p}_{[t]}^k\}$ *generated by*

$$\mathbf{p}_{[t]}^{k+1} = (\mathbf{\Lambda}_{[t]} + \alpha\overline{\mathbf{\Omega}}_{[t]})^{-1}(\alpha\overline{\mathbf{\Omega}}_{[t]} - \mathbf{\Omega}_{[t]})\mathbf{p}_{[t]}^k + (\mathbf{\Lambda}_{[t]} + \alpha\overline{\mathbf{\Omega}}_{[t]})^{-1}(-\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{g}_{[t]}) \tag{44}$$

converges to the solution of (31) as $k \to \infty$*.*

By Lemmas 11 and 12, the key to proving Proposition 13 is to verify that both the sum and difference of the two components in the splitting scheme are strictly diagonally dominant. We relegate the proof details to Appendix E.

Remark 3. The matrix splitting scheme in Proposition 13 generalizes the matrix splitting scheme in [8]. In both matrix splitting schemes, the goal is to construct a diagonal nonsingular matrix ($\mathbf{\Lambda}_{[t]} + \alpha\overline{\mathbf{\Omega}}_{[t]}$ in our paper) for which the inverse can be separated and easily computed by each node (as in our case) or each link (as in [8]). However, our matrix splitting scheme differs from that in [8] in the following aspects. First, since $(\mathbf{\Lambda}_{[t]} + \alpha\overline{\mathbf{\Omega}}_{[t]})$ is not element-wise non-negative (c.f. [8]), the definition of the matrix $\overline{\mathbf{\Omega}}_{[t]}$ in this work is different from that in [8], which also leads to a different proof. Second, we parameterize the splitting scheme (using α) to allow for tuning the convergence speed in (44), where the scheme in [8] is a special case of our scheme when $\alpha = 1$.

Several remarks on the parameter α are in order. It can be seen from (40) that the solution error shrinks in magnitude approximately by a factor of $\rho(\mathbf{K}_1^{-1}\mathbf{K}_2)$. Thus, the smaller $\rho(\mathbf{K}_1^{-1}\mathbf{K}_2)$, the faster the convergence rate of the iterative scheme. The following result [9,10] for the selection of the parameter α states the above observation:

Proposition 14. *Consider two alternative matrix splitting schemes with parameters α_1 and α_2, respectively, satisfying $\frac{1}{2} < \alpha_1 \leq \alpha_2$. Let ρ_{α_1} and ρ_{α_2} be their spectral radii, respectively. Then, $\rho_{\alpha_1} \leq \rho_{\alpha_2}$.*

Proposition 14 indicates that we should choose a smaller α in order to make the matrix splitting scheme converge faster, i.e., we can let $\alpha = \frac{1}{2} + \epsilon$, where $\epsilon > 0$ is small. The proof of Proposition 14 makes use of the comparison theorem in [21].

The next theorem shows that the matrix splitting scheme in Theorem 13 can be implemented in a distributed fashion. We first define two types of link sets as follows:

$$\Phi(n) \triangleq \mathcal{I}(n) \cup \mathcal{O}(n), \qquad \Psi(n,f) \triangleq \{l \in \mathcal{I}(n) \cup \mathcal{O}(n) : \mathrm{Tx}(l) = \mathrm{Dst}(f) \text{ or } \mathrm{Rx}(l) = \mathrm{Dst}(f)\}.$$

We let $\mathbb{1}_S(a)$ denote the set indicator function, which takes value 1 if $a \in S$ and 0 otherwise. Then, we have the following result:

Theorem 15. *Given a primal solution $\mathbf{y}$, the update of the dual variable $p_n^{(f)}$ can be iteratively computed using local information at each node. More specifically, $p_n^{(f)}$ can be computed as:*

$$p_n^{(f)}[k+1] = \frac{1}{U_{n,[t]}^f[k]}(V_{n,1}^{(f)}[k] + V_{n,2}^{(f)}[k] - W_n^f[k]), \tag{45}$$

where $U_n^{(f)}[k]$, $V_n^{(f)}[k]$, and $W_n^{(f)}[k]$ are, respectively, defined as

$$U_n^{(f)}[k] \triangleq \begin{cases} \sum_{l\in\Phi(n)}[1+\alpha(1-\mathbb{1}_{\Psi(n,f)}(l))](x_l^{(f)})^2\Big(1-\frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big)+ \\ \quad \frac{1}{p_n^{(f)}[t]}\Big[\sum_{l\in\mathcal{O}(n)} x_l^{(f)} - s_f\mathbb{1}_f(n) - \sum_{l\in\mathcal{I}(n)} x_l^{(f)}\Big]+ \\ \quad \sum_{f'=1,\neq f}^{F}\Big(\sum_{l\in\Psi(n,f')}(1+\mathbb{1}_{\Psi(n,f')}(l))\frac{\alpha(x_l^{(f)}x_l^{(f')})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big) & \text{if } n \neq \mathrm{Src}(f), \\ \sum_{l\in\Phi(n)}[1+\alpha(1-\mathbb{1}_{\Psi(n,f)}(l))](x_l^{(f)})^2\Big(1-\frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big)+ \\ \quad \frac{1}{p_n^{(f)}[t]}\Big[\sum_{l\in\mathcal{O}(n)} x_l^{(f)} - s_f\mathbb{1}_f(n) - \sum_{l\in\mathcal{I}(n)} x_l^{(f)}\Big]+ \\ \quad \sum_{f'=1,\neq f}^{F}\Big(\sum_{l\in\Psi(n,f')}(1+\mathbb{1}_{\Psi(n,f')}(l))\frac{\alpha(x_l^{(f)}x_l^{(f')})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big) + \frac{1}{-\mu U_f''(s_f)+\frac{1}{(s_f)^2}} & \text{if } n = \mathrm{Src}(f), \end{cases} \tag{46}$$

$$\begin{aligned} V_{n,1}^{(f)}[k] \triangleq & \sum_{l\in\mathcal{I}(n)\setminus\Psi(n,f)} (x_l^{(f)})^2\Big(1-\frac{((x_l^{(f)})^2)}{\|\widehat{\mathbf{x}}_l\|^2}\Big)(p_{\mathrm{Tx}(l)}^{(f)} + \alpha p_{\mathrm{Rx}(l)}^{(f)})+ \\ & \sum_{l\in\mathcal{O}(n)\setminus\Psi(n,f)} (x_l^{(f)})^2\Big(1-\frac{((x_l^{(f)})^2)}{\|\widehat{\mathbf{x}}_l\|^2}\Big)(p_{\mathrm{Rx}(l)}^{(f)} + \alpha p_{\mathrm{Tx}(l)}^{(f)})- \\ & \sum_{f'=1,\neq f}^{F}\Big(\sum_{l\in\Phi(n)}(1+\mathbb{1}_{\Psi(n,f')}(l))\frac{\alpha(x_l^{(f)}x_l^{(f')})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big)p_n^f, \end{aligned} \tag{47}$$

$$V_{n,2}^{(f)}[k] \triangleq \sum_{f'=1,\neq f}^{F} \Big(\Big(\sum_{l\in\mathcal{I}(n)} \frac{(x_l^{(f)} x_l^{(f')})^2}{\|\widehat{\mathbf{x}}_l\|^2} - \sum_{l\in\mathcal{O}(n)} \frac{(x_l^{(f)} x_l^{(f')})^2}{\|\widehat{\mathbf{x}}_l\|^2} \Big) (p_{\mathrm{Rx}(l)}^{(f')} - p_{\mathrm{Tx}(l)}^{(f')}) \Big), \tag{48}$$

$$W_n^{(f)}[k] \triangleq \begin{cases} \Big(1 - \frac{x_l^{(f)}}{\delta_l}\Big) \Big[\sum_{l\in\mathcal{O}(n)} \Big(1 - \sum_{f'=1}^{F} \frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2} x_l^{(f')}\Big) - \\ \quad \sum_{l\in\mathcal{I}(n)} \Big(1 - \sum_{f'=1}^{F} \frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2} x_l^{(f')}\Big) \Big] + \frac{1}{p_n^{(f)}} & \text{if } n \neq \mathrm{Src}(f), \\ \sum_{l\in\mathcal{O}(n)} \Big(1 - \sum_{f'=1}^{F} \frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2} x_l^{(f')}\Big) - \\ \quad \sum_{l\in\mathcal{I}(n)} \Big(1 - \sum_{f'=1}^{F} \frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2} x_l^{(f')}\Big) \Big] + \frac{1}{p_n^{(f)}} + \frac{s_f(1+\mu s_f U_f'(s_f))}{\mu s_f^2 U_f''(s_f) - 1} & \text{if } n = \mathrm{Src}(f). \end{cases} \tag{49}$$

Theorem 15 can be proved by computing the element-wise expansion of (44). We relegate the proof details to Appendix F.

Remark 4. Several interesting remarks pertaining to Theorem 15 are in order. First, from (46), (47), (48), and (49), we can observe that all the information needed to update $w_n^{(f)}$ are either locally available at node n or at links that incident at node n. This not only shows that the matrix splitting scheme can be implemented in a distributed fashion, but it also means that the information exchange scale is at most *one-hop*. Second, although the dual update scheme within a second-order method is more complex at each node, the more rapid convergence rate of a second-order method, with its accompanying less information exchange, outweigh this local computational cost increase.

5.3.2 A More Efficient Approach Based on Sherman-Morrison-Woodbury Matrix Inversion Lemma

Although the matrix-splitting scheme only requires one-hop local information exchange, the main drawback is that the obtained solution is an approximation and only converges asymptotically to $\widetilde{\mathbf{p}}_{[t+1]}$, and thereby to the true dual Newton direction $\mathbf{p}_{[t+1]}$. Here, we propose a more efficient scheme to compute $\widetilde{\mathbf{p}}_{[t+1]}$, but at the expense of a greater information exchange scale. Our basic idea to compute $\widetilde{\mathbf{p}}_{[t+1]}$ is that, instead of splitting $\mathbf{G}$ and computing its inverse implicitly, we directly update $\mathbf{G}^{-1}$ by using the Sherman-Morrison-Woodbury (SMW) matrix inversion lemma. For notational simplicity, in what follows, we omit the time-slot index "$[t]$".

More specifically, consider

$$\mathbf{G}^{-1} = \Big(\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}\Big)^{-1} = \left[\Big(\mathbf{B}\mathbf{S}_{[t]}^{-1}\mathbf{B} - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}\Big) + \sum_{l=1}^{L} \mathbf{A}_l \mathbf{X}_l^{-1} \mathbf{A}_l^T\right]^{-1}. \tag{50}$$

From Lemma 7, we have that $\mathbf{B}\mathbf{S}_{[t]}^{-1}\mathbf{B}-\mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}$ is a diagonal matrix, which can be written as follows:

$$\mathbf{B}\mathbf{S}_{[t]}^{-1}\mathbf{B}-\mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}= \\ \mathrm{Diag}\left\{\frac{1}{p_n^{(f)}}\left(\sum_{l\in\mathcal{O}(n)}x_l^{(f)}-\sum_{l\in\mathcal{I}(n)}x_l^{(f)}-s_f\mathbb{1}_f(n)\right)+\left(-\mu U_f''(s_f)+\frac{1}{s_f^2}\right)\mathbb{1}_f(n),\quad f=1,\ldots,F\right\}. \tag{51}$$

Due to this diagonal structure, $(\mathbf{B}\mathbf{S}_{[t]}^{-1}\mathbf{B}-\mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]})^{-1}$ can be readily computed at each node in a distributed fashion. We let $\overline{\mathbf{D}}=\mathrm{Diag}\left\{\overline{\mathbf{D}}_1,\ldots,\overline{\mathbf{D}}_F\right\}$ denote this diagonal matrix.

