Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

On the Taylor Expansion of Value Functions

Anton Braverman, Itai Gurvich, Junfei Huang

To cite this article:
Anton Braverman, Itai Gurvich, Junfei Huang (2020) On the Taylor Expansion of Value Functions. Operations Research 68(2):631-654. https://doi.org/10.1287/opre.2019.1903

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 68, No. 2, March–April 2020, pp. 631–654
ISSN 0030-364X (print), ISSN 1526-5463 (online)

On the Taylor Expansion of Value Functions

Anton Braverman,[a] Itai Gurvich,[b] Junfei Huang[c]

[a] Kellogg School of Management, Northwestern University, Evanston, Illinois 60208; [b] Cornell School of Operations Research and Information Engineering and Cornell Tech, New York, New York 10044; [c] Department of Decision Sciences and Managerial Economics, CUHK Business School, Chinese University of Hong Kong, Shatin, Hong Kong

Contact: anton.braverman@kellogg.northwestern.edu, http://orcid.org/0000-0003-4030-3172 (AB); gurvich@cornell.edu, http://orcid.org/0000-0001-9746-7755 (IG); junfeih@cuhk.edu.hk, http://orcid.org/0000-0002-3764-354X (JH)

Received: April 13, 2018
Revised: December 25, 2018; June 6, 2019
Accepted: June 28, 2019
Published Online in Articles in Advance: March 4, 2020

Subject Classifications: dynamic programming/optimal control: Markov: finite state, infinite state; queues: diffusion models
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2019.1903

Abstract. We introduce a framework for approximate dynamic programming that we apply to discrete-time chains on $\mathbb{Z}_+^d$ with countable action sets. The framework is grounded in the approximation of the (controlled) chain's generator by that of another Markov process. In simple terms, our approach stipulates applying a second-order Taylor expansion to the value function, replacing the Bellman equation with one in continuous space and time in which the transition matrix is reduced to its first and second moments. In some cases, the resulting equation can be interpreted as a Hamilton–Jacobi–Bellman equation for a Brownian control problem. When tractable, the "Taylored" equation serves as a useful modeling tool. More generally, it is a starting point for approximation algorithms. We develop bounds on the optimality gap—the suboptimality introduced by using the control produced by the Taylored equation. These bounds can be viewed as a conceptual underpinning, analytical rather than relying on weak convergence arguments, for the good performance of controls derived from Brownian approximations. We prove that under suitable conditions and for suitably "large" initial states, (1) the optimality gap is smaller than a $1-\alpha$ fraction of the optimal value, with which $\alpha \in (0, 1)$ is the discount factor, and (2) the gap can be further expressed as the infinite-horizon discounted value with a "lower-order" per-period reward. Computationally, our framework leads to an "aggregation" approach with performance guarantees. Although the guarantees are grounded in partial differential equation theory, the practical use of this approach requires no knowledge of that theory.

Funding: The work of Itai Gurvich is supported by the National Science Foundation [CMMI-1662294].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2019.1903.

Keywords: Taylor expansion • Markov decision process • Hamilton–Jacobi–Bellman equation • approximate dynamic programming • approximate policy iteration

1. Introduction

Dynamic programming is the primary tool for solving optimization problems in which decisions are subject to dynamic changes in the system state. It is used in the study and practice of a variety of applications.

Deriving structural insights is typically a challenge, and computationally, the number of calculations grows exponentially with the size of the state space (the infamous "curse of dimensionality"). As in other classes of optimization problems—combinatorial, stochastic, and so forth—approximations are often the only way to gain modeling and computational tractability for large problems. The computational challenge has motivated the development of approximate dynamic programming methods (see, e.g., the books by Powell (2007) and Bertsekas (2007)). As a modeling tool, Brownian approximations have made inroads across multiple disciplines, notably in economics, operations management, and electrical engineering. They often capture structural relationships that are inaccessible in the original, "too" detailed dynamic programming problem. Yet, for a variety of reasons, these have not been widely used as a way to reduce computational complexity.

What we add is an approximation to dynamic programs that is inspired by *perturbation* techniques that were recently developed for the approximation of stationary queues by "Brownian queues" (see Gurvich (2014), Braverman and Dai (2017), Huang and Gurvich (2018), and the additional discussion as follows). The seeds of the idea for extending these methods from performance analysis to optimal control appear in Ata and Gurvich (2012) and Huang and Gurvich (2018). This paper seeks to expand those ideas—applied to queues in heavy traffic—into an accessible and generalizable framework. The initial step in our approach is intuitively straightforward: we formally replace the value function in the optimality (a.k.a. Bellman) equation with its second-order Taylor expansion to obtain an equation considered over a

continuous state space. As an example, consider a discrete time and space Markov chain on $\mathbb{Z}$ collecting a reward $r(x)$ when visiting state x and making transitions following the stochastic matrix $P \equiv P_{x,y}$. Fixing $\alpha \in (0,1)$, the infinite-horizon discounted reward

$$V(x) = \mathbb{E}_x\left[\sum_{t=0}^{\infty} \alpha^t r(X_t)\right], \quad x \in \mathbb{Z},$$

satisfies the functional equation

$$V(x) = r(x) + \alpha \sum_y P_{x,y} V(y), \quad x \in \mathbb{Z},$$

which can be rewritten as

$$0 = r(x) + \alpha \sum_y P_{x,y}(V(y) - V(x)) - (1-\alpha)V(x), \quad x \in \mathbb{Z}.$$

Applying (formally) a second-order Taylor expansion $V(y) \approx V(x) + V'(x)(y-x) + \frac{1}{2}V''(x)(y-x)^2$, we obtain the *differential* equation

$$0 = r(x) + \alpha\mu(x)V'(x) + \alpha\frac{1}{2}\sigma^2(x)V''(x) - (1-\alpha)V(x), \quad x \in \mathbb{R},$$

where $\mu(x) := \mathbb{E}_x[X_1 - x] = \sum_y P_{x,y}(y-x)$ and $\sigma^2(x) := \mathbb{E}_x[(X_1 - x)^2] = \sum_y P_{x,y}(y-x)^2$.

When it exists, the solution $\widehat{V}$ to this Taylored equation can be interpreted as corresponding to the infinite-horizon discounted reward of a diffusion process with drift $\alpha\mu(x)$, diffusion coefficient $\alpha\sigma^2(x)$, and exponential discounting $e^{-(1-\alpha)t}$. Such an interpretation, although conceptually useful, is not mathematically necessary. Second-order Tayloring leads naturally to bounds in terms of the third derivative of $\widehat{V}$:

$$\left|\widehat{V}(x) - V(x)\right| \le \bar{\jmath}^3 \mathbb{E}_x\left[\sum_{t=0}^{\infty} \alpha^t \left|D^3\widehat{V}(X_t)\right|^*_{X_t \pm \bar{\jmath}}\right],$$

where $|D^3\widehat{V}(X_t)|^*_{X_t \pm \bar{\jmath}}$ is the maximum of the third derivative in a neighborhood of radius $\bar{\jmath}$ around X_t, and $\bar{\jmath}$ is the maximal jump of the Markov chain (see Theorem 1 and Remark 1).

This analysis of performance *evaluation* suggests an approach for *optimization*. Applying the second-order Taylor expansion to the Bellman equation

$$V(x) = \max_{u \in \mathcal{U}(x)}\left\{r(x,u) + \alpha \sum_y P^u_{x,y} V(y)\right\},$$

we obtain a Hamilton–Jacobi–Bellman (HJB) equation (see Section 2). We refer to this equation as a "Taylored" control problem (TCP) to underscore its origins in Tayloring. Formulating the TCP is the first step. The next steps are (1) to translate the Tayloring-induced error into bounds on optimality gaps and (2) to build on Tayloring to propose solution algorithms. In this paper, we focus mostly on step 1. For step 2, we provide a strong starting point: a conceptual framework (TCP equivalence) and initial implications.

For the development of optimality-gap bounds, we draw on the theory of partial differential equations (PDEs) to prove a vanishing-discount and an order-optimality result, both under suitable "smoothness" conditions on the primitives μ, σ^2, and r. For suitably "large" initial conditions, we have the following: (a) as $\alpha \uparrow 1$, the optimality gap shrinks in relative terms proportionally to $(1-\alpha)$, and (b) the gap can be bounded by the infinite-horizon discounted reward with an immediate-reward function that is of a lower polynomial order. It should not come as a surprise that our approach "inherits" some of the challenges and subtleties of PDE theory. This is reflected in the bounds in Theorem 2, which depend on the amount of time that the chain spends in "corners" of the state space.

From a computational perspective, because Tayloring collapses the transition matrices into μ and σ, multiple chains can induce the same TCP; they are TCP-equivalent. We can rely on the TCP to "translate" the original chain to another, more tractable one. The TCP "couples" the two chains and supplies bounds on the approximation error (see Equation (11)).

What we are about to introduce in this paper has intimate connections to and creates a bridge between two somewhat disparate streams of the literature.

Asymptotic Optimality in Queues and Generator Comparisons. Asymptotic optimality arguments in queueing theory typically rely on the machinery of weak convergence to produce a so-called *diffusion approximation*. One starts from the renewal processes (arrival and service completions), which are the building blocks, and applies central limit theorem scaling to the state process and "embeds" the queueing system being studied within a sequence of such. Heavy traffic is imposed by assuming suitable convergence of the arrival rate to infinity or/and the utilization to 100%. One then interprets, in the context of the original system, the policy arising from the "limit" diffusion-control problem. Near optimality is shown by means of convergence arguments along the sequence of queues in heavy traffic.

Our approach is motivated by recent developments in queueing theory pertaining to Stein's method and offers (in applicable cases) a simple alternative with explicit bounds.[1] In performance analysis (i.e., for a given control), Stein's method allows us to bound directly—without resorting to convergence arguments—the (impressive) "proximity" between the stationary distribution of a queueing system and its Brownian approximation by comparing their transition probabilities (or, more precisely, their *generators*)—that of the

Markov chain and that of a suitable diffusion process (see Gurvich 2014, Braverman and Dai 2017). Although the use of the language of generators is mathematically natural, it is simpler and conceptually useful to view this as Taylor expansion applied to equations that characterize stationary performance and/or optimality conditions.

Tayloring *does not go through diffusion approximations*. It applies to the Markov chain as given; no space scaling is used. It is also applied at the level of the value function rather than in the level of the stochastic process. It is these properties that make it relevant in settings in which there is no natural notion of scaling. The absence of scaling simplifies the very construction of the TCP and, in turn, the derivation of optimality-gap bounds (see Section 2.3).

Nevertheless, specializing our results to queueing examples and relating the discount factor to the utilization does shed some light on the nature of our results (see Section 5).

Transitioning from performance analysis, as considered in earlier papers, to *controlled* chains, as we do here, is like considering a family of generators ("indexed" by the control) instead of a single generator. One can interpret the Taylored equation as the HJB equation for a suitable Brownian control problem. The relation we seek to uncover is based not on process-limit theory but rather on first principles, namely, the Tayloring of the value function.

Approximate Dynamic Programming (ADP). Approximate value or policy iteration typically starts with the choice of a function family (a base) from which to construct a candidate value function. The queueing-approximations literature teaches us that as a heuristic, the value function of a suitable Brownian control problem is a good candidate for a base function; such an approach is taken, for example, in Chen et al. (2009). Our analysis supports this approach: we establish that the TCP solution, even taken as the sole item in the base, yields an approximation whose performance is related to properties of a closely related differential equation.

Algorithmically, our *Taylored approximate policy iteration* (TAPI) algorithm is a modification of policy iteration in which the policy evaluation portion of iteration k requires solving a *linear* PDE to get an approximate value function $V^{(k)}$, which is subsequently plugged into a policy improvement step (an optimization problem that does not require the solution of a PDE) to produce $u^{(k+1)}$ and so on. The linear PDE can be solved via finite difference (FD) or other PDE discretization-based solution methods. The coarser the discrete grid, the more efficient is the computation.

An alternative to FD in the implementation of TAPI is to build on the Taylored equation as an intermediate step—a translator—between Bellman equations corresponding to two *TCP-equivalent* chains, that is, that induce the same TCP. Given a controlled chain, one possible construction of a TCP-equivalent one is inspired by the transformation put forth in Kushner and Dupuis (2013) and Dupuis and James (1998). Their construction relates the differential equation to a control problem for a Markov chain, henceforth referred to as the "K-D chain"—one with a smaller state space and a simpler transition structure. In contrast with the infinitesimal view inherent to the K-D approach (in which one takes the discretization to zero to approximate continuous state space), we use it with *coarse discretization* so that the new Bellman equation can be viewed as an aggregation method in which the state space is reduced to a coarser grid of "super states" (for existing aggregation ideas see, e.g., Bertsekas (2007, chapter 6)). Concurrent work (Zhang and Gurvich 2018) builds on the observed connection to aggregation to develop scalable algorithms based on Tayloring (see Remark 8).

Our Tayloring approach to approximate dynamic programming stands on strong mathematical footing. The gap introduced by using a TCP-equivalent chain can be bounded via the (suitably integrated) third derivative of the PDE solution. From a computational viewpoint, although the algorithm that we propose is not entirely immune to the curse of dimensionality, it pushes computational barriers. Ultimately, we believe that the analysis put forth here can enhance existing ADP algorithms by facilitating a rigorous (rather than ad hoc) choice of "design parameters" (Remark 8 hints at the plausibility of this pursuit).

This paper introduces the framework and provides analytical support and initial numerical evidence. Extensions to continuous chains and other criteria (e.g., long-run average) as well as a full account of algorithms and computational benefits are left for future work (see Section 6 and Remark 8).

1.1. Notation

We use the standard notation $\mathbb{R}^d_+$ for the positive orthant in $\mathbb{R}^d$ and $\mathbb{R}^d_{++}$ for its interior—the space of strictly positive d-dimensional vectors. We use $d(x,y)$ to denote the Euclidean distance between two points x and y and $d(x,\Omega) = \inf_{y\in\Omega} d(x,y)$ to denote the distance from x to a set $\Omega \subseteq \mathbb{R}^d_+$. For a set $\Omega \subseteq \mathbb{R}^d_+$, $\partial\Omega$ denotes its boundary. In particular, $\partial\mathbb{R}^d_+ = \mathbb{R}^d_+ \setminus \mathbb{R}^d_{++} = \cup_{i=1}^d \mathcal{B}_i$, where $\mathcal{B}_i := \{x \geq 0 : x_i = 0\}$. The standard Euclidean norm is denoted by $|\cdot|$, and for $x \in \mathbb{R}^d_+$ and $\epsilon > 0$, we denote by $x \pm \epsilon$ the *set* $\{y \in \mathbb{R}^d_+ : |y - x| \leq \epsilon\}$. For a function $f : \mathbb{R}^d_+ \to \mathbb{R}$ and a subset $\Omega \subseteq \mathbb{R}^d_+$, we let $|f|^*_\Omega = \sup_{y\in\Omega} |f(y)|$ and for $\beta \in (0,1]$,

$$[f]^*_{\beta,\Omega} = \sup_{y,z\in\Omega} \frac{|f(y) - f(z)|}{|y - z|^\beta}.$$

If f is twice continuously differentiable, we write $f_i(x) = \frac{\partial}{\partial x_i} f(x)$ and $f_{ij}(x) = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$. We use $Df(x)$ for the gradient vector (whose elements are $f_i(x)$) and $D^2 f(x)$ for the Hessian matrix (whose elements are $f_{ij}(x)$). We use the standard notation $\mathscr{C}^2(\Omega)$ for the family of twice continuously differentiable functions over Ω, and $\mathscr{C}^{2,\beta}(\Omega)$ is the subset of $\mathscr{C}^2(\Omega)$ whose members have a second derivative that is Hölder continuous on Ω with exponent β; $\beta = 1$ corresponds to Lipschitz continuity on Ω. In this paper, when we speak of a *solution* to a differential equation, we mean that in the classical sense.

