Achieving Low-Delay and Fast-Convergence in Stochastic Network Optimization: A Nesterovian Approach*

Jia Liu
Dept. of Electrical and Computer Engineering
The Ohio State University, Columbus, OH 43210, USA
liu@ece.osu.edu

ABSTRACT

Due to the rapid growth of mobile data demands, there have been significant interests in stochastic resource control and optimization for wireless networks. Although significant advances have been made in stochastic network optimization theory, to date, most of the existing approaches are plagued by either slow convergence or unsatisfactory delay performances. To address these challenges, in this paper, we develop a new stochastic network optimization framework inspired by the Nesterov accelerated gradient method. We show that our proposed Nesterovian approach offers utility-optimality, fast-convergence, and significant delay reduction in stochastic network optimization. Our contributions in this paper are three-fold: i) we propose a Nesterovian joint congestion control and routing/scheduling framework for both single-hop and multi-hop wireless networks; ii) we establish the utility optimality and queueing stability of the proposed Nesterovian method, and analytically characterize its delay reduction and convergence speed; and iii) we show that the proposed Nesterovian approach offers a three-way performance control between utility-optimality, delay, and convergence.

CCS Concepts

•Networks → Network resources allocation; Network control algorithms; Network performance modeling;

1. INTRODUCTION

Fueled by the massive amounts of mobile data and the rapid integration of new devices, recent years have witnessed an active research on stochastic resource control and optimization for wireless networks (see, e.g., [1–5], and [6] for a survey). The large body of work in this area has given rise to a beautiful queue-length-based control algorithmic framework (QCA), based on which congestion control, scheduling, and routing algorithms are naturally coupled by queueing states. Such algorithms do not need statistical knowledge of either the arrivals or the channel fading distributions. Rather, they only require instantaneous queue-lengths and channel state information (CSI) to make control decisions. Further, these algorithms can be interpreted by the Lagrangian dual decomposition framework plus the subgradient method in convex optimization [1, 2], where queue-lengths can be viewed as Lagrangian dual variables and the queue-length evolutions play the role of subgradient updates.

However, the QCA approaches suffer from several notable limitations. First, for the existing QCA approaches, it is well-known that an $O(1/K)$ utility-optimality gap is achieved at the expense of an $O(K)$ penalty in steady-state queue-length, where $K > 0$ is a system parameter. Hence, a small utility-optimality gap implies a large K and results in large queueing delay. To alleviate this queueing delay problem, there have been significant recent efforts (e.g., [4, 7–9], etc.) on improving the utility-optimality and delay tradeoff scaling law (see Section 2 for more in-depth discussions). Second, due to the subgradient nature of the queue-length-based weight adjustment, the QCA framework is oblivious to the curvature of the objective function [1–4]. The resultant "zigzagging" phenomenon [10] entails unsatisfactory convergence speed. To overcome this limitation, several second-order congestion control and routing/scheduling algorithms have recently been proposed to enhance the convergence speed (see, e.g., [11, 12]). However, due to the high complexity in Hessian inverse computation, these second-order designs require large information exchange overhead and may not work well for large-scale networks. The limitations of these existing works motivate us to pursue a new *Nesterovian* approach in this paper.

More specifically, in this work, our goal is to develop a low-complexity weight adjustment scheme based on the much simpler first-order Nesterov's accelerated gradient descent (AGD) method [13, 14] to reduce the queueing delay and increase the convergence speed of the QCA approaches, while without affecting their utility-optimality performance and without increasing their algorithmic complexity. Our fundamental rationale behind this approach is that the AGD method, first appeared in Nesterov's seminal work [13], is known for being an order-optimal first-order optimization method in terms of convergence rate [14]. Our key idea is to separate the weights and queue-lengths in the QCA framework, which then allows us to develop a weight updating scheme based on a Nesterovian scheme to accelerate the algorithm's convergence speed in the dual domain. Surpris-

*The work of Jia Liu has been supported in part by NSF grants CNS-1527078 and CNS-1446582.

SIGMETRICS '16, June 14-18, 2016, Antibes Juan-Les-Pins, France.

DOI: http://dx.doi.org/10.1145/2896377.2901474

ingly, it turns out that the theoretical findings of adopting a Nesterovian approach in network optimization are far *richer* than just convergence acceleration.

However, due to a number of technical challenges, developing a Nesterovian solution for stochastic network optimization is highly non-trivial. First, since the original Nesterov's AGD method was developed for unconstrained deterministic convex optimization, it is unclear how to modify the algorithm for wireless network utility maximization, which is not only constrained but also a stochastic optimization problem with a far more complex structure. Second, unlike the clear relationship between Lagrangian dual variables and queue-lengths in the QCA approaches, the relationship between the Nesterov's AGD method and the network states (e.g., queue-lengths, channel states, etc.) is unknown, which will cause difficulty in the trade-off analysis between delay and network utility. Third, due to the memory of past iterations' values, the structure of a Nesterovian approach is quite different from that of the QCA approaches. Hence, many existing analytical techniques used in QCA for utility optimality and delay tradeoff are not applicable.

The main contribution of this paper is that we develop a Nesterovian wireless network utility optimization framework that addresses the aforementioned technical challenges. This framework entails a series of new theoretical results on delay reduction and convergence speed, while maintaining utility-optimality. The main results and technical contributions of this paper are as follows:

- Motivated by the Nesterov's AGD idea, we propose a new weight adjustment scheme for joint congestion control and routing/scheduling in wireless networks. Our work not only establishes a connection between Nesterov's AGD method and the queue-length and channel state information to allow simple implementation in practice, it also extends the classical Nesterov's AGD method from unconstrained determinstic optimization to constrained stochastic network optimization.
- We establish the utility optimality and the queuing stability of the proposed Nesterovian approach. More precisely, under our Nesterovian congestion control and scheduling scheme with a β-parameterized memory term ($\beta \in [0,1)$ is a system parameter), we show that a utility-optimality gap $O(1/K)$ can be achieved with an $O((1-\beta)K)+O((1+\beta)\sqrt{K})$ cost in queueing-delay, where K is the same parameter as used in the traditional QCA framework. Moreover, in the asymptotic regime of K with β being chosen as $\beta = 1 - O(1/\sqrt{K})$, our Nesterovian approach achieves an $[O(1/K), O(\sqrt{K})]$ utility-delay trade-off, which is a much stronger result compared to the $[O(1/K), O(K)]$ trade-off scaling obtained by the QCA approaches.
- We investigate the choices of system parameters β and K, and their impacts on convergence. More specifically, we characterize the linear convergence rate factor of our Nesterovian approach and further show that it achieves superior convergence performance even compared with some state-of-the-art momentum-based schemes (e.g., [15]) in certain cases. Further, integrating with the results in the previous bullet, our Nesterovian approach offers a three-way performance control between utility, delay, and convergence speed. We offer insights on how to implement the proposed Nesterovian approach in a *distributed* fashion for multi-hop networks. We show that the distributed Nesterovian method only requires one-hop local message exchange, which is identical to the QCA schemes and hence does *not* incur any additional information exchange overhead in practical implementations.

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 our Nesterovian approach and the performance analysis of the proposed algorithm. In Section 5, we extend the proposed algorithm to multi-hop networks. Section 6 presents numerical results and Section 7 concludes this paper.

Notation: In this paper, we use boldface to denote matrices/vectors. We let $\mathbf{A}^\top$ be the transpose of $\mathbf{A}$. We let $\mathbf{I}_N$ and $\mathbf{O}_N$ denote the $N \times N$ identity and all-zero matrices, respectively. We let $\mathbf{1}_N$ and $\mathbf{0}_N$ denote the N-dimensional all-one and all-zero vectors, respectively. We will often omit "N" for brevity if the dimension is clear from the context. We use $\|\cdot\|$ and $\|\cdot\|_1$ to denote L^2- and L^1-norms, respectively. $\mathbf{A} \succeq 0$ (or $\mathbf{A} \preceq 0$) means that $\mathbf{A}$ is positive (resp., negative) semidefinite. $\mathbf{A} \succeq \mathbf{B}$ means $\mathbf{A} - \mathbf{B} \succeq 0$.

2. RELATED WORK

In this section, we first provide a synopsis of the state-of-the-art on delay reduction in the QCA literature that is most related to our work. Then, we provide a brief overview of the Nesterov's AGD method and its applications.

Delay Reduction for QCA: As mentioned in Section 1, there have been significant efforts on reducing the delay of the QCA approaches. For example, in [4, 16–19], virtual queues are used to reduce delay, where the virtual queue-lengths evolve based on service rates that are a fraction of the actual service rates. The key finding in these earlier efforts is that a slight sacrifice of throughput can lead to notable improvements in delay. More recent research on delay-reduction can be found in [7–9]. In [8], a virtual backlog mechanism with place-holder bits instead of real data was proposed. It was shown that, by accepting some non-zero packet dropping probability, this approach achieves an $[O(1/K), O(\log^2(K))]$ utility-delay trade-off. An exponential Lyapunov virtual backlog method combined with a threshold-based packeting-dropping scheme was proposed in [7] to achieve an even stronger $O(\log(K))$ delay. Although enjoying a log-type delay scaling, a common limitation of [7, 8] is that choosing the size of place-holder bits in [8] and the threshold value in [7] require *non-causal* global arrival and channel statistics (cf. [7, Eq. (17)], [8, Eq. (45)]), which is usually hard to obtain. If the parameters are not set appropriately, these schemes may result in non-negligible packet dropping probability. To address this problem, a per-iteration learning step was proposed in [9] to learn the optimal size of place-holder bits. However, the per-iteration learning component significantly increases the complexity. We note that, in some sense, all these delay reduction schemes can be viewed as sacrificing throughput-optimality (reflected in reduced service rates or packet dropping) for delay reduction, which is undesirable in practice. In contrast, *without sacrificing any throughput-optimality and without requiring any non-causal statistical knowledge*, our Nesterovian approach achieves an $O(\sqrt{K})$ delay scaling. We also note that although attempts to get rid of the back-pressure nature of the QCA framework have also been proposed in the liter-

ature (see, e.g., [20, 21]), convergence performance was not addressed in these works.

The Nesterov's AGD Method: A Primer: For this paper to be self-contained, here we provide an overview of the Nesterov's AGD method. Historically, the AGD method was proposed by Nesterov in 1983 for unconstrained convex optimization [13]. Specifically, consider a general unconstrained optimization problem $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$, where $f(\cdot)$ is twice continuously differentiable and convex. Let $\mathbf{H}(\mathbf{x})$ denote the Hessian matrix evaluated at $\mathbf{x}$. We assume that there exist two constants $0 \leq \phi \leq \Phi$ such that $\phi\mathbf{I} \preceq \mathbf{H}(\mathbf{x}) \preceq \Phi\mathbf{I}$, $\forall\mathbf{x}$ (thus Φ is a Lipschitz constant of the gradient of $f(\cdot)$). Then, the Nesterov's AGD method operates as follows [14, Chap. 2.2]:

- Choose an initial point $\mathbf{x}[0]$. Define an auxiliary variable $\mathbf{y}$ and let $\mathbf{y}[0] = \mathbf{x}[0]$. Choose $\alpha_0 \in (0, 1)$. Let $\kappa = \Phi/\phi$ be the condition number of $f(\cdot)$.
- In iteration $t \geq 0$: a) Compute $f'(\mathbf{x}[k])$ and let:

$$\mathbf{y}[t+1] = \mathbf{x}[t] - (1/\Phi)f'(\mathbf{x}[t]). \tag{1}$$

b) Compute $\alpha[t+1] \in (0, 1)$ by solving the quadratic equation $\alpha^2[t+1] + [\alpha^2[t] - (1/\kappa)]\alpha[t+1] - \alpha^2[t] = 0$. Let:

$$\beta[t] = \frac{\alpha[t](1-\alpha[t+1])}{(\alpha^2[t]+\alpha[t+1])}, \text{ and} \tag{2}$$

$$\mathbf{x}[t+1] = \mathbf{y}[t+1] + \beta[t](\mathbf{y}[t+1] - \mathbf{y}[t]). \tag{3}$$

It can be shown that the Nesterov's AGD method achieves: i) order-optimal $O(1/t^2)$ convergence rate for weakly convex problems ($\phi = 0$); and ii) an $O(1/\sqrt{\kappa})$ linear convergence factor for strongly convex problems (see [14] for details). In each iteration, the Nesterov's AGD method first performs a basic gradient descent step to move from $\mathbf{x}[t]$ to $\mathbf{y}[t+1]$ (cf. (1)), and then "slides" a little bit further (controlled by $\alpha[t]$ and $\beta[t]$) from $\mathbf{y}[t+1]$ in the direction with respect to $\mathbf{y}[t]$ (cf. (3)). Unfortunately, the intuition behind the Nesterov updates in (1) and (3) is difficult to grasp. Over the years, researchers have made numerous attempts to explain why the acceleration works, rather than settling on the mysterious (yet beautiful) algebraic manipulations in the original proof. Recent efforts in this area include, e.g., viewing AGD as a linear coupling of gradient descent and mirror descent [22], interpretation through discretization of certain second-order ordinary differential equation (ODE) in physics [23], and a geometric explanation inspired by the ellipsoid method [24]. Although interpreting the Nesterov's AGD method is beyond the scope of this paper, we hope that our Nesterovian network optimization approach could also provide some insights from a networking angle.

We note that the Nesterov's AGD method exploits past memory (cf. (3)), which is similar to another family of first-order methods termed "multi-step methods" (or called "heavy ball") that also leverage memory for convergence acceleration [25–28]. However, the key difference is that the Nesterov's method exploits both past iterates and gradients, while the multi-step methods only use past iterates. Also, the Nesterov's method is convergence-rate order-optimal for general convex problems, while multi-step methods only work for strongly convex problems ($\phi > 0$). We also note that, since its inception, the Nesterov's method has found applications in signal processing (e.g., [29] and references therein). However, to our knowledge, the Nesterov's AGD idea remains unexplored in network system optimization.