On the other hand, from (39), we can see that $\sum_{l=1}^{L}\mathbf{A}_l\mathbf{X}_l^{-1}\mathbf{A}_l^T$ can be written as:

$$\sum_{l=1}^{L}\mathbf{A}_l\mathbf{X}_l^{-1}\mathbf{A}_l^T=\sum_{l=1}^{L}\left\{\begin{bmatrix}(x_l^{(1)})^2\mathbf{a}_l^{(1)}(\mathbf{a}_l^{(1)})^T & & \\ & \ddots & \\ & & (x_l^{(F)})^2\mathbf{a}_l^{(F)}(\mathbf{a}_l^{(F)})^T\end{bmatrix}\right. \tag{52}$$

$$\left.-\frac{1}{\|\widehat{\mathbf{x}}_l\|^2}\begin{bmatrix}(x_l^{(1)})^2\mathbf{a}_l^{(1)}\\ \vdots\\ (x_l^{(F)})^2\mathbf{a}_l^{(F)}\end{bmatrix}\begin{bmatrix}(x_l^{(1)})^2(\mathbf{a}_l^{(1)})^T & \cdots & (x_l^{(F)})^2(\mathbf{a}_l^{(F)})^T\end{bmatrix}.\right\} \tag{53}$$

Now, note that the term in (52) is block-diagonal and can be merged with $\mathbf{B}\mathbf{S}_{[t]}^{-1}\mathbf{B}-\mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}$, and the resultant matrix is exactly the $\mathbf{D}$ matrix in Lemma 7, which is also block-diagonal. Moreover, each diagonal block has the following form:

$$\overline{\mathbf{D}}_f+\sum_{l=1}^{L}(x_l^{(f)})^2\mathbf{a}_l^{(f)}(\mathbf{a}_l^{(f)})^T,$$

which can be thought of as applying L rank-1 updates on $\overline{\mathbf{D}}_f$. Hence, we can start from $(\mathbf{B}\mathbf{S}_{[t]}^{-1}\mathbf{B}-\mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]})^{-1}$ and apply the SMW matrix inversion lemma L times to compute $\mathbf{D}_f^{-1}$. More specifically, let $\mathbf{D}_{f,[l-1]}^{-1}$ denote the intermediate result we have before applying the l-th SMW-correction. Also, let $\mathbf{D}_{f,[0]}^{-1}=\overline{\mathbf{D}}_f^{-1}$. Then, we have the following computational scheme:

$$\mathbf{D}_{f,[l]}^{-1}=\mathbf{D}_{f,[l-1]}^{-1}-\frac{\mathbf{D}_{f,[l-1]}^{-1}(x_l^{(f)})^2\mathbf{a}_l^{(f)}(\mathbf{a}_l^{(f)})^T\mathbf{D}_{f,[l-1]}^{-1}}{1+(x_l^{(f)})^2(\mathbf{a}_l^{(f)})^T\mathbf{D}_{f,[l-1]}^{-1}\mathbf{a}_l^{(f)}},\quad l=1,\ldots,L. \tag{54}$$

After L times of SMW-corrections, we would have achieved $\mathbf{D}^{-1}=\mathrm{Diag}\left\{\mathbf{D}_1^{-1},\ldots,\mathbf{D}_F^{-1}\right\}$.

Further, we note that

$$\mathbf{G}=\left(\begin{bmatrix}\mathbf{D}_1 & & \\ & \ddots & \\ & & \mathbf{D}_F\end{bmatrix}-\frac{1}{\|\widehat{\mathbf{x}}_l\|^2}\begin{bmatrix}(x_l^{(1)})^2\mathbf{a}_l^{(1)}\\ \vdots\\ (x_l^{(F)})^2\mathbf{a}_l^{(F)}\end{bmatrix}\begin{bmatrix}(x_l^{(1)})^2(\mathbf{a}_l^{(1)})^T & \cdots & (x_l^{(F)})^2(\mathbf{a}_l^{(F)})^T\end{bmatrix}.\right), \tag{55}$$

Algorithm 2 SMW-based approach for computing $\mathbf{G}^{-1}$

Initialization:

1. For each node, compute the corresponding components in $\overline{\mathbf{D}} = \text{Diag}\{\mathbf{D}_1, \ldots, \mathbf{D}_F\}$ using (51), and send the result to the starting link.

Main Iteration:

2. For all $f = 1, \ldots, F$, let $\mathbf{D}_{f,[0]}^{-1} = \overline{\mathbf{D}}_f^{-1}$. For links $l = 1, \ldots, L$, update $\mathbf{D}_{f,[l]}$ using (54). Let $\mathbf{D}^{-1} = \text{Diag}\left\{\mathbf{D}_{1,[L]}^{-1}, \ldots, \mathbf{D}_{F,[L]}^{-1}\right\}$.
3. Let $\mathbf{K}_{[0]}^{-1} = \mathbf{D}^{-1}$. For links $l = 1, \ldots, L$, update $\mathbf{K}_{[l]}^{-1}$ using (56). Let $\mathbf{G}^{-1} = \mathbf{K}_{[L]}^{-1}$ and stop.

which again can be thought of as applying rank-1 updates L times on $\mathbf{D}$. Since $\mathbf{D}^{-1}$ has been computed, we can again apply the SWM matrix inversion lemma L times to compute $\mathbf{G}^{-1}$. Compared to the computation of $\mathbf{D}^{-1}$, however, applying SMW-corrections L times to compute $\mathbf{G}^{-1}$ is slightly more complex. The reason is that each rank-1 update in (55) is a dense matrix. Therefore, each SMW-correction cannot be done in a block-wise fashion and needs to be performed over the entire matrix. Fortunately, each SWM-correction still only involves information locally available at link l, and hence can be done locally. Toward this end, let $\mathbf{K}_{[l-1]}^{-1}$ denote the intermediate result we have before applying the l-th SMW-correction. Also, let $\mathbf{K}_{[0]}^{-1} = \mathbf{D}^{-1}$. For notational convenience, let

$$\mathbf{u}_l = \frac{1}{\|\widehat{\mathbf{x}}_l\|^2} \begin{bmatrix} (x_l^{(1)})^2 \mathbf{a}_l^{(1)} \\ \vdots \\ (x_l^{(F)})^2 \mathbf{a}_l^{(F)} \end{bmatrix}.$$

Then, we have the following computational scheme:

$$\mathbf{K}_{[l]}^{-1} = \mathbf{K}_{[l-1]}^{-1} + \frac{\mathbf{K}_{[l-1]}^{-1} \mathbf{u}_l (\mathbf{u}_l)^T \mathbf{K}_{[l-1]}^{-1}}{\|\widehat{\mathbf{x}}_l\|^2 - (\mathbf{u}_l)^T \mathbf{K}_{[l-1]}^{-1} \mathbf{u}_l}, \quad l = 1, \ldots, L. \tag{56}$$

Finally, after L SMW-corrections, we achieve $\mathbf{G}^{-1}$, which can in turn be used to compute $\widetilde{\mathbf{p}}_{[t+1]}$ and $\Delta\mathbf{p}_{[t]}$. To conclude the discussion, we summarize the SMW-based approach in Algorithm 2.

There is one important remark pertaining to implementing the SMW-based approach in Algorithm 2 in a distributed fashion. Noting that each SMW-correction only involves information locally available at each link, the scheme can proceed following *any* pre-determined link ordering. Here, unlike the matrix-splitting based approach that only converges asymptotically, we require *exactly* $2L$ SWM-corrections to compute *precise* value of $\mathbf{G}^{-1}$. Thus, the SMW-based approach is much more efficient. However, since the SMW-based approach involves all L links in the network, the scale of information exchange is clearly larger than the 1-hop scale required by the matrix-splitting approach, and is determined by the network diameter. Fortunately, many communication networks in practice

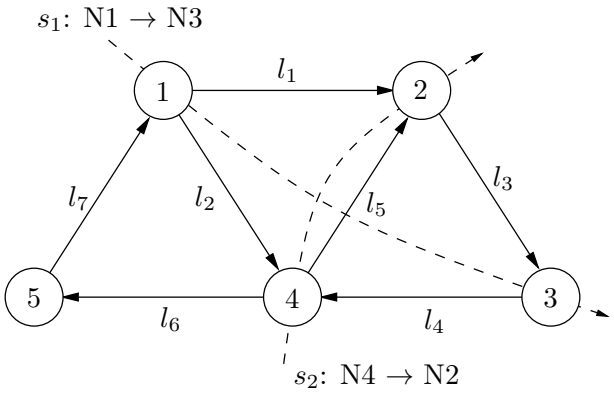

Figure 4: A five-node two-session network.

(e.g., the Internet, data centers, etc.) are constructed in such a way that the network diameter is usually small.

6 Numerical Results

In this section, we use a 5-node 2-session network example as shown in Figure 4 to illustrate the performance of our proposed second-order joint congestion control and routing algorithm. There are two sessions in the network: N1 to N3 and N4 to N2. Each link in the network has unit capacity. We use $\log(s_f)$ as the utility function, which represents the proportional fairness [23]. In the simulation, we set $\mu = 1000$, meaning that a point $(\mathbf{y}, \mathbf{p})$ that satisfies the perturbed KKT system implies $-\text{Diag}\{\mathbf{My}\}\,\mathbf{p}/\mu = \frac{1}{\mu} = 0.001$, i.e., the accuracy of the CS condition is on the order of 10^{-3}. The (feasible) primal and dual initial points are summarized in Table 2.

Table 2: The primal and dual initial points.

s_1	$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$	$x_4^{(1)}$	$x_5^{(1)}$	$x_6^{(1)}$	$x_7^{(1)}$
0.45	0.3	0.35	0.7	0.1	0.35	0.12	0.15
s_2	$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$	$x_4^{(1)}$	$x_5^{(1)}$	$x_6^{(1)}$	$x_7^{(1)}$
0.25	0.23	0.22	0.1	0.15	0.26	0.38	0.4
$p_1^{(1)}$	$p_2^{(1)}$	$p_4^{(1)}$	$p_5^{(1)}$	$p_1^{(2)}$	$p_3^{(2)}$	$p_4^{(2)}$	$p_5^{(2)}$
22.61	18.62	20.43	21.72	6.17	1.64	7.52	7.06

The convergence behavior is illustrated in Figure 5. It can be seen from Figure 5 that the source rates (not just the average source rates) rapidly converge to the following pair ($s_1 = 0.9634$, $s_2 = 1.0247$) in approximately 15 iterations. This shows the efficiency of our proposed second-order al-

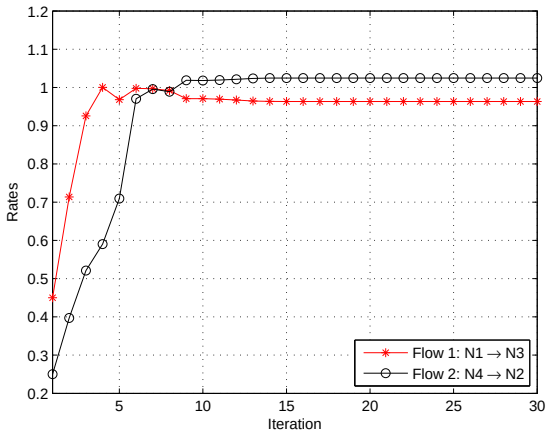

Figure 5: Convergence behavior of the proposed second-order algorithm for the network in Figure 4.

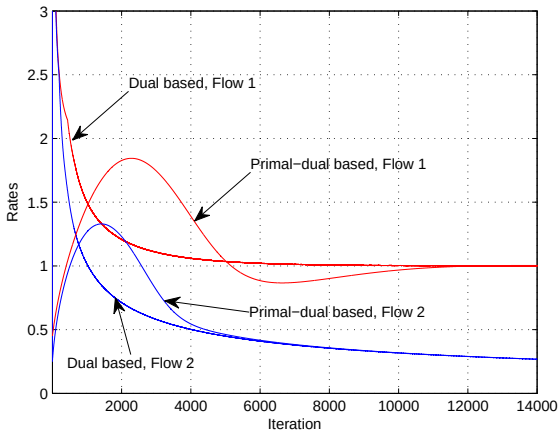

Figure 6: Convergence behavior of the first-order schemes for the network in Figure 4.