For $j \in \{1, 2, \dots, \}$, we use $[j]$ to denote the set of integers $\{1, \dots, j\}$. Throughout, to simplify notation, we use γ, Γ to denote Hardy-style constants that may change from one line to the next and that do not depend on the discount factor α or on the state x.

1.2. Outline of the Paper

Section 2 comprises the mathematical portion of this paper. In it, we introduce the Taylored control equation and state the key results in Theorems 1 and 2 and Corollaries 1 and 2. Sections 3.1 and 4 discuss computation and include numerical experiments. Section 5 briefly explores the connection to the heavy-traffic approximations of queues. All the proofs can be found in the e-companion.

2. Tayloring the Bellman Equation

Consider an infinite-horizon discounted Markov decision process (MDP) on $\mathbb{Z}_+^d$:

$$V_*^\alpha(x) := \max_U \mathbb{E}_x^U \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, U(X_t)) \right], \ x \in \mathbb{Z}_+^d,$$

where $r(x,u)$ is the reward collected at state x under a control u. A stationary policy $U = \{U(x), x \in \mathbb{Z}_+^d\}$ has the property that $U(x) \in \mathcal{U}(x)$, where $\mathcal{U}(x)$ is the set of actions allowed in state x. We assume that $\mathcal{U}(x)$ is *discrete* (possibly countably infinite), and it is an intersection of a polyhedron (that can depend on x) and a discrete set $\mathbb{D}$ that does not depend on x; that is, for some d_a, $\mathcal{U}(x) = \{u \in \mathbb{R}^{d_a} : Au \le b(x)\} \cap \mathbb{D}$, where $b(x)$ is defined for all $x \in \mathbb{R}_+^d$. We let $\mathbb{U} = \times_{x \in \mathbb{Z}_+^d} \mathcal{U}(x)$.

Given $x \in \mathbb{Z}_+^d$ and $u \in \mathcal{U}(x)$, we write $r_u(x) = r(x,u)$ and let $P_{x,y}^u$ be the probability of transitioning from x to y under an action $u \in \mathcal{U}(x)$. We write $\mathbb{E}_x^U[\cdot]$ for the expectation with respect to the law of the U-controlled Markov chain $(X_t, t \ge 0)$ with the initial state x; $\mathbb{E}_x^u[\cdot]$ is the expectation with respect to the law $P_{x,\cdot}^u$.

Under standard conditions (e.g., Bertsekas 2007, section 1.4), $V_*^\alpha(x)$ solves the Bellman equation

$$\begin{aligned} V(x) &= \max_{u \in \mathcal{U}(x)} \{ r_u(x) + \alpha P^u V(x) \} \\ &= \max_{u \in \mathcal{U}(x)} \{ r_u(x) + \alpha \mathbb{E}_x^u [V(X_1)] \}, \end{aligned} \tag{1}$$

where we use the operator notation $P^u V(x) = \sum_y P_{x,y}^u V(y) = \mathbb{E}_x^u[V(X_1)]$. Subtracting $V(x)$ on both sides of the Bellman equation, we have

$$0 = \max_{u \in \mathcal{U}(x)} \{ r_u(x) + \alpha(P^u V(x) - V(x)) - (1-\alpha) V(x) \}, \quad x \in \mathbb{Z}_+^d. \tag{2}$$

Pretending that V is extendable to $\mathbb{R}_+^d$ and twice-continuously differentiable there, we have

$$\begin{aligned} ((P^u - I)V)(x) &= \sum_y P_{x,y}^u V(y) - V(x) \\ &\approx \sum_y P_{x,y}^u \left(\sum_i V_i(x)(y_i - x_i) \right. \\ &\qquad \left. + \frac{1}{2} \sum_{i,j} V_{ij}(x)(y_i - x_i)(y_j - x_j) \right). \end{aligned}$$

Defining

$$\begin{aligned} (\mu_u)_i(x) &:= \mathbb{E}_x^u [(X_1)_i - x_i] \\ &= \sum_y P_{x,y}^u (y_i - x_i), \ i \in [d], \quad \text{and} \end{aligned} \tag{3}$$

$$\begin{aligned} (\sigma_u^2)_{ij}(x) &:= \mathbb{E}_x^u \left[((X_1)_i - x_i) \left((X_1)_j - x_j \right) \right] \\ &= \sum_y P_{x,y}^u (y_i - x_i)(y_j - x_j), \ i,j \in [d], \end{aligned} \tag{4}$$

for all $x \in \mathbb{Z}_+^d$ and $u \in \mathcal{U}(x)$ (and extending these to $\mathbb{R}_+^d$; see the discussion after Assumption 1), we arrive at

$$\begin{aligned} &r_u(x) + \alpha(P^u V(x) - V(x)) - (1-\alpha)V(x) \\ &\quad \approx r_u(x) + \alpha \mathcal{L}_u V(x) - (1-\alpha)V(x), \ x \in \mathbb{R}_+^d, \end{aligned}$$

(2nd order Taylor)

where

$$\begin{aligned} \mathcal{L}_u V(x) &= \sum_i (\mu_u)_i(x) V_i(x) + \frac{1}{2} \sum_{i,j} (\sigma_u^2)_{ij}(x) V_{ij}(x) \\ &= \mu_u(x)' DV(x) + \frac{1}{2} \operatorname{trace}(\sigma_u^2(x)' D^2 V(x)). \end{aligned}$$

This suggests, heuristically at this stage, replacing (1) with

$$0 = \max_{u \in \mathcal{U}(x)} \{ r_u(x) + \alpha \mathcal{L}_u V(x) - (1-\alpha) V(x) \}, \ x \in \mathbb{R}_+^d. \tag{5}$$

A solution to (5) is a pair $(\widehat{U}_*^\alpha(x), \widehat{V}_*^\alpha(x))$, where $\widehat{U}_*^\alpha(x)$ is the maximizer. The restriction of the maximizer $\widehat{U}_*^\alpha(x)$ to $\mathbb{Z}_+^d$ gives a feasible control for the original chain, allowing us to refer to the $\widehat{U}_*^\alpha$-controlled chain.

Implicit in this derivation is an extension of r_u, μ_u, and σ_u^2 from $\mathbb{Z}_+^d$ to $\mathbb{R}_+^d$. We require that the primitives have natural extensions from $\{(x,u) : x \in \mathbb{Z}_+^d, u \in \mathcal{U}(x)\}$ to $\{(x,u) : x \in \mathbb{Z}_+^d, u \in \mathbb{D}\}$.

Assumption 1 (Primitives). *There exist functions*

$$f_r(x,u), \ f_\mu(x,u), \ f_\sigma(x,u), \ x \in \mathbb{Z}_+^d, u \in \mathbb{D} \tag{6}$$

such that r_u, μ_u, and σ_u^2 are the restrictions of these functions to $x \in \mathbb{Z}_+^d$ and $u \in \mathcal{U}(x)$ and satisfy the following properties: (a) *f_r is locally Lipschitz in $\mathbb{Z}_+^d$ (uniformly in u), and* (b) *the functions f_μ and f_σ are globally bounded and Lipschitz uniformly in u; that is, there exists $L > 0$ (not depending on u) such that*

$$\left|f_\mu(\cdot,u)\right|^*_{\mathbb{Z}_+^d} + \left[f_\mu(\cdot,u)\right]^*_{1,\mathbb{Z}_+^d},\ \left|f_\sigma(\cdot,u)\right|^*_{\mathbb{Z}_+^d} + \left[f_\sigma(\cdot,u)\right]^*_{1,\mathbb{Z}_+^d} \le L.$$

Finally, (c) *$f_\sigma(x,u)$ (and, in turn, its restriction $\sigma_u^2(x)$) satisfies the ellipticity condition: there exists $\lambda > 0$ (not depending on u) such that*

$$\lambda^{-1}|\xi|^2 \ge \sum_{i,j} \xi_i \xi_j \left(f_\sigma\right)_{ij}(x,u) \ge \lambda|\xi|^2,$$
$$\text{for all } \xi \in \mathbb{R}^d,\ x \in \mathbb{Z}_+^d, u \in \mathbb{D}. \qquad \text{(elliptic)}$$

Under the Lipschitz requirement in Assumption 1, the McShane–Whitney extension theorem (McShane 1934) constructs an explicit extension to $\mathbb{R}_+^d$ that is itself Lipschitz continuous with the same constant L (or locally Lipschitz in the case of f_r). It is sometimes convenient to leave a discontinuity at the boundary (see the oblique-derivative (OD) boundary condition as follows and Example 1). Importantly, the computational algorithm in Sections 3.1 and 4 relies on the extension to $\mathbb{R}_+^d$ only (if at all) on the boundary. Continuity properties of these extensions do matter for our analytical results. *Henceforth, $f_r(\cdot,u), f_\mu(\cdot,u)$, and $f_\sigma(\cdot,u)$ are the extensions to $\mathbb{R}_+^d$.*

Finally, because every discrete state space can be embedded in $\mathbb{Z}_+$, it is fair to ask what the requirements that we impose on the original problem are. As in Assumption 1, these requirements are stated as constraints on μ and σ^2. Our optimality-gap bounds require, for example, that the optimally controlled chain has bounded jumps (see Theorems 1 and 2). The bound's magnitude, in turn, depends on the maximal jump size as it depends on the Lipschitz constant L in Assumption 1. The embedding of a two-dimensional chain into one dimension might induce μ_u and/or a maximal jump size that are significantly larger than in the original two-dimensional model.

Our approach is thus relevant to settings in which (1) there is a natural meaningful metric on the state space so that μ can be interpreted as the average step size starting at x, (2) one can speak of large and small initial states, and (3) boundaries have physical meaning. Thus, for example, inventory and queuing problems are natural candidates for this approach, but a Markov chain in which the states are colors or letters might not be.

2.1. Boundary Conditions

Equation (5), although well defined, poses a challenge insofar as we want to apply existing PDE theory as collected, for example, in Gilbarg and Trudinger (2001) and Lieberman (2013). The theory covers mostly first-order conditions on the boundary, that is, those in which either DV or V appear but not D^2V. We consider two such conditions: (1) first-order Taylor-ing (FOT) and (2) an oblique-derivative condition that supports second-order Tayloring on the boundary.

FOT Boundary. Applying first-order Tayloring in boundary states, that is, replacing $V(y) - V(x) \leftarrow DV(x)'(y-x)$ for $x \in \partial\mathbb{R}_+^d$, leads to

$$0 = \max_{u \in \mathcal{U}(x)} \{r_u(x) + \alpha \mathcal{L}_u V(x) - (1-\alpha)V(x)\},\ x \in \mathbb{R}_{++}^d,$$
$$0 = \max_{u \in \mathcal{U}(x)} \{r_u(x) + \alpha \mu_u(x)' DV(x) - (1-\alpha)V(x)\},\ x \in \partial\mathbb{R}_+^d.$$

We say that the FOT boundary condition is *control independent* if $\mu_u(x) \equiv \mu(x)$ for all $x \in \partial\mathbb{R}_+^d$. In that case, the maximizer on the boundary $\widehat{U}_*(x), x \in \partial\mathbb{R}_+^d$ does not depend on the value of $\widehat{V}_*$ and $D\widehat{V}_*$ there.

OD Boundary. Under certain assumptions on the behavior of μ near the boundary, certain first-order boundary conditions imply that (5) also holds (as a second-order equation) on the boundary. Informally, suppose that there exists a vector $\eta(x)$ such that for y close to a boundary point $x \in \partial\mathbb{R}_+^d$ and all $u \in \mathbb{D}$,

$$f_\mu(y,u) - f_\mu(x,u) \propto \eta(x),$$

that is, that the boundary change in the drift is approximately proportional to η ($f_\mu(y,u) - f_\mu(x,u) \approx \varphi(u)\eta(x)$ for some real-valued function $\varphi(u)$). Then $(\widehat{U}_*, \widehat{V}_*)$ with $\widehat{V}_* \in \mathcal{C}^2(\mathbb{R}_+^d)$ that solves the *OD-boundary TCP*

$$0 = \max_{u \in \mathcal{U}(x)} \{r_u(x) + \alpha \mathcal{L}_u V(x) - (1-\alpha)V(x)\},\ x \in \mathbb{R}_{++}^d, \quad (7)$$
$$0 = \eta(x)' DV(x),\ x \in \partial\mathbb{R}_+^d, \quad (8)$$

also solves the second-order TCP (5). See Lemma EC.1 in the e-companion for the formal statement.

These mapping alternatives emphasize, in particular, the flexibility there is in constructing the TCP. Because the Markov chain and its *discrete* state space are fixed, we have some freedom in designing extensions near the boundary and, in turn, determining the boundary conditions. In all our examples, the reader will notice, OD boundary conditions arise naturally.

An advantage of the TCP with OD boundary condition is its interpretability as the HJB of a control problem for a reflected diffusion (see, e.g., Borkar and Budhiraja 2004). The FOT boundary, in contrast, imposes fewer structural requirements. Although queuing settings provide an intuitive way to identify η (see Example 2 and Section 4.3), FOT is more direct

and requires less context-specific expertise. It does come, however, at the cost of weaker bounds (see Remark 3).

Example 1 (A Discrete-Time Single-Server Queue)**.** Consider a controlled random walk on $\mathbb{Z}_+$, where, for $x \geq 1$, $P^u_{x,x-1} = u, P^u_{x,x+1} = 1-u$ and $P^u_{0,1} \equiv 1$. We take $\mathcal{U}(x) = \mathbb{D} = [0,1] \cap \mathbb{Q}$ ($\mathbb{Q}$ denotes the rational numbers) for all $x \in \mathbb{Z}_+$. Then $\mu_u(x) = 1-2u =: f_\mu(x,u)$ for $x \geq 1$ and $f_\mu(0,u) = 1$. Also, $\sigma^2_u(x) \equiv 1 =: f_\sigma(x,u)$. We use a reward function that penalizes for large states (holding cost) and for speedy service (effort cost) $r_u(x) = -x^4 - \frac{c_s}{1-u}$, where $c_s > 0$.