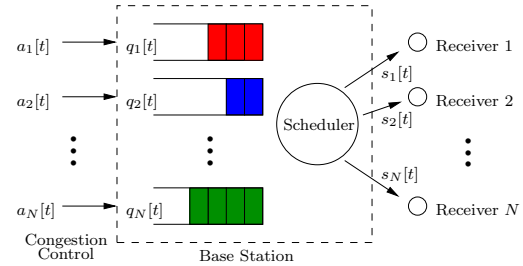

Figure 1: An illustration of the single-hop cellular downlink.

3. NETWORK MODEL AND PROBLEM FORMULATION

From this section to Section 4, we focus on a single-hop model with N links, which could model a cellular downlink/uplink channel with N users, or a set of communicating pairs in an ad hoc network. We will discuss in Section 5 how to extend the results to multi-hop networks.

Network model: In the single-hop case, we use a cellular downlink system as shown in Figure 1 to facilitate discussions. We assume a time-slotted system with time being indexed by $t = 0, 1, 2, \ldots$. Suppose that the channel fading is characterized by a total of M states and denoted by vectors $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_M$, where each $\boldsymbol{\pi}_m \in \mathbb{R}^N$, $m = 1, \ldots, M$, corresponds to the N links' channel states under state m. Let $\mathcal{C}_{\boldsymbol{\pi}_m}$ denote the achievable rate region for $\boldsymbol{\pi}_m$, which is defined as $\mathcal{C}_{\boldsymbol{\pi}_m} \triangleq Conv\{x_1^{(m)}, \ldots, x_N^{(m)}\}$, where $Conv\{\cdot\}$ represents convex hull and $x_n^{(m)}$ denotes a feasible rate of link n that can be scheduled under channel state m. We assume that, for each link n and channel state m, $x_n^{(m)} \leq s^{\max} < \infty$. We use a vector $\mathbf{x}^{(m)} = [x_1^{(m)}, \ldots, x_N^{(m)}]^\top \in \mathbb{R}^N$ to denote the feasible rates of all receivers under channel state m. We assume that the channel states are independent and identically distributed across time-slots[1]. Let $\boldsymbol{\pi}[t]$ denote the channel state vector in time-slot t and let $p_m \triangleq \Pr\{\boldsymbol{\pi}[t] = \boldsymbol{\pi}_m\}$ be the stationary distribution of the channel state being in state m. We let $\bar{\mathcal{C}}$ denote the mean achievable rate region, which can be computed as:

$$\bar{\mathcal{C}} \triangleq \left\{ \mathbf{x} \,\middle|\, \mathbf{x} = \sum_{m=1}^{M} p_m \mathbf{x}^{(m)},\ \forall \mathbf{x}^{(m)} \in \mathcal{C}_{\boldsymbol{\pi}_m} \right\}. \tag{4}$$

We note that, in this paper, neither the channel state statistics nor $\bar{\mathcal{C}}$ is assumed to be known at the base station.

Queue-stability: As shown in Figure 1, each link n is associated with a queue. We denote the queue-length in time-slot t as $q_n[t]$. In every time-slot t, the controller observes the current channel state $\boldsymbol{\pi}[t]$ and then chooses a service rate vector $\mathbf{s}[t] \triangleq [s_1[t], \ldots, s_N[t]]^\top \in \mathcal{C}_{\boldsymbol{\pi}[t]}$ and a congestion control rate vector $\mathbf{a}[t] \triangleq [a_1[t], \ldots, a_N[t]]^\top \in \mathbb{R}_+^N$. Clearly, the queue-length process $\{q_n[t]\}$ evolves as:

$$q_n[t+1] = \{q_n[t] - s_n[t] + a_n[t]\}^+, \quad \forall n, \tag{5}$$

where $\{\cdot\}^+ \triangleq \max\{0, \cdot\}$. We let $\mathbf{q}[t] \triangleq [q_1[t], \ldots, q_N[t]]^\top$ denote the queue-length vector in time-slot t. Same as in [2, 3], in this paper, we say that a network is *stable* if the

[1]Following the same arguments as in [8, 30], our results can be generalized to Markovian channel state processes.

steady-state total queue-length is finite, i.e.,

$$\limsup_{t\to\infty} \mathbb{E}\left\{\|\mathbf{q}[t]\|_1\right\} < \infty. \tag{6}$$

Problem formulation: Let $\bar{a}_n \triangleq \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} a_n[t]$ denote the average arrival rate of link n under congestion control. Each receiver n is associated with a utility function $U_n(\bar{a}_n)$, which represents the utility gained by receiver n when its data is injected at rate $\bar{a}_n$. We assume that $U_n(\cdot)$, $\forall n$, is concave, monotonically increasing, and twice continuously differentiable, and strongly concave, i.e., there exist constants $0 < \phi \leq \Phi < \infty$ such that

$$\phi \leq -U_n''(a_n) \leq \Phi, \quad \forall a_n \in [0, a^{\max}], \tag{7}$$

where $a^{\max}$ is an arrival rate upper bound for burst control. As an example, the function $\log(\epsilon + a_n)$ with some constant $\epsilon > 0$ satisfies (7). Our goal is to maximize $\sum_{n=1}^{N} U_n(\bar{a}_n)$, subject to achievable rate region $\mathcal{C}_{\boldsymbol{\pi}[t]}$ in each time-slot and the queue-stability constraint. Putting together the models presented above, we have the following joint congestion control and scheduling (CCS) optimization problem:

$$\begin{aligned}
\textbf{CCS: Maximize} \quad & \sum_{n=1}^{N} U_n(\bar{a}_n) \\
\text{subject to} \quad & \text{Queue-length stability constraint in (6),} \\
& s_n[t] \in \mathcal{C}_{\boldsymbol{\pi}[t]},\ a_n[t] \in [0, a^{\max}]\ \forall n, t.
\end{aligned}$$

4. A NESTEROVIAN APPROACH FOR STOCHASTIC NETWORK OPTIMIZATION

In this section, we first introduce the Nesterovian algorithm in Section 4.1. In Section 4.2, we will present the main theoretical results. Then, in Section 4.3, we will discuss the key insights and intuition of the theoretical results. Section 4.4 provides the proofs for the main theorems.

4.1 The Nesterovian Algorithm

Our proposed Nesterovian joint congestion control and scheduling algorithm is described in Algorithm 1:

Algorithm 1: A Nesterovian Approach for Joint Congestion Control and Scheduling.

Initialization:

1. Choose parameters $K > 0$ and $\beta \in [0, 1)$. Set $t = 0$.
2. Let queue-states $q_n[0] = 0$ and $\Delta q_n[-1] = 0$, $\forall n$.
3. Associate each link n with a non-negative weight w_n and set the initial weights $w_n[0] = w_n[-1] = 0$, $\forall n$.

Main Loop:

4. *MaxWeight Scheduler:* In time-slot $t \geq 0$, observe the current weight vector $\mathbf{w}[t] \triangleq [w_1[t], \ldots, w_N[t]]^\top$ and the current channel state $\boldsymbol{\pi}[t]$. Then, the scheduler chooses a service rate vector $\mathbf{s}[t]$ as follows:

$$\mathbf{s}[t] = \arg\max_{\mathbf{x}\in\mathcal{C}_{\boldsymbol{\pi}[t]}} (\mathbf{w}[t])^\top \mathbf{x}. \tag{8}$$

5. *Congestion Controller:* For each link n, given its current weight $w_n[t]$, the data injection rate $a_n[t]$ is an integer-valued random variable that satisfies:

$$\mathbb{E}\{a_n[t]|w_n[t]\} = \min\left\{U_n'^{-1}\left(\frac{w_n[t]}{K}\right), a^{\max}\right\}, \tag{9}$$

$$\mathbb{E}\{a_n^2[t]|w_n[t]\} \leq A < \infty, \quad \forall w_n[t], \tag{10}$$

where $U_n'^{-1}(\cdot)$ represents the inverse function of the first-order derivative of $U_n(\cdot)$. In (9), $a^{\max}$ is a positive constant satisfying $a^{\max} > 2s^{\max}$. In (10), the second moment bound A will be used in establishing subsequent theoretical results.

6. *Queue-Length and Nesterovian Weight Updates:* Update the queue-lengths following (5). Let $\Delta q_n[t] \triangleq q_n[t+1] - q_n[t]$, $\forall n$, be the resultant queue-length changes. Next, for all links, update the weights in the following (projected) **Nesterovian** manner:

$$\begin{aligned}
w_n[t+1] = \{&w_n[t] + \Delta q_n[t] + \beta[(w_n[t] + \Delta q_n[t]) - \\
&(w_n[t-1] + \Delta q_n[t-1])]\}^+,\ \forall n.
\end{aligned} \tag{11}$$

Let $t = t + 1$. Go to Step 4 and repeat the scheduling and congestion control processes.

Some important remarks on Algorithm 1 are in order:

1) Relation to QCA: We can see that the congestion control and scheduling components in Algorithm 1 are similar to those in the QCA schemes (see, e.g., [2, 3, 30]). However, in both scheduling and congestion control components, the weights in (8) and (9) are *not* based on current queue-lengths (or a direct function of current queue-lengths). We will see later that this *separation* of weights and queue-lengths leads to significant delay reductions. Note also that when $\beta = 0$, our Nesterovian algorithm reduces to the traditional QCA approach. Hence, the QCA approach can be viewed as a special case of our Nesterovian algorithm.

2) Nesterovian weight update: The weight update idea in (11) is motivated by the Nesterov updates in (1) and (3). To see this, one only needs to do the following: i) Let $\beta[t] \equiv \beta$, $\forall t$; ii) Let $\mathbf{x}[t] = \mathbf{w}[t]$ and $f'(\mathbf{x}[t]) = \Delta\mathbf{q}[t]$ in (1) (ignoring the scaling factor $1/\Phi$); and lastly iii) Substitute (1) into (3) and apply the non-negative projection. We note that changing $\beta[t]$ to a constant step-size β not only simplifies the computation to allow easy implementations in practice, it also leads to an elegant three-way performance control relationship between utility-optimality, delay, and convergence speed, which will be presented later.

3) First-order memory: One can see that the weight update in (11) integrates a β-parameterized first-order memory of the weights $w_n[t-1]$ and queue-length changes $\Delta q_n[t-1]$ from the previous time-slot. By contrast, the weight updates in traditional QCA approaches are of zero-order memory in the sense that queue-lengths only inherit the absolute weight values from the current time-slot. As will be seen shortly, this algorithmic structural difference necessitates new proof techniques in establishing our theoretical results.

4) Zero-valued initial states: We note that, in Algorithm 1, the zero-valued initial states of the w_n- and q_n-variables are necessary for the delay reduction and queue-length scaling results to be established in Theorem 1. If the initial queueing buffers are non-empty, these zero-valued initial states can still be met by turning off the injection rates $a_n[t]$ and let the scheduler drain all the existing packets. As long as the number of existing packets is finite, this evacuation period must also be finite. Hence, this finite period of injection rate shutdown will not affect the utility-optimality of the average injection rates over an infinite time horizon.

4.2 Main Theoretical Results

Our first main result is on the queue-length reduction performance of the proposed Nesterovian algorithm:

THEOREM 1 (QUEUE-LENGTH REDUCTION). *Given $\beta \in [0,1)$, the scaling of the steady-state total queue-length:*

$$\limsup_{t\to\infty} \mathbb{E}\{\|\mathbf{q}[t]\|_1\} = O\big((1-\beta)K\big) + O\big((1+\beta)\sqrt{K}\big). \quad (12)$$

Further, if $\beta \uparrow 1$ at a speed faster than $\beta = 1 - O(1/\sqrt{K})$, then Eq. (12) implies that $\limsup_{t\to\infty} \mathbb{E}\{\|\mathbf{q}[t]\|_1\} = O(\sqrt{K})$.

Several important remarks on Theorem 1 are in order: i) If β is fixed and $K \to \infty$, the second term on the right-hand-side of (12) is dominated by the first term. Hence, $\limsup_{t\to\infty} \mathbb{E}\{\|\mathbf{q}[t]\|_1\} \approx O\big((1-\beta)K\big)$. Note that in the traditional K-parameterized QCA algorithm (see, e.g., [3, 4]), the total queue-length grows as $O(K)$. Hence, Theorem 1 shows that the steady-state total queue-length under a β-parameterized Nesterovian scheme is approximately $(1-\beta)$–fraction of that of the traditional QCA approaches.

ii) Rather than fixing β, if we allow β to vary in relation to K, then Theorem 1 implies that if β approaches 1 fast enough as $K \to \infty$, the total queue-length scales as $O(\sqrt{K})$. This significantly outperforms the $O(K)$ delay of the QCA algorithms, yet *without sacrificing any throughput and without requiring non-causal global statistics as in [7, 8].*

iii) Incorporating the weight and queue-length changes from the previous time-slot in (11) can be loosely viewed as a way to memorize the queue states' and weights' evolution in history. Remarkably, Theorem 1 shows that simply learning from "immediate past" makes a big difference in steady-state queue-length reduction.

Next, we let $U(\mathbf{a}) = \sum_{n=1}^{N} U_n(a_n)$ denote the utility sum of Problem CCS and let $\mathbf{a}^*$ be the optimal solution. Also, we let $a_n^\infty \triangleq \mathbb{E}\{\min\{U_n^{'-1}(w_n^\infty/K), a^{\max}\}\}$, $\forall n$, be the mean steady-state congestion control rates achieved by our algorithm (the existence of steady-state will be proved later in Section 4.4). Further, we let $\mathbf{a}^\infty \triangleq [a_1^\infty, \ldots, a_N^\infty]^\top$. Then, the second key result in this paper can be stated as follows:

THEOREM 2 (UTILITY-OPTIMALITY). *The mean of the stationary rate vector $\mathbf{a}^\infty$ under Algorithm 1 tsatisfies $\|\mathbf{a}^\infty - \mathbf{a}^*\| = O(1/\sqrt{K})$. Meanwhile, the acheived utility objective value can be bounded as $U(\mathbf{a}^*) - U\big(\mathbf{a}^\infty\big) = O(1/K)$. These imply that $\mathbf{a}^\infty \to \mathbf{a}^*$ asymptotically as K increases.*

Theorem 2 says that our Nesterovian approach is utility-optimal and the optimality is *independent* of β. In other words, the new addition of the β-parameterized memory term in (11) does *not* affect the utility-optimality of the traditional QCA methods, which exactly achieves our goal.