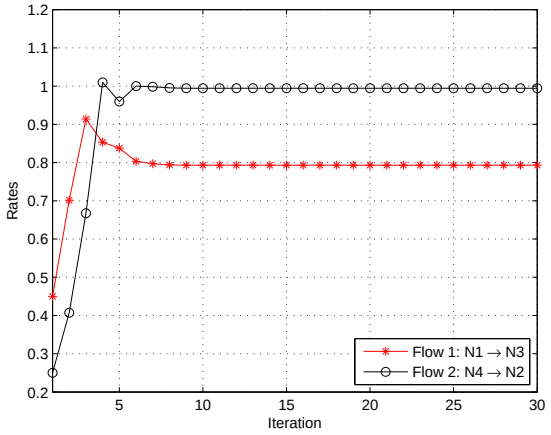

Figure 7: Convergence behavior of the proposed second-order algorithm for the network in Figure 4.

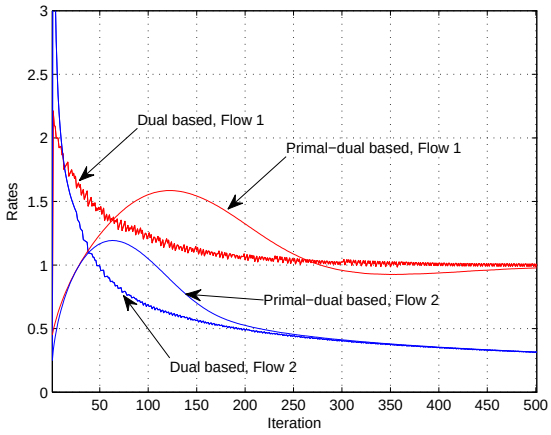

Figure 8: Convergence behavior of the first-order schemes for the network in Figure 4.

gorithm. To compare the convergence performance with the first-order back-pressure algorithm, we also used the same network example in Figure 4 to experiment with both primal-dual [3] and dual based first-order schemes [1, 2]. For a fair comparison, both first-order back-pressure based schemes were started from the same primal and dual initial points. Targeting approximately the same level of accuracy, we set the step-size scaling factor, denoted as V, as $V = 1000$ (see the discussions in Section 4.4). The convergence performances of both primal-dual and dual based first-order schemes

are illustrated in Figure 6. We can see from Figure 6 that in order to achieve high accuracy solutions, the first-order schemes converge very slowly: in both primal-dual and dual based schemes, s_2 shows no signs of convergence even after 14000 iterations. This shows that our second-order scheme converges at least *two orders of magnitude faster* than the first-order schemes. We can also observe that the iterates of the primal-dual based scheme in the first-order domain evolve less abruptly compared to the dual-based scheme, but also converge more slowly. To see the impacts of μ and V on the second-order and first-order methods, we let $\mu = V = 50$ and run another experiment on the network in Figure 4. As shown in Figures 7, we can see that when μ is smaller, our second-order scheme converges even faster (less than 10 iterations) but at the cost of a larger optimality gap. On the other hand, as shown in Figure 8, with $V = 50$, the convergence of the first-order methods can be made faster but also exhibits much larger fluctuations. Again, we can observe that the iterates in the primal-dual based scheme evolves less abruptly with less fluctuations, but converges slower. However, regardless of which first-order scheme and what choice of V, the obtained solutions under both first-order schemes are far from being optimal since the obtained objective value is much smaller than that of the second-order scheme.

Next, we verify whether the obtained solution under our second-order scheme is indeed optimal (or the accuracy of the obtained solution). First, we illustrate the routing solutions for Sessions 1 and 2 in Figure 9 and Figure 10, respectively. The obtained dual solutions are summarized in Table 3. It can be readily verified that the obtained solutions are strictly primal and dual feasible. Further, we list the components of $\mathbf{g} + \mathbf{M}^T\mathbf{p}$ and $-\text{Diag}\{\mathbf{My}\}\,\mathbf{p}$ (i.e., the μ-ST and μ-CS conditions) in Table 4. We can see that (up to MATLAB's numerical accuracy) both μ-ST and μ-CS conditions are satisfied, which confirms the optimality of the obtained solution.

Table 3: The optimal dual initial solution.

$p_1^{(1,*)}$	$p_2^{(1,*)}$	$p_4^{(1,*)}$	$p_5^{(1,*)}$	$p_1^{(2,*)}$	$p_3^{(2,*)}$	$p_4^{(2,*)}$	$p_5^{(2,*)}$
1039.1	63.23	1037.9	1038.5	974.05	641.72	9976.87	975.48

Table 4: The μ-ST and μ-CS evaluation of the obtained solution.

$\mathbf{g} + \mathbf{M}^T\mathbf{p}$							
0	0	-4.9E-11	-4.9E-11	1E-13	3E-14	-2E-14	0
0	0	-5.6E-11	-5.6E-11	6E-13	0	-1.6E-13	-3E-14
$-\text{Diag}\{\mathbf{My}\}\,\mathbf{p}$							
1	1	1	1	1	1	1	1

Lastly, the simulation results of average total queue-length vs. mean arrive rates is illustrated

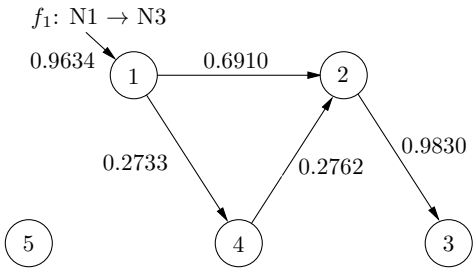

Figure 9: The routing solutions for session N1 → N3.

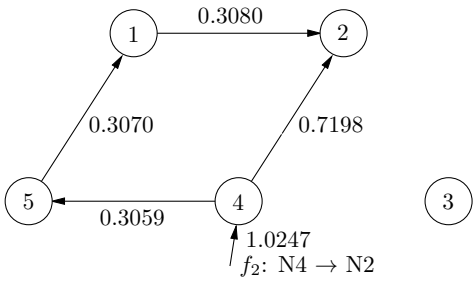

Figure 10: The routing solutions for session N4 → N2.

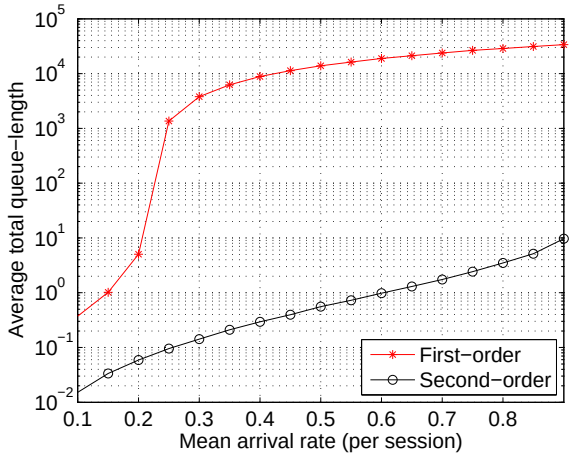

Figure 11: Average total queue-length vs. mean arrival rate ($\mu = V = 1000$).

in Fig. 11, where we can see that the delay performance of our second-order scheme significantly outperforms that of the first-order methods (more than *three orders of magnitude lower*). This large delay performance gap is a direct consequence of the slow convergence of the first-order methods.

7 Conclusion

In this paper, we have developed a new second-order algorithmic framework for joint congestion control and routing optimization. Unlike most joint congestion control and routing methods in the literature, our proposed algorithmic framework fundamentally deviates from the classical back-pressure idea to offer not only rate optimality and queuing stability, but also fast convergence and high accuracy. Our main contributions in this paper are three-fold: i) We have proposed a second-order joint congestion control and routing framework based on a *primal-dual* interior-point approach that is

well-suited for implementation in practical network systems; ii) we have rigorously established the rate optimality and queuing stability of the proposed second-order joint congestion control and routing framework; and iii) we have proposed several novel approaches for the distributed implementation of our second-order joint congestion control and routing optimization algorithm. These results serve as an exciting first step toward an analytical foundation for a second-order joint congestion control and optimization theory that offers fast convergence performance. Collectively, these results serve as a first building block of a new second-order theoretical framework for cross-layer optimization for network systems. Second-order cross-layer optimization for network system sis an important and yet under-explored area. Future research topics may include extending and generalizing our proposed second-order algorithmic framework to applications in other network systems, such as wireless networks with stochastic channel models, cloud computing resource allocations, and energy production scheduling in the smart electric power grid.

A Proof of Lemma 4

We prove Lemma 4 result by induction. For $t = 0$, the result is trivially true by assumption. Suppose that at time slot we have t $\|\mathbf{p}_{[t]}\| < B < \infty$, we will show that $\|\mathbf{p}_{[t+1]}\|$ is also bounded. We let $\widetilde{\mathbf{p}}_{[t+1]} \triangleq \mathbf{p}_{[t]} + \Delta\mathbf{p}_{[t]}$, i.e., we let $\pi[t] = 1$. After some algebraic derivations, we have:

$$\widetilde{\mathbf{p}}_{[t+1]} = \left(\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}\right)^{-1}\left[\mathbf{M}\mathbf{H}_{[t]}^{-1}(-\mathbf{g}_{[t]}) + \mathbf{P}_{[t]}^{-1}\mathbf{1}\right].$$

Now, we claim that $\|\widetilde{\mathbf{p}}_{[t+1]}\|$ is bounded. This is true because

$$\begin{aligned}
\|\widetilde{\mathbf{p}}_{[t+1]}\| &\leq \left\|\left(\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}\right)^{-1}\left[\mathbf{M}\mathbf{H}_{[t]}^{-1}(-\mathbf{g}_{[t]}) + \mathbf{P}_{[t]}^{-1}\mathbf{1}\right]\right\| \\
&\overset{(a)}{\leq} \left\|(\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T)^{-1}\left[\mathbf{M}\mathbf{H}_{[t]}^{-1}(-\mathbf{g}_{[t]}) + \mathbf{P}_{[t]}^{-1}\mathbf{1}\right]\right\| \\
&\overset{(b)}{\leq} \lambda_{\min}^{-1}\left\{\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T\right\}\left(\left\|\mathbf{M}\mathbf{H}_{[t]}^{-1}(-\mathbf{g}_{[t]})\right\| + \left\|\mathbf{P}_{[t]}^{-1}\mathbf{1}\right\|\right) \\
&\overset{(c)}{\leq} \lambda_{\min}^{-1}\left\{\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T\right\}\left(\lambda_{\min}^{-1}\left\{\mathbf{H}[t]\right\}\left\|\mathbf{M}(-\mathbf{g}_{[t]})\right\| + \left\|\mathbf{P}_{[t]}^{-1}\mathbf{1}\right\|\right),
\end{aligned} \tag{57}$$

where (a) holds because of the strict feasibility of $\mathbf{y}_{[t]}$ and $\mathbf{p}_{[t]}$ (and hence $-\mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}$ is a positive definite diagonal matrix, which can only increase the eigenvalues of $\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T$); (b) follows from triangular inequality and taking the smallest eigenvalue of $\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T$ and factoring it outside the norm; and (c) follows from factoring $\lambda_{\min}^{-1}\{\mathbf{H}[t]\}$ outside the norm. By assumption, since $\mathbf{g}_{[t]}$ is Lipschitz continuous, implying the spectral radius $\rho(\mathbf{H}_{[t]})$ is bounded. Also, since $\mathbf{M}$ is constructed by the node-arc incidence matrix of a connected graph, $\rho(\mathbf{H}_{[t]})$ is also finite. As a result, $\lambda_{\min}^{-1}\left\{\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T\right\}$ must be finite. Also, since $\mathbf{s}_f[t]$ and $\mathbf{x}_l^{(f)}[t]$ are strictly bounded away from zero (due to the step-size

selection rule), we have that $\|\mathbf{g}_{[t]}\|$ is bounded. Likewise, since $\mathbf{p}_{[t]}$ is also strictly bounded away from $\mathbf{0}$, we have that $\left\|\mathbf{P}_{[t]}^{-1}\mathbf{1}\right\|$ is bounded from above. Therefore, we can conclude that the RHS of (c) in (57) is bounded, i.e., $\|\widetilde{\mathbf{p}}_{[t+1]}\|$ is bounded.