We use the discontinuous extension for $f_\mu(x,u)$ that has $f_\mu(x,u) = 1-2u$ for all $x > 0$ and $f_\mu(0,u) = 1$ so that $f_\mu(0+,u) - f_\mu(0,u) = -2u \propto -1$ and the OD boundary condition is $V'(0) = 0$. This condition—familiar from performance equations for reflected Brownian motion (Harrison 2013, section 6.3)—finds a natural justification in Lemma EC.1 in the e-companion: if a solution $(\widehat{V}^\alpha_*, \widehat{U}_*)$ to

$$0 = \max_{u \in \mathcal{U}(x)} \Big\{ r_u(x) + \alpha(1-2u)V'(x) + \frac{\alpha}{2}V''(x) - (1-\alpha)V(x) \Big\}, \quad x > 0,$$
$$0 = V'(0),$$

has $\widehat{U}_*$ that is continuous at $x = 0$, then this solution satisfies (5) at $x = 0$.

In this example, the FOT boundary condition reduces to the (control-independent) equation

$$0 = \max_{u \in \mathbb{D}} \{ r_u(0) + \alpha V'(0) - (1-\alpha)V(0) \} = -c_s + \alpha V'(0) - (1-\alpha)V(0).$$

2.1.1. State-Space Truncation and Boundary Conditions.

The discussion of boundary conditions is unnecessary if the state space is $\mathbb{Z}^d$—as in the inventory example in Section 4. But even in these cases, computation requires truncating the state space, making boundary conditions relevant.

We impose the truncation of $\mathbb{Z}^d_+$ to a square $\mathbb{S}_M = \{x \in \mathbb{Z}^d_+ : \max_i x_i \leq M\}$.[2] The boundary conditions for the TCP depend on the way in which we define the transition probabilities on these artificial boundaries. It is natural to define the transition probabilities for $x \in \mathbb{S}_M$ by

$$\tilde{P}^u_{x,y} = \begin{cases} 0, & \text{for } y \notin \mathbb{S}_M, \\ \frac{P^u_{x,y}}{\sum_{z \in \mathbb{S}_M} P^u_{x,z}}, & \text{otherwise.} \end{cases}$$

In the random walk of Example 1, this simply means $\tilde{P}^u_{M,M+1} = 0$ and $\tilde{P}^u_{M,M-1} = 1$, which leads naturally to the OD boundary condition $V'(M) = 0$.

2.2. The Initial Tayloring Bound

In what follows, for a fixed stationary policy U and a function $f : \mathbb{Z}^d_+ \to \mathbb{R}$, we write

$$V^\alpha_U[f](x) = \mathbb{E}^U_x \left[\sum_{t=0}^\infty \alpha^t f(X_t) \right].$$

We drop the argument f when the immediate reward function is $r_u(x)$ and clear from the context. Thus, for example, $V^\alpha_{\widehat{U}_*}(x)$ is the value under the policy $\widehat{U}_*$ with the reward function $r_u(x)$.

Given a stationary policy U, we define j_U to be the smallest integer (allowing for infinity) such that for all $x, y \in \mathbb{Z}^d_+$ with $|y-x| > \mathrm{j}_U$, $\mathbb{P}^{U(x)}_{x,y} = 0$. We say that the chain has uniformly bounded jumps if

$$\bar{\mathrm{j}} := \sup_{U \in \mathbb{U}} \mathrm{j}_U < \infty.$$

The controls $\widehat{U}_*$ and U_* are likely to depend on α, but for notational convenience, we do not make this dependence explicit. For the following result, recall that $V^\alpha_*(x)$ is the (exact) optimal value that solves the Bellman equation (1) and U_* is the optimal control; V^α_U is the value under a fixed (not necessarily optimal) control U.

Theorem 1 (Initial Bound with Second-Order Tayloring at the Boundary)**.** *Fix $\alpha \in (0,1)$ and suppose that there exists a solution $(\widehat{U}_*, \widehat{V}_*)$ to (5) with $\widehat{V}_* \in \mathcal{C}^{2,\beta}(\mathbb{R}^d_+)$ for some $\beta \in (0,1]$. Suppose further that $\mathrm{j}_{\widehat{U}_*}, \mathrm{j}_{U_*} < \infty$ and that $|\widehat{V}_*(x)| \leq \Gamma(1+|x|^m)$ for some m and Γ (that can depend on α). Then, for $x \in \mathbb{Z}^d_+$,*

$$\Big(\Big| \widehat{V}_*(x) - V^\alpha_*(x) \Big| \vee \Big| V^\alpha_{\widehat{U}_*}(x) - V^\alpha_*(x) \Big| \Big) \leq \mathrm{j}^{2+\beta}_{\widehat{U}_*} \vee \mathrm{j}^{2+\beta}_{U_*} \left(\mathbb{E}^{\widehat{U}_*}_x \left[\sum_{t=0}^\infty \alpha^t \Big[D^2 \widehat{V}_* \Big]^*_{\beta, X_t \pm \mathrm{j}_{\widehat{U}_*}} \right] + \mathbb{E}^{U_*}_x \left[\sum_{t=0}^\infty \alpha^t \Big[D^2 \widehat{V}_* \Big]^*_{\beta, X_t \pm \mathrm{j}_{U_*}} \right] \right). \quad (9)$$

This first theorem states that using the control derived from the TCP produces an optimality gap that is bounded by a suitable "integrated" higher derivative of the TCP solution. For the bound to take explicit meaning, two things must be addressed: (1) the theorem assumes the existence of a smooth solution and leaves unexplored the dependence of the Hölder coefficient $[D^2 \widehat{V}_*]^*_{\beta, X_t \pm \mathrm{j}_{U_*}}$ on the state x and the discount factor α, and (2) the right-hand side of (9) depends on the optimal control U_*, which is the very thing we want to avoid computing. We address these issues in Theorem 2 and its Corollaries 1 and 2. There we develop explicit bounds that relate the right-hand side directly to V^α_* and establish, roughly speaking, that $|V^\alpha_{\widehat{U}_*}(x) - V^\alpha_*(x)| = o(V^\alpha_*(x))$.

Remark 1 (Performance Approximation)**.** We make the obvious observation that Theorem 1 applies as well to the performance analysis of a given control. Fixing a control U is the same as taking control sets $\mathcal{U}(x)$ that contain the single action $U(x)$. Equation (9) reduces to

$$\left|\widehat{V}(x) - V_U^\alpha(x)\right| \le \mathrm{j}_U^{2+\beta}\mathbb{E}_x^U\left[\sum_{t=0}^{\infty}\alpha^t\left[D^2\widehat{V}\right]^*_{\beta,X_t\pm\mathrm{j}_U}\right]. \tag{10}$$

In this case, the TCP is a linear PDE. □

Remark 2 (Unbounded Jumps)**.** The bound can be easily adjusted to unbounded jumps with suitable finite moments. In this case, the right-hand side of (9) takes the form

$$\mathbb{E}_x^{\widehat{U}_*}\left[\sum_{t=0}^{\infty}\alpha^t\mathbb{E}_{X_t}\left[|\Delta_{X_t}|^{2+\beta}\left[D^2\widehat{V}\right]^*_{\beta,X_t\pm|\Delta_{X_t}|}\right]\right] + \mathbb{E}_x^{U_*}\left[\sum_{t=0}^{\infty}\alpha^t\mathbb{E}_{X_t}\left[|\Delta_{X_t}|^{2+\beta}\left[D^2\widehat{V}\right]^*_{\beta,X_t\pm|\Delta_{X_t}|}\right]\right],$$

where $\Delta_{X_t} = X_{t+1} - X_t$ (see the proof of Theorem 1). □

Remark 3 (FOT Boundary)**.** With first-order Tayloring on the boundary, (9) is replaced with

$$\left(\left|\widehat{V}_*(x) - V_*^\alpha(x)\right| \vee \left|V_{\widehat{U}_*}^\alpha(x) - V_*^\alpha(x)\right|\right) \le \mathrm{j}^2_{\widehat{U}_*} \vee \mathrm{j}^2_{U_*}\left(\mathbb{E}_x^{\widehat{U}_*}\left[\sum_{t=0}^{\infty}\alpha^t\left|\mathfrak{e}\left[\widehat{V}_*,\mathrm{j}_{\widehat{U}_*}\right]\right|^*_{\beta,X_t\pm\mathrm{j}_{\widehat{U}_*}}\right] + \mathbb{E}_x^{U_*}\left[\sum_{t=0}^{\infty}\alpha^t\left|\mathfrak{e}\left[\widehat{V}_*,\mathrm{j}_{U_*}\right]\right|^*_{\beta,X_t\pm\mathrm{j}_{U_*}}\right]\right),$$

where

$$|\mathfrak{e}[f,z]|^*_{\beta,\Omega} = z^\beta\left[D^2 f\right]^*_{\beta,\Omega} + \sum_i \mathbb{1}\{x\in\mathcal{B}_i\}\sum_{j\ne i}|f_{ij}|^*_\Omega.$$

Relative to (9), the second derivative on the boundary factors into the optimality gap. In addition, the Hölder bounds for the second derivative are somewhat weaker in the case of the FOT boundary condition (compare Lemmas 1 (for OD boundary) and EC.2 in the e-companion (for FOT boundary). □

2.2.1. Toward Computability: TCP-Equivalent Chains.

The primitives of the MDP are the reward function(s) r_u, the transition matrices P^u—from which we build μ_u and σ_u^2—and the discount factor $\alpha\in(0,1)$. There are multiple MDPs (or primitives) that induce the same TCP. Specifically, consider an MDP for a controlled chain $\widetilde{X}$ with the same state and action spaces. Let $\{\widetilde{P}^u\}$ be a family of transition matrices and $\tilde{\alpha}(x)\in(0,1)$ be a (possibly state-dependent) discount factor that jointly satisfy the constraints

$$\sum_y \widetilde{P}^u_{x,y}(y_i - x_i) = \frac{\alpha(1-\tilde{\alpha}(x))}{\tilde{\alpha}(x)(1-\alpha)}(\mu_u)_i(x),$$

$$\sum_y \widetilde{P}^u_{x,y}(y_i - x_i)(y_j - x_j) = \frac{\alpha(1-\tilde{\alpha}(x))}{\tilde{\alpha}(x)(1-\alpha)}(\sigma_u^2)_{ij}(x),$$

and take the reward function $\tilde{r}_u(x) = \frac{1-\tilde{\alpha}(x)}{1-\alpha}r_u(x)$.

These "tilde" primitives then induce *the same* TCP as the original primitives. The two chains X and $\widetilde{X}$ are *TCP-equivalent*. Let $\widetilde{U}_*$ be the optimal policy for this new optimal control problem (generating the optimal value $\widetilde{V}_*^\alpha$). It then follows that

$$\left|V_*^\alpha(x) - \widetilde{V}_*^\alpha(x)\right| \le \Gamma\left(\mathbb{E}_x^{\widehat{U}_*}\left[\sum_{t=0}^{\infty}\alpha^t\left[D^2\widehat{V}_*\right]^*_{\beta,X_t\pm\mathrm{j}_{\widehat{U}_*}}\right] + \mathbb{E}_x^{U_*}\left[\sum_{t=0}^{\infty}\alpha^t\left[D^2\widehat{V}_*\right]^*_{\beta,X_t\pm\mathrm{j}_{U_*}}\right] + \mathbb{E}_x^{\widehat{U}_*}\left[\sum_{t=0}^{\infty}\bar{\alpha}^t\left[D^2\widehat{V}_*\right]^*_{\beta,\widetilde{X}_t\pm\mathrm{j}_{\widehat{U}_*}}\right] + \mathbb{E}_x^{\widetilde{U}_*}\left[\sum_{t=0}^{\infty}\bar{\alpha}^t\left[D^2\widehat{V}_*\right]^*_{\beta,\widetilde{X}_t\pm\mathrm{j}_{\widetilde{U}_*}}\right]\right), \tag{11}$$

where $\mathrm{j}_{\widetilde{U}_*}$ is the maximal jump of the chain $\widetilde{X}$ under the policy $\widetilde{U}_*$, $\bar{\alpha} = \sup_{x\in\mathbb{Z}_+^d}\tilde{\alpha}(x)$, and Γ is an appropriate constant that depends on $\mathrm{j}_{\widehat{U}^*}$, j_{U^*}, $\mathrm{j}_{\widetilde{U}^*}$, and β.

Among all TCP-equivalent chains, it is reasonable to look for one that introduces significant computational benefits. There are substantial degrees of freedom in making this choice. The K-D chain that we use for illustration in Section 3.1 has, for example, a state space that is a *strict* subset of that of X. Within so-called *soft aggregation* (see, e.g., Bertsekas 2019), TCP-equivalence can support the choice of an algorithm's design parameters (see Remark 8).

This transition from one Markov chain to a different but TCP-equivalent one does not require solving any continuous state-and-time control problem. The TCP merely serves as the basis for optimality-gap guarantees. What we pursue next is making these guarantees more explicit.

2.3. Explicit Bounds

We open the pursuit of explicit bounds with a simple example that serves to motivate and illustrate two notions of near optimality that we generalize in this section's main results.