The third main result in this work is on the convergence speed performance. In this paper, the convergence speed metric is defined in terms of the number of time-slots required by the sequence $\{\mathbb{E}\{\mathbf{a}[t]|\mathbf{w}[t]\}\}$ to reach the $O(1/\sqrt{K})$-neighborhood of $\mathbf{a}^*$ as stated in Theorem 2.

THEOREM 3 (CONVERGENCE RATE). *Let $K \in (\Phi, \infty)$ and $\beta \in \big[0,1\big)$. Let $\mathbf{H}^* \in \mathbb{R}^{N\times N}$ denote the Hessian matrix of the objective function of Problem CCS evaluated at $\mathbf{a}^*$. Then, $\{\mathbb{E}\{\mathbf{a}[t]|\mathbf{w}[t]\}\}$ converges linearly[2] with a factor $R_{(K,\beta)}$ satisfying:*

$$R_{(K,\beta)} \le \frac{1}{2}\max_{\lambda_i^*,\forall i}\Bigg\{\Bigg|(1+\beta)\Big(1-\frac{\lambda_i^*}{K}\Big) \pm \sqrt{(1+\beta)^2\Big(1-\frac{\lambda_i^*}{K}\Big)^2 - 4\beta\Big(1-\frac{\lambda_i^*}{K}\Big)}\Bigg|\Bigg\} < 1, \quad (13)$$

where λ_i^*, $i = 1, \ldots, N$, denote the eigenvalues of $\mathbf{H}^*$. Further, minimizing $R_{(K,\beta)}$ over K and β yields the smallest convergence factor bound: $R^* = (\sqrt{\Phi} - \sqrt{\phi})/\sqrt{\Phi}$, which is achieved by letting $K \to \Phi$ and $\beta^* = (\sqrt{\Phi} - \sqrt{\phi})/(\sqrt{\Phi} + \sqrt{\phi})$.

Theorem 3 implies that K and β can be optimized to achieve $R^* = (\sqrt{\kappa} - 1)/\sqrt{\kappa}$, where $\kappa \triangleq \Phi/\phi$ is the condition number of the Hessian $\mathbf{H}$. Note that R^* is always *smaller* than that of the QCA approaches, where $R_{\text{QCA}} = (\kappa - 1)/(\kappa + 1)$ (cf. e.g., [2]). This implies that our Nesterovian approach converges faster than QCA, particularly when κ is large (i.e., when the problem is ill-conditioned).

Further, based on Theorem 3 and by comparing the convergence factor bounds, it can be shown that when $\beta > \sqrt{1 - \phi/K}$, the Nesterovian approach converges faster than our previous heavy-ball approach in [15]. This shows that the proposed Nesterovian approach further improves the convergence speed of the heavy-ball approach in cases where β is chosen close to 1 (implying low-delay).

The proofs of Theorems 1–3 can be found in Section 4.4. In what follows, we will further discuss some insights of the theoretical results.

4.3 Discussions

1) Intuition behind delay reduction: From the connection between the QCA approaches and the Lagrangian dual decomposition (see, e.g., [2–4,31]), the QCA algorithms can be interpreted as using queue-lengths as the dual variables to solve Problem CCS. One drawback of this approach is that a large amount of packets need to be kept in each queue *only* to maintain the "right amount of pressure," which we denote as $w_n^*(K)$, $\forall n$ here. As will be seen in Section 4.4, $w_n^*(K)$ scales as $O(K)$ as K increases to approach optimal network utility. However, this "queue-lengths as dual" approach is *not* necessary since the $w_n^*(K)$ is merely a mathematical construct and needs not be associated with queue-lengths. In theory, one has the freedom to choose any quantity to play the role of the dual variables.

Indeed, there are several existing works based on the above intuition to reduce delay. To our knowledge, the idea that is most related to ours is [8], where the authors proposed to use $\mathcal{W}_n \in \big[0, w_n^*(K)\big)$ amount of "place-holder bits" in each queueing buffer n and use the weight $w_n[t] = q_n[t] + \mathcal{W}_n$ to conduct scheduling, and the weights evolve as:

$$w_n[t+1] = [w_n[t] + \Delta q_n[t]]^+ = [q_n[t] + \Delta q_n[t] + \mathcal{W}_n]^+, \forall n. \quad (14)$$

However, as mentioned in Section 2, determining an appropriate value for each $\mathcal{W}_n$ is a non-trivial task since doing so requires non-causal global knowledge $w_n^*(K)$, $\forall n$, which is unavailable at the initial state. Now, consider the Nesterovian weight update in (11), which is re-stated below:

$$w_n[t+1] = \{w_n[t] + \Delta q_n[t] + \beta[(w_n[t] + \Delta q_n[t]) - (w_n[t-1] + \Delta q_n[t-1])]\}^+, \ \forall n. \quad (15)$$

[2]We say that $\{x_k\}_{k=1}^\infty$ converges linearly to x^* if there exists a factor $R \in (0,1)$ such that $\|x_{k+1} - x^*\| \le R\|x_k - x^*\|$, $\forall k$.

Table 1: Three-way performance control.

	Type I	Type II	Type III
Utility Optimality Gap	Small	Small	Large
Queueing Delay	Low	High	Low
Convergence Speed	Slow	Fast	Fast
K	Large	Large	Optimized for convergence
β	Close to 1	Optimized for convergence	Close to 1

Comparing (14) and (15), we can see that the memory term $\beta[(w_n[t]+\Delta q_n[t])-(w_n[t-1]+\Delta q_n[t-1])]$ plays a similar role as the place-holder bits $\mathcal{W}_n$ in the sense that it effectively reduces the size of required $\Delta q_n[t]$ to maintain the pressure level. Moreover, the use of memory term has the following advantages: i) Unlike the artificial notion of place-holder bits whose initial value is difficult to set, the memory term only requires two time-slots of weight and queue-length update history, which not only have a real physical meaning but also render easy implementations; ii) The memory term can also be viewed as a simple way to implicitly learn and adapt to the unknown $\mathbf{w}_n^*(K)$, thus eliminating the need for an explicit per-iteration learning step as in [9]; and (iii) Unlike the possibility of using an overly aggressive $\mathcal{W}_n$-value that results in packet dropping, our Nesterovian weight update scheme evolves gracefully and does *not* incur packet dropping and thus retaining full throughput-optimality.

2) Thee-way performance trade-offs: Collectively, Theorems 1–3 imply a three-way control between utility-optimality, delay, and convergence. By appropriately choosing K and β, one can simultaneously improve *two* out of the three performance metrics by trading-off the third. We summarize the three-way performance control in Table 1.

As shown in Table 1, Control Type I corresponds to achieving utility-optimality and low-delay at the cost of slower convergence, by setting K large and β close to 1. To see this, we note from Theorem 2 that a large K implies small utility-optimality gap $O(1/K)$. Also, by setting β close to 1, Theorem 1 indicates that a big $(1-\beta)$–fraction delay reduction. However, as $K\to\infty$ and $\beta\to 1$, it can be verified from Theorem 3 that $\lim_{\beta\to 1, K\to\infty} R_{(K,\beta)}\to 1$, which implies an increasingly slower convergence. Likewise, Control Type II corresponds to achieving utility-optimality and fast-convergence at the cost of less delay reduction, by setting K large and optimizing β for convergence. To see this, from Theorem 2, we have that a large K implies a small $O(1/K)$ utility-optimality gap. Also, by Theorem 3, β can be optimized under a given K to minimize $R_{(K,\beta)}$ to increase the convergence. However, the obtained β may or may not be close to 1. Thus, the delay performance gain may not be dramatic. Control Type III can be verified similarly as Trade-off Type II, so we omit the details for brevity.

4.4 Proofs of the Main Theorems

In this subsection, we provide the proofs for the main theorems stated in Section 4.2.

Proof of Theorem 1. Since the proof of Theorem 1 is lengthy, we structure the proof of into several key steps.

Step 1): A K-Parameterized Deterministic Problem: Consider a deterministic problem where the channel state process is not random but fixed at its mean level, i.e., $\mathcal{C}_{\boldsymbol{\pi}[t]}=\bar{\mathcal{C}}$, $\forall t$. The congestion control and scheduling variables are not time-varying and denoted as a_n and s_n, $\forall n$. The K-parameterized deterministic problem can be written as:

$$K\textbf{-DCCS: Maximize}\quad K\sum\nolimits_{n=1}^{N} U_n(a_n)$$
$$\text{subject to}\quad a_n-s_n\le 0,\ \forall n,\ s_n\in\bar{\mathcal{C}},\ a_n\in[0,a^{\max}],\ \forall n.$$

We associate dual variables $w_n\ge 0$, $\forall n$ with the constraints $a_n-s_n\le 0$, $\forall n$ to obtain the Lagrangian as follows:

$$\Theta_K(\mathbf{w})\triangleq\max_{\mathbf{a},\mathbf{s}}\left\{K\sum_{n=1}^{N}U_n(a_n)+\sum_{n=1}^{N}w_n(s_n-a_n)\right\},\quad(16)$$

where the vector $\mathbf{w}\triangleq[w_1,\ldots,w_N]^\top\in\mathbb{R}_+^N$ contains all dual variables. Then, the Lagrangian dual problem of Problem K-DCCS can be written as:

$$\textbf{D-}K\textbf{-DCCS: Minimize}\quad\Theta_K(\mathbf{w})$$
$$\text{subject to}\ \mathbf{w}\in\mathbb{R}_+^N.$$

It can be verified that Problem K-DCCS is convex and satisfies the Slater condition [10]. Hence, the optimal objective value of Problem K-LD-CCS is the same as that of Problem D-K-DCCS, i.e., strong duality holds. Let $\mathbf{w}^*(K)$ be the optimal dual solution to Problem D-K-DCCS. Due to the strict convexity of Problem K-LD-CCS, $\mathbf{w}^*(K)$ is unique. Further, we have the following result of $\mathbf{w}^*(K)$:

Lemma 1 (Linear scaling of $\mathbf{w}^*(K)$). *For a given K, $\mathbf{w}^*(K)=K\mathbf{w}^*(1)$, or equivalently, $\mathbf{w}^*(K)=O(K)$.*

Proof. Dividing K on both sides of (16), we have

$$\frac{1}{K}\Theta_K(\mathbf{w})=\max_{\mathbf{a},\mathbf{s}}\left\{\sum_{n=1}^{N}U_n(a_n)+\sum_{n=1}^{N}\widehat{w}_n(s_n-a_n)\right\},\quad(17)$$

where $\widehat{w}_n=w_w/K$. Note that the right hand side (RHS) of (17) is precisely $\Theta_1(\mathbf{w})$, for which the maximizer is $\widehat{\mathbf{w}}=\mathbf{w}^*(1)$. Hence, we have $\Theta_K(\mathbf{w})$ is maximized at $K\mathbf{w}^*(1)$. □

Step 2): A Special-Structured Block Matrix: Now, we define a β-parameterized block matrix $\boldsymbol{\Gamma}(\beta)$ as follows:

$$\boldsymbol{\Gamma}(\beta)\triangleq\begin{bmatrix}(1+\beta)\mathbf{I}_N & -\beta\mathbf{I}_N\\ \mathbf{I}_N & \mathbf{0}_N\end{bmatrix}\in\mathbb{R}^{2N\times 2N},\quad(18)$$

where $0<\beta<1$. Next, we prove a lemma about the eigenvalues of $\boldsymbol{\Gamma}(\beta)$ that will be useful in proving Theorem 1.

Lemma 2 (Eigen-spectrum of $\boldsymbol{\Gamma}(\beta)$). *$\boldsymbol{\Gamma}(\beta)$ only has two distinct eigenvalues β and 1, and both eigenvalues are of algebraic multiplicity N. Hence, $\boldsymbol{\Gamma}(\beta)$ is a non-expansive linear transformation in $\mathbb{R}^{2N}$.*

Proof. Let λ denote an eigenvalue of $\boldsymbol{\Gamma}(\beta)$ and consider the characteristic equation $\det(\boldsymbol{\Gamma}(\beta)-\lambda\mathbf{I}_{2N})=0$, which can be written in block-wise fashion as:

$$\det\begin{bmatrix}(1+\beta-\lambda)\mathbf{I}_N & -\beta\mathbf{I}_N\\ \mathbf{I}_N & -\lambda\mathbf{I}_N\end{bmatrix}=0.\quad(19)$$

We claim that $\lambda\ne 1+\beta$ and thus the block $(1+\beta-\lambda)\mathbf{I}_N$ in (19) is invertible. To verify this, suppose on the contrary that $\lambda=1+\beta$ and (19) holds. In this case, we have:

$$\det\begin{bmatrix}\mathbf{0}_N & -\beta\mathbf{I}_N\\ \mathbf{I}_N & -(1+\beta)\mathbf{I}_N\end{bmatrix}=\det(\beta\mathbf{I}_N)=-\beta^N\ne 0,$$

contradicting to the assumption that (19) holds. Given that the block $(1+\beta-\lambda)\mathbf{I}_N$ is invertible, it follows from the Schur complements determinantal formulae [32] that

$$\begin{aligned}(19) =& \det\big[(1+\beta-\lambda)\mathbf{I}_N\big]\times \\ & \det\big[(-\lambda\mathbf{I}_N)-\mathbf{I}_N((1+\beta-\lambda)\mathbf{I}_N)^{-1}(-\beta\mathbf{I}_N)\big] \\ =&(1+\beta-\lambda)^N\det\Big[\Big(-\lambda+\frac{\beta}{1+\beta-\lambda}\Big)\mathbf{I}_N\Big] \\ =&[\lambda^2-(1+\beta)\lambda+\beta]^N=(\lambda-1)^N(\lambda-\beta)^N=0. \quad (20)\end{aligned}$$

Hence, the result stated in the lemma follows. □

The spectral result in Lemma 2 implies that $\mathbf{\Gamma}(\beta)$ is a *non-expansive* linear operator. This result will play a key role in our subsequent analysis of the proposed Nesterov's method.