Finally, note that

$$\begin{aligned}\|\mathbf{p}_{[t+1]}\| &= \left\|(1-\pi[t])\mathbf{p}_{[t]} + \pi[t]\widehat{\mathbf{p}}_{[t+1]}\right\| \\ &\overset{(a)}{\leq} (1-\pi[t])\|\mathbf{p}_{[t]}\| + \pi[t]\|\widehat{\mathbf{p}}_{[t+1]}\|,\end{aligned}$$

where (a) follows from triangular inequality. Hence, we can conclude that $\|\mathbf{p}_{[t+1]}\|$ is also bounded. This completes the proof.

B Proof of Theorem 1

The main idea and the key steps for proving Theorem 1 are based on Lyapunov drift analysis. First, we analyze the one-slot drift of the following quadratic Lyapunov function:

$$V\left(\mathbf{y}_{[t]}, \mathbf{p}_{[t]}\right) \triangleq \frac{1}{2\pi}\left\|\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right\|^2 + \frac{1}{2\mu^3\pi}\left\|\mathbf{p}_{[t]} - \mathbf{p}^*\right\|^2,$$

which can be interpreted as measuring the (unscaled) distance between a primal-dual iterate $(\mathbf{y}_{[t]}, \mathbf{p}_{[t]})$ and a perturbed KKT point $(\bar{\mathbf{y}}^*, \mathbf{p}^*)$ satisfying (14)–(17). The one-slot drift analysis reveals the following key relationship:

$$\begin{aligned}\Delta V\left(\mathbf{y}_{[t]}, \mathbf{p}_{[t]}\right) =& V\left(\mathbf{y}_{[t+1]}, \mathbf{p}_{[t+1]}\right) - V\left(\mathbf{y}_{[t]}, \mathbf{p}_{[t]}\right) \\ \leq& -R\|\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\| + \frac{1}{\mu}B,\end{aligned}$$

where R and B are both some positive finite quantities independent of μ. Based on this relationship, the result stated in Theorem 1 follows from telescoping T one-slot drifts and then letting T go to infinity.

We begin with evaluating the one-slot Lyapunov drift $\Delta V\left(\mathbf{y}_{[t]}, \mathbf{p}_{[t]}\right)$:

$$\begin{aligned}\Delta V\left(\mathbf{y}_{[t]}, \mathbf{p}_{[t]}\right) =& \frac{1}{2\pi}\left\|\mathbf{y}_{[t+1]} - \bar{\mathbf{y}}^*\right\|^2 + \frac{1}{2\mu^3\pi}\left\|\mathbf{p}_{[t+1]} - \mathbf{p}^*\right\|^2 \\ &- \frac{1}{2\pi}\left\|\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right\|^2 - \frac{1}{2\mu^3\pi}\left\|\mathbf{p}_{[t]} - \mathbf{p}^*\right\|^2 \\ =& \frac{1}{2\pi}\left(\mathbf{y}_{[t+1]} + \mathbf{y}_{[t]} - 2\bar{\mathbf{y}}^*\right)^T\left(\mathbf{y}_{[t+1]} - \mathbf{y}_{[t]}\right) \qquad (58) \\ &+ \frac{1}{2\mu^3\pi}\left(\mathbf{p}_{[t+1]} + \mathbf{p}_{[t]} - 2\mathbf{p}^*\right)^T\left(\mathbf{p}_{[t+1]} - \mathbf{p}_{[t]}\right). \qquad (59)\end{aligned}$$

In what follows, we will bound the two expressions in (58) and (59) in Section B.1 and B.2, respectively.

B.1 One-slot Lyapunov drift of (58)

Note that (58) can be further expanded as:

$$\begin{aligned}(58) = &\left[\frac{1}{\pi}\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right) - \frac{1}{2}\left(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M}\right)^{-1}\left(\mathbf{g}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}\right)\right]^T \times \\ &\left[-\pi\left(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M}\right)^{-1}\left(\mathbf{g}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}\right)\right] \\ = &-\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right)^T\left(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M}\right)^{-1}\left(\mathbf{g}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}\right) \qquad (60) \\ &+\frac{1}{2}\pi\left(\mathbf{g}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}\right)^T\left(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M}\right)^{-2}\left(\mathbf{g}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}\right). \qquad (61)\end{aligned}$$

We first examine (60), which can be computed as follows:

$$\begin{aligned}(60) = &-\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right)^T\left(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M}\right)^{-1}\left(\mathbf{g}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}\right) \\ \overset{(a)}{=} &-\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right)^T\left(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M}\right)^{-1}\left(\mathbf{g}_{[t]} - \mathbf{g}^* - \mathbf{M}^T\mathbf{p}^* - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}\right) \\ \overset{(b)}{=} &-\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right)^T\left(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M}\right)^{-1}\left(\mathbf{g}_{[t]} - \mathbf{g}^* + \mathbf{M}^T\mathbf{Q}_*^{-1}\mathbf{1} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}\right) \\ = &-\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right)^T\left(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M}\right)^{-1}\left(\mathbf{g}_{[t]} - \mathbf{g}^*\right) \qquad (62) \\ &-\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right)^T\left(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M}\right)^{-1}\mathbf{M}^T\left(\mathbf{Q}_*^{-1} - \mathbf{Q}_{[t]}^{-1}\right)\mathbf{1}, \qquad (63)\end{aligned}$$

where (a) follows from the fact that $\mathbf{g}^* + \mathbf{M}^T\mathbf{p}^* = \mathbf{0}$ (i.e., the μ-ST condition) and (b) follows from the fact that $\mathbf{p}^* = -\mathbf{Q}_*^{-1}\mathbf{1}$ (i.e., the μ-CS condition).

For notational convenience, we let $\mathbf{F} = \left(\mathbf{H}_{[t]} - \mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{P}_{[t]}\mathbf{M}\right)$ and note that the following relationship follows from (62) and the convexity of $\mathbf{f}_\mu(\cdot)$:

$$\begin{aligned}(62) &= -\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right)^T\mathbf{F}^{-1}\left(\mathbf{g}_{[t]} - \mathbf{g}^*\right) \\ &\leq -\frac{1}{\lambda_{\min}\{\mathbf{F}\}}\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right)^T\left(\mathbf{g}_{[t]} - \mathbf{g}^*\right). \qquad (64)\end{aligned}$$

By the Mean-Value Theorem, we have the following pair of relationships:

$$f_\mu\left(\mathbf{y}_{[t]}\right) = f_\mu(\bar{\mathbf{y}}^*) + (\mathbf{g}^*)^T\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right) + \frac{1}{2}\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right)^T\mathbf{H}[\widetilde{\mathbf{y}}_1]\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right), \qquad (65)$$

$$f_\mu(\bar{\mathbf{y}}^*) = f_\mu\left(\mathbf{y}_{[t]}\right) + \left(\mathbf{g}_{[t]}\right)^T\left(\bar{\mathbf{y}}^* - \mathbf{y}_{[t]}\right) + \frac{1}{2}\left(\bar{\mathbf{y}}^* - \mathbf{y}_{[t]}\right)^T\mathbf{H}[\widetilde{\mathbf{y}}_2]\left(\bar{\mathbf{y}}^* - \mathbf{y}_{[t]}\right). \qquad (66)$$

In (65) and (66), $\mathbf{H}[\widetilde{\mathbf{y}}_1]$ and $\mathbf{H}[\widetilde{\mathbf{y}}_2]$ represent the matrices evaluated at points $\widetilde{\mathbf{y}}_1$ and $\widetilde{\mathbf{y}}_2$, where $\widetilde{\mathbf{y}}_1 = (1-\alpha_1)\mathbf{y}_{[t]} + \alpha_1\bar{\mathbf{y}}^*$ and $\widetilde{\mathbf{y}}_2 = (1-\alpha_2)\mathbf{y}_{[t]} + \alpha_2\bar{\mathbf{y}}^*$, for some $0 \leq \alpha_1, \alpha_2 \leq 1$. Next, adding (65) and (66) yields:

$$\left(\mathbf{g}_{[t]} - \mathbf{g}^*\right)^T\left(\bar{\mathbf{y}}^* - \mathbf{y}_{[t]}\right) + \frac{1}{2}\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right)^T\left(\mathbf{H}[\widetilde{\mathbf{y}}_1] + \mathbf{H}[\widetilde{\mathbf{y}}_2]\right)\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right) = 0,$$

which implies that

$$\begin{aligned}\left(\mathbf{g}_{[t]}-\mathbf{g}^*\right)^T\left(\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right)&=\frac{1}{2}\left(\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right)^T\left(\mathbf{H}[\widetilde{\mathbf{y}}_1]+\mathbf{H}[\widetilde{\mathbf{y}}_2]\right)\left(\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right)\\&\geq\lambda_{\min}(\mathbf{H})\left\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right\|^2.\end{aligned}\tag{67}$$

Combining (64) and (67), we can conclude that

$$(62)\leq -R\left\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right\|^2,\tag{68}$$

where we let $R\triangleq\frac{\lambda_{\min}\{\mathbf{H}\}}{\lambda_{\min}\{\mathbf{F}\}}$. Noting that the μ-factors in $\lambda_{\min}\{\mathbf{H}\}$ and $\lambda_{\min}\{\mathbf{F}\}$ cancel each other, we have that R is independent of μ.

Now, we evaluate the term in (63), which is non-positive because:

$$\begin{aligned}(63)=&-\left(\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right)^T\mathbf{F}^{-1}\mathbf{M}^T\left(\mathbf{Q}_*^{-1}-\mathbf{Q}_{[t]}^{-1}\right)\mathbf{1}\\\leq&\frac{1}{\lambda_{\min}(\mathbf{F})\Gamma}\left(\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right)^T\mathbf{M}^T\mathrm{Diag}\left\{\mathbf{M}\left(\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right)\right\}\mathbf{1}\\=&\frac{1}{\lambda_{\min}(\mathbf{F})\Gamma}\left\|\mathrm{Diag}\left\{\mathbf{M}\left(\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right)\right\}\mathbf{1}\right\|^2\leq 0,\end{aligned}\tag{69}$$

where Γ is defined as

$$\Gamma=\inf_t\left\{\left(\sum_{l\in\mathcal{I}(n)}x_{l,[t]}^{(f)}+s_{f,[t]}\mathbb{1}_f(n)-\sum_{l\in\mathcal{O}(n)}x_{l,[t]}^{(f)}\right)\left(\sum_{l\in\mathcal{I}(n)}\bar{x}_l^{(f),*}+\bar{s}_f^*\mathbb{1}_f(n)-\sum_{l\in\mathcal{O}(n)}\bar{x}_l^{(f,*)}\right)\right\}.$$

By combining (68) and (69), we have that

$$(60)\leq -R\left\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right\|^2.\tag{70}$$

Next, we analyze the quadratic term (61), for which we have:

$$\begin{aligned}(61)&\leq\frac{1}{2}\pi\left(\mathbf{g}_{[t]}-\mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}\right)^T\mathbf{F}^{-2}\left(\mathbf{g}_{[t]}-\mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}\right)\\&\leq\frac{\pi}{2\lambda_{\min}^2\{\mathbf{F}\}}\left\|\mathbf{g}_{[t]}-\mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}\right\|^2\\&\overset{(a)}{=}\frac{\pi}{2\lambda_{\min}^2\{\mathbf{F}\}}\left\|\mathbf{g}_{[t]}-\mathbf{g}^*-\mathbf{M}^T\mathbf{p}^*-\mathbf{M}^T\mathbf{Q}_{[t]}^{-1}\mathbf{1}\right\|^2\\&\overset{(b)}{=}\frac{\pi}{2\lambda_{\min}^2\{\mathbf{F}\}}\left\|\mathbf{g}_{[t]}-\mathbf{g}^*-\mathbf{M}^T\left(\mathbf{Q}_{[t]}^{-1}-\mathbf{Q}_*^{-1}\right)\mathbf{1}\right\|^2\\&\overset{(c)}{\leq}\frac{\pi}{2\lambda_{\min}^2\{\mathbf{F}\}}\left[\left\|\mathbf{g}_{[t]}-\mathbf{g}^*\right\|^2+\left\|\mathbf{M}^T\left(\mathbf{Q}_{[t]}^{-1}-\mathbf{Q}_*^{-1}\right)\right\|^2\right],\end{aligned}\tag{71}$$

where inequality (a) utilizes the μ-ST condition $\mathbf{g}^*+\mathbf{M}^T\mathbf{p}^*=\mathbf{0}$ (cf. (14)); equality (b) utilizes the μ-CS condition $\mathbf{p}^*=-\mathbf{Q}_*^{-1}\mathbf{1}$ (cf. (17)); and inequality (c) follows from the same argument

in (69). Note that the μ-factors in (71) cancel each other. Also, due to the boundedness of the primal variables $\mathbf{y}_{[t]}$ under the algorithmic design and the assumption that the utility function $U_f(\cdot)$ is Lipschtiz continuous, we can conclude that (71) is upper-bounded by some constant. By letting $B_1 \triangleq \sup_t \left\{ \frac{1}{2\lambda_{\min}^2\{\mathbf{F}\}} \left[\left\| \mathbf{g}_{[t]} - \mathbf{g}^* \right\|^2 + \left\| \mathbf{M}^T \left(\mathbf{Q}_{[t]}^{-1} - \mathbf{Q}_*^{-1} \right) \right\|^2 \right] \right\}$, we have

$$(61) \leq \pi B_1. \tag{72}$$

So far, we have analyzed the term (58) in the one-slot Lyapunov drift.