Example 2 (The Discrete Queue Revisited)**.** In the setting of Example 1, let us *fix the control* to $U(x)\equiv 1/2$ (see Remark 6 (at the end of this section) and Example EC.1

in the e-companion for the full control version). The OD-boundary TCP is given by

$$0 = -x^4 - \frac{c_s}{1-U(x)} + \alpha(1-2U(x))V'(x) + \alpha\frac{1}{2}V''(x) - (1-\alpha)V(x), \quad x>0,$$
$$0 = V'(0),$$

and admits the unique solution[3]

$$\widehat{V}_U(x) = -\frac{x^4}{1-\alpha} - \frac{6\alpha x^2}{(1-\alpha)^2} - \frac{6\alpha^2}{(1-\alpha)^3} - \frac{2c_s}{1-\alpha}, \ x\geq 0, \tag{12}$$

so that

$$\left[D^2\widehat{V}_U\right]^*_{1,[0,x]} \leq \left|D^3\widehat{V}_U\right|^*_{[0,x]} \leq \frac{24x}{1-\alpha}, \ x\geq 0.$$

Because the maximal jump is 1 ($\bar{\mathrm{j}}=1$), Theorem 1 (with $\beta=1$) and Equation (10) imply that

$$\left|\widehat{V}_U(x) - V^\alpha_U(x)\right| \leq \mathbb{E}^U_x\left[\sum_{t=0}^\infty \alpha^t \left|D^3\widehat{V}_U\right|^*_{X_t\pm 1}\right] \leq 24\mathbb{E}^u_x\left[\sum_{t=0}^\infty \alpha^t \frac{X_t+1}{1-\alpha}\right], \ x\in\mathbb{Z}_+,$$

where, recall, $V^\alpha_U(x)$ is the infinite-horizon discounted reward with the immediate reward r_u and under the policy U. For all $x\geq 0$, $\frac{x}{1-\alpha}\leq (1-\alpha)x^4 + \frac{1}{(1-\alpha)^2}$ so that

$$\begin{aligned}\left|\widehat{V}_U(x) - V^\alpha_U(x)\right| &\leq 24\mathbb{E}^U_x\left[\sum_{t=0}^\infty \alpha^t \frac{X_t+1}{1-\alpha}\right] \\ &\leq \mathbb{E}^U_x\left[\sum_{t=0}^\infty \alpha^t\left((1-\alpha)X_t^4 + \frac{1}{(1-\alpha)^2}\right)\right] + \frac{1}{(1-\alpha)^2} \\ &\leq (1-\alpha)\left|V^\alpha_U(x)\right| + \frac{2}{(1-\alpha)^3}. \end{aligned} \tag{13}$$

We claim that $|V^\alpha_U(x)|\geq \frac{\gamma}{(1-\alpha)^4}$ for all $x\geq \frac{1}{1-\alpha}$ so that

$$\left|\widehat{V}_U(x) - V^\alpha_U(x)\right| \leq \Gamma(1-\alpha)\left|V^\alpha_U(x)\right|, \text{ for all } x\geq \frac{1}{1-\alpha} \tag{14}$$

(see Corollary 1 to Theorem 2 and its proof). Furthermore, because $\frac{x}{1-\alpha}\leq x^3 + \frac{1}{(1-\alpha)^{\frac{3}{2}}}$ for all $x\geq 0$, we have

$$\begin{aligned}\left|\widehat{V}_U(x) - V^\alpha_U(x)\right| &\leq 24\mathbb{E}^u_x\left[\sum_{t=0}^\infty \alpha^t \frac{X_t+1}{1-\alpha}\right] \\ &\leq \mathbb{E}^u_x\left[\sum_{t=0}^\infty \alpha^t\left(X_t^3 + \frac{1}{(1-\alpha)^{\frac{3}{2}}}\right)\right] + \frac{1}{(1-\alpha)^2} \\ &\leq V^\alpha_U[f_3](x) + \frac{2}{(1-\alpha)^{\frac{5}{2}}}, \end{aligned} \tag{15}$$

where $V^\alpha_U[f_3](x)$ is the value under the control U with the "lower-order" cost function $f_3(x)=x^3$ replacing $x^4 + \frac{c_s}{1-u}$. We claim that $V^\alpha_U[f_3](x)\geq \frac{1}{(1-\alpha)^{5/2}}$ for all $x\geq \frac{1}{(1-\alpha)^{5/8}}$, leading to the *order-optimality* result

$$\left|\widehat{V}_U(x) - V^\alpha_U(x)\right| \leq \Gamma V^\alpha_U[f_3](x), \text{ for all } x\geq \frac{1}{(1-\alpha)^{5/8}} \tag{16}$$

(see Corollary 2). The arguments in this example are not the tightest, but they illustrate the generalizable arguments in Section 2.3.

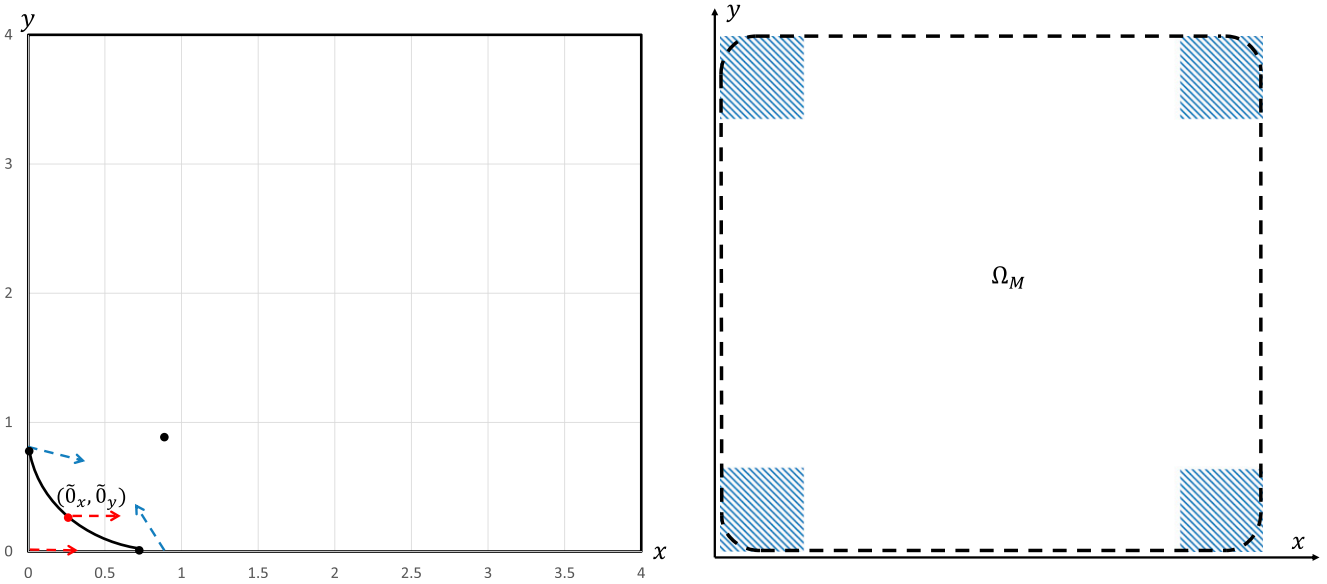
Remark 4 (Impact of Initial State x and Discount Factor α)**.** Let us reconsider (14). First, it is evident that for small values of α, for example, $\alpha=1/2$, this bound becomes weak (see Section EC.1 in the e-companion). Furthermore, when the initial state x is small, the constant portion of the bound, $2/(1-\alpha)^3$, dominates the bound (especially if α is close to 1). It is only when the initial state is sufficiently large—specifically larger than $1/(1-\alpha)$—that the relative error is *proved* to be small relative to the value function. This additive constant term does not make a strong appearance in our numerical control examples. The reason for this might be that in the context of control, the optimal policy U^* and the TCP policy $\widehat{U}^\alpha_*$ are insensitive to constant shifts of the value function.

The order-optimality bound in (16) is, in contrast, also useful for small values of α. It is somewhat weak if x is small because, in that case, $V^\alpha_U[f_3]$ and $V^\alpha_U(x)$ might be of the same order of magnitude. □

Some preliminary construction and definitions are needed for the statement of our bounds.

Smoothing the State Space. PDEs do not, in general, admit classical solutions in domains with corners (see Dupuis and Ishii 1990). Fortunately, the discreteness of the state space facilitates a smoothing of the domain without compromising the bounds. Consider a two-dimensional controlled chain on the "square" state space $\{x\in\mathbb{Z}^2_+ : x_1, x_2\leq M\}$ in Figure 1. We can replace the point 0 with a point $\tilde{0}$—through which we can "pass" a smooth boundary while preserving the transition probabilities and the reward function (see Figure 1). This *does not change* the value function V^α_* or the optimal control U_* of the original chain. It changes the extensions of $r_u(x)$ as well as the values (and extensions) of μ_u and σ^2_u, but notice that these change only at states that connect to the new origin $\tilde{0}$. Notice how the discreteness of the state space facilitates this smoothing by replacing the point 0 with the point $\tilde{0}$ without changing the true value function. This may not be feasible with continuous state space (see the discussion in Section 6).

Figure 1. (Color online) A Two-Dimensional Example of Truncating the State Space and "Curving" the Boundary



Notes. The right-hand side displays the smoothed and truncated domain. The shaded squares are the points outside of Ω_-^{ℓ}.

The boundary of the truncated and smoothed state space is illustrated for $d = 2$ on the right-hand side of Figure 1. We refer to Ω_M as the open subset of $\mathbb{R}_+^d$ defined by this boundary, and $\overline{\Omega_M}$ is its closure. This is the domain on which we consider the PDE (see Remark 9 for a concrete example of domain smoothing).

Corners. We distinguish between corners, such as the point $\tilde{0}$, and one-dimensional boundaries. Recall that $\mathcal{B}_i = \{x \in \mathbb{R}_+^d : x_i = 0\}$. Let us also define $\mathcal{B}_i^{out} = \{x \in \mathbb{R}_+^d : x_i = M\}$. Given a strictly positive function $\varrho(x)$, define the set

$$\Omega_-^{\varrho(\cdot)} := \{x \in \overline{\Omega_M} : d(x, \mathcal{B}_i \cup \mathcal{B}_i^{out}) \le \varrho(x), \text{ for at most one } i\}.$$

The set $\Omega_-^{\varrho(\cdot)}$ includes states that, although possibly close to an axis, are far from points at which two or more axes meet. In the two-dimensional case, taking a constant $\varrho(x) = \ell$ is the same as carving out squares from the corners of Ω_M. Our bounds distinguish between points in $\Omega_-^{\varrho(x)}$ and points outside of this set. In the one-dimensional case ($d = 1$), $\Omega_-^{\rho(\cdot)} = \overline{\Omega_M} = [0, M]$.

Actions Space. Lemma 1 covers action sets that do not depend on the state x. We can accommodate $\mathcal{U}(x) = \{u : Au \le b(x)\} \cap \mathbb{D}$ by introducing a penalty of the form $H[Au - b(x)]^+$ for suitably large H. Notice that the updated reward function that incorporates the penalty—$\bar{r}_u(x) \leftarrow r_u(x) - H[Au - b(x)]^+$—has[4]

$$[\bar{r}_u]_{1,\Omega}^* \le [r_u]_{1,\Omega}^* + H[b]_{1,\Omega}^*.$$

The following is an indirect corollary of the PDE literature. All statements focus on the case of the OD boundary condition. Analogues for the FOT boundary condition appear in the e-companion (see the discussion that precedes Theorem EC.1 there).

Lemma 1. *Let Ω_M be an open-bounded domain as earlier with boundary in $\mathcal{C}^{2,1}$, and suppose that $\mathcal{U}(x) \equiv \mathcal{U} = \mathbb{D}$ and that Assumption 1 holds. Suppose further that $\eta \in \mathcal{C}^{1,1}(\partial\Omega_M)$ with $|\eta|_{\partial\Omega_M}^* + |D\eta|_{\partial\Omega_M}^* + [D\eta]_{1,\partial\Omega_M}^* \le L$ and that there exists $\nu_0 > 0$ such that $\eta(x) \cdot \theta(x) \ge \nu_0 |\eta(x)|$, where $\theta(x)$ is the outward normal at $x \in \partial\Omega_M$. Finally, suppose that $\sup_{u\in\mathcal{U}} |r_u|_{\overline{\Omega_M}}^* + \sup_{u\in\mathcal{U}} [r_u]_{1,\overline{\Omega_M}}^* < \infty$. Then the TCP with the OD boundary condition has a unique solution $\widehat{V}_* \in \mathcal{C}^{2,\beta}(\Omega_M)$ for some $\beta \in (0,1)$ (that does not depend on α), and given a function $\varrho : \mathbb{Z}_+^d \to \mathbb{R}_{++}$, we have the bound*

$$\left[D^2 \widehat{V}_*\right]_{\beta, x \pm \frac{\varrho(x)}{2}}^* \le \Gamma \left(\frac{\left|\widehat{V}_*\right|_{x\pm\varrho(x)}^*}{\varrho^{2+\beta}(x)} + \max_{u\in\mathcal{U}} [r_u]_{\beta, x\pm\varrho(x)}^* \right), \quad x \in \Omega_-^{\varrho(\cdot)}. \tag{17}$$

We also have the global bound

$$\left[D^2 \widehat{V}_*\right]_{\beta, x \pm \frac{\varrho(x)}{2}}^* \le \Gamma \Theta_M, \quad x \in \overline{\Omega_M},$$

where $\Gamma \equiv \Gamma(d, \lambda, L, \nu_0, \partial\Omega_M)$, $\Theta_M := |\widehat{V}_*|_{\overline{\Omega_M}}^* + \max_{u\in\mathbb{D}} [r_u]_{\beta,\overline{\Omega_M}}^*$, *and λ is as in the requirement* (elliptic).

Two comments are due: (1) the bound can depend on α only through $\widehat{V}_*$ and $\varrho(x)$ (if the latter is chosen to depend on α), and (2) in the one-dimensional case, $\Omega_-^{\varrho(\cdot)} = \overline{\Omega_M}$, so the bound in (17) holds in fact in all of $\overline{\Omega_M}$.

Theorem 2 uses Lemma 1 to translate the bounds in Theorem 1—given in terms of infinite-horizon "integration" of the higher-order derivative—to meaningful explicit bounds. For its statement, we define $f_k(x) = |x|^k$ for $k \in \mathbb{N}$ and $x \in \mathbb{R}^d_+$ and let

$$\mathcal{T}_{CO} = \left\{t \geq 0 : X_t \notin \Omega_-^{\varrho(\cdot)}\right\}$$

be the set of times the chain spends close to the corners of Ω_M. This set is, by definition, empty in the one-dimensional case.