Step 3): Mean Weight Deviation: Our next key step toward proving Theorem 1 is to establish the following mean deviation result of the weights:

THEOREM 4 (MEAN WEIGHT DEVIATION BOUND). *For a given K, there exists a constant C that depends on Φ, $s^{\max}$, and $a^{\max}$, such that $\mathbb{E}\{\|\mathbf{w}^\infty-\mathbf{w}^*(K)\|\}\le C\sqrt{K}$, where $\mathbf{w}^\infty$ denotes the weights $\mathbf{w}[t]$ in steady-state.*

PROOF. We start by rewriting the Nesterovian weight update in (11) in the following equivalent vector form:

$$\begin{aligned}\mathbf{w}[t+1] =&\mathbf{w}[t]+\Delta\mathbf{q}[t]+\beta[(\mathbf{w}[t]+\Delta\mathbf{q}[t]) \\ &-(\mathbf{w}[t-1]+\Delta\mathbf{q}[t-1])]+\mathbf{u}[t], \quad (21)\end{aligned}$$

where $\mathbf{u}[t]$ is the projection term representing unused services: $\mathbf{u}[t]\triangleq\big\{\beta(\mathbf{w}[t-1]+\Delta\mathbf{q}[t-1])-(1+\beta)(\mathbf{w}[t]+\Delta\mathbf{q}[t])\big\}^+$. Note that the memory term in (21) depends on two consecutive time-slots of memory (i.e., $\mathbf{w}[t]$, $\mathbf{w}[t-1]$, $\Delta\mathbf{q}[t]$, and $\Delta\mathbf{q}[t-1]$), which is challenging in subsequent analysis. To overcome this challenge, we define a $2N$-dimensional vector $\mathbf{z}[t]$ as follows (we simplify the notation of $\mathbf{w}^*(K)$ to $\mathbf{w}^*$):

$$\mathbf{z}[t]\triangleq\begin{bmatrix}\mathbf{w}[t]-\mathbf{w}^* \\ \mathbf{w}[t-1]-\mathbf{w}^*\end{bmatrix}. \quad (22)$$

Then, it can be readily verified that (21) can be rewritten in terms of $\mathbf{z}[t]$ as follows:

$$\mathbf{z}[t+1]=\mathbf{\Gamma}(\beta)\mathbf{z}[t]+\mathbf{\Gamma}'(\beta)\begin{bmatrix}\Delta\mathbf{q}[t] \\ \Delta\mathbf{q}[t-1]\end{bmatrix}+\begin{bmatrix}\mathbf{u}[t] \\ \mathbf{0}_N\end{bmatrix}, \quad (23)$$

where $\mathbf{\Gamma}'(\beta)\triangleq\begin{bmatrix}(1+\beta)\mathbf{I} & -\beta\mathbf{I} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}$. We define a quadratic Lyapunov function $V(\mathbf{z}[t])\triangleq\frac{1}{2}\|\mathbf{z}[t]\|^2$ and consider its conditional expectation of one-slot drift:

$$\mathbb{E}\big\{\Delta V(\mathbf{z}[t])\big|\mathbf{z}[t]\big\}\triangleq\frac{1}{2}\mathbb{E}\big\{\|\mathbf{z}[t+1]\|^2-\|\mathbf{z}[t]\|^2\big|\mathbf{z}[t]\big\}. \quad (24)$$

Let $\mathbb{1}_{\mathcal{A}}(\mathbf{x})$ be the indicator function that takes value 1 if $\mathbf{x}\in\mathcal{A}$ and 0 otherwise. After some algebraic derivations and upper-bounding (see Appendix A for details), we have:

PROPOSITION 1. *Let $B\triangleq N[A+(s^{\max})^2]$. There exist constants $\delta,\eta>0$ such that*

$$\begin{aligned}&\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]=\mathbf{z}\}\le-\frac{\delta}{\sqrt{K}}\big(\big\|\mathbf{w}[t]-\mathbf{w}^*\big\|\mathbb{1}_{\mathcal{B}^c}(\mathbf{w}[t])+ \\ &\big\|\mathbf{w}[t-1]-\mathbf{w}^*\big\|\mathbb{1}_{\mathcal{B}^c}(\mathbf{w}[t-1])\big)+\eta\big(\mathbb{1}_{\mathcal{B}}(\mathbf{w}[t])+\mathbb{1}_{\mathcal{B}}(\mathbf{w}[t-1])\big),\end{aligned}$$

where $\mathcal{B}\triangleq\{\mathbf{w}:\big\|\mathbf{w}-\mathbf{w}^*\big\|\le\sqrt{(1/2)B\Phi K}\}$, and $\mathcal{B}^c$ denotes the complement of $\mathcal{B}$.

Note that $\{\mathbf{z}[t]\}$ is a continuous Markov chain in $\mathbb{R}^{2N}$ and Proposition 1 ensures the Foster-Lyapunov criterion for positive Harris-recurrence. Thus, a steady-state exists [33]. Now, we consider the T-step conditional mean Lyapunov drift. For notational simplicity, we define the following sets:

$$\begin{aligned}\Omega_1&\triangleq\big\{\mathbf{z}\in\mathbb{R}^{2N}:\|\mathbf{w}[t]-\mathbf{w}^*\|\in\mathcal{B},\|\mathbf{w}[t-1]-\mathbf{w}^*\|\in\mathcal{B}\big\}, \\ \Omega_2&\triangleq\big\{\mathbf{z}\in\mathbb{R}^{2N}:\|\mathbf{w}[t]-\mathbf{w}^*\|\in\mathcal{B},\|\mathbf{w}[t-1]-\mathbf{w}^*\|\notin\mathcal{B}\big\}, \\ \Omega_3&\triangleq\big\{\mathbf{z}\in\mathbb{R}^{2N}:\|\mathbf{w}[t]-\mathbf{w}^*\|\notin\mathcal{B},\|\mathbf{w}[t-1]-\mathbf{w}^*\|\in\mathcal{B}\big\}, \\ \Omega_4&\triangleq\big\{\mathbf{z}\in\mathbb{R}^{2N}:\|\mathbf{w}[t]-\mathbf{w}^*\|\notin\mathcal{B},\|\mathbf{w}[t-1]-\mathbf{w}^*\|\notin\mathcal{B}\big\}.\end{aligned}$$

By telescoping (24) from $t=0$ to $T-1$, we have that

$$\begin{aligned}&\mathbb{E}\{V(\mathbf{z}[T])|\mathbf{z}[0]\}-V(\mathbf{z}[0])\overset{(a)}{=}\sum_{t=0}^{T-1}\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[0]\} \\ &=\sum_{t=0}^{T-1}\int_{\mathbb{R}^{2N}}p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]=\mathbf{z}\}d\mathbf{z} \\ &=\sum_{i=1}^{4}\sum_{t=0}^{T-1}\int_{\Omega_i}p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]=\mathbf{z}\}d\mathbf{z}, \quad (25)\end{aligned}$$

where (a) follows from the fact that $\mathbf{z}[t]$ is a continuous state Markov chain in $\mathbb{R}^{2N}$. It then follows from Proposition 1 and the definitions of Ω_i, $i=1,\ldots,4$, that each term in (25) can be respectively upper-bounded as:

$$\sum_{t=0}^{T-1}\int_{\Omega_1}\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]=\mathbf{z}\}d\mathbf{z}\le2\eta\int_{\Omega_1}\sum_{t=0}^{T-1}p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})d\mathbf{z}. \quad (26)$$

$$\begin{aligned}&\sum_{t=0}^{T-1}\int_{\Omega_2}\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]=\mathbf{z}\}d\mathbf{z}\le\eta\int_{\Omega_2}\sum_{t=0}^{T-1}p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})d\mathbf{z} \\ &-\frac{\delta}{\sqrt{K}}\int_{\Omega_2}\sum\nolimits_{t=0}^{T-1}p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})\|\mathbf{w}[t-1]-\mathbf{w}^*\|d\mathbf{z}. \quad (27)\end{aligned}$$

$$\begin{aligned}&\sum_{t=0}^{T-1}\int_{\Omega_3}\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]=\mathbf{z}\}d\mathbf{z}\le\eta\int_{\Omega_3}\sum_{t=0}^{T-1}p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})d\mathbf{z} \\ &\quad-\frac{\delta}{\sqrt{K}}\int_{\Omega_3}\sum\nolimits_{t=0}^{T-1}p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})\|\mathbf{w}[t]-\mathbf{w}^*\|d\mathbf{z}. \quad (28)\end{aligned}$$

$$\begin{aligned}&\sum_{t=0}^{T-1}\int_{\Omega_4}\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]=\mathbf{z}\}d\mathbf{z}\le-\frac{\delta}{\sqrt{K}}\int_{\Omega_4}\sum_{t=0}^{T-1}(\\ &\quad\|\mathbf{w}[t]-\mathbf{w}^*\|+\|\mathbf{w}[t-1]-\mathbf{w}^*\|)p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})d\mathbf{z}. \quad (29)\end{aligned}$$

Adding (26) to (29), we have that

$$\begin{aligned}&\mathbb{E}\{V(\mathbf{z}[T])|\mathbf{z}[0]\}-V(\mathbf{z}[0])\le\eta\int_{\cup_{i=1}^3\Omega_i}\sum_{t=0}^{T-1}p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})d\mathbf{z} \\ &-\frac{\delta}{\sqrt{K}}\Big[\int_{\Omega_2}\sum_{t=0}^{T-1}\|\mathbf{w}[t-1]-\mathbf{w}^*\|p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})d\mathbf{z}+\end{aligned}$$

$$\int_{\Omega_3}\sum_{t=0}^{T-1}\|\mathbf{w}[t]-\mathbf{w}^*\|p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})d\mathbf{z}+\int_{\Omega_4}\sum_{t=0}^{T-1}\big(\|\mathbf{w}[t]-\mathbf{w}^*\|+\|\mathbf{w}[t-1]-\mathbf{w}^*\|\big)p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})d\mathbf{z}\Bigg]. \quad (30)$$

Note that for any $\mathbf{z}\in\mathbb{R}^{2N}$, $\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}p_{\mathbf{z}[t]|\mathbf{z}[0]}=p_{\mathbf{z}}^{\infty}$ for all $\mathbf{z}[0]$, where $p_{\mathbf{z}}^{\infty}$ denotes the stationary distribution of the continuous state Markov chain $\mathbf{z}[t]$. Moving $V(\mathbf{z}[0])$ to the right hand side (RHS) of (30), dividing both sides of (30) by T, and letting $T\to\infty$ yields:

$$0\le-\frac{\delta}{\sqrt{K}}\Bigg[\int_{\Omega_2\cup\Omega_3}\|\mathbf{w}^{\infty}-\mathbf{w}^*\|p_{\mathbf{z}}^{\infty}d\mathbf{z}+\int_{\Omega_4}2\|\mathbf{w}^{\infty}-\mathbf{w}^*\|p_{\mathbf{z}}^{\infty}d\mathbf{z}\Bigg]+\eta\int_{\cup_{i=1}^{3}\Omega_i}p_{\mathbf{z}}^{\infty}d\mathbf{z}. \quad (31)$$

Rearranging terms and adding $\frac{\delta}{\sqrt{K}}\int_{\Omega_2\cup\Omega_3}\|\mathbf{w}^{\infty}-\mathbf{w}^*\|p_{\mathbf{z}}^{\infty}d\mathbf{z}+\frac{\delta}{\sqrt{K}}\int_{\Omega_1}2\|\mathbf{w}^{\infty}-\mathbf{w}^*\|p_{\mathbf{z}}^{\infty}d\mathbf{z}$ to both sides of (31) yields:

$$\frac{\delta}{\sqrt{K}}\int_{\mathbb{R}^{2N}}2\|\mathbf{w}^{\infty}-\mathbf{w}^*\|p_{\mathbf{z}}^{\infty}d\mathbf{z}\le\eta\int_{\cup_{i=1}^{3}\Omega_i}p_{\mathbf{z}}^{\infty}d\mathbf{z}+\frac{\delta}{\sqrt{K}}\Bigg[\int_{\Omega_2\cup\Omega_3}\|\mathbf{w}^{\infty}-\mathbf{w}^*\|p_{\mathbf{z}}^{\infty}d\mathbf{z}+\int_{\Omega_1}2\|\mathbf{w}^{\infty}-\mathbf{w}^*\|p_{\mathbf{z}}^{\infty}d\mathbf{z}\Bigg]$$
$$\overset{(a)}{\le}\eta\int_{\cup_{i=1}^{3}\Omega_i}p_{\mathbf{z}}^{\infty}d\mathbf{z}+4\delta\sqrt{(1/2)B\Phi}\le\eta+\delta\sqrt{8B\Phi}, \quad (32)$$

where (a) follows from the definitions of Ω and $\mathcal{B}$. Note that the left-hand-side (LHS) of (32) is exactly $\frac{2\delta}{\sqrt{K}}\mathbb{E}\{\|\mathbf{w}^{\infty}-\mathbf{w}^*\|\}$. Multiplying both sides of (32) by $\frac{\sqrt{K}}{2\delta}$ yields:

$$\mathbb{E}\{\|\mathbf{w}^{\infty}-\mathbf{w}^*\|\}\le\Big(\frac{\eta}{2\delta}+\sqrt{2B\Phi}\Big)\sqrt{K}=O(\sqrt{K}). \quad (33)$$

This completes the proof of Theorem 4. □

Step 4): Final Step to Prove Theorem 1: Recall that the Nesterovian update can be written as:

$$\mathbf{w}[t+1]=\mathbf{w}[t]+\Delta\mathbf{q}[t]+\beta[(\mathbf{w}[t]+\Delta\mathbf{q}[t])-(\mathbf{w}[t-1]+\Delta\mathbf{q}[t-1])]+\mathbf{u}[t],$$