B.2 One-slot Lyapunov drift of (59)

Next, we move on to analyzing the other term (59) in the one-slot drift, which can be further expanded as follows:

$$\begin{aligned}
(59) = & \left[\frac{1}{\mu^3 \pi} \left(\mathbf{p}_{[t]} - \mathbf{p}^* \right) - \frac{1}{2\mu^3} \left(\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]} \right)^{-1} \left(\mathbf{M}\mathbf{H}_{[t]}^{-1} \left(\mathbf{g}_{[t]} + \mathbf{M}^T \pi_{[t]} \right) - \left(\mathbf{Q}_{[t]} + \mathbf{P}_{[t]}^{-1} \right) \mathbf{1} \right) \right]^T \\
& \times \left[-\pi \left(\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]} \right)^{-1} \left(\mathbf{M}\mathbf{H}_{[t]}^{-1} \left(\mathbf{g}_{[t]} + \mathbf{M}^T \mathbf{p}_{[t]} \right) - \left(\mathbf{Q}_{[t]} + \mathbf{P}_{[t]}^{-1} \right) \mathbf{1} \right) \right] \\
= & - \frac{1}{\mu^3} \left(\mathbf{p}_{[t]} - \mathbf{p}^* \right)^T \left[\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]} \right]^{-1} \left[\mathbf{M}\mathbf{H}_{[t]}^{-1} \left(\mathbf{g}_{[t]} + \mathbf{M}^T \mathbf{p}_{[t]} \right) - \left(\mathbf{Q}_{[t]} + \mathbf{P}_{[t]}^{-1} \right) \mathbf{1} \right] \quad (73) \\
& + \frac{\pi}{2\mu^3} \left[\mathbf{M}\mathbf{H}_{[t]}^{-1} \left(\mathbf{g}_{[t]} + \mathbf{M}^T \mathbf{p}_{[t]} \right) - \left(\mathbf{Q}_{[t]} + \mathbf{P}_{[t]}^{-1} \right) \mathbf{1} \right]^T \left[\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]} \right]^{-2} \\
& \times \left[\mathbf{M}\mathbf{H}_{[t]}^{-1} \left(\mathbf{g}_{[t]} + \mathbf{M}^T \mathbf{p}_{[t]} \right) - \left(\mathbf{Q}_{[t]} + \mathbf{P}_{[t]}^{-1} \right) \mathbf{1} \right]. \quad (74)
\end{aligned}$$

For convenience, we let $\mathbf{G}_{[t]} \triangleq \left[\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]} \right]$. Note that due to the $\mathbf{H}_{[t]}^{-1}$ term in $\mathbf{G}_{[t]}$, $\mathbf{G}_{[t]}^{-1}$ scales as $O(\mu)$. We first analyze (73), which can be further decomposed as follows:

$$\begin{aligned}
(73) = & - \frac{1}{\mu^3} \left(\mathbf{p}_{[t]} - \mathbf{p}^* \right)^T \mathbf{G}_{[t]}^{-1} \left[\mathbf{M}\mathbf{H}_{[t]}^{-1} \left(\mathbf{g}_{[t]} + \mathbf{M}^T \mathbf{p}_{[t]} \right) - \left(\mathbf{Q}_{[t]} + \mathbf{P}_{[t]}^{-1} \right) \mathbf{1} \right] \\
\overset{(a)}{=} & - \frac{1}{\mu^3} \left(\mathbf{p}_{[t]} - \mathbf{p}^* \right)^T \mathbf{G}_{[t]}^{-1} \left[\mathbf{M}\mathbf{H}_{[t]}^{-1} \left(\mathbf{g}_{[t]} - \mathbf{g}^* - \mathbf{M}^T \mathbf{p}^* + \mathbf{M}^T \mathbf{p}_{[t]} \right) - \left(\mathbf{Q}_{[t]} + \mathbf{P}_{[t]}^{-1} \right) \mathbf{1} \right] \\
= & - \frac{1}{\mu^3} \left(\mathbf{p}_{[t]} - \mathbf{p}^* \right)^T \mathbf{G}_{[t]}^{-1}\mathbf{M}\mathbf{H}_{[t]}^{-1} \left(\mathbf{g}_{[t]} - \mathbf{g}^* \right) - \frac{1}{\mu^2} \left(\mathbf{p}_{[t]} - \mathbf{p}^* \right)^T \mathbf{G}^{-1}\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T \left(\mathbf{p}_{[t]} - \mathbf{p}^* \right) \\
& + \frac{1}{\mu^3} \left(\mathbf{p}_{[t]} - \mathbf{p}^* \right)^T \mathbf{G}_{[t]}^{-1} \left(\mathbf{Q}_{[t]} + \mathbf{P}_{[t]}^{-1} \right) \mathbf{1} \\
\overset{(b)}{\leq} & - \frac{1}{\mu^3} \left(\mathbf{p}_{[t]} - \mathbf{p}^* \right)^T \mathbf{G}_{[t]}^{-1}\mathbf{M}\mathbf{H}_{[t]}^{-1} \left(\mathbf{g}_{[t]} - \mathbf{g}^* \right) \quad (75) \\
& + \frac{1}{\mu^3} \left(\mathbf{p}_{[t]} - \mathbf{p}^* \right)^T \mathbf{G}_{[t]}^{-1} \left(\mathbf{Q}_{[t]} + \mathbf{P}_{[t]}^{-1} \right) \mathbf{1}, \quad (76)
\end{aligned}$$

where (a) follows from subtracting the μ-ST condition $\mathbf{g}^* + \mathbf{M}^T\mathbf{p}^* = \mathbf{0}$; while (b) holds because $\mathbf{G}_{[t]}^{-1}\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T$ is positive semidefinite, which implies that

$$-\frac{1}{\mu^3} \left(\mathbf{p}_{[t]} - \mathbf{p}^* \right)^T \mathbf{G}_{[t]}^{-1}\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T \left(\mathbf{p}_{[t]} - \mathbf{p}^* \right) \leq 0.$$

As a result, to study the boundedness of (73) we only need to focus on the remaining two terms in (75) and (76). We begin with (76), which can be further computed as follows:

$$\begin{aligned}(76) =& \frac{1}{\mu^3}\left(\mathbf{p}_{[t]}-\mathbf{p}^*\right)^T\mathbf{G}_{[t]}^{-1}\left(\mathbf{Q}_{[t]}+\mathbf{P}_{[t]}^{-1}\right)\mathbf{1}\\ \overset{(a)}{=}& \frac{1}{\mu^3}\left(\mathbf{p}_{[t]}-\mathbf{p}^*\right)^T\mathbf{G}_{[t]}^{-1}\left[\mathbf{Q}_{[t]}-\mathbf{Q}_*-\mathbf{P}_*^{-1}+\mathbf{P}_{[t]}^{-1}\right]\mathbf{1}\\ \leq& \frac{1}{\mu^3}\left(\mathbf{p}_{[t]}-\mathbf{p}^*\right)^T\mathbf{G}_{[t]}^{-1}\mathbf{M}\left(\mathbf{y}_{[t]}-\mathbf{y}^*\right)+\frac{1}{\mu^2}\left(\mathbf{p}_{[t]}-\mathbf{p}^*\right)^T\mathbf{G}^{-1}\left(\mathbf{P}_{[t]}^{-1}-\mathbf{P}_*^{-1}\right)\mathbf{1}\\ \overset{(b)}{\leq}& \frac{1}{\mu^3}\left(\mathbf{p}_{[t]}-\mathbf{p}^*\right)^T\mathbf{G}_{[t]}-1\mathbf{M}\left(\mathbf{y}_{[t]}-\mathbf{y}^*\right),\end{aligned} \tag{77}$$

where equality (a) utilizes the μ-CS condition $\mathbf{Q}_*\mathbf{P}_* = -\mathbf{I}$ (i.e., $\mathbf{Q}_* = -\mathbf{P}_*^{-1}$) and inequality (b) holds because:

$$\begin{aligned}&\frac{1}{\mu^3}\left(\mathbf{p}_{[t]}-\mathbf{p}^*\right)^T\mathbf{G}_{[t]}^{-1}\left(\mathbf{P}_{[t]}^{-1}-\mathbf{P}_*^{-1}\right)\mathbf{1}\\ \leq& -\frac{1}{\mu^3\Phi\lambda_{\min}\{\mathbf{G}_{[t]}\}}\left(\mathbf{p}_{[t]}-\mathbf{p}^*\right)^T\left(\mathbf{p}_{[t]}-\mathbf{p}^*\right)\leq 0,\end{aligned}$$

where we let $\Phi \triangleq \inf_{t,n,f}\{p_{n,[t]}^{(f)}p_n^{(f),*}\}$. Next, combining (77) with (75), we have

$$-\frac{1}{\mu^3}\left(\mathbf{p}_{[t]}-\mathbf{p}^*\right)^T\mathbf{G}_{[t]}^{-1}\mathbf{M}\left[\mathbf{H}_{[t]}^{-1}\left(\mathbf{g}_{[t]}-\mathbf{g}^*\right)-\left(\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right)\right]. \tag{78}$$

By the vector-valued Taylor expansion of $\mathbf{g}$ [24], we have

$$\mathbf{g}^* = \mathbf{g}_{[t]}+\mathbf{H}_{[t]}\left(\bar{\mathbf{y}}^*-\mathbf{y}_{[t]}\right)+o(\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\|)\mathbf{1},$$

which further implies that

$$\mathbf{H}_{[t]}^{-1}\left(\mathbf{g}_{[t]}-\mathbf{g}^*\right)-\left(\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right) = o(\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\|)\mathbf{1}. \tag{79}$$

Therefore, we have

$$\begin{aligned}(78) \leq& -\frac{1}{\mu^3}\left(\mathbf{p}_{[t]}-\mathbf{p}^*\right)^T\mathbf{G}_{[t]}^{-1}\mathbf{M}O\left(\left\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right\|^2\right)\mathbf{1}\\ \overset{(a)}{\leq}& -\frac{O\left(\left\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right\|^2\right)}{\mu^3\lambda_{\min}\{\mathbf{G}_{[t]}\}}\left\|\mathbf{p}_{[t]}-\mathbf{p}^*\right\|\|\mathbf{M}\mathbf{1}\|,\end{aligned} \tag{80}$$

where inequality (a) follows from Cauchy-Schwarz inequality. From the boundedness result of $\mathbf{p}$ in Lemma 4, we have that $\left\|\mathbf{p}_{[t]}-\mathbf{p}^*\right\|$ is bounded. Also, from the control scheme itself, we know that the entries in $\mathbf{y}_{[t]}$ is fundamentally bounded by the link capacities. Hence, we can conclude that (80) is upper-bounded by some constant. By letting

$$B_2 \triangleq \frac{\|\mathbf{M}\mathbf{1}\|}{\mu^2\lambda_{\min}\{\mathbf{G}\}}\sup_t\left\{\left\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right\|\left\|\mathbf{p}_{[t]}-\mathbf{p}^*\right\|\right\},$$