Theorem 2 (Explicit Bounds). *Let $(\widehat{U}_*, \widehat{V}_*)$ be a solution to the TCP as in Lemma 1. In addition to the requirements in Lemma 1, suppose that $\bar{\mathrm{j}} < \infty$ and that, for all $x \in \mathbb{R}^d_+$,*

$$\max_{u\in\mathcal{U}}[r_u]^*_{\beta,0\pm|x|} \leq \Gamma\left(1+|x|^{k-\beta}\right), \text{ and}$$
$$\left|\widehat{V}_*\right|^*_{0\pm|x|} \leq \Gamma\left(\frac{1}{(1-\alpha)^m} + \frac{|x|^k}{1-\alpha}\right). \tag{18}$$

Then

$$\left[D^2\widehat{V}_*\right]^*_{\beta,x\pm\bar{\mathrm{j}}} \leq \Gamma\left((1+|x|)^{k-\beta} + \frac{1}{(1-\alpha)^{\frac{m(k-\beta)}{k+2}}} + \frac{1}{(1-\alpha)^{\frac{1}{2}(k-\beta)}}\right),$$

for all $x \in \Omega_-^{\varrho(\cdot)}$. In turn, for a stationary policy $U \in \mathbb{U}$ and all $x \in \overline{\Omega_M}$,

$$\mathbb{E}^U_x\left[\sum_{t\notin\mathcal{T}_{CO}} \alpha^t \left[D^2\widehat{V}_*\right]^*_{\beta,X_t\pm\bar{\mathrm{j}}}\right] \leq \Gamma\left(V^\alpha_U[f_{k-\beta}](x) + \frac{1}{(1-\alpha)^{\frac{m(k-\beta)}{k+2}+1}} + \frac{1}{(1-\alpha)^{\frac{1}{2}(k+2-\beta)}}\right),$$

so

$$\begin{aligned}\left|V^\alpha_{\widehat{U}_*}(x) - V^\alpha_*(x)\right| &\leq \Gamma\Bigg(\max\left\{V^\alpha_{U_*}[f_{k-\beta}](x), V^\alpha_{\widehat{U}_*}[f_{k-\beta}](x)\right\} \\ &\quad + \frac{1}{(1-\alpha)^{\frac{m(k-\beta)}{k+2}+1}} + \frac{1}{(1-\alpha)^{\frac{1}{2}(k+2-\beta)}} \\ &\quad + \Theta_M\mathbb{E}^{\widehat{U}_*}_x\left[\sum_{t\in\mathcal{T}_{CO}}\alpha^t\right] + \Theta_M\mathbb{E}^{U_*}_x\left[\sum_{t\in\mathcal{T}_{CO}}\alpha^t\right]\Bigg),\end{aligned} \tag{19}$$

where $\Theta_M \leq \Gamma\left(\frac{1}{(1-\alpha)^m} + \frac{|M|^k}{1-\alpha}\right)$.

The term in the second line of (19) is the contribution of the corners to the optimality gap. The magnitude of the boundary effect depends on how much time the chain spends near corners. We revisit this point in Section 5.

Remark 5 (A Priori Requirements on $\widehat{V}_*$ and the Value of m). We can gain some insight into the value of m in requirement (18) by considering the one-dimensional case. Suppose that $(\widehat{U}_*, \widehat{V}_*)$ satisfies

$$0 = r_{\widehat{U}_*}(x) + \alpha\mathcal{L}_{\widehat{U}_*}\widehat{V}_*(x) - (1-\alpha)\widehat{V}_*(x), \ x > 0,$$

and $\widehat{V}'_*(0) = 0$. Suppose, further, that (a) $\mu_{\widehat{U}_*}(x) \leq 0$ for all $x > 0$ and that (b) uniformly in x, $0 < \underline{\sigma} \leq \sigma_{\widehat{U}_*}(x) \leq \bar{\sigma} < \infty$. It follows from basic arguments (see Lemma EC.3 in the e-companion) that for $k > 1$,

$$\widehat{V}_*(x) \leq \Gamma\left(\frac{x^k}{1-\alpha} + \frac{1}{(1-\alpha)^{\frac{k+2}{2}}}\right),$$

so we can take $m = (k+2)/2$.

If there exists $\kappa > 0$ such that $\mu_{\widehat{U}_*}(x) \leq -\kappa$ for all $x > 0$, then $m = 1$; that is, $\widehat{V}_*(x) \leq \Gamma((x^k+1)/(1-\alpha))$. If $\mu_{\widehat{U}_*}(x)$ is not necessarily negative but is bounded—$|\mu_{\widehat{U}_*}(x)| \leq \kappa$ (not dependent on α or x)—then we can take $m = k+1$. □

If $r_u(x) \geq \gamma f_k(x)$ for some k and all u, then it trivially holds that $|V^\alpha_U(x)| \geq \gamma V^\alpha_U[f_k](x)$. When this "super-polynomial" property holds, the following corollary shows that the difference between the optimal value and the value induced by the TCP control is at most $(1-\alpha)$ of the true optimal value; that is, the relative gap is $(1-\alpha)$.

Corollary 1 (Vanishing Discount Optimality). *Let $(\widehat{U}_*, \widehat{V}_*)$ be a solution to the TCP, and suppose that the assumptions of Theorem 2 hold with $m \leq k+1$. Then, for every stationary policy U,*

$$\mathbb{E}^U_x\left[\sum_{t\notin\mathcal{T}_{CO}} \alpha^t\left[D^2\widehat{V}_*\right]^*_{\beta,X_t\pm\bar{\mathrm{j}}}\right] \leq \Gamma(1-\alpha)^\beta V^\alpha_U[f_k](x),$$

for all $x : |x| \geq 1/(1-\alpha)$. Consequently, if

$$\begin{aligned}\left|V^\alpha_*(x)\right| = \left|V^\alpha_{U_*}(x)\right| &\geq \gamma V^\alpha_{U_*}[f_k](x) \text{ and} \\ \left|V^\alpha_{\widehat{U}_*}(x)\right| &\geq \gamma V^\alpha_{\widehat{U}_*}[f_k](x), \ x \in \mathbb{Z}^d_+,\end{aligned} \tag{20}$$

for some $\gamma > 0$ that does not depend on α, then

$$\begin{aligned}\left|V^\alpha_{\widehat{U}_*}(x) - V^\alpha_*(x)\right| &\leq \Gamma(1-\alpha)^\beta \max\left\{\left|V^\alpha_*(x)\right|, \left|V^\alpha_{\widehat{U}_*}(x)\right|\right\} \\ &\quad + \Gamma\Theta_M\left(\mathbb{E}^{\widehat{U}_*}_x\left[\sum_{t\in\mathcal{T}_{CO}}\alpha^t\right] + \mathbb{E}^{U_*}_x\left[\sum_{t\in\mathcal{T}_{CO}}\alpha^t\right]\right).\end{aligned}$$

Our second corollary states that the optimality gap is proportional to the value function with a lower-order reward function. Roughly speaking, if the immediate reward function is bounded (in absolute value) by a polynomial of order k, the error is bounded by the infinite-horizon discounted value with an immediate reward that is polynomial of order $k-1$ (recall Example 2).

Corollary 2 (Order Optimality). *Suppose that the assumptions of Theorem 2 hold, and let* $\zeta(m,k) = \frac{1}{k+1-\beta} \cdot \max\left\{\frac{m(k-\beta)}{k+2} + 1, \frac{1}{2}(k+2-\beta)\right\}$. *Then, for all* x *such that* $|x| \geq \frac{1}{(1-\alpha)^{\zeta(m,k)}}$,

$$\left|V^{\alpha}_{\widehat{U}_*}(x) - V^{\alpha}_{*}(x)\right| \leq \Gamma \max\left\{V^{\alpha}_{U_*}[f_{k-\beta}](x), V^{\alpha}_{\widehat{U}_*}[f_{k-\beta}](x)\right\} + \Gamma\Theta_M\left(\mathbb{E}^{\widehat{U}_*}_x\left[\sum_{t\in\mathcal{T}_{CO}} \alpha^t\right] + \mathbb{E}^{U_*}_x\left[\sum_{t\in\mathcal{T}_{CO}} \alpha^t\right]\right).$$

Remark 6 (Example 2 Revisited). The bounds in Example 2 (in which we consider a fixed control) are consistent with Corollaries 1 and 2 taking $m = 3 = k-1$ and $\beta = 1$ so that $\frac{1}{2}(k+2-\beta) = m(k-\beta)/(k+2) + 1 = 2.5$, and the terms depending on $(1-\alpha)$ in the first line of (19) would correspond to $1/(1-\alpha)^{2.5}$.

Although Theorem 2 and its corollaries cover only TCP solutions with $\beta \in (0,1)$, the direct derivation in Example 2 gave us a solution with $\beta = 1$.

Because $r_u(x) \leq -x^4$, the requirements of both Corollaries 1 and 2 are satisfied, and their conclusions apply, allowing us to extend the bounds in Example 2 from performance analysis to optimization. Also, there are no corners in this one-dimensional example, so the second line of (19) is zero. □

3. Computation: Tayloring-Based Approximate Dynamic Programming

The two staple algorithms for solving MDPs are the value and policy iteration algorithms. The curse of dimensionality renders both incapable of solving large-scale problems and motivates the development of approximation algorithms. In this section, we offer a direct algorithmic interpretation of Tayloring. The resulting approximate policy iteration algorithm already offers a reduction in computational effort compared with solving the original MDP; in Section 4.3, for example, we solve a problem in which Taylored approximate policy iteration (TAPI) takes less than 10 minutes for some instances in which the exact solution takes more than 15 hours. This provides evidence for the feasibility of scalable algorithms with optimality-gap bounds that are grounded in our analysis of Tayloring (see Remark 8).

Whereas the state space of the MDP is unbounded, truncating is inevitable for computation. We use $\mathbb{S}$ for the truncated state space, $\widetilde{\mathbb{S}}$ for its continuation, and $\partial\widetilde{\mathbb{S}}$ for the boundary of $\widetilde{\mathbb{S}}$. Mostly, we restrict attention to the case in which the original state space is $\mathbb{Z}^d_+$ and introduce the truncated state space $\mathbb{S} = [0,M]^d \cap \mathbb{Z}^d_+$ and its continuation $\widetilde{\mathbb{S}} = [0,M]^d$.

Because $\mathbb{S}$ is finite, the Bellman equation

$$V_*(x) = \max_{u\in\mathcal{U}(x)}\{r_u(x) + \alpha\mathbb{E}^u_x[V_*(X_1)]\}, \quad x \in \mathbb{S},$$

has a unique solution, and the policy iteration (PI) algorithm is guaranteed to converge to this solution in finitely many iterations.

Algorithm 1 (Standard PI)

1. Start with some initial stationary policy $U^{(0)} \in \mathbb{U}$.
2. For $k = 0, 1, \ldots,$
   (a) *Policy evaluation:* solve for the infinite horizon discounted performance under $U^{(k)}$; that is, find $V^{(k)}(x)$ that satisfies

$$V^{(k)}(x) = r\left(x, U^{(k)}(x)\right) + \alpha\mathbb{E}^{U^{(k)}(x)}_x\left[V^{(k)}(X_1)\right], \quad x \in \mathbb{S}. \tag{21}$$

   (b) *Policy improvement:* Find

$$U^{(k+1)}(x) = \operatorname*{argmax}_{u\in\mathcal{U}(x)}\left\{r_u(x) + \alpha\mathbb{E}^u_x\left[V^{(k)}(X_1)\right]\right\}, \quad x \in \mathbb{S}. \tag{22}$$

The computational bottlenecks of PI are well understood:

Value-function storage: We require $\mathbb{O}(|\mathbb{S}|)$ space ($|\mathbb{S}|$ is the size of the set $\mathbb{S}$) to store $V^{(k)}(x)$, and $|\mathbb{S}|$ may grow exponentially with the dimension of the problem.

Transition-matrix storage: In the kth step of the PI algorithm, we invert $P^{(k)} - I$, where $P^{(k)}$ is the transition probability matrix associated with policy $U^{(k)}(x)$. Depending on density (or sparsity) of this matrix, storing it may require as much as $\mathbb{O}(|\mathbb{S}|^2)$. The value-iteration algorithm does not require such storage.

Optimization complexity: Each iteration includes a policy improvement step: finding (greedy) optimal actions relative to the value-function approximation $V^{(k)}(x)$. For each state $x \in \mathbb{S}$ and action $u \in \mathcal{U}(x)$, computing the expectation $\mathbb{E}^u_x[V^{(k)}(X_1)]$ may require as many as $\mathbb{O}(|\mathbb{S}|)$ function evaluations, depending on the transition structure of the Markov chain. Furthermore, the optimization may require exhaustive search over the discrete action space $\mathcal{U}(x)$ (whose size is denoted by $|\mathcal{U}(x)|$). The total cost of the policy improvement step therefore can be up to $\mathbb{O}(|\mathcal{U}|_{max}|\mathbb{S}|^2)$, where $|\mathcal{U}|_{max} = \sup_{x\in\mathbb{S}}|\mathcal{U}(x)|$ is an upper bound on the number of feasible actions. Our example in Section 4.3 is one in which this worst-case cost is realized.

3.1. TAPI

The basic idea in approximate policy iteration (API) is to produce an approximation of $V^{(k)}(x)$ for the value function at iteration k and then use it in the policy improvement step. Linear architecture is a popular approximation scheme that uses an element of the space $\{\Phi r | r \in \mathbb{R}^F\}$, that is, functions of the form $\hat{V}(x) = \sum_{i=1}^F r_i\Phi_i(x)$, where Φ is the $|\mathbb{S}| \times F$ matrix that collects the so-called feature vectors $\Phi_i : \mathbb{S} \to \mathbb{R}$. The feature vectors capture preselected properties of each state $x \in \mathbb{S}$. Features of a particular state can be generated

on the fly, so there is no need to store the entire matrix Φ; it suffices to store r to represent all elements of $\{\Phi r | r \in \mathbb{R}^F\}$. This produces computational benefits when F is significantly smaller than $|\mathbb{S}|$. The optimality gaps of API depend on the "richness" of the feature vectors, which are typically chosen based on structural insight into the problem at hand (see Bertsekas (2011) for a survey of API methods).

Tayloring offers a generalizable way to approximate $V^{(k)}(x)$ that requires little ad hoc intuition. In this scheme, the intermediate solution $V^{(k)}(x)$ is replaced in the policy evaluation step by the solution to the associated PDE. In addition to approximating the policy evaluation step, we can also approximate the policy improvement step. The details of TAPI are presented in Algorithm 2.

Algorithm 2 (TAPI)

1. Start with initial stationary policy $U^{(0)}(x)$.
2. For $k = 0, 1, \ldots,$

(a) *Approximate policy evaluation:* approximate, using the Taylored equation, the infinite-horizon discounted performance under $U^{(k)}$; that is, find $V^{(k)}(x)$ that satisfies

$$r\left(x, U^{(k)}(x)\right) + \alpha \mathcal{L}_{U^{(k)}(x)} V^{(k)}(x) - (1-\alpha) V^{(k)}(x) = 0, \quad x \in \widetilde{\mathbb{S}}, \tag{23}$$

$$\eta(x)' D V^{(k)}(x) = 0, \quad x \in \partial\widetilde{\mathbb{S}}.$$

(b) *Approximate policy improvement:* Let $U^{(k+1)}(x)$ be the greedy policy associated with $V^{(k)}(x)$ in the Taylored equation

$$U^{(k+1)}(x) = \operatorname*{argmax}_{u \in \mathcal{U}(x)} \left\{ r_u(x) + \alpha \mathcal{L}_u V^{(k)}(x) - (1-\alpha) V^{(k)}(x) \right\}, \quad x \in \widetilde{\mathbb{S}}. \tag{24}$$

In the kth step, assuming that the linear PDE (23) has a solution, it can be numerically approximated by any of a number of solution methods. The most standard of these is the FD method (see, e.g., Larsson and Thomée 2008). FD returns a solution defined on a suitably spaced grid. In our experiments, this grid is a subset of the state space $\mathbb{S}$. The efficiency gains of TAPI cover all three of the previously identified computation bottlenecks of PI:

Value-function storage: Any method to solve (23) involves either a discretized grid or some other state-space-partitioning scheme (as in the finite element method). As the discretization gets finer, the approximation converges, under suitable conditions, to the true solution of the PDE. Choosing a *coarser* grid reduces the cost of storing the value function estimates.