Rearranging terms and noting that $\mathbf{u}[t]\ge\mathbf{0}$, we have

$$\Delta\mathbf{q}[t]\le\big(\mathbf{w}[t+1]-\mathbf{w}[t]\big)-\beta\big(\mathbf{w}[t]-\mathbf{w}[t-1]\big)-\beta(\Delta q[t]-\Delta q[t-1]). \quad (34)$$

Telescoping the inequality in (34) from $t=0$ to $T-1$ yields:

$$\sum_{t=0}^{T-1}\Delta\mathbf{q}[t]\le\big(\mathbf{w}[T]-\mathbf{w}[0]\big)-\beta\big(\mathbf{w}[T-1]-\mathbf{w}[-1]\big)$$
$$=-\beta(\Delta q[T-1]-\Delta q[-1])=\mathbf{w}[T]-\beta\mathbf{w}[T-1]-\beta\Delta\mathbf{q}[T-1],$$

where the last equality holds because, by assumption, $\mathbf{w}[0]=\mathbf{w}[-1]=\mathbf{q}[-1]=\mathbf{0}$. Also, since $\mathbf{q}[0]=\mathbf{0}$, we have

$$\|\mathbf{q}[T]\|_1=\|\mathbf{q}[0]+\sum_{t=0}^{T-1}\Delta\mathbf{q}[t]\|_1\le\|\mathbf{w}[T]-\beta\mathbf{w}[T-1]-\beta\Delta\mathbf{q}[T-1]\|_1.$$

Taking expectation on both sides, letting $T\to\infty$, and noting that in steady-state $\mathbb{E}\{\Delta\mathbf{q}[\infty]\}=\mathbf{0}$, we have:

$$\limsup_{T\to\infty}\mathbb{E}\big\{\|\mathbf{q}[T]\|_1\big\}\le\mathbb{E}\big\{\mathbf{w}^{\infty}-\beta\mathbf{w}^{\infty}\big\}\overset{(a)}{\le}\mathbf{w}^*+O(\sqrt{K})-\beta\big(\mathbf{w}^*-O(\sqrt{K})\big)=(1-\beta)\mathbf{w}^*+(1+\beta)O(\sqrt{K}), \quad (35)$$

where (a) follows from Theorem 4 and $\|\cdot\|_1\le\sqrt{N}\|\cdot\|$ (simplifying $\mathbf{w}^*(K)$ to $\mathbf{w}^*$). This proves the first part of Theorem 1. Moreover, in the asymptotic regime where $1-\beta=O(\frac{1}{\sqrt{K}})$, it follows from (35) that

$$\limsup_{T\to\infty}\mathbb{E}\big\{\|\mathbf{q}[t]\|_1\big\}=O(\sqrt{K}). \quad (36)$$

This completes the proof of Theorem 1.

Proof of Theorem 2. We first prove the optimality gap for the stationary rates $a_n^{\infty}\triangleq\mathbb{E}\{\min\{U_n'^{-1}(w_n^{\infty}/K),a^{\max}\}\}$. Note that $\mathbb{E}\{a_n[t]|w_n[t]\}=\min\{U_n'^{-1}(w_n[t]/K),a^{\max}\}$ and $a_n^*=U_n'^{-1}(w_n^*/K)$, $\forall n$. Thus, we have

$$\|\mathbf{a}^{\infty}-\mathbf{a}^*\|^2=\sum_{n=1}^{N}\Bigg[\mathbb{E}\Bigg\{\min\Bigg\{U_n'^{-1}\Big(\frac{w_n^{\infty}}{K}\Big),M\Bigg\}\Bigg\}-U_n'^{-1}\Big(\frac{w_n^*}{K}\Big)\Bigg]^2$$
$$\overset{(a)}{\le}\sum_{n=1}^{N}\mathbb{E}\Bigg\{\Bigg[\min\Bigg\{U_n'^{-1}\Big(\frac{w_n^{\infty}}{K}\Big),M\Bigg\}-U_n'^{-1}\Big(\frac{w_n^*}{K}\Big)\Bigg]^2\Bigg\}$$
$$\overset{(b)}{\le}\sum_{n=1}^{N}\mathbb{E}\Bigg\{\Bigg[U_n'^{-1}\Big(\frac{w_n^{\infty}}{K}\Big)-U_n'^{-1}\Big(\frac{w_n^*}{K}\Big)\Bigg]^2\Bigg\}$$
$$\overset{(c)}{=}\sum_{n=1}^{N}\mathbb{E}\Bigg\{\Bigg[\Bigg[U_n'^{-1}\Big(\frac{\tilde{w}_n}{K}\Big)\Bigg]'\Big(\frac{w_n^{\infty}}{K}-\frac{w_n^*}{K}\Big)\Bigg]^2\Bigg\}$$
$$\overset{(d)}{=}\sum_{n=1}^{N}\mathbb{E}\Bigg\{\Bigg[\Bigg[\frac{1}{U_n''\big(\frac{\tilde{w}_n}{K}\big)}\Bigg]^2\Big(\frac{w_n^{\infty}}{K}-\frac{w_n^*}{K}\Big)\Bigg]^2\Bigg\}$$
$$\overset{(e)}{\le}\sum_{n=1}^{N}\mathbb{E}\Bigg\{\frac{1}{\phi^2}\big(w_n^{\infty}-w_n^*\big)^2\frac{1}{K^2}\Bigg\}\overset{(f)}{=}\frac{1}{\phi^2K^2}\mathbb{E}\Bigg\{\sum_{n=1}^{N}\big(w_n^{\infty}-w_n^*\big)^2\Bigg\}$$
$$=\frac{1}{\phi^2K^2}\mathbb{E}\left\{\|\mathbf{w}^{\infty}-\mathbf{w}^*\|^2\right\}, \quad (37)$$

where (a) follows from the convexity of quadratic function and Jensen's inequality; (b) follows from the non-expansion property of the $\min\{\cdot\}$ function; (c) follows by using mean value theorem for some $\tilde{w}_n\in\big[\min\{w_n^{\infty},w_n^*\},\max\{w_n^{\infty},w_n^*\}\big]$; (d) follows from inverse function lemma; (e) follows from the strong convexity assumption in (7); and (f) follows from exchanging the order or summation and expectation.

Consider the term $\mathbb{E}\left\{\|\mathbf{w}^{\infty}-\mathbf{w}^*\|^2\right\}$ in (37). From the proof of Proposition 1, we have (cf. Eq. (62) in Appendix A) the following one-slot mean Lyapunov drift bound:

$$\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]\}\le-\frac{1}{\Phi K}(\|\mathbf{w}[t]-\mathbf{w}^*\|^2+\|\mathbf{w}[t-1]-\mathbf{w}^*\|^2)+B. \quad (38)$$

Following the same argument in the proof of Theorem 4, we telescope the inequality in (38) from $t=0$ to $T-1$ to obtain:

$$\mathbb{E}\{V(\mathbf{z}[T])|\mathbf{z}[0]\}-V(\mathbf{z}[0])=\sum_{t=0}^{T-1}\mathbb{E}\{V(\mathbf{z}[t+1])-V(\mathbf{z}[t])|\mathbf{z}[0]\}$$
$$=\sum_{t=0}^{T-1}\int_{\mathbb{R}^{2N}}p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z})\mathbb{E}\{V(\mathbf{z}[t+1])-V(\mathbf{z}[t])|\mathbf{z}[t]=\mathbf{z}\}d\mathbf{z}$$

$$= \sum_{t=0}^{T-1} \int_{\mathbb{R}^{2N}} p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z}) \mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]\} d\mathbf{z} \leq -\frac{1}{\Phi K} \sum_{t=0}^{T-1} \int_{\mathbb{R}^{2N}} p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z}) (\|\mathbf{w}[t]-\mathbf{w}^*\|^2 + \|\mathbf{w}[t-1]-\mathbf{w}^*\|^2) d\mathbf{z} + TB. \quad (39)$$

Dividing both sides by $\frac{2T}{\Phi K}$, rearranging terms, and letting $T \to \infty$, we have

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \int_{\mathbb{R}^{2N}} p_{\mathbf{z}[t]|\mathbf{z}[0]}(\mathbf{z}) \|\mathbf{w} - \mathbf{w}^*\|^2 d\mathbf{z} \leq \frac{1}{2} B\Phi K. \quad (40)$$

Note that the LHS of (40) is precisely $\mathbb{E}\left\{\|\mathbf{w}^\infty - \mathbf{w}^*\|^2\right\}$. Hence, it follows that

$$\|\mathbf{a}^\infty - \mathbf{a}^*\|^2 \leq \frac{1}{\phi^2 K^2} \mathbb{E}\left\{\|\mathbf{w}^\infty - \mathbf{w}^*\|^2\right\} \leq \frac{B\Phi}{2\phi^2} \frac{1}{K}. \quad (41)$$

Taking square root on both sides of (41) yields:

$$\|\mathbf{a}^\infty - \mathbf{a}^*\| \leq \frac{\sqrt{B\Phi}}{\phi} \frac{1}{\sqrt{K}} = O\Big(\frac{1}{\sqrt{K}}\Big),$$

i.e., the result in the first half of Theorem 2.

Next, we prove the optimality gap result for the objective value, i.e., $U\big(\mathbf{a}^\infty_{(K)}\big) \geq U(\mathbf{a}^*) - O(1/K)$. To this end, similar to the proof of Theorem 4, we define an augmented vector and its quadratic Lyapunov function as follows:

$$\mathbf{y}[t] \triangleq \begin{bmatrix} \mathbf{w}[t] \\ \mathbf{w}[t-1] \end{bmatrix} \quad \text{and} \quad \tilde{V}(\mathbf{y}[t]) = \frac{1}{2} \|\mathbf{y}[t]\|^2.$$

Following the same steps in the proof of Theorem 4, one can verify that

$$\mathbf{y}[t+1] = \mathbf{\Gamma}(\beta)\mathbf{y}[t] + \mathbf{\Gamma}'(\beta) \begin{bmatrix} \Delta\mathbf{q}[t] \\ \Delta\mathbf{q}[t-1] \end{bmatrix} + \begin{bmatrix} \mathbf{u}[t] \\ \mathbf{0}_N \end{bmatrix}.$$

where $\mathbf{\Gamma}(\beta)$ and $\mathbf{\Gamma}'(\beta)$ are the same as defined earlier. Then, following the same steps as in the proof of Proposition 1, we can show that the conditional expectation of the one-slot Lyapunov drift can be bounded as follows:

$$\mathbb{E}\big\{\Delta\tilde{V}(\mathbf{y}[t])\big|\mathbf{y}[t]\big\} \leq -\big(\mathbf{y}[t]\big)^\top \mathbb{E}\{\tilde{\mathbf{a}}[t] - \tilde{\mathbf{s}}[t]\big|\mathbf{y}[t]\} + B. \quad (42)$$

Note that (42) is in the same form as in [3, Eq. (24)]. Then, following the same arguments in [3], we have that $U\big(\mathbf{a}^\infty_{(K)}\big) \geq U(\mathbf{a}^*) - O(1/K)$. This completes the proof. □

Proof of Theorem 3. We first show the ranges of K and β that suffice for convergence. Due to the one-to-one mapping between $\mathbb{E}\{\mathbf{a}[t]|\mathbf{w}[t]\}$ and $\mathbf{w}[t]$, the convergence of $\mathbb{E}\{\mathbf{a}[t]|\mathbf{w}[t]\}$ can be equivalently analyzed by examining the convergence of $\mathbf{w}[t]$. Using the fact that $\Delta\mathbf{q}[t] = \frac{1}{K}\mathbf{H}^*(\mathbf{w}[t] - \mathbf{w}^*) + o(\|\mathbf{w}[t] - \mathbf{w}^*\|)$, where $\mathbf{H}^*$ denotes the Hessian matrix of $\Theta_K(\mathbf{w})$ evaluated at $\mathbf{w}^*$, we can rewrite (11) in a local neighborhood of $\mathbf{w}^*$ as:

$$\mathbf{z}[t+1] \leq \begin{bmatrix} (1+\beta)(\mathbf{I}_N - \frac{1}{K}\mathbf{H}^*) & -\beta(\mathbf{I}_N - \frac{1}{K}\mathbf{H}^*) \\ \mathbf{I}_N & \mathbf{O}_N \end{bmatrix} \mathbf{z}[t].$$

where $\mathbf{z}[t]$ is defined the same as in the proof of Theorem 1. For convenience, we let $\widehat{\mathbf{\Gamma}}$ denote the coefficient matrix and consider the eigenvalue equation:

$$\begin{bmatrix} (1+\beta)(\mathbf{I}_N - \frac{1}{K}\mathbf{H}^*) & -\beta(\mathbf{I}_N - \frac{1}{K}\mathbf{H}^*) \\ \mathbf{I}_N & \mathbf{O}_N \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} = \lambda_{\widehat{\mathbf{\Gamma}}} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}.$$

Noting from the second row that $\mathbf{v}_1 = \lambda_{\widehat{\mathbf{\Gamma}}}\mathbf{v}_2$, we have that

$$\Big[\lambda^2_{\widehat{\mathbf{\Gamma}}}\mathbf{I} - \lambda_{\widehat{\mathbf{\Gamma}}}(1+\beta)\Big(\mathbf{I} - \frac{1}{K}\mathbf{H}^*\Big) + \beta\Big(\mathbf{I} - \frac{1}{K}\mathbf{H}^*\Big)\Big]\mathbf{v}_2 = \mathbf{0}.$$

Let λ_i^*, $i = 1, \ldots, N$, denote the eigenvalues of $\mathbf{H}^*$. Through eigenvalue decomposition, it can be verified that $\lambda_{\widehat{\mathbf{\Gamma}}}$ satisfies the following quadratic equations:

$$\lambda^2_{\widehat{\mathbf{\Gamma}},i} - (1+\beta)\Big(1 - \frac{\lambda_i^*}{K}\Big)\lambda_{\widehat{\mathbf{\Gamma}},i} + \beta\Big(1 - \frac{\lambda_i^*}{K}\Big) = 0, \quad \forall i. \quad (43)$$

It then follows from (43) that

$$R_{(K,\beta)} \leq \max_i \lambda_{\widehat{\mathbf{\Gamma}},i} = \frac{1}{2} \max_{\lambda_i^*, \forall i} \left\{ \left| (1+\beta)\Big(1 - \frac{\lambda_i^*}{K}\Big) \pm \sqrt{\Delta_i} \right| \right\}, \quad (44)$$

where $\Delta_i \triangleq (1+\beta)^2(1 - \frac{\lambda_i^*}{K})^2 - 4\beta(1 - \frac{\lambda_i^*}{K})$. Next, we consider (44) for the following two cases, where we let $\alpha_i \triangleq 1 - \lambda_i^*/K$ for notational convenience:

- $\Delta_i \geq 0$: Then, $|\lambda_{\widehat{\mathbf{\Gamma}},i}| < 1$ is equivalent to $(1+\beta)^2\alpha_i^2 - 4\beta\alpha_i \geq 0$ and $-2 < (1+\beta)\alpha_i \pm \sqrt{\Delta_i} < 2$, which, after simplifications, yields $K \geq [(1+\beta)/(1-\beta)]^2\lambda_i^*$, $\forall i$
- $\Delta_i < 0$: Then, $|\lambda_{\widehat{\mathbf{\Gamma}},i}| < 1$ is equivalent to $(1+\beta)^2\alpha_i^2 - 4\beta\alpha_i < 0$ and $0 \leq \frac{1}{4}[(1+\beta)^2\alpha_i^2 - \Delta_i] < 1$, which, after simplifications, yields: $\lambda_i^* < K < [(1+\beta)/(1-\beta)]^2\lambda_i^*$, $\forall i$.