(cf. B_2 in (23)) where we leave a μ^2-factor inside the denominator to cancel out the μ-factors in $\left\|\mathbf{p}_{[t]}-\mathbf{p}^*\right\|$ and $\frac{1}{\lambda_{\min}\{\mathbf{G}_{[t]}\}}$, we have

$$-\frac{1}{\mu^3}\left(\mathbf{p}_{[t]}-\mathbf{p}^*\right)^T\mathbf{G}_{[t]}^{-1}\mathbf{M}\left[\mathbf{H}_{[t]}^{-1}\left(\mathbf{g}_{[t]}-\mathbf{g}^*\right)-\left(\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right)\right]\leq(78)\leq\frac{1}{\mu}B_2. \tag{81}$$

Based on the above derivations, we can finally bound (73) as:

$$(73)\leq(75)+(76)\leq\frac{1}{\mu}B_2. \tag{82}$$

Lastly, we evaluate (74), for which we have

$$\begin{aligned}
(74)\leq&\frac{1}{2\mu^3}\left[\mathbf{M}\mathbf{H}_{[t]}^{-1}\left(\mathbf{g}_{[t]}+\mathbf{M}^T\mathbf{p}_{[t]}\right)-\left(\mathbf{Q}_{[t]}+\mathbf{P}_{[t]}^{-1}\right)\mathbf{1}\right]^T\mathbf{G}_{[t]}^{-2}\\
&\times\left[\mathbf{M}\mathbf{H}_{[t]}^{-1}\left(\mathbf{g}_{[t]}+\mathbf{M}^T\mathbf{p}_{[t]}\right)-\left(\mathbf{Q}_{[t]}+\mathbf{P}_{[t]}^{-1}\right)\mathbf{1}\right]\\
\leq&\frac{1}{2\mu^3\lambda_{\min}^2\{\mathbf{G}_{[t]}\}}\left\|\mathbf{M}\mathbf{H}_{[t]}^{-1}\left(\mathbf{g}_{[t]}+\mathbf{M}^T\mathbf{p}_{[t]}\right)-\left(\mathbf{Q}_{[t]}+\mathbf{P}_{[t]}^{-1}\right)\mathbf{1}\right\|^2\\
=&\frac{1}{2\mu^3\lambda_{\min}^2\{\mathbf{G}_{[t]}\}}\left\|\mathbf{M}\left(\mathbf{H}_{[t]}^{-1}\mathbf{g}_{[t]}-\mathbf{y}_{[t]}\right)+\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T\mathbf{p}_{[t]}-\mathbf{P}_{[t]}^{-1}\mathbf{1}\right\|^2\\
\overset{(a)}{\leq}&\frac{1}{2\mu^3\lambda_{\min}^2\{\mathbf{G}_{[t]}\}}\left[\left\|\mathbf{M}\left(\mathbf{H}_{[t]}^{-1}\mathbf{g}_{[t]}-\mathbf{y}_{[t]}\right)\right\|+\left\|\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T\mathbf{p}_{[t]}\right\|+\left\|\mathbf{P}_{[t]}^{-1}\mathbf{1}\right\|\right]^2\\
=&\frac{1}{2\mu^3\lambda_{\min}^2\{\mathbf{G}_{[t]}\}}\left[\left\|\mathbf{M}\left(\mathbf{H}_{[t]}^{-1}\left(\mathbf{g}_{[t]}-\mathbf{g}^*\right)-\left(\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right)\right)+\mathbf{M}\left(\mathbf{H}_{[t]}^{-1}\mathbf{g}^*-\bar{\mathbf{y}}^*\right)\right\|\right.\\
&\left.+\left\|\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T\mathbf{p}_{[t]}\right\|+\left\|\mathbf{P}_{[t]}^{-1}\mathbf{1}\right\|\right]^2\\
\overset{(b)}{\leq}&\frac{1}{2\mu^3\lambda_{\min}^2\{\mathbf{G}_{[t]}\}}\left[O\left(\left\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right\|^2\right)\|\mathbf{M}\mathbf{1}\|+\left\|\mathbf{M}\left(\mathbf{H}_{[t]}^{-1}\mathbf{g}^*-\bar{\mathbf{y}}^*\right)\right\|\right.\\
&\left.+\left\|\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T\mathbf{p}_{[t]}\right\|+\left\|\mathbf{P}_{[t]}^{-1}\mathbf{1}\right\|\right]^2,
\end{aligned}\tag{83}$$

where (a) is due to triangular inequality and (b) follows from (79). Note that in (83), $O\left(\left\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\right\|^2\right)\|\mathbf{M}\mathbf{1}\|$ is upper-bounded since, by our algorithmic design, $\|\mathbf{y}_{[t]}\|$ is bounded; $\left\|\mathbf{M}\left(\mathbf{H}_{[t]}^{-1}\mathbf{g}^*-\bar{\mathbf{y}}^*\right)\right\|$ is upper-bounded due to the Lipschitz continuity of $\mathbf{g}$ as well as the μ-factor cancellation between $\mathbf{H}_{[t]}^{-1}$ and $\mathbf{g}^*$; and $\left\|\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T\mathbf{p}_{[t]}\right\|$ is upper-bounded due to: i) the boundedness of $\|\mathbf{p}_{[t]}\|$ from Lemma 4, and ii) the μ factors cancellation between $\mathbf{H}_{[t]}^{-1}$ and $\mathbf{p}_{[t]}$. Also, $\left\|\mathbf{P}_{[t]}^{-1}\mathbf{1}\right\|$ is a diminishing term when μ is large. Therefore, from the above discussions, we can conclude that (83) is upper-bounded. By letting

$$\begin{aligned}
B_3\triangleq&\frac{1}{2\mu^2\lambda_{\min}^2\{\mathbf{G}\}}\sup_t\left\{\left[\|\mathbf{y}_{[t]}-\bar{\mathbf{y}}^*\|\|\mathbf{M}\mathbf{1}\|\right.\right.\\
&\left.\left.+\|\mathbf{M}(\mathbf{H}_{[t]}^{-1}\mathbf{g}^*-\bar{\mathbf{y}}^*)\|\|\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T\mathbf{p}_{[t]}\|+\|\mathbf{P}_{[t]}^{-1}\mathbf{1}\|\right]\right\},
\end{aligned}$$

we have

$$(74)\leq\frac{1}{\mu}B_3. \tag{84}$$

Finally, combining all the results in (70), (72), (82), and (84), we arrive at the following result for the one-slot drift analysis:

$$\begin{aligned} & V\left(\mathbf{y}_{[t+1]}, \mathbf{p}_{[t+1]}\right) - V\left(\mathbf{y}_{[t]}, \mathbf{p}_{[t]}\right) \\ & \leq -R\|\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\|^2 + \pi B_1 + \frac{1}{\mu} B_2 + \frac{1}{\mu} B_3. \end{aligned} \tag{85}$$

B.3 Final Telescoping Step

Clearly, we can see that if π scales as $O(\frac{1}{\mu})$, (85) implies that the following relationship holds:

$$V\left(\mathbf{y}_{[t+1]}, \mathbf{p}_{[t+1]}\right) - V\left(\mathbf{y}_{[t]}, \mathbf{p}_{[t]}\right) \leq -R\|\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\|^2 + \frac{1}{\mu}\widehat{B}, \tag{86}$$

where $\widehat{B} = \alpha B_1 + B_2 + B_3$ for some $\alpha > 0$. Writing this one-slot drift expressions for $t = 0, \ldots, T-1$, we have a telescoping series. Summing all the terms in this series yields:

$$V\left(\mathbf{y}_{[T]}, \mathbf{p}_{[T]}\right) - V\left(\mathbf{y}_{[0]}, \mathbf{p}_{[0]}\right) \leq -R\sum_{t=0}^{T-1} \|\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\|^2 + \frac{T}{\mu}\widehat{B}.$$

Dividing both sides by T and rearranging terms, we have

$$\frac{R}{T}\sum_{t=0}^{T-1} \|\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\|^2 \leq \frac{\widehat{B}}{\mu} - \frac{1}{T}\left[V\left(\mathbf{y}_{[T]}, \mathbf{p}_{[T]}\right) - V\left(\mathbf{y}_{[0]}, \mathbf{p}_{[0]}\right)\right].$$

Dividing both sides by R and taking the limit as T goes to infinity, we have

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \|\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\|^2 \leq \frac{B^2}{\mu}, \tag{87}$$

where we let $B^2 \triangleq \widehat{B}/R$. Therefore, as T gets large, we have

$$\left|\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right)\right| \overset{(a)}{\leq} \frac{1}{T}\sum_{t=0}^{T-1}\left|\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\right| \overset{(b)}{\leq} \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\|\mathbf{y}_{[t]} - \bar{\mathbf{y}}^*\|^2} \leq \frac{B}{\sqrt{\mu}}, \tag{88}$$

where (a) follows from triangular inequality and (b) is due to the relationship between l_1 and l_2 norms. Then, the result stated in Theorem 1 follows by taking lim sup and lim inf, respectively. This completes the proof of Theorem 1.

C Proof of Lemma 5

Recall that in each time-slot t, the primal and dual Newton directions are obtained via the following linear equation system:

$$\begin{bmatrix} \mathbf{H}_{[t]} & \mathbf{M}^T \\ -\mathbf{P}_{[t]}\mathbf{M} & -\mathbf{Q}_{[t]} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{y}_{[t]} \\ \Delta\mathbf{p}_{[t]} \end{bmatrix} = -\begin{bmatrix} \mathbf{g}_{[t]} + \mathbf{M}^T\mathbf{p}_{[t]} \\ -(\mathbf{P}_{[t]}\mathbf{Q}_{[t]} + \mathbf{I})\mathbf{1} \end{bmatrix}.$$

From the first row, we have

$$\mathbf{H}_{[t]}\Delta\mathbf{y}_{[t]} + \mathbf{M}^T\Delta\mathbf{p}_{[t]} = -\mathbf{g}_{[t]} - \mathbf{M}^T\mathbf{p}_{[t]}. \tag{89}$$

Moving the term $\mathbf{M}^T\Delta\mathbf{p}_{[t]}$ to the RHS and noting that $\widetilde{\mathbf{p}}_{[t+1]} = \mathbf{p}_{[t]} + \Delta\mathbf{p}_{[t]}$, we have $\mathbf{H}_{[t]}\Delta\mathbf{y}_{[t]} = -\left(\mathbf{g}_{[t]} + \mathbf{M}^T\widetilde{\mathbf{p}}_{[t+1]}\right)$, which implies that

$$\Delta\mathbf{y}_{[t]} = -\mathbf{H}_{[t]}^{-1}\left(\mathbf{g}_{[t]} + \mathbf{M}^T\widetilde{\mathbf{p}}_{[t+1]}\right), \tag{90}$$

which is exactly the expression in (29).