Transition-matrix storage: In (23), the transition probability matrix "collapses" into the lower-dimensional $\mu_u(x)$ and $\sigma_u^2(x)$. In contrast to the standard PI algorithm, we are not inverting the full matrix $P^{(k)}$ here.

Optimization complexity: The computational benefit of the approximate policy improvement in TAPI comes from the fact that $\mathcal{L}_u V^{(k)}(x)$ depends on u only through $\mu_u(x)$ and $\sigma_u^2(x)$, and $V^{(k)}(x)$ and its derivatives do not depend on u. For finite action and state spaces, the quantities $\mu_u(x)$ and $\sigma_u^2(x)$ can be precomputed once in advance (or computed on the fly and kept in memory). Contrast this with PI, in which the term $\mathbb{E}_x^u[V^{(k)}(X_1)]$ has to be recomputed for each u and x at *each* iteration k, and computing this expectation requires going over all the "neighbors" of x. The computational cost of the approximate policy improvement is, consequently, $\mathbb{O}(|\mathcal{U}|_{max}|\mathbb{S}|)$ per iteration compared with $\mathbb{O}(|\mathcal{U}|_{max}|\mathbb{S}|^2)$ for exact policy improvement. This cost may be further reduced if, given x, one has tractable expressions for the dependence of $\mu_u(x)$ and $\sigma_u^2(x)$ on u (see the examples in the next section).

Given the discrete nature of the controls, the exact policy improvement step in (22) can be computationally expensive. Our example in Section 4.3 is one in which it is difficult to avoid exhaustive search. In Moallemi et al. (2008, p. 7), the authors show how to leverage an "affine-expectations" assumption to approximate the solution of this problem by that of a linear program. The approximate policy improvement step in TAPI is a generalizable way to simplify this step.

3.1.1. Convergence and Error Bounds. As stated earlier, the PDE in (23) might not be mathematically meaningful; $U^{(k)}(x)$ could be such that a solution does not exist to the PDE in the policy evaluation step. One implementation we propose—developed for the specific PDEs arising from diffusion control problems—is put forth in Kushner and Dupuis (2013). Roughly speaking, applying certain finite difference schemes to (23) leads back to discrete (time and space) MDPs. The goal in Kushner and Dupuis (2013) is to solve the continuous control problem by taking the discretization to be increasingly finer. We take the opposite approach and choose a coarse grid to reduce the computational effort relative to the original Bellman equation. The construction of Kushner and Dupuis (2013) generates one concrete TCP-equivalent chain. The approximation error that it introduces is related to the third derivative of the PDE solution multiplied by a number that captures the coarseness of the grid (recall the discussion closing Section 2.2 and see (30)).

With the method of Kushner and Dupuis (2013), the approximate policy evaluation and improvement are replaced with exact evaluation and improvement for the (newly constructed) chain on a finite state space.

The convergence to the optimal policy then follows from standard results for policy iteration.

Example 3 (K-D Construction in One Dimension)**.** Consider the one-dimensional TCP on the truncated state space $[0, M]$:

$$0 = \max_{u\in\mathcal{U}(x)}\left\{r_u(x) + \alpha\left(\mu_u(x)V'(x) + \frac{1}{2}\sigma_u^2(x)V''(x)\right) - (1-\alpha)V(x)\right\}, \quad x \in (0, M), \tag{25}$$

with the boundary condition

$$V'(0) = V'(M) = 0.$$

Fix $h > 0$, and let $\mathbb{S}_h = \{0, h, 2h, \ldots, M\}$ be the discretized space. Let us make the assumption that M is divisible by h.

The K-D chain construction is most intuitive under the "small drift" assumption

$$\sigma_u^2(x) \geq |\mu_u(x)|h, \quad x \in \mathbb{S}_h,\ u \in \mathcal{U}(x). \tag{26}$$

For each $x \in \mathbb{S}_h\backslash\{0, M\}$, let us replace $V'(x)$ and $V''(x)$ with the appropriate "central" differences

$$V'(x) \leftarrow \frac{V(x+h) - V(x-h)}{2h}, \text{and}$$
$$V''(x) \leftarrow \frac{V(x+h) - 2V(x) + V(x-h)}{h^2},$$

to get

$$\begin{aligned}(1-\alpha)V(x) &= \max_{u\in\mathcal{U}(x)}\left\{r_u(x) + \alpha\left(\mu_u(x)\frac{V(x+h) - V(x-h)}{2h} + \frac{1}{2}\sigma_u^2(x)\frac{V(x+h) - 2V(x) + V(x-h)}{h^2}\right)\right\}\\ &= \max_{u\in\mathcal{U}(x)}\left\{r_u(x) + \alpha\left(\frac{\mu_u(x)h + \sigma_u^2(x)}{2h^2}V(x+h) + \frac{-\mu_u(x)h + \sigma_u^2(x)}{2h^2}V(x-h) - \frac{\sigma_u^2(x)}{h^2}V(x)\right)\right\}.\end{aligned} \tag{27}$$

Let $\Sigma(x) = \sup_{u\in\mathcal{U}(x)}\sigma_u^2(x) > 0$. Multiplying both sides of (27) by $h^2/\alpha\Sigma(x)$, we arrive at

$$\begin{aligned}V(x) &= \max_{u\in\mathcal{U}(x)}\left\{\frac{\alpha_h(x)h^2 r_u(x)}{\alpha\Sigma(x)} + \alpha_h(x)\left(\frac{\mu_u(x)h + \sigma_u^2(x)}{2\Sigma(x)}V(x+h) + \frac{-\mu_u(x)h + \sigma_u^2(x)}{2\Sigma(x)}V(x-h) + \left(1 - \frac{\sigma_u^2(x)}{\Sigma(x)}\right)V(x)\right)\right\},\end{aligned} \tag{28}$$

where

$$\alpha_h(x) := \left(1 + \frac{h^2}{\Sigma(x)}\left(\frac{1}{\alpha} - 1\right)\right)^{-1}.$$

Let, for $x \in \mathbb{S}_h\backslash\{0, M\}$,

$$P^{u,h}_{x,x+h} = \frac{\mu_u(x)h + \sigma_u^2(x)}{2\Sigma(x)}, P^{u,h}_{x,x-h} = \frac{-\mu_u(x)h + \sigma_u^2(x)}{2\Sigma(x)}, \quad P^{u,h}_{x,x} = 1 - P^{u,h}_{x,x+h} - P^{u,h}_{x,x-h}, \tag{29}$$

and $\tilde{r}_h(x, u) = \alpha_h(x)h^2 r_u(x)/(\alpha\Sigma(x))$. Notice that these are well-defined probabilities because of assumption (26). Also notice that $\tilde{r}_h(x, u) = \frac{1-\alpha_h(x)}{1-\alpha}r(x, u)$. We arrive at the equation

$$V(x) = \max_{u\in\mathcal{U}(x)}\left\{\tilde{r}_h(x, u) + \alpha_h(x)\left(P^{u,h}_{x,x+h}V(x+h) + P^{u,h}_{x,x-h}V(x-h) + P^{u,h}_{x,x}V(x)\right)\right\}.$$

This equation, in the interior, is recognizable as a Bellman equation for a new Markov chain with state space $\mathbb{S}_h$, transition probabilities and reward function as specified, and *state-dependent* discount factor $\alpha_h(x)$.

For the boundary points $x = 0$ and $x = M$, we cannot use central differences because the points $-h$ and $M + h$ are not available. We can use instead the forward difference $V'(0) \leftarrow \frac{V(h)-V(0)}{h}$ at $x = 0$ and the backward difference $V'(M) \leftarrow \frac{V(M)-V(M-h)}{h}$, which then lead to the added equations $V(h) = V(0)$ and $V(M-h) = V(M)$. No discount factor is associated with these boundary states. A thorough treatment of reflecting boundaries appears in Kushner and Dupuis (2013, chapter 5.7). This construction can be easily modified to have FOT on the boundary instead of the oblique-derivative condition.

The K-D chain is but one concrete construction of a TCP-equivalent chain. A nice property of this construction is the sparsity of neighbors—that, from each state, one can only transition to at most 2^d neighbors. In fact, it follows from Kushner and Dupuis (2013) that in the setting of the oblique-derivative boundary condition, there *always* exists a TCP-equivalent construction on the coarser grid. Although this construction need not be as simple as in the one-dimensional case, it always maintains the desirable properties of sparsity of neighbors.

In the multidimensional case, the state space of the K-D chain is

$$\mathbb{S}_h = \times_{i=1}^d\{[0, M_i] \cap \{h\mathbb{Z}_+ \cup \{M_i\}\}\},$$

where $h\mathbb{Z}_+ = \{0, h, 2h, 3h, \ldots\}$. We denote by $X^h = \{X^h_t, t = 1, 2, \ldots\}$ the (controlled) Markov chain on the state space $\mathbb{S}_h$ arising from the K-D construction. As in Example 3, we let $V^h_*(x)$ be the solution to the Bellman equation for the K-D chain and denote by U^h_* the optimal stationary policy for this chain.

Then, under the requirements on $\widehat{V}_*$ in Theorem 1, Equation (11) implies the bound

$$\left|\widehat{V}_*(x) - V_*^h(x)\right| \le h^{2+\beta}\left(\mathbb{E}_x^{\widehat{U}_*}\left[\sum_{t=0}^{\infty}\bar{\alpha}_h^t\left[D^2\widehat{V}_*\right]^*_{\beta,X_t^h\pm h}\right] + \mathbb{E}_x^{U_*^h}\left[\sum_{t=0}^{\infty}\bar{\alpha}_h^t\left[D^2\widehat{V}_*\right]^*_{\beta,X_t^h\pm h}\right]\right),\ x\in\mathbb{S}_h. \quad (30)$$

This bound is similar in spirit to and inspired by Dupuis and James (1998). A challenge here is that $V_*^h(x)$ and U_*^h are only defined on the coarse grid $\mathbb{S}_h$. To borrow a term from the ADP literature, we must now "disaggregate" these to generate a policy for the original chain.

One way to achieve this is via one-step improvement relative to an extended $V_*^h(x)$. Assume that we have an extension of $V_*^h(x)$, denoted by $\widetilde{V}_*^h(x)$, that is defined for all $x\in\mathbb{S}$. Then we can use, in the original chain, a greedy policy relative to $\widetilde{V}_*^h$:

$$U^h(x) = \operatorname*{argmax}_{u\in\mathcal{U}(x)}\left\{r_u(x) + \alpha\mathbb{E}^u\left[\widetilde{V}_*^h(X_1)\right]\right\}.$$

The error introduced by doing so can be suitably bounded (see Remark EC.1 in the e-companion). Alternative aggregation methods (see Remark 8) do produce direct controls that are defined for all states in the detailed state space $\mathbb{S}$.

Remark 7 (Exact Policy Improvement)**.** Algorithmically, one can replace the approximate policy improvement step in (24) with an *exact* policy improvement step in which we let $U^{(k+1)}(x)$ be the greedy policy associated with $V^{(k)}(x)$; that is,

$$U^{(k+1)}(x) = \operatorname*{argmax}_{u\in\mathcal{U}(x)}\left\{r_u(x) + \alpha\mathbb{E}_x^u\left[V^{(k)}(X_1)\right]\right\},\quad x\in\mathbb{S}.$$

Because the policy improvement is done exactly—using the transitions and state space of the Markov chain—we must extend $V^{(k)}(x)$ to the state space $\mathbb{S}$ (say, by interpolation). In our examples, we find that although this version of TAPI has no convergence guarantees, it may result in a smaller optimality gap. □

Remark 8 (Aggregation as a Basis for Scalable Algorithms)**.** The construction of a TCP-equivalent chain on a coarser grid can be viewed as an aggregation procedure. Bounds on the optimality gaps introduced by aggregation methods are often stated in terms of oscillations of the true value function over the coarser grid (see, e.g., Bertsekas and Tsitsiklis 1996, section 6.7). The bounds depend on the same quantity that we are trying to avoid computing. In contrast, our construction of the approximate coarse chain is grounded in the TCP, which also provides a grounding for optimality-gap analysis. In the cases in which our bounds apply, they are stated in terms of the approximation $\widehat{V}_*$ rather than by the value function itself. At the same time, existing (flexible) aggregation algorithms may offer scalability that far exceeds our relatively simple discretization-based aggregation.

It seems feasible to piggyback on existing aggregation approaches to develop algorithms that simultaneously (1) make use of the insights (and bounds) that we developed and (2) preserve the scalability of the algorithm on which we piggyback; this direction is explored in Zhang and Gurvich (2018). To convey this feasibility, we briefly outline this approach here.

Recall (see Equation (11) and the discussion there) that two controlled chains on the same state space $\mathbb{S}$ and with matrices P and $\widetilde{P}$ have a common TCP if they share the (induced) drift vectors and diffusion matrices. We can then apply the control derived from the tilde chain to the original chain with the optimality guarantees stated in this paper.

Soft aggregation (see Bertsekas 2007) is a flexible ADP algorithm that is perfectly suited for the creation of such a Markov chain $\widetilde{P}$. One (probabilistically) groups states into a small number of "metastates." There is a set $\mathcal{M} = \{1,\dots,m\}$ of metastates. To be useful, m must be smaller than the size of the detailed state space $\mathbb{S}$. Instead of solving the original Bellman equation, one solves an *aggregated* Bellman equation on the metastates:

$$R(k) = \sum_{x\in\mathbb{S}} q_{k,x}\min_{u\in\mathcal{U}(x)}\sum_{y\in\mathbb{S}}P_{x,y}^u\left(r_u(x) + \alpha\sum_{l\in\mathcal{M}}\phi_{y,l}R(l)\right),\quad k\in\mathcal{M}, \quad (31)$$

where the $m\times n$ matrix $Q = \{q_{l,x}\}$ contains the disaggregation probabilities and the $n\times m$ matrix $\Phi = \{\phi_{x,l}\}$ contains the aggregation probabilities.