Combining both cases and noting that they should hold for all λ_i^*, we can conclude that $K > \Phi$. This completes the proof of the first half of the theorem.

To prove the second half of the theorem, we minimize the upper bound in (44), i.e.,

$$\underset{K\in(\Phi,\infty),\beta\in[0,1)}{\text{Minimize}} \left\{ \frac{1}{2} \max_{\lambda_i^*, \forall i} \left\{ \left| (1+\beta)\Big(1 - \frac{\lambda_i^*}{K}\Big) \pm \sqrt{\Delta_i} \right| \right\} \right\}. \quad (45)$$

First, we claim that the upper bound in (44) is monotonically decreasing with respect to the λ_i^*-variables. To see this, we again consider two cases.

- $\Delta_i \geq 0$: In this case, $\alpha \in [4\beta/(1+\beta)^2, 1)$. In this interval, Δ_i monotonically increases as λ_i decreases. Hence, the upper bound in (44) is monotonically decreasing.
- $\Delta_i < 0$: From (44), we have $\lambda_{\widehat{\mathbf{\Gamma}},i} = 2\sqrt{\beta(1 - \lambda_i^*/K)}$, which is monotonically increasing as λ_i^* decreases.

Combining both cases, we have that (45) can be written as:

$$\min_{\beta\in[0,1)} \left\{ \min_{K\in(\Phi,\infty)} \left\{ \left| (1+\beta)\Big(1 - \phi/K\Big) \pm \sqrt{(1+\beta)^2(1-\phi/K)^2 - 4\beta(1-\phi/K)} \right| \right\} \right\}. \quad (46)$$

For notational simplicity, we let $\psi(K,\beta) \triangleq |(1+\beta)(1 - \phi/K) \pm \sqrt{(1+\beta)^2(1-\phi/K)^2 - 4\beta(1-\phi/K)}|$. Then, based on the positivity of the discriminant in (46), the search domain in the optimization problem in (46) can be divided into three subdomains:

- $\beta \in [0, (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1))$, $K \in (\Phi, \infty)$,
- $\beta \in [(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1), 1)$, $K \in (\Phi, ((1+\beta)/(1-\beta))^2\,\phi)$,
- $\beta \in [(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1), 1)$, $K \in [((1+\beta)/(1-\beta))^2\,\phi, \infty)$,

where $\kappa \triangleq \Phi/\phi$ is the condition number. It can be readily verified that (46) is minimized to $(\sqrt{\kappa}-1)/\sqrt{\kappa}$ by $K^* = \Phi$ and $\beta^* = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$. This completes the proof of the second half of the theorem. □

5. EXTENSION TO MULTI-HOP NETWORKS

In this section, we will generalize our Nesterovian algorithmic framework to multi-hop networks. In the multi-hop setting, the utility optimization problem becomes the joint congestion control and routing optimization as in [1–3]. We first state the network model and problem formulation.

Network Model and Problem Formulation: *1) Network model:* Consider a time-slotted communication network system as in the single-hop case. 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 flows in the network, indexed by $f = 1, \ldots, F$. Each flow 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 flow f travel from $\mathrm{Src}(f)$ to $\mathrm{Dst}(f)$ through the network, possibly via multi-hop and multi-path routing.

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

3) Routing: We let $x_l^{(f)}[t] \geq 0$ denote the rate offered to route session f's data in time-slot t at link l. We let $\bar{x}_l^{(f)} \triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} x_l^{(f)}[t]$ represent the time-average routing rate of session f at link l. The time-varying model of the channel states at each link remains the same as in the single-hop case. As in [2, 31, 34], we define the *network capacity region* as the largest set of congestion control rates such that there exists a routing policy for which the time-average routing rates satisfy the following constraints:

$$\sum_{l\in\mathcal{O}(n)} \bar{x}_l^{(f)} \geq \sum_{l\in\mathcal{I}(n)} \bar{x}_l^{(f)} + \bar{a}_f \mathbb{1}_f(n), \ \forall f, \forall n \neq \mathrm{Dst}(f), \quad (47)$$

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.

4) Queue-stability: We assume that each node has a separate queue for each flow f. Let $q_n^{(f)}[t] \geq 0$ represent the queue-length of flow f at node n at time t. Since data leave the network upon reaching destinations, we have $q_{\mathrm{Dst}(f)}^{(f)}[t] = 0$, $\forall t$. Then, $q_n^{(f)}[t]$, $n \neq \mathrm{Dst}(f)$ evolves as:

$$q_n^{(f)}[t+1] = \left(q_n^{(f)}[t] - \sum_{l\in\mathcal{O}(n)} x_l^{(f)}[t]\right)^+ + \sum_{l\in\mathcal{I}(n)} \widehat{x}_l^{(f)}[t] + a_f[t]\mathbb{1}_f(n), \quad (48)$$

where $(\cdot)^+ \triangleq \max\{0, \cdot\}$ and $\widehat{x}_l^{(f)}[t]$ is the *actual* routing rate. Note that $\widehat{x}_l^{(f)}[t] \leq x_l^{(f)}[t]$ since $\mathrm{Tx}(l)$ may have less than $x_{l,[t]}^{(f)}$ amount of data to transmit. Let $\mathbf{q}[t] \triangleq [q_n^{(f)}[t], \forall f, \forall n \neq \mathrm{Dst}(f)]^T$ group all queue lengths at time t. Similar to the single-hop case, under a congestion control and routing scheme, we say that the network is stable if the norm of steady-state queue-lengths is finite, i.e., $\limsup_{t\to\infty} \mathbb{E}\{\|\mathbf{q}[t]\|_1\} < \infty$.

5) Problem formulation: In the multi-hop case, our goal is again to develop an optimal joint congestion control and routing scheme to maximize the total utility $\sum_{f=1}^{F} U_f(\bar{a}_f)$, subject to the network capacity region and stability constraints. Putting together the models presented above yields the following joint congestion control and routing (CCR) optimization problem:

$$\textbf{CCR: } \text{Max} \quad \sum_{f=1}^{F} U_f(\bar{a}_f)$$

s.t. Routing constr. in (47); Stability of all queues,

$$x_l^{(f)}[t] \in \mathcal{C}_{\boldsymbol{\pi}[t]}, \ \forall l, t, f, \quad a_f[t] \geq 0, \ \forall f, t.$$

5.1 A Nesterovian Joint Congestion and Routing Optimization Algorithm

Similar to the the QCA schemes in the multi-hop case, in our multi-hop Nesterovian algorithm, the weights are replaced by *weight differentials* for dynamic routing. Let $\mathcal{E}(l)$ denote the two end nodes of link l. The Nesterovian congestion control and routing algorithm is stated as follows:

Algorithm 2: A Nesterovian Approach for Joint Congestion Control and Routing.

Initialization:

1. Choose parameters $K > 0$ and $\beta \in [0, 1)$. Set $t = 0$.
2. Let queue-states: $q_n^{(f)}[0], \Delta q_n^{(f)}[-1] = 0$, $\forall n, f$.
3. Associate each link l with a weight $w_n^{(f)} \geq 0$ and set the initial weights $w_n^{(f)}[0] = w_n^{(f)}[-1] = 0, \forall n, f$.

Main Loop:

4. *Weight Differentials:* In time-slot $t \geq 1$, we let $\Delta w_l^{(f)}[t] = \max\left\{w_n^{(f)}[t] - w_{\mathcal{E}(l)\setminus n}^{(f)}[t], 0\right\}$ denote the *weight differential of each flow* f, $\forall n, l \in \mathcal{O}(n)$. Let $\Delta w_l^*[t] = \max_f \Delta w_l^{(f)}[t]$ and let $f_l^*[t] = \arg\max_f \Delta w_l^{(f)}[t]$ (breaking ties arbitrarily). Let $\Delta\mathbf{w}^*[t] \triangleq [w_1^*[t], \ldots, w_L^*[t]]^\top$ be the maximum weight differentials vector over all links.
5. *Routing and MaxWeight Scheduling:* Given $\Delta\mathbf{w}^*[t]$ and channel state $\boldsymbol{\pi}[t]$, the scheduler chooses a service rate vector $\mathbf{x}[t] \in \mathbb{R}^L$ to transmit flow $f_l^*[t]$ at each link l:
$$\mathbf{x}[t] = \arg\max_{\mathbf{x}\in\mathcal{C}_{\boldsymbol{\pi}[t]}} (\Delta\mathbf{w}^*[t])^\top \mathbf{x}. \quad (49)$$
6. *Congestion Controller:* For each flow f and in each time-slot t, let w be the value of $w_{\mathrm{Src}(f)}[t]$ that the source node $\mathrm{Src}(f)$ observes. Then, $\mathrm{Src}(f)$ sets $a_f[t]$ to be an integer-valued random variable that satisfies:
$$\mathbb{E}\{a_f[t]|w\} = \min\left\{U_f'^{-1}\left(\frac{w}{K}\right), a^{\max}\right\}, \quad (50)$$
$$\mathbb{E}\{a_n^2[t]|w\} \leq A < \infty, \quad (51)$$
where $U_f'^{-1}(\cdot)$ represents the inverse function of first-order derivative of $U_f(\cdot)$. In (50) and (51), $a^{\max}$ and A are positive constants with $a^{\max} > 2s^{\max}$.
7. *Queue-Length and Nesterovian Weight Updates:* Update queue-lengths following (48). Let $\Delta q_n^{(f)}[t] \triangleq q_n^{(f)}[t+1] - q_n^{(f)}[t]$, $\forall n$, be the queue-length change of flow f at node

n following the update in (48). Next, update the weights in the following (projected) **Nesterovian** manner:

$$w_n^{(f)}[t+1] = \{w_n^{(f)}[t] + \Delta q_n[t] + \beta[(w_n^{(f)}[t] + \Delta q_n^{(f)}[t]) - (w_n^{(f)}[t-1] + \Delta q_n^{(f)}[t-1])]\}^+, \ \forall n, f. \quad (52)$$

Let $t = t+1$. Go to Step 4 and repeat the whole dynamic routing, scheduling and congestion control processes.

Distributed Implementation: We note that, as in the QCA algorithms, the congestion control and dynamic routing components in the multi-hop version of the Nesterovian algorithm only require weight information either locally or from one-hop neighbors. Also, the Nesterovian weight updates in (52) only require two time-slots of local history. Thus, the congestion control and dynamic routing naturally lend themselves to *distributed* implementation in exactly the *same* fashion as that in the QCA algorithms. In other words, compared to QCA algorithms, our multi-hop Nesterovian algorithm does *not* incur any additional complexity in terms of messaging passing between nodes.

On the other hand, also same as in the QCA algorithms, we can see that the scheduling problem in (49) require global weight information from all links and could be challenging for developing distributed solutions (depending on the structure of $\mathcal{C}_{\boldsymbol{\pi}[t]}$). In many cases, it can be shown that the scheduling problem in (49) is NP-hard and even developing centralized solutions is difficult. Fortunately, due to the same messaging passing requirement, many distributed algorithms (e.g., [35, 36]) developed for the QCA framework can be applied in the scheduling component in our Nesterovian algorithm. One example is to adopt our weight adjustment scheme to Q-CSMA [36] by setting the attempt probability of each link in the form of $e^{w[t]}/(e^{w[t]}+1)$ (cf. [36, Eq. (10)]). Then, under the time-scale separation assumption, one can follow the same line of arguments in [36] to establish the throughput optimality with fewer iterations to update weights in the outer time-scale.

Performance Analysis: The delay reduction at source nodes, convergence rates, and utility-optimality results in Theorems 1–3 in the single-hop case can be generalized to the multi-hop cases. Also, their proofs follow the same steps and arguments but with more complicated notation. Due to space limitation, we omit these results and their proofs in this paper for brevity.