Next, from the second row, we have

$$\begin{aligned}
&-\mathbf{P}_{[t]}\mathbf{M}\Delta\mathbf{y}_{[t]} - \mathbf{Q}_{[t]}\Delta\mathbf{p}_{[t]} = \mathbf{Q}_{[t]}\mathbf{p}_{[t]} + \mathbf{1} \\
\Rightarrow &-\mathbf{Q}_{[t]}\left(\mathbf{p}_{[t]} + \Delta\mathbf{p}_{[t]}\right) = \mathbf{P}_{[t]}\mathbf{M}\Delta\mathbf{y}_{[t]} + \mathbf{1} \\
\overset{(a)}{\Rightarrow} &-\mathbf{Q}_{[t]}\widetilde{\mathbf{p}}_{[t+1]} = \mathbf{P}_{[t]}\mathbf{M}\left[-\mathbf{H}_{[t]}^{-1}\left(\mathbf{g}_{[t]} + \mathbf{M}^T\widetilde{\mathbf{p}}_{[t+1]}\right)\right] + \mathbf{1} \\
\Rightarrow &-\mathbf{Q}_{[t]}\widetilde{\mathbf{p}}_{[t+1]} = -\mathbf{P}_{[t]}\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{g}_{[t]} - \mathbf{P}_{[t]}\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T\widetilde{\mathbf{p}}_{[t+1]} + \mathbf{1} \\
\Rightarrow &\left(\mathbf{P}_{[t]}\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{Q}_{[t]}\right)\widetilde{\mathbf{p}}_{[t+1]} = -\mathbf{P}_{[t]}\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{g}_{[t]} + \mathbf{1} \\
\overset{(b)}{\Rightarrow} &\left(\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}\right)\widetilde{\mathbf{p}}_{[t+1]} = -\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{g}_{[t]} + \mathbf{P}_{[t]}^{-1}\mathbf{1} \\
\Rightarrow &\widetilde{\mathbf{p}}_{[t+1]} = \underbrace{\left(\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}\right)^{-1}}_{=\mathbf{G}^{-1}}\left[-\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{g}_{[t]} + \mathbf{P}_{[t]}^{-1}\mathbf{1}\right],
\end{aligned} \tag{91}$$

where (a) utilizes $\widetilde{\mathbf{p}}_{[t]} = \mathbf{p}_{[t]} + \Delta\mathbf{p}_{[t]}$ and $\Delta\mathbf{y}_{[t]} = -\mathbf{H}_{[t]}^{-1}\left(\mathbf{g}_{[t]} + \mathbf{M}^T\widetilde{\mathbf{p}}_{[t+1]}\right)$; and (b) follows from multiplying $\mathbf{P}_{[t]}^{-1}$ on both sides. This completes the proof.

D Proof of Lemma 7

First, consider the diagonal entries in $\mathbf{D}_f$. Note that $\mathbf{D}_f = \frac{s_f^2}{\mu}\mathbf{b}^{(f)}(\mathbf{b}^{(f)})^T + \sum_{l=1}^{L}(x_l^{(f)})^2\mathbf{a}_l^{(f)}(\mathbf{a}_l^{(f)})^T$. From [9, Lemma 2], the i-th diagonal entry in $\mathbf{a}_l^{(f)}(\mathbf{a}_l^{(f)})^T$ is equal to 1 if the corresponding node of the i-th entry, say n, is either $\mathrm{Tx}(l)$ or $\mathrm{Rx}(l)$. Thus, when summing over all l, the number of ones is precisely given by the number of links that have node n either as its transmitting node or receiving node, i.e., the links that are in either $\mathcal{O}(n)$ and $\mathcal{I}(n)$. Thus, we have $(\sum_{l=1}^{L}(x_l^{(f)})^2\mathbf{a}_l^{(f)}(\mathbf{a}_l^{(f)})^T)_{ii} = \sum_{l\in\mathcal{I}(n)\cup\mathcal{O}(n)}(x_l^{(f)})^2$. Also, from [9, Lemma 1], we have that the i-th diagonal entry is equal to 1 if $n = \mathrm{Src}(f)$. Hence, we have

$$(\mathbf{D}_f)_{ii} = \begin{cases} \sum_{l\in\mathcal{O}(n)\cup\mathcal{I}(n)}(x_l^{(f)})^2 + \frac{s_f^2}{\mu} + \frac{1}{p_n^{(f)}}\left[\sum_{l\in\mathcal{O}(n)}x_l^{(f)} - s_f\mathbb{1}_f(n) - \sum_{l\in\mathcal{I}(n)}x_l^{(f)}\right] \\ \quad \text{if row } i \text{ corresponds to node } n \text{ and } n = \mathrm{Src}(f), \\ \sum_{l\in\mathcal{O}(n)\cup\mathcal{I}(n)}(x_l^{(f)})^2 + \frac{1}{p_n^{(f)}}\left[\sum_{l\in\mathcal{O}(n)}x_l^{(f)} - s_f\mathbb{1}_f(n) - \sum_{l\in\mathcal{I}(n)}x_l^{(f)}\right] \\ \quad \text{otherwise,} \end{cases}$$

which is the same expression as in Lemma 7.

Next, consider the off-diagonal entries in $\mathbf{D}_f$. Again, from [9, Lemma 2], we know that the (i,j)-th entry in $\mathbf{a}_l^{(f)}(\mathbf{a}_l^{(f)})^T$ is equal to -1 if the corresponding nodes of the (i,j)-th entry, say n_1 and n_2, are $\mathrm{Tx}(l)$ and $\mathrm{Rx}(l)$, or vice versa. Thus, when summing over all l, the number of -1 entries is precisely given by the number of links that have nodes n_1 and n_2 either as their transmitting node and receiving node, i.e., the links that are in $\Gamma(n_1,n_2)$. Hence, we have

$$(\mathbf{D}_f)_{ij} = \begin{cases} -\sum_{l\in\Gamma(n_1,n_2)}(x_l^{(f)})^2 & \text{if row } i \text{ and column } j \text{ correspond to two connected nodes } n_1 \text{ and } n_2, \\ 0 & \text{otherwise,} \end{cases}$$

which is the same expression as in Lemma 7, and the proof is complete.

E Proof of Proposition 13

First, note that, if $\mathbf{y}_{[t]}$ and $\mathbf{p}_{[t]}$ are primal and dual feasible, $\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]} \succ 0$ because $f(\mathbf{y})$ is convex. Hence, $(\mathbf{\Lambda}_{[t]} + \alpha\overline{\mathbf{\Omega}}) - (\alpha\overline{\mathbf{\Omega}} - \mathbf{\Omega}_{[t]})$ is positive definite. Next, we check the positive definiteness of $(\mathbf{\Lambda}_{[t]} + \alpha\overline{\mathbf{\Omega}}) + (\alpha\overline{\mathbf{\Omega}} - \mathbf{\Omega}_{[t]})$. Note that

$$(\mathbf{\Lambda}_{[t]} + \alpha\overline{\mathbf{\Omega}}) + (\alpha\overline{\mathbf{\Omega}} - \mathbf{\Omega}_{[t]}) = \mathbf{\Lambda}_{[t]} + 2\alpha\overline{\mathbf{\Omega}}_{[t]} - \mathbf{\Omega}_{[t]}. \tag{92}$$

From the definition of $\mathbf{\Lambda}_k$, Lemma 7, and Lemma 8, we have that all diagonal entries in $\mathbf{\Lambda}_k$ are positive. Hence, $\mathbf{\Lambda}_{[t]} \succ 0$. On the other hand, by the definitions of $\overline{\mathbf{\Omega}}_{[t]}$ and $\mathbf{\Omega}_{[t]}$, we have that the entries of each row in $2\alpha\overline{\mathbf{\Omega}}_{[t]} - \mathbf{\Omega}_{[t]}$ satisfy

$$\begin{aligned} &(2\alpha\overline{\mathbf{\Omega}}_{[t]} - \mathbf{\Omega}_{[t]})_{ii} - \sum_{j\neq i} |(2\alpha\overline{\mathbf{\Omega}}_{[t]} - \mathbf{\Omega}_{[t]})_{ij}| \\ &= (2\alpha - 1)\sum_{j\neq i} |(\mathbf{\Omega}_{[t]})_{ij}| > 0, \quad \text{for } \alpha > \frac{1}{2}. \end{aligned}$$

Also, it is clear from the definitions of $\overline{\mathbf{\Omega}}_{[t]}$ and $\mathbf{\Omega}_{[t]}$ that $(2\alpha\overline{\mathbf{\Omega}}_{[t]} - \mathbf{\Omega}_{[t]})_{ii} > 0$. Thus, $2\alpha\overline{\mathbf{\Omega}}_{[t]} - \mathbf{\Omega}_{[t]}$ is diagonally dominant and hence positive definite. Therefore, $\mathbf{\Lambda}_{[t]} + 2\alpha\overline{\mathbf{\Omega}}_{[t]} - \mathbf{\Omega}_{[t]}$ is also positive definite, and the proof is complete.

F Proof of Theorem 15

The expressions in Theorem 15 can be derived by computing the element-wise expansion of (44). First, since $(\mathbf{\Lambda}_{[t]} + \alpha\overline{\mathbf{\Omega}}_{[t]})$ is diagonal, and its inverse can be easily computed by taking the inverse of each diagonal entry. Thus, we begin with computing each diagonal entry in $(\mathbf{\Lambda}_{[t]} + \alpha\overline{\mathbf{\Omega}}_{[t]})$. Toward

this end, we first define the following index function $\beta_f(n)$, $n \neq \mathrm{Dst}(f)$:

$$\beta_f(n) \triangleq \begin{cases} n & \text{if } n < \mathrm{Dst}(f), \\ n-1 & \text{if } n > \mathrm{Dst}(f). \end{cases} \tag{93}$$

Since $\mathbf{\Lambda}_{[t]}$ contains the main diagonal of $\mathbf{G} = \mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}$, from Theorem 10, we obtain that

$$(\mathbf{\Lambda}_{[t]})_{ii} = \begin{cases} \sum_{\Phi(n)}(x_l^{(f)})^2\Big(1 - \frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big) + \frac{1}{p_n^{(f)}}\Big[\sum_{l\in\mathcal{O}(n)} x_l^{(f)} - s_f \mathbb{1}_f(n) - \sum_{l\in\mathcal{I}(n)} x_l^{(f)}\Big] + \\ \quad \frac{1}{-\mu U_f''(s_f) + \frac{1}{(s_f)^2}}, \quad \text{if } n = \mathrm{Src}(f), \\ \sum_{\Phi(n)}(x_l^{(f)})^2\Big(1 - \frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big) + \frac{1}{p_n^{(f)}}\Big[\sum_{l\in\mathcal{O}(n)} x_l^{(f)} - s_f \mathbb{1}_f(n) - \sum_{l\in\mathcal{I}(n)} x_l^{(f)}\Big], \quad \text{if } n \neq \mathrm{Src}(f), \end{cases} \tag{94}$$

where the index i satisfies $i = (f-1)(N-1) + \beta_f(n)$.

Next, note that each diagonal entry in $\overline{\mathbf{\Omega}}_{[t]}$ is the row sum of non-diagonal entries in $\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{M}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}$. Therefore, from Theorem 10, we have that

$$(\overline{\mathbf{\Omega}}_{[t]})_{ii} = \sum_{l\in\Phi(n)\setminus\Psi(n,f)} (x_l^{(f)})^2\Big(1 - \frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big) + \sum_{f'=1,\neq f}^{F} \sum_{l\in\Psi(n,f')} \frac{(x_l^{(f)} x_l^{(f')})^2}{\|\widehat{\mathbf{x}}_l\|^2}. \tag{95}$$

Then, using the indicator function $\mathbb{1}_{\Psi(n,f)}$ and combining (94) and (95), we have that

$$(\mathbf{\Lambda}_k + \alpha\overline{\mathbf{\Omega}}_{[t]})_{ii} = \begin{cases} \sum_{l\in\Phi(n)}[1+\alpha(1-\mathbb{1}_{\Psi(n,f)}(l))](x_l^{(f)})^2\Big(1 - \frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big) + \\ \quad \frac{1}{p_n^{(f)}}\Big[\sum_{l\in\mathcal{O}(n)} x_l^{(f)} - s_f \mathbb{1}_f(n) - \sum_{l\in\mathcal{I}(n)} x_l^{(f)}\Big] + \\ \quad \sum_{f'=1,\neq f}^{F}\Big(\sum_{l\in\Psi(n,f')} \frac{\alpha(x_l^{(f)} x_l^{(f')})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big) & \text{if } n \neq \mathrm{Src}(f), \\ \sum_{l\in\Phi(n)}[1+\alpha(1-\mathbb{1}_{\Psi(n,f)}(l))](x_l^{(f)})^2\Big(1 - \frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big) + \\ \quad \frac{1}{p_n^{(f)}}\Big[\sum_{l\in\mathcal{O}(n)} x_l^{(f)} - s_f \mathbb{1}_f(n) - \sum_{l\in\mathcal{I}(n)} x_l^{(f)}\Big] + \\ \quad \sum_{f'=1,\neq f}^{F}\Big(\sum_{l\in\Psi(n,f')} \frac{\alpha(x_l^{(f)} x_l^{(f')})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big) + \frac{1}{-\mu U_f''(s_f) + \frac{1}{(s_f)^2}} & \text{if } n = \mathrm{Src}(f), \end{cases}$$

which is the same as the definition of $U_n^{(f)}[k]$ in (46).