The optimal policy in (31) is identical—it is shown in Zhang and Gurvich (2018)—to that of a Markov chain on the detailed state space $\mathbb{S}$ with the transition matrix

$$\widetilde{P}_{x,y}^u = \sum_{z\in\mathbb{S},l\in\mathcal{M}}P_{x,z}^u\phi_{z,l}q_{l,y}.$$

In other words, the optimal control in the Bellman equation

$$\widetilde{V}(x) = \max_{u\in\mathcal{U}(x)}\left\{r_u(x) + \alpha\sum_{y\in\mathbb{S}}\widetilde{P}_{x,y}^u\widetilde{V}(y)\right\}$$

can be derived by solving the *lower*-dimensional aggregate equation (31). It is precisely here that Zhang and Gurvich (2018) piggyback on aggregation. If we can choose the matrix aggregation's design variables Φ, Q (of a suitable rank smaller than n) such that the *local moment matching* holds, then (1) the two chains P and $\widetilde{P}$ are "coupled" via the TCP, and (2) computation

requires solving only the aggregate equation (31). Matching (at least closely) the moments is doable and can be done with great efficiency (see Zhang and Gurvich 2018). □

4. Examples

The three examples we study are intended to illustrate the performance of the proposed algorithm. The first two examples have a one-dimensional state space and are, hence, computationally cheap even for an exact solution. We use them because visualization is easier in $d = 1$ and supports useful observations.

4.1. Service-Rate Control

This is a variant of Example 2. We consider the holding cost $c(x) = x^2$ when there are x customers in the system and the control cost $f(u) = 1/(1-u)$. The cost minimization problem is equivalent to a reward maximization problem with the negative reward $-x^2 - 1/(1-u)$. The control set consists of the rational numbers (denoted by $\mathbb{Q}$) in $[0,1]$. The Taylored equation is

$$0 = \min_{u\in[0,1]\cap\mathbb{Q}}\left\{x^2 + \frac{1}{1-u} + \alpha\left((1-2u)V'(x) + \frac{1}{2}V''(x)\right) - (1-\alpha)V(x)\right\},$$

$$0 = V'(0).$$

Per Lemma 1, this equation has a solution $\widehat{V}_* \in \mathcal{C}^{2,\beta}$ for some $\beta \in (0,1)$, and the policy $\widehat{U}^*$ derived from this equation induces an optimality gap $|V^{\alpha}_{\widehat{U}^*} - V^{\alpha}_*|$ that obeys the bound in Theorem 2.

We use TAPI based on the K-D chain to obtain the optimal control $\widehat{U}^h_*$ for this chain. We build a control for the original chain by extending to $\mathbb{Z}_+$ in a piecewise constant manner: the control at point mh is kept constant for all points $mh, \dots, (m+1)h-1$. We denote this control by U^h. We try $h \in \{1,2\}$. In our computations, we allow the control to be any number in $[0,1]$, which allows us, in this example, to write the control as an explicit function of the value in neighboring states.

In Figure 2, we plot the absolute (rather than relative) optimality gap $V_{U^h}(x) - V_*(x)$ for $\alpha = 0.99$ and discretization h equal to 1 and 2. It is important that even in the case of $h = 2$, an optimality gap of 30 (at $x = 100$) is negligible relative to the optimal value at that state, which is of the order of 3×10^5. More impressive is the performance after one-step policy improvement. The greedy policy achieves an optimality gap that is indistinguishable from zero. This result is explained by Figure 3, in which we report the comparison of *actions*. The plot also includes the control after one-step policy improvement starting at the K-D chain interpolated value.

Per Theorem 1, the error should be of the order of the integrated third derivative. If this derivative is uniformly bounded by Γ, the optimality gap must be smaller than (or equal to) $\Gamma/(1-\alpha)$. The central differences proxy with $h = 1$ for the third derivative of $\widehat{V}_*(x)$ is given by

$$\widehat{V}'''_*(x) \approx \frac{1}{2}\left(V^h_*(x+2) - 2V^h_*(x+1) + 2V^h_*(x-1) - V^h_*(x-2)\right).$$

The peak of this proxy for $h = 1$ and $\alpha = 0.99$ is 1.8, generating in our theorem an error bound of $1.8/(1-\alpha) = 180$, which is 0.0006 of the value of more than 270,000 for the state $x = 1/(1-\alpha) = 100$.

4.2. An Inventory Problem

In the inventory problem we study next, the optimal policy is a so-called order-up-to level policy, implying large jumps in some states. This problem seems to pose a challenge to our bounds, which depend on the size of the maximal jump.

Period t demand D_t is drawn from a Poisson distribution with mean $\mathbb{E}[D_t] = \lambda$. Demand is independent across periods. There is a backlog cost of b, a per-period cost H for holding a unit in inventory, and a per-unit order cost of c. The lead time is zero.

The state is the inventory position. The per-period cost is given by

$$r_u(x) = cu + H\mathbb{E}[(x+u-D)^+] + b\mathbb{E}[(D-(x+u))^+],$$

where the action u is the amount ordered. Orders are placed (and received), and then demand is realized, and backorder and holding costs are incurred. Transitions have the form

$$X_t \to (X_t + U_t - D_t),$$

where U_t is the order quantity in period t. The Bellman equation is given by

$$V(x) = \min_{u\in\mathbb{Z}_+}\{r_u(x) + \alpha\mathbb{E}[V(x+u-D)]\}.$$

The drift and diffusion coefficients are given by

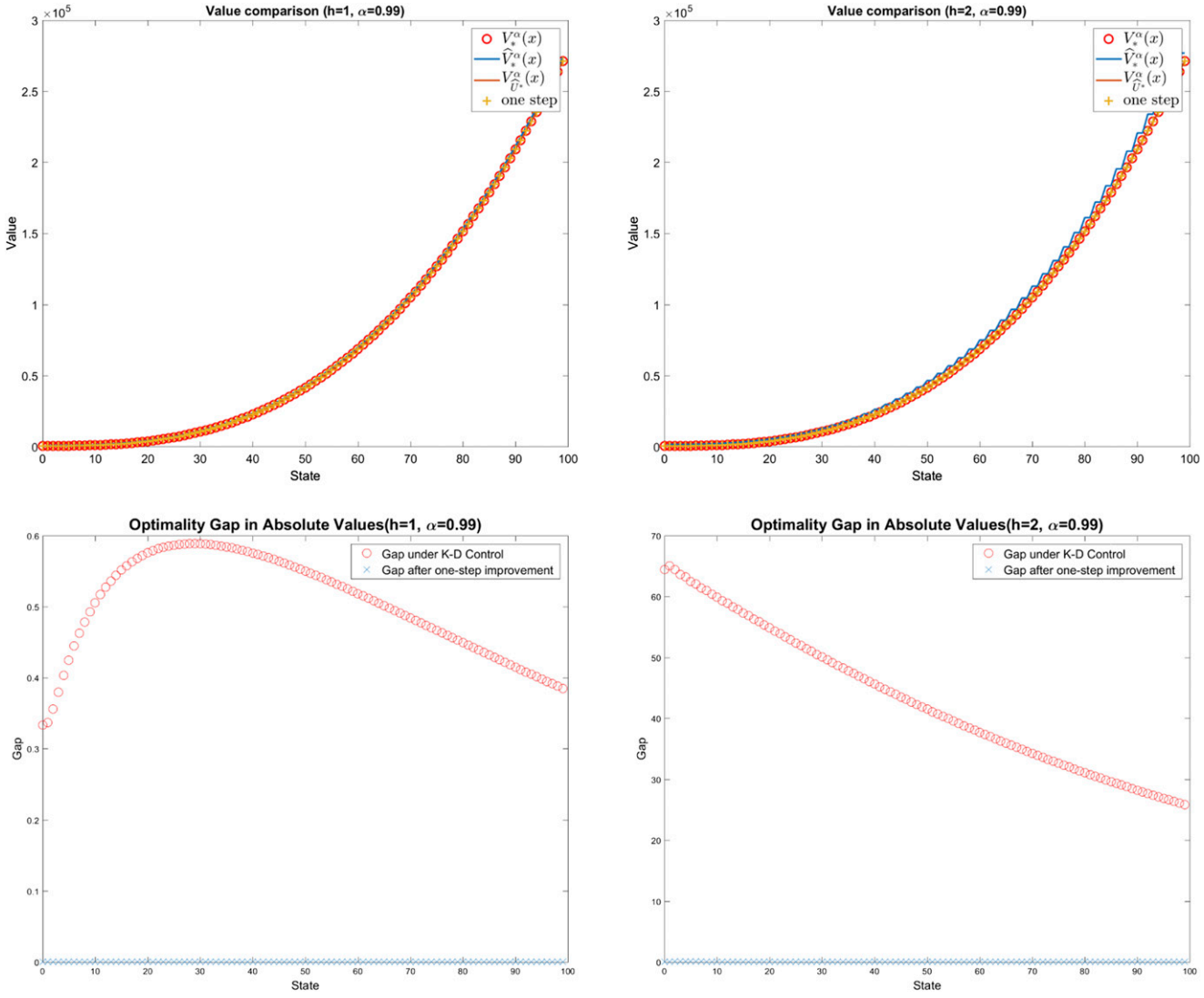
$$\mu_u \equiv \mu_u(x) = \mathbb{E}[X_1 - x] = \mathbb{E}[x+u-D-x] = u-\lambda \text{ and}$$
$$\sigma_u^2 \equiv \sigma_u^2(x) = \mathbb{E}\left[(X_1-x)^2\right] = \mathbb{E}\left[(x+u-D-x)^2\right] = (u-\lambda)^2 + \lambda,$$

so the TCP is

$$0 = \min_{u\in\mathbb{Z}_+}\left\{r_u(x) + \alpha\left(\mu_u V'(x) + \frac{1}{2}\sigma_u V''(x)\right) - (1-\alpha)V(x)\right\}, \quad x \in \mathbb{R}.$$

Notice that $\sigma_u^2 \geq \lambda$ for all x and $u \in \mathbb{Z}$ so that strict ellipticity holds. Because the state space includes all the integers, there are no boundary conditions here.

Figure 2. (Color online) Algorithm Performance for the Service-Rate Control Problem



Notes. Top: The optimal value compared against the approximate values. The lines are indistinguishable at the scale of the optimal value. Bottom: The absolute optimality gap: the difference between the cost under the proposed policy and the optimal cost. The '×' series is the result after one-step improvement. It is indistinguishable from zero.

These are artificially introduced in our numerical computation to make the state space finite. Specifically, we truncate the space at state M and $-M$, where $\mu_u(-M) = u$, $\sigma_u^2(-M) = u^2$, and $\mu_u(M) = -\lambda$, $\sigma_u^2(M) = \lambda + \lambda^2$. We then introduce the boundary condition $V'(M) = V'(-M) = 0$. Notice that, for constant u, the drift and diffusion coefficient are constant in x and in particular Lipschitz. The cost function is, as well, Lipschitz in x uniformly in u. The existence of a solution follows from Lemma 1.

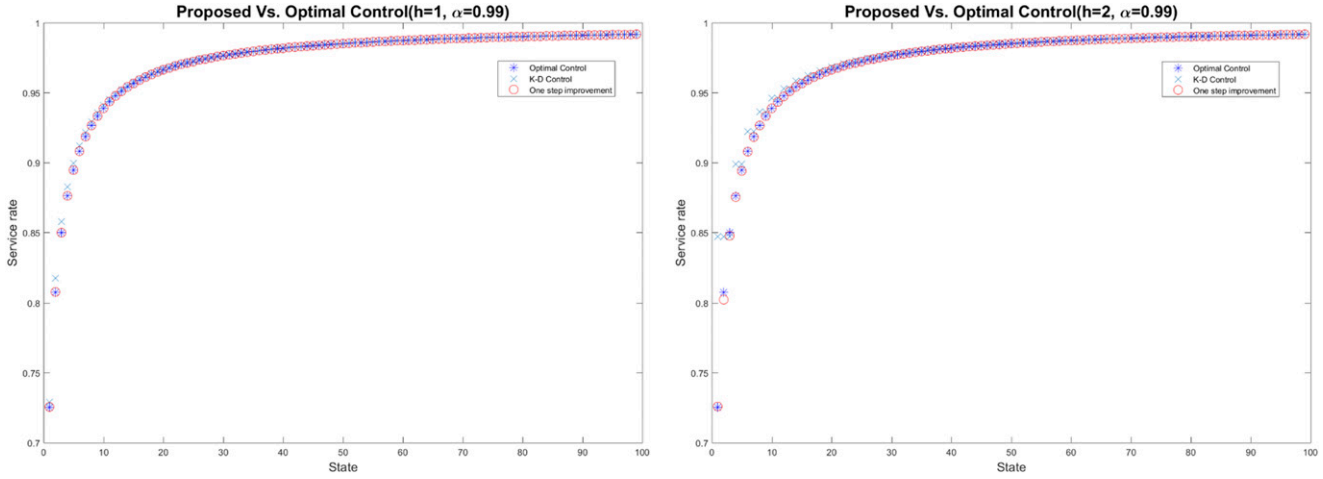
We use the K-D chain for value of coarseness $h \in \{1, 3\}$. For each h, we use (with some abuse of notation) $\widehat{V}_*^{\alpha}(x)$ for the value from the K-D approximation, which is a proxy for the TCP value. For $h = 3$, we extend the value function to the integers in a piecewise constant manner. We also take the control $\widehat{U}_h^*$ and interpolate it to the whole state space in a piecewise constant manner. Denote by $V_{\widehat{U}_*}(x)$ the value in the original chain when using this control. Finally, the value after one-step improvement is the infinite-horizon discounted reward under U^h—the greedy control relative to $V_*^{h,\alpha}(x)$. Figure 4 displays the computational results.

4.3. A Routing Problem

This example is based on the inpatient-management queuing model studied in Dai and Shi (2017). The task is to optimally route patients from dedicated queues to internal hospital wards so as to minimize the aggregate cost of holding and routing.