6. NUMERICAL RESULTS

In this section, we conduct some numerical studies to verify the theoretical results presented in Section 4. To better visualize the insights of our theoretical results and not being blurred by random noises, we first use a 4-link non-fading cellular network as an example. We assume that each link has unit capacity and only one link can be activated in each time-slot. We use $\log(0.001 + a)$ as the utility function for each link, i.e., the well-known proportional fairness metric [6]. Due to the symmetry of the setting, the optimal congestion control rates are $\bar{a}_1^* = \bar{a}_2^* = \bar{a}_3^* = \bar{a}_4^* = \frac{1}{4}$. To see the impact of β on delay and convergence, we let $K = 25$ and increase β from 0 to 0.99 ($\beta = 0$ corresponds to QCA). Due to the symmetry of the setting, we only plot the results of link 1. As shown in Figure 2, as β increases, the average queue-lengths are 100.1, 50.1, 20.2, and 1.1, respectively, which confirm the $(1-\beta)$–fraction reduction result in Theorem 1. We see from Figure 3 that, for all choices of β, the congestion control rates with different β's all converge to the same optimal solution, confirming Theorem 2 that utility-optimality is not affected byβ. However, varying β significantly affects the convergence speed: As β increases from 0 to 0.99, the convergence speed initially increases and then decreases. Particularly, we see from Figure 2 and Fig. 3 that, by letting $\beta = 0.99$, we achieve both utility-optimality and low-delay at the expense of slower convergence, hence confirming the three-way performance trade-off. Next, we increase K from 25 to 100 and conduct another set of experiments on the same network. The results are shown in Figure 4 and Figure 5, respectively. With a larger K, the congestion control rates again converge to the same optimal solution with a smaller variance, but at the cost of larger delay and slower convergence, again confirming Theorems 1–3.

Now, we test our Nesterovian algorithm in a larger 15-user cellular downlink system with quasi-static block fading (channel states vary from one slot to the next but remain constant in each slot). Again, we assume that only one user can be activated in each time-slot. First, we let $K = 100$ and set *beta* to 0 (i.e., QCA), 0.35, 0.65, 0.95, respectively. For fewer random noise, we only plot the congestion control rate and queue-length of user 1 in Figure 6 and Figure 7, respectively. In Figure 6, as β grows, the queue-length again decreases monotonically and follows the $(1-\beta)$-fraction reduction in Theorem 1. In Figure 7, the congestion control rates under different β's all converge to the same optimal solution, which is approximately $\frac{5}{15} = \frac{1}{3}$. Next, we increases K to 300 and conduct another set of experiments. The results are illustrated in Figure 8 and Figure 9, respectively. We can see that, with a larger K, the congestion control rates again converge to the same optimal solution with a smaller variance, but at the cost of larger delay and longer convergence time, again confirming Theorems 1 and 2.

Finally, we compare the steady-state queue-length scalings with respect to K under QCA and our Nesterovian algorithm, respectively. We let $\beta \uparrow 1$ as $\beta = 1 - \frac{1}{\sqrt{K}}$ as K increases. In Figure 10, we can see that the total queue-length of QCA exhibits the expected $O(K)$ linear scaling and is much larger than that of our Nesterovian approach. Figure 11 plots the zoom-in view of the Nesterovian curve in Figure 11. We can see that the total queue-length scales as $8\sqrt{K}$, perfectly matching the $O(\sqrt{K})$ result in Theorem 1.

7. CONCLUSION

In this paper, we have developed a Nesterovian algorithmic framework for stochastic network optimization. Compared to the traditional queue-length-based approaches, our Nesterovian algorithmic framework offers not only utility-optimality and queue-stability, but also dramatic delay reduction and fast convergence. Further, our proposed Nesterovian algorithmic framework that is well-suited for implementation in practice. We rigorously proved the utility-optimality of the proposed Nesterovian algorithmic framework and characterized the delay reduction and convergence speed performances; Also, we offered design rules for optimal selection of systems parameters, as well as insights on a three-way trade-off between utility-optimality, delay, and convergence. Collectively, these results shed lights on a new cross-layer network optimization theory that based on the

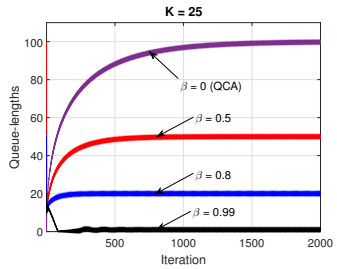

Figure 2: The impact of β on queueing delay.

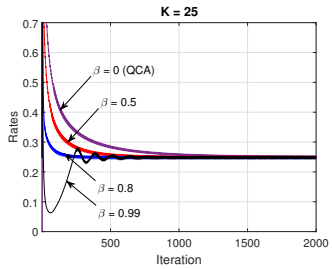

Figure 3: The impact of β on convergence speed.

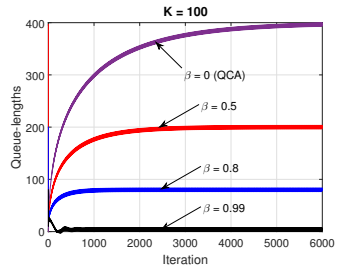

Figure 4: The impact of K on queueing delay.

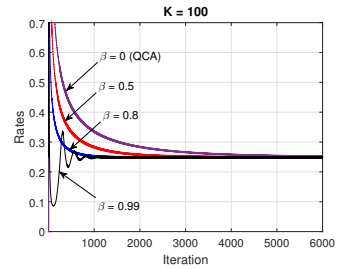

Figure 5: The impact of K on convergence speed.

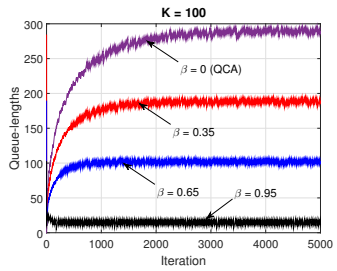

Figure 6: The impact of β on queueing delay for a 15-user cellular downlink with fading ($K = 100$).

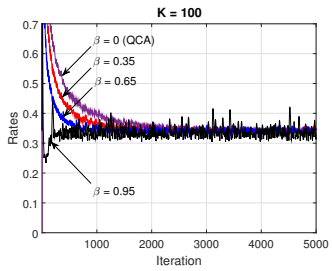

Figure 7: The impact of β on convergence for a 15-user cellular downlink with fading ($K = 100$).

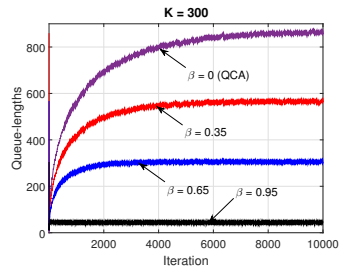

Figure 8: The impact of β on queueing delay for a 15-user cellular downlink with fading ($K = 300$).

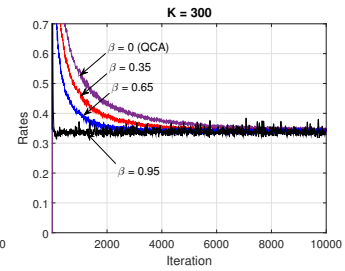

Figure 9: The impact of β on convergence for a 15-user cellular downlink with fading ($K = 300$).

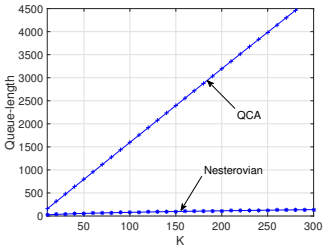

Figure 10: Steady-state queue lengths of QCA and the Nesterovian approaches.

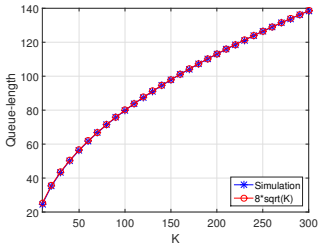

Figure 11: Zoom-in view of the Nesterovian approach in Fig. 10, showing the $O(\sqrt{K})$ scaling.

Nesterov's AGD method. Nesterovian cross-layer network optimization is an exciting and yet under-explored area. In our future research, we will further explore how to incorporate other Nesterovian variants into stochastic network control and optimization and investigate tight queue-length upper bounds rather than Big-O characterizations.

8. REFERENCES

[1] 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.

[2] 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.

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

[4] A. Eryilmaz and R. Srikant, "Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1333–1344, Dec. 2007.

[5] A. L. Stolyar, "Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm," *Queueing Systems*, vol. 50, no. 4, pp. 401–457, 2005.

[6] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1452–1463, Aug. 2006.

[7] M. J. Neely, "Super-fast delay tradeoffs for utility optimal fair scheduling in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1489–1501, Aug. 2006.

[8] L. Huang and M. J. Neely, "Delay reduction via lagrange multipliers in stochastic network optimization," *IEEE Trans. Autom. Control*, vol. 56, no. 4, pp. 842–857, Apr. 2011.

[9] L. Huang, X. Liu, and X. Hao, "The power of online learning in stochastic network optimization," in *Proc. ACM Sigmetrics*, Austin, TX, Jun.16-20, 2014, pp. 153–165.

[10] 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.

[11] 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.

[12] J. Liu, N. B. Shroff, C. H. Xia, and H. D. Sherali, "Joint congestion control and routing optimization:

An efficient second-order distributed approach," *IEEE/ACM Trans. Netw.*, 2015, accepted, to appear.

[13] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Soviet Math. Doklady*, vol. 27, no. 2, pp. 372 – 376, 1983.

[14] ——, *Introductory Lectures on Convex Programming.* Boston/Dordrecht/London: Kluwer Academic Publishers, 2004.

[15] J. Liu, A. Eryilmaz, N. B. Shroff, and E. S. Bentley, "Heavy-ball: A new approach to tame delay and convergence in wireless network optimization," in *Proc. IEEE INFOCOM*, 2016.

[16] R. J. Gibbens and F. P. Kelly, "Resource pricing and the evolution of congestion control," *Automatica*, vol. 35, pp. 1969–1985, 1999.

[17] S. Kunniyur and R. Srikant, "Analysis and design of an adaptive virtual queue algorithm for active queue management," in *Proc. ACM SIGCOMM*, San Diego, CA, Aug. 2001, pp. 123–134.

[18] A. Laksmikantha, C. Beck, and R. Srikant, "Robustness of real and virtual queue-based active queue management schemes," *IEEE/ACM Trans. Netw.*, vol. 13, no. 1, pp. 81–93, Feb. 2005.

[19] E. Athanasopoulou, L. Bui, T. Ji, R. Srikant, and A. Stolyar, "Back-pressure-based packet-by-packet adaptive routing in communication networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 1, pp. 244–257, Feb. 2013.

[20] N. Walton, "Concave switching in single and multihop networks," in *Proc. ACM Sigmetrics*, Austin, TX, Jun. 11-14, 2014, pp. 139 – 151.

[21] M. Bramson, B. D'Auria, and N. Walton, "Proportional switching in FIFO networks." [Online]. Available: http://arxiv.org/abs/1412.4390

[22] Z. A. Zhu and L. Orecchia, "Linear coupling: An ultimate unification of gradient and mirror descent," MIT CSAIL, Tech. Rep., January 2015. [Online]. Available: http://arxiv.org/pdf/1407.1537v4.pdf

[23] W. Su, S. Boyd, and E. J. Candes, "A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights," *Journal of Machine Learning Research*, 2015, accepted, to appear. [Online]. Available: http://arxiv.org/abs/1503.01243

[24] S. Bubeck, Y. T. Lee, and M. Singh, "A geometric alternative to Nesterov's accelerated gradient descent," June 2015. [Online]. Available: http://arxiv.org/abs/1506.08187

[25] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[26] ——, *Introduction to Optimization.* New York, NY: Optimization Software, Inc., May 1987.

[27] E. Ghadimi, I. Shames, and M. Johansson, "Multi-step gradient methods for networked optimization," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5417–5429, Nov. 2013.

[28] P. Ochs, T. Brox, and T. Pock, "iPiasco: Inertial proximal algorithm for strongly convex optimization," *Journal of Mathematical Imaging and Vision (JMIV)*, 2015.

[29] D. Jakovetic, J. M. F. Xavier, and J. M. F. Moura, "Convergence rates of distributed nesterov-like gradient methods on random networks," *IEEE Trans. Signal Process.*, vol. 62, no. 4, pp. 868 – 882, Feb. 2014.

[30] A. Eryilmaz, R. Srikant, and J. R. Perkins, "Stable scheduling policies for fading wireless channels," *IEEE/ACM Trans. Netw.*, vol. 13, no. 2, pp. 411–424, Apr. 2005.

[31] 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.

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

[33] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge, UK: Cambridge University Press, 2009.

[34] 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.

[35] L. Jiang and J. Walrand, "A distributed CSMA algorithm for throughput and utility maximization in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 18, no. 3, pp. 960–972, Jun. 2010.

[36] J. Ni, B. Tan, and R. Srikant, "Q-CSMA: Queue-length-based CSMA/CA algorithms for achieving maximum throughput and low delay in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 3, pp. 825–836, Jun. 2012.

APPENDIX

A. PROOF OF PROPOSITION 1

As stated in Section 4.4, in $\mathbb{R}^{2N}$, the Nesterovian update in (11) can be rewritten as:

$$\mathbf{z}[t+1] = \mathbf{\Gamma}(\beta)\mathbf{z}[t] + \mathbf{\Gamma}'(\beta)\begin{bmatrix} \Delta\mathbf{q}[t] \\ \Delta\mathbf{q}[t-1] \end{bmatrix} + \begin{bmatrix} \mathbf{u}[t] \\ \mathbf{0}_N \end{bmatrix}. \quad (53)$$

For convenience, we define $\Delta\tilde{\mathbf{q}}[t] \triangleq \left[\Delta\mathbf{q}^\top[t] + \mathbf{u}[t],\ \Delta\mathbf{q}^\top[t-1]\right]^\top$. Then, the right-hand-side of (53) can be written as:

$$\mathbf{z}[t+1] = \mathbf{\Gamma}(\beta)\mathbf{z}[t] + \mathbf{\Gamma}'(\beta)\Delta\tilde{\mathbf{q}}[t]. \quad (54)$$

Also, based on the queueing evolution in (5), we can further explicitly express $\Delta\tilde{\mathbf{q}}[t]$ as follows:

$$\Delta\tilde{\mathbf{q}}[t] = \begin{bmatrix} \mathbf{a}[t] - \mathbf{s}[t] + \mathbf{u}'[t] + \mathbf{u}[t] \\ \mathbf{a}[t-1] - \mathbf{s}[t-1] + \mathbf{u}'[t-1] \end{bmatrix}, \quad (55)$$

where $\mathbf{a}[t] \triangleq [a_1[t], \ldots, a_N[t]]^\top$, $\mathbf{s}[t] \triangleq [s_1[t], \ldots, s_N[t]]^\top$, $\forall t$, and $\mathbf{u}'[t] \triangleq [u'_1[t], \ldots, u'_N[t]]^\top$, $\forall t$, represents the projection term in (5), meaning the unused services in each time-slot.