Next, consider the entries in $(\alpha\overline{\mathbf{\Omega}}_{[t]} - \mathbf{\Omega}_{[t]})\mathbf{p}_{[t]}$. Recall from Theorem 10 that the matrix $\mathbf{G} = \widetilde{\mathbf{M}}\widetilde{\mathbf{H}}_k^{-1}\widetilde{\mathbf{M}}^T - \mathbf{P}_{[t]}^{-1}\mathbf{Q}_{[t]}$ has a partitioned matrix structure. Thus, the vector $(\alpha\overline{\mathbf{\Omega}}_{[t]} - \mathbf{\Omega}_{[t]})\mathbf{p}_{[t]}$ can be partitioned into F blocks, where each block is of the form

$$((\alpha\overline{\mathbf{\Omega}}_{[t]} - \mathbf{\Omega}_{[t]})\mathbf{p}_{[t]})_f = -\mathbf{R}_f \mathbf{p}_{[t]}^f + \sum_{f'=1,\neq f}^{F} \mathbf{G}_{ff'}\mathbf{p}_{[t]}^{(f')}, \quad f = 1, \ldots, F, \tag{96}$$

where $\mathbf{R}_f$ is obtained by replacing the main diagonal of $\mathbf{D}_f - \widehat{\mathbf{D}}_f$ with the corresponding entries in $-\alpha\overline{\mathbf{\Omega}}_{[t]}$. Then, by computing the entries in $-\mathbf{R}_f\mathbf{p}_{[t]}^f$ and noticing the special structure in $\mathbf{R}_f$ (only

containing entries 1, -1, and 0), we have

$$(-\mathbf{R}_f \mathbf{p}_{[t]}^f)_n = \sum_{l\in\mathcal{I}(n)} (x_l^{(f)})^2 \Big(1 - \frac{((x_l^{(f)})^2)}{\|\widehat{\mathbf{x}}_l\|^2}\Big)(p_{\mathrm{Tx}(l)}^{(f)} - \alpha p_{\mathrm{Rx}(l)}^{(f)}) + \sum_{l\in\mathcal{O}(n)\setminus\Psi(n,f)} (x_l^{(f)})^2 \Big(1 - \frac{((x_l^{(f)})^2)}{\|\widehat{\mathbf{x}}_l\|^2}\Big)(p_{\mathrm{Rx}(l)}^{(f)} - \alpha p_{\mathrm{Tx}(l)}^{(f)}) - \sum_{f'=1,\neq f}^{F} \Big(\sum_{l\in\Psi(n,f')} \frac{\alpha (x_l^{(f)} x_l^{(f')})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big) p_n^f,$$

which is exactly the definition of $V_{n,1}^{(f)}(k)$ in (47).

Likewise, by computing the entries in $\sum_{f'=1,\neq f}^{F} \mathbf{G}_{ff'}\mathbf{p}_{[t]}^{(f')}$, we have

$$\Big(\sum_{f'=1,\neq f}^{F} \mathbf{G}_{ff'}\mathbf{p}_{[t]}^{(f')}\Big)_n = \sum_{f'=1,\neq f}^{F} \Big(\Big(\sum_{l\in\mathcal{O}(n)} \frac{(x_l^{(f)} x_l^{(f')})^2}{\|\widehat{\mathbf{x}}_l\|^2} - \sum_{l\in\mathcal{I}(n)} \frac{(x_l^{(f)} x_l^{(f')})^2}{\|\widehat{\mathbf{x}}_l\|^2}\Big)(p_{\mathrm{Tx}(l)}^{(f')} - p_{\mathrm{Rx}(l)}^{(f')})\Big),$$

which is the same as the definition of $V_{n,2}^{(f)}(k)$ in (48).

Finally, consider the term $\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{g}_{[t]}$. Note that $\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{g}_{[t]}$ can be decomposed into

$$\mathbf{M}\mathbf{H}_{[t]}^{-1}\mathbf{g}_{[t]} = \mathbf{B}\mathbf{S}^{-1}\nabla_{\mathbf{s}} f(\mathbf{y}_{[t]}) + \sum_{l=1}^{L} -\mathbf{A}_l \mathbf{X}_l^{-1} \nabla_{\mathbf{x}_l} f(\mathbf{y}_{[t]}),$$

where $\mathbf{s} \triangleq [s_1, \ldots, s_F]^T$ and $\mathbf{x}_l \triangleq [x_l^{(1)}, \ldots, x_l^{(F)}]^T$. Now, first consider the term $\mathbf{B}\mathbf{S}^{-1}\nabla_{\mathbf{s}} f(\mathbf{y}_{[t]})$. Using the diagonal structure of $\mathbf{B}$ and $\mathbf{S}$, it can be verified that

$$(\mathbf{B}\mathbf{S}^{-1}\nabla_{\mathbf{s}} f(\mathbf{y}_{[t]}))_n^{(f)} = \begin{cases} \frac{s_f(1+\mu s_f U_f'(s_f))}{\mu s_f^2 U_f''(s_f) - 1} & \text{if } n = \mathrm{Src}(f), \\ 0 & \text{otherwise.} \end{cases}$$

Recall that $\mathbf{H}_{[t]}^{-1}$ can be decomposed into a diagonal matrix and a rank-one update matrix. Hence, we have

$$-\mathbf{A}_l \mathbf{X}_l^{-1} \nabla_{\mathbf{x}_l} f(\mathbf{y}_{[t]}) = -\mathbf{A}_l \mathrm{Diag}\Big\{(x_l^{(1)})^2, \ldots, (x_l^{(F)})^2\Big\} \begin{bmatrix} \frac{1}{\delta_l} - \frac{1}{x_l^{(1)}} \\ \vdots \\ \frac{1}{\delta_l} - \frac{1}{x_l^{(F)}} \end{bmatrix} + \frac{1}{\|\widehat{\mathbf{x}}_l\|^2} \mathbf{A}_l \begin{bmatrix} (x_l^{(1)})^4 & \cdots & (x_l^{(1)} x_l^{(F)})^2 \\ \vdots & \ddots & \vdots \\ (x_l^{(F)} x_l^{(1)})^2 & \cdots & (x_l^{(F)})^4 \end{bmatrix} \begin{bmatrix} \frac{1}{\delta_l} - \frac{1}{x_l^{(1)}} \\ \vdots \\ \frac{1}{\delta_l} - \frac{1}{x_l^{(F)}} \end{bmatrix}.$$

Hence, computing each term in the above decomposition, then adding $\mathbf{B}\mathbf{S}^{-1}\nabla_{\mathbf{s}} f(\mathbf{y}_{[t]})$, and then summing over all l, we obtain that

$$(\mathbf{M}\mathbf{H}_{[t]}^{-1}\nabla f(\mathbf{y}_{[t]}))_n^{(f)} = \begin{cases} \left(1-\frac{x_l^{(f)}}{\delta_l}\right)\Big[\sum_{l\in\mathcal{O}(n)}\left(1-\sum_{f'=1}^{F}\frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2}x_l^{(f')}\right)- \\ \qquad \sum_{l\in\mathcal{I}(n)}\left(1-\sum_{f'=1}^{F}\frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2}x_l^{(f')}\right)\Big]+\frac{1}{p_n^{(f)}} & \text{if } n\neq \mathrm{Src}(f), \\ \sum_{l\in\mathcal{O}(n)}\left(1-\sum_{f'=1}^{F}\frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2}x_l^{(f')}\right)- \\ \qquad \sum_{l\in\mathcal{I}(n)}\left(1-\sum_{f'=1}^{F}\frac{(x_l^{(f)})^2}{\|\widehat{\mathbf{x}}_l\|^2}x_l^{(f')}\right)\Big]+\frac{1}{p_n^{(f)}}+\frac{s_f(1+\mu s_f U_f'(s_f))}{\mu s_f^2 U_f''(s_f)-1} & \text{if } n=\mathrm{Src}(f), \end{cases}$$

which is the same as the definition of $W_n^{(f)}[k]$ as in (49). Finally, the result in (45) simply follows from Proposition 13, and the proof is complete.

References

[1] X. Lin and N. B. Shroff, "Joint rate control and scheduling in multihop wireless networks," in *Proc. IEEE CDC*, Atlantis, Paradise Island, Bahamas, Dec. 2006, pp. 1484–1489.

[2] M. J. Neely, E. Modiano, and C.-P. Li, "Faireness and optimal stochastic control for heterogeneous networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 2, pp. 396–409, Apr. 2008.

[3] A. Eryilmaz and R. Srikant, "Joint congestion control, routing, and MAC for stability and fairness in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1514–1524, Aug. 2006.

[4] X. Lin and N. B. Shroff, "The impact of imperfect scheduling on cross-layer congestion control in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 14, no. 2, pp. 302–315, Apr. 2006.

[5] L. Tassiulas and A. Ephremides, "Stability properties of constrained queuing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.

[6] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3rd ed. New York, NY: John Wiley & Sons Inc., 2006.

[7] A. Eryilmaz and R. Srikant, "Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control," in *Proc. IEEE INFOCOM*, Miami, FL, Mar. 2005, pp. 1804–1814.

[8] E. Wei, A. Ozdaglar, and A. Jadbabaie, "A distributed Newton method for network utitlity maximization," in *Proc. IEEE Conference on Decision and Control (CDC)*, Atlanta, GA, Dec. 15-17, 2010.

[9] J. Liu and H. D. Sherali, "A distributed Newton's method for joint multi-hop routing and flow control: Theory and algorithm," in *Proc. IEEE INFOCOM*, Orlando, FL, Mar. 25-30, 2012, pp. 2489–2497.

[10] J. Liu, C. H. Xia, N. B. Shroff, and H. D. Sherali, "Distributed cross-layer optimization in wireless networks: A second-order approach," in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 14-19, 2013.

[11] A. Jadbabaie, A. Ozdaglar, and M. Zargham, "A distributed Newton method for network optimization," in *Proc. IEEE Conference on Decision and Control (CDC)*, Shanghai, China, Dec. 16-18, 2009.

[12] A. Forsgren, P. E. Gill, and M. H. Wright, "Interior methods for nonlinear optimization," *SIAM Review*, vol. 44, no. 4, pp. 525–597, Oct. 2002.

[13] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time varying wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 89–103, Jan. 2005.

[14] A. Zymnis, N. Trichakis, S. Boyd, and D. ONeill, "An interior-point method for large scale network utility maximization," in *Proc. Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 26-28, 2007.

[15] D. Bickson, Y. Tock, O. Shental, and D. Dolev, "Polynomial linear programming with Gaussian belief propagation," in *Proc. Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 23-26, 2008, pp. 895–901.

[16] D. Bickson, Y. Tock, A. Zymnis, S. Boyd, and D. Dolev, "Distributed large scale network utility maximization," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Seoul, Korea, Jun.28–Jul.3, 2009, pp. 829–833.

[17] D. Bickson, "Gaussian belief propagation: Theory and application," Ph.D. dissertation, Hebrew University of Jerusalem, 2009.

[18] M. J. Neely, E. Modiano, and C. E. Rohrs, "Power allocation and routing in multibeam satellites with time-varying channels," *IEEE/ACM Trans. Netw.*, vol. 11, no. 2, pp. 138–152, Feb. 2003.

[19] M. S. Bazaraa, J. J. Jarvis, and H. D. Sherali, *Linear Programming and Network Flows*, 4th ed. New York: John Wiley & Sons Inc., 2010.

[20] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[21] Z. I. Woznicki, "Matrix splitting principles," *International Journal of Mathematics and Mathematical Sciences*, vol. 28, no. 5, pp. 251–284, May 2001.

[22] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York, NY: Cambridge University Press, 1990.

[23] F. P. Kelly, A. K. Malullo, and D. K. H. Tan, "Rate control in communications networks: Shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, pp. 237–252, 1998.

[24] W. Rudin, Ed., *Principles of Mathematical Analysis*. New York, NY: McGraw-Hill, 1976.