Also, it can be readily verified that $\mathbf{\Gamma}'(\beta)$ has only two distinct eigenvalues 0 and $(1+\beta)$. Hence, the largest eigenvalues of $\mathbf{\Gamma}'(\beta)$ is bounded by 2 since $\beta \in [0, 1)$. Due to the non-expansive properties of the projection terms $\mathbf{u}'[t]$ and $\mathbf{u}[t]$, we further have that

$$\|\Delta\tilde{\mathbf{q}}[t]\| \leq \|\tilde{\mathbf{a}}[t] - \tilde{\mathbf{s}}[t]\|, \quad (56)$$

where $\tilde{\mathbf{a}}[t] \triangleq \left[\mathbf{a}^\top[t],\, \mathbf{a}^\top[t-1]\right]^\top$ and $\tilde{\mathbf{s}}[t] \triangleq \left[\mathbf{s}^\top[t],\, \mathbf{s}^\top[t-1]\right]^\top$.

Now, consider the conditional expectation of the one-slot Lyapunov drift in (24), for which we have:

$$\begin{aligned}
&\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]\} = \frac{1}{2}\mathbb{E}\big\{\|\mathbf{z}[t+1]\|^2 - \|\mathbf{z}[t]\|^2|\mathbf{z}[t]\big\} \\
&= \frac{1}{2}\mathbb{E}\big\{(\mathbf{z}[t+1]-\mathbf{z}[t])^\top(\mathbf{z}[t+1]+\mathbf{z}[t])|\mathbf{z}[t]\big\} \\
&\overset{(a)}{=} \frac{1}{2}\mathbb{E}\big\{[(\mathbf{\Gamma}(\beta)-\mathbf{I})\mathbf{z}[t]+\mathbf{\Gamma}'(\beta)\Delta\tilde{\mathbf{q}}[t]]^\top \\
&\qquad \big\{[(\mathbf{\Gamma}(\beta)+\mathbf{I})\mathbf{z}[t]+\mathbf{\Gamma}'(\beta)\Delta\tilde{\mathbf{q}}[t]]|\mathbf{z}[t]\big\} \\
&= \frac{1}{2}\mathbb{E}\big\{\mathbf{z}^\top[t](\|\mathbf{\Gamma}(\beta)\|^2-\mathbf{I})\mathbf{z}[t]+\mathbf{z}^\top[t](2\mathbf{\Gamma}(\beta)\mathbf{\Gamma}'(\beta))\Delta\tilde{\mathbf{q}}[t] \\
&\qquad + \|\mathbf{\Gamma}'(\beta)\Delta\tilde{\mathbf{q}}[t]\|^2|\mathbf{z}[t]\big\} \\
&\overset{(b)}{\leq} \frac{1}{2}\mathbb{E}\big\{4\mathbf{z}^\top[t]\Delta\tilde{\mathbf{q}}[t]+4\|\Delta\tilde{\mathbf{q}}[t]\|^2|\mathbf{z}[t]\big\} \\
&\overset{(c)}{\leq} \mathbb{E}\left\{2\|\tilde{\mathbf{a}}[t]-\tilde{\mathbf{s}}[t]\|^2+2\big\langle(\tilde{\mathbf{a}}[t]-\tilde{\mathbf{s}})[t],\mathbf{z}[t]\big\rangle\Big|\mathbf{z}[t]\right\} \\
&= 2\mathbf{z}^\top[t]\mathbb{E}\{\tilde{\mathbf{a}}[t]-\tilde{\mathbf{s}}[t]\big|\mathbf{z}[t]\}+2\mathbb{E}\left\{\|\tilde{\mathbf{a}}[t]\|^2+\|\tilde{\mathbf{s}}[t]\|^2\big|\mathbf{z}[t]\right\}, \quad (57)
\end{aligned}$$

where (a) follows from (54), (b) follows the spectral property of $\mathbf{\Gamma}(\beta)$ in Lemma 2 (i.e., the eigenvalues of $\mathbf{\Gamma}(\beta)$ are less or equal to 1), and and (c) from the non-expansive property of the projection terms. Further, from the second moment constraint of congestion control in (10) in our Nesterovian algorithm, we have $\mathbb{E}\{\|\tilde{\mathbf{a}}[t]\|^2|\ \mathbf{z}[t]\} \leq 2AN$. From the assumption that $s_n[t] \leq s^{\max}$, we can conclude that $\mathbb{E}\{\|\tilde{\mathbf{s}}[t]\|^2|\mathbf{z}[t]\} \leq 2N(s^{\max})^2$. Hence, by defining $B \triangleq 2N[A+(s^{\max})^2]$ (Notice here that B is independent of K), we have:

$$\begin{aligned}
&\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]\} \leq 2\mathbf{z}^\top[t]\mathbb{E}\{\tilde{\mathbf{a}}[t]-\tilde{\mathbf{s}}[t]\big|\mathbf{z}[t]\}+B \overset{(a)}{=} 2\mathbf{z}^\top[t] \\
&\times\big(\mathbb{E}\{\tilde{\mathbf{a}}[t]|\mathbf{z}[t]\}-\tilde{\mathbf{s}}^*\big)+\mathbb{E}\{2\mathbf{z}^\top[t]\big(\tilde{\mathbf{s}}^*-\tilde{\mathbf{s}}[t]\big)\big|\mathbf{z}[t]\}+B, \qquad (58)
\end{aligned}$$

where $\tilde{\mathbf{s}}^* \triangleq [(\mathbf{s}^*)^\top,(\mathbf{s}^*)^\top]^\top$ and $(\mathbf{s}^*,\mathbf{w}^*)$ is a pair of optimal primal and dual solutions to Problem D-CCS. In (58), (a) follows from adding and subtracting $\tilde{\mathbf{s}}^*$ as well as the fact that $\tilde{\mathbf{a}}[t]$ is independent of the channel state and determined solely by $\mathbf{w}[t]$. Consider the term $\mathbf{z}^\top[t]\big(\tilde{\mathbf{s}}^*-\tilde{\mathbf{s}}[t]\big)$ in (58). From the design of the scheduler in (8), we have the pair of relationships holding true:

$$\big(\mathbf{w}^*\big)^\top\mathbf{s}^* \geq \big(\mathbf{w}^*\big)^\top\mathbf{s}[t], \quad \big(\mathbf{w}[t]\big)^\top\mathbf{s}[t] \geq \big(\mathbf{w}[t]\big)^\top\mathbf{s}^*.$$

Adding these two inequalities and rearranging terms yields: $\big(\mathbf{w}[t]-\mathbf{w}^*\big)^\top\big(\mathbf{s}^*-\mathbf{s}[t]\big) \leq 0$. It then follows that

$$\begin{aligned}
&\mathbf{z}^\top[t]\big(\tilde{\mathbf{s}}^*-\tilde{\mathbf{s}}[t]\big) = [(\mathbf{w}[t]-\mathbf{w}^*)^\top,(\mathbf{w}[t-1]-\mathbf{w}^*)^\top]\times \\
&\left\{\begin{bmatrix}\mathbf{s}^*\\ \mathbf{s}^*\end{bmatrix}-\begin{bmatrix}\mathbf{s}[t]\\ \mathbf{s}[t-1]\end{bmatrix}\right\} = \sum_{\tau=t-1}^{t}\big(\mathbf{w}[\tau]-\mathbf{w}^*\big)^\top\big(\mathbf{s}^*-\mathbf{s}[\tau]\big) \leq 0
\end{aligned}$$

Hence, we can further rewrite (58) as:

$$\begin{aligned}
&\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]\} \leq 2\mathbf{z}^\top[t]\big(\mathbb{E}\{\tilde{\mathbf{a}}[t]|\mathbf{z}[t]\}-\tilde{\mathbf{s}}^*\big)+B = \sum_{\tau=t-1}^{t}\sum_{n=1}^{N} \\
&2\big(w_n[\tau]-w_n^*\big)\left[U_n'^{-1}\left(\frac{w_n[\tau]}{K}\right)-U_n'^{-1}\left(\frac{w_n^*}{K}\right)\right]+B. \quad (59)
\end{aligned}$$

Since $U_n(\cdot)$ is concave and increasing, $\forall n$, we have

$$\big(w_n[\tau]-w_n^*\big)^\top\left[U_n'^{-1}\left(\frac{w_n[\tau]}{K}\right)-U_n'^{-1}\left(\frac{w_n^*}{K}\right)\right] \leq 0.$$

Thus, by using Cauchy-Schwatz inequality, we can equivalently rewrite (59) as:

$$\begin{aligned}
\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]\} \leq -\sum_{\tau=t-1}^{t}\sum_{n=1}^{N} 2\,|w_n[\tau]-w_n^*| \times \\
\left|U_n'^{-1}\left(\frac{w_n[\tau]}{K}\right)-U_n'^{-1}\left(\frac{w_n^*}{K}\right)\right|+B. \quad (60)
\end{aligned}$$

By the strong convexity of $-U_n(\cdot)$ (cf. Eq. (7)) and the Lipschitz continuity of $U_n'(\cdot)$, we have

$$\phi\,|a_{n,1}-a_{n,2}| \leq \left|U_n'\left(a_{n,1}\right)-U_n'\left(a_{n,2}\right)\right| \leq \Phi\,|a_{n,1}-a_{n,2}|\,.$$

Therefore, by the inverse function lemma, we have

$$\begin{aligned}
\frac{1}{\Phi}\left|\frac{w_n[t]}{K}-\frac{w_n^*}{K}\right| &\leq \left|U_n'^{-1}\left(\frac{w_n[t]}{K}\right)-U_n'^{-1}\left(\frac{w_n^*}{K}\right)\right| \\
&\leq \frac{1}{\phi}\left|\frac{w_n[t]}{K}-\frac{w_n^*}{K}\right|. \quad (61)
\end{aligned}$$

Hence, we can further upper-bound (60) as:

$$\begin{aligned}
\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]\} &\leq -\frac{2}{\Phi K}\sum_{\tau=t-1}^{t}\sum_{n=1}^{N}\left(w_n[\tau]-w_n^*\right)^2+B \\
&= -\frac{2}{\Phi K}\sum_{\tau=t-1}^{t}\|\mathbf{w}[\tau]-\mathbf{w}^*\|^2+B. \quad (62)
\end{aligned}$$

Now, suppose that $\big\|\mathbf{w}[\tau]-\mathbf{w}^*\big\| \geq c\sqrt{K}$, $\tau=t-1,t$, where c will be specified shortly. Then, we have

$$\frac{1}{\big\|\mathbf{w}[\tau]-\mathbf{w}^*\big\|} \leq \frac{1}{c\sqrt{K}}, \quad \tau=t-1,t.$$

From (62), we that

$$\begin{aligned}
&\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]\} \leq -\frac{2}{\Phi\sqrt{K}}\sum_{\tau=t-1}^{t}\Big[\big\|\mathbf{w}[\tau]-\mathbf{w}^*\big\|\frac{\big\|\mathbf{w}[\tau]-\mathbf{w}^*\big\|}{\sqrt{K}}+\frac{B}{2}\Big] \\
&= -\frac{2}{\Phi\sqrt{K}}\sum_{\tau=t-1}^{t}\big\|\mathbf{w}[\tau]-\mathbf{w}^*\big\|\left[\frac{\big\|\mathbf{w}[\tau]-\mathbf{w}^*\big\|}{\sqrt{K}}+\frac{B\Phi\sqrt{K}}{2\big\|\mathbf{w}[\tau]-\mathbf{w}^*\big\|}\right] \\
&\leq -\frac{2}{\Phi\sqrt{K}}\sum_{\tau=t-1}^{t}\big\|\mathbf{w}[\tau]-\mathbf{w}^*\big\|\left(c-\frac{B\Phi}{2c}\right). \quad (63)
\end{aligned}$$

Therefore, by choosing $c > \sqrt{(1/2)BL}$, we have

$$\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]\} \leq -\frac{2\hat{\delta}}{\Phi\sqrt{K}}\sum_{\tau=t-1}^{t}\big\|\mathbf{w}[\tau]-\mathbf{w}^*\big\| \qquad (64)$$

for some $\hat{\delta} = c-\frac{BL}{2c} > 0$. Plugging in $c > \sqrt{(1/2)B\Phi}$ to define a ball $\mathcal{B} \triangleq \{\mathbf{w} : \big\|\mathbf{w}-\mathbf{w}^*\big\| \leq \sqrt{(1/2)B\Phi K}\}$, we have

$$\begin{aligned}
\mathbb{E}\{\Delta V(\mathbf{z}[t])|\mathbf{z}[t]\} \leq -\frac{\delta}{\sqrt{K}}\Big[\big\|\mathbf{w}[t]-\mathbf{w}^*\big\|+ \\
\|\mathbf{w}[t-1]-\mathbf{w}^*\|\Big], \text{ if } \mathbf{w}[t] \in \mathcal{B}^c. \quad (65)
\end{aligned}$$

where $\delta \triangleq \frac{2\hat{\delta}}{\Phi}$. On the other hand, when $\big\|\mathbf{w}[\tau]-\mathbf{w}^*\big\| \leq \sqrt{(1/2)B\Phi K}$, $\tau=t-1,t$, it trivially holds that

$$-\frac{\delta}{\sqrt{K}}\big\|\mathbf{w}[t]-\mathbf{w}^*\big\| \leq \eta, \quad \tau=t-1,t, \qquad (66)$$

for some $\eta > 0$. Combining (65) and (66) yields the result stated in Proposition 1. This completes the proof.