The dynamics of the queues are modeled via a discrete-time queuing model with J server pools (the internal wards), in which pool j has N_j servers (beds),

Figure 3. (Color online) Control Comparison for the Service-Rate Control Problem



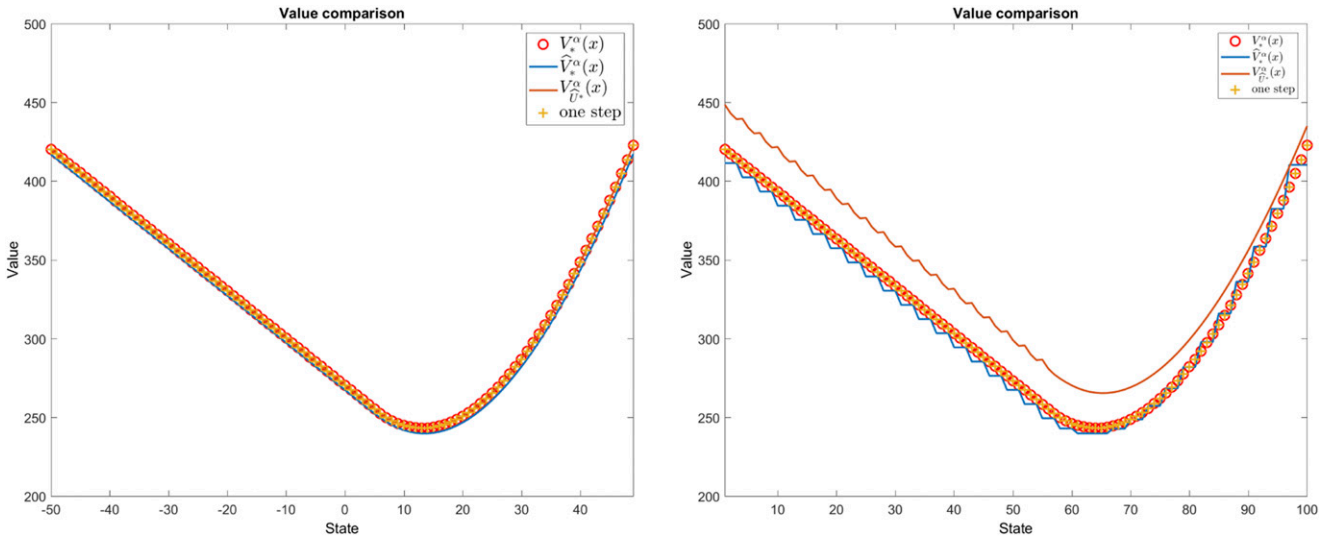
Notes. Left: $h = 1$. Right: $h = 2$. The control obtained from one-step improvement is almost identical (at all states) to the optimal control.

a dedicated inflow of customers, and an infinite-sized dedicated buffer. This buffer is truncated for the numerical experiments. Customers from the jth inflow are referred to as type-j customers. These customers can be served by their dedicated pool j and also by other pools.

We let $X_j(t)$ be the number of customers of type j in the system at time $t = 0, 1, 2, \ldots$ and let $X(t) \in \mathbb{Z}_+^d$ be the vector whose components are $X_j(t)$. A customer in the system can either be in service or waiting in a buffer to be served.

The (controlled) chain's evolution is as follows: at the start of time period t, customers waiting in buffer j enter service in pool j until the buffer is emptied or all idle servers are taken. If any customers remain in buffer j, we proceed to the *overflow* decision. This is the overflow control. At a cost of B_{ij} per customer, we can choose to assign a customer waiting in buffer i to immediately enter service in pool $j \neq i$ if that pool has an idle server available. We can also decide not to overflow any customers. We let $U_{ij}(t) = U_{ij}(X(t))$ be the number of customers overflown from buffer i to pool j in time period t. After overflows are executed, a holding cost H_i per customer waiting in buffer i is incurred. Next, departures are resolved: a type j customer in service completes service and leaves the system with probability p_j (service time is geometric with mean $1/p_j$). Otherwise, the customer remains in service until the next period.

Figure 4. (Color online) Algorithm Performance for the Inventory Problem



Notes. Left: $h = 1$. Right: $h = 3$. The performance of the K-D *control* is not as good, but one-step improvement relative to the K-D value function $V_*^{h,\alpha}$ generates an extremely accurate value for all states.

After departures are resolved, new arrivals occur: the number of type-j customers to arrive per period is Poisson distributed with mean λ_j. Arrivals are independent across types and across time periods. An incoming customer either occupies an idle server in the customer's dedicated pool or, if there are no such servers, enters the buffer and waits for service.

Under a stationary control u, $X(t)$ satisfies the dynamics

$$X_i(t) = X_i^P(t-1) + A_i(t-1) - D_i\big(X_i^P(t-1)\big), \quad i \in [d], \tag{32}$$

where

$$X_i^P(t-1) = X_i(t-1) + \sum_{j\neq i} U_{ji}(X(t-1)) - \sum_{j\neq i} U_{ij}(X(t-1))$$

is the post-action state in period $t-1$,

$$D_i(x) \sim \text{Binomial}\big(x \wedge N_i, p_i\big)$$

is the number of departures from pool i, and

$$A_i(t) \sim \text{Poisson}(\lambda_i)$$

is the number of new unblocked type-i arrivals, independent across periods.

The state space is $\mathbb{Z}_+^d$, and the action space is

$$\mathcal{U}(x) = \left\{ u \in \mathbb{Z}_+^{I\times J} \;\middle|\; \sum_{j\neq i} u_{ji} \le (N_i - x_i)^+ \text{ and } \sum_{j\neq i} u_{ij} \le (x_i - N_i)^+, \quad i \in [d] \right\}.$$

The first constraint guarantees that the overflow to pool i does not exceed the number of idle servers, and the second constraint guarantees that the number of overflowed type-i customers does not exceed the number of customers waiting in buffer i. The action space satisfies the structure $\mathcal{U}(x) = \mathbb{D} \cap \{u : Au \le b(x)\}$. Finally, the per-period cost includes the cost of overflow and linear cost of holding and is given by

$$r_u(x) = \sum_i \sum_{j\neq i} B_{ij} u_{ij} + \sum_i H_i \times \left(x_i - \sum_{j\neq i} u_{ij} - N_i \right)^+. \tag{33}$$

The goal is to make overflow decisions that minimize the expected infinite-horizon discounted cost.

The Bellman equation for this dynamic program is computationally challenging. A modest system with $J = 3$, $N_i \equiv 40$, and $M = 60$ has more than 1 million states. Moreover, depending on the policy $U(x)$, a state can have many "neighboring" states, making the transition probability matrix dense and expensive to store. Finally, because actions are discrete, the only option is exhaustive search over the very large action space: we have to decide how many customers to overflow from *each* buffer into *each* pool.

We next construct the TCP's ingredients. For $x \in \mathbb{S}$ and $u \in \mathcal{U}(x)$, let $x_i^P(u) = x_i + \sum_{j\neq i}(u_{ji} - u_{ij})$. Then

$$\begin{aligned}\big(\mu_u(x)\big)_i &= \sum_{j\neq i} u_{ji} - \sum_{j\neq i} u_{ij} + \mathbb{E}[A_i(t)] - \mathbb{E}\big[D_i\big(x_i^P(u)\big)\big] \\ &= \sum_{j\neq i}\big(u_{ji} - u_{ij}\big) + \lambda_i - p_i\left(\left(\lceil x_i\rceil + \sum_{j\neq i}\big(u_{ji} - u_{ij}\big)\right) \wedge N_i\right) \\ &=: \big(f_\mu(u,x)\big)_i, \quad i \in [d],\end{aligned}$$

and

$$\begin{aligned}\big(\sigma_u^2(x)\big)_{ij} = \; & \mathbb{E}\left[\left(\sum_{k\neq i}(u_{ki} - u_{ik}) + A_i(t) - D_i\big(x_i^P(u)\big)\right)\right. \\ & \left.\left(\sum_{k\neq j}\big(u_{kj} - u_{jk}\big) + A_j(t) - D_j\Big(x_j^P(u)\Big)\right)\right] \\ =: \; & \big(f_\sigma(u,x)\big)_{ij}, \quad i,j \in [d].\end{aligned}$$

We use the oblique derivative condition

$$\eta(x)' D\widehat{V}(x) = 0, \quad x \in \partial\mathbb{R}_+^d,$$

where $\eta_i(x) = p_i$ if $x_i = 0$ and is zero otherwise.

This is grounded in intuition about the "pushback" at zero but is also mathematically supported by choosing the suitable extension of μ_u and σ_u^2. The function $f_\mu(u,x)$ is well defined for all $x \in \mathbb{Z}_+^d$ and $u \in \mathbb{D}$ and can be continuously extended to $x \in \mathbb{R}_{++}^d$ in multiple ways. We choose to extend $\lceil x_i\rceil$ so that it is constant (and equal to 1) for all $x_i \in (0,1]$. We extend $f_\sigma(u,x)$ so that it is continuous on all of (not just the interior of) $\mathbb{R}_+^d$. For x with $x_i < N_i$, $\mathcal{U}(x)$ contains only u with $u_{ij} = 0$ for all $j \neq i$. If $\widehat{U}_*$ is piecewise constant and continuous at the boundary,[5] then $(\eta(x))_i = (\mu_{\widehat{U}_*}(x_{-i}, 0+))_i - (\mu_{\widehat{U}_*}(x_{-i}, 0))_i = -p_i$.

We also tried the more direct FOT boundary conditions for this example, obtaining similar performance to what is reported in Table 1, panels (A)–(C).

In our computational experiments, we truncate the state space by using finite buffers and truncating arrivals in an intuitive way. We use exhaustive search over $u \in \mathcal{U}(x)$ in the policy improvement step rather than relaxing the integrality constraints (see, e.g., Moallemi et al. 2008). We do so because we wish to capture the error induced by the Taylor expansion without confounding it by approximations to the action space. Still, the computational savings of TAPI are significant: a single iteration of TAPI took a few

Table 1. Applying TAPI to an instance of the model with ($J = 3$)

Panel A

		$\lambda_i = 0.7N_ip_i$		$\lambda_i = 0.8N_ip_i$	
α	h	Maximum relative error	Mean relative error	Maximum relative error	Mean relative error
0.9	2	0.058	0.001	0.064	0.001
	4	0.043	0.0004	0.030	0.0005
	8	0.038	0.0002	0.037	0.0009
0.99	2	0.023	0.001	0.019	0.002
	4	0.016	0.0004	0.017	0.0005
	8	0.019	0.0004	0.014	0.005
0.999	2	0.004	0.001	0.004	0.002
	4	0.003	0.0004	0.003	0.0005
	8	0.003	0.0004	0.007	0.006

Panel B

		$\lambda_i = 0.7N_ip_i$		$\lambda_i = 0.8N_ip_i$	
α	h	Maximum relative error	Mean relative error	Maximum relative error	Mean relative error
0.9	2	0.685	0.013	0.516	0.011
	4	0.685	0.012	0.45	0.009
	8	0.312	0.028	0.186	0.022
0.99	2	0.206	0.011	0.096	0.005
	4	0.184	0.012	0.120	0.005
	8	0.110	0.032	0.055	0.011
0.999	2	0.037	0.007	0.028	0.002
	4	0.036	0.010	0.014	0.003
	8	0.051	0.039	0.016	0.011

Panel C

		$\lambda_i = 0.7N_ip_i$		$\lambda_i = 0.8N_ip_i$	
α	h	Maximum relative error	Mean relative error	Maximum relative error	Mean relative error
0.9	2	0.114	0.016	0.104	0.011
	4	0.079	0.006	0.075	0.004
	8	0.071	0.018	0.053	0.014
0.99	2	0.077	0.021	0.053	0.009
	4	0.060	0.009	0.039	0.005
	8	0.069	0.042	0.025	0.014
0.999	2	0.034	0.024	0.016	0.009
	4	0.019	0.010	0.009	0.004
	8	0.059	0.055	0.017	0.016

Notes. Panel A: $N_1 = N_2 = N_3 = 10, M = 14, (p_1, p_2, p_3) = (0.8, 0.8, 0.8), (H_1, H_2, H_3) = (1, 2, 3), (B_{12}, B_{13}) = (1, 1), (B_{21}, B_{23}) = (4, 1)$, and $(B_{31}, B_{32}) = (2, 1)$. Panel B: $N_1 = N_2 = N_3 = 10, M = 14, (p_1, p_2, p_3) = (0.4, 0.6, 0.1), (H_1, H_2, H_3) = (10, 2, 6), (B_{12}, B_{13}) = (5, 2), (B_{21}, B_{23}) = (3, 7)$, and $(B_{31}, B_{32}) = (7, 9)$. Panel C: $N_1 = N_2 = N_3 = 10, M = 14, (p_1, p_2, p_3) = (0.2, 0.7, 0.5), (H_1, H_2, H_3) = (1, 1, 4), (B_{12}, B_{13}) = (5, 2), (B_{21}, B_{23}) = (7, 1)$, and $(B_{31}, B_{32}) = (7, 9)$.

minutes compared with about three hours for a PI iteration, leading to a reduction of total running time from more than 15 hours to less than 10 minutes.

Table 1, panels (A)–(C), presents the results of applying TAPI to multiple three-dimensional ($J = 3$) instances of the model. The panels show the difference between the value function under the proposed policies $V^{\alpha}_{U^h_*}(x)$ and the actual optimal value $V^{\alpha}_{*}(x)$. The maximal relative error is computed by $\max_{x\in\mathbb{S}} |V_{U^h_*}(x) - V_*(x)|/V_*(x)$, where U^h_* is the policy suggested by the approximation algorithm. The mean relative error column reports $\frac{1}{|\mathbb{S}|}\sum_{x\in\mathbb{S}} |V_{U^h_*}(x) - V_*(x)|/V_*(x)$.

In all three panels, we consider an increasing sequence of discount factors in which $\alpha = 0.9$ is considered small (corresponding to an effective horizon of length $1/(1-\alpha) = 10$). Per the discussion in Remark 4, one expects that (1) the gap will be larger for small values of α (vanishing-discount optimality) and also that (2) even for small values of α, the optimality gap will be small for states x with large $|x|$ (order optimality). In other words, the biggest errors are

confined to a small set of states, so the mean relative error should be small.

We see both effects in our numbers. For $\alpha = 0.9$, the maximal (over the state space) relative gap can be fairly large (and decreases as α increases). The mean gap is small for all α values because the big errors are confined to a relatively small number of states.

In Figure 5, we use a two-dimensional example to visualize this fact. Even in cases in which the *maximal* error is as large as 3.7%, such errors are confined to a very small portion of the state space (close to a boundary) and are much smaller in most of the state space. This is captured in Figure 5, in which we plot the relative error in a two-dimensional ($d = 2$) case.

Finally, in reference to Remark 7, we compare the performance of TAPI with a heuristic in which the policy improvement is executed exactly rather than approximately. Introducing exact improvement, although having no convergence guarantees, can result in better performance (see the "+Exact improve." column in Table 2). This performance is, however, matched by using TAPI as is (with approximation improvement) and adding at the very end a *single* exact policy improvement step (see the column "One step" in the same table).

Remark 9 (Smoothing the State Space)**.** We use this example, with $d = 2$, to illustrate the domain smoothing described in Section 2.3. We replace the corner point $(0,0)$ in the Markov chain's state space with the point $\tilde{0} = (\epsilon, \epsilon)$ so that we can pass a smooth boundary through this point. Let us denote the drift and diffusion coefficients by $\tilde{\mu}_u$ and $\tilde{\sigma}_u^2$. Because $P^u_{\tilde{0},y} = P^u_{0,y}$, we have

$$\begin{aligned}\big(\tilde{\mu}_u\big)_i(\tilde{0}) &= \sum_y P^u_{0,y}\big(y_i - \epsilon\big) = \sum_y P^u_{0,y}\big(y_i - 0\big) - \epsilon \\ &= \mu_i(0) - \epsilon = \lambda_i - \epsilon, \quad \text{for} \quad i \in [d],\end{aligned}$$

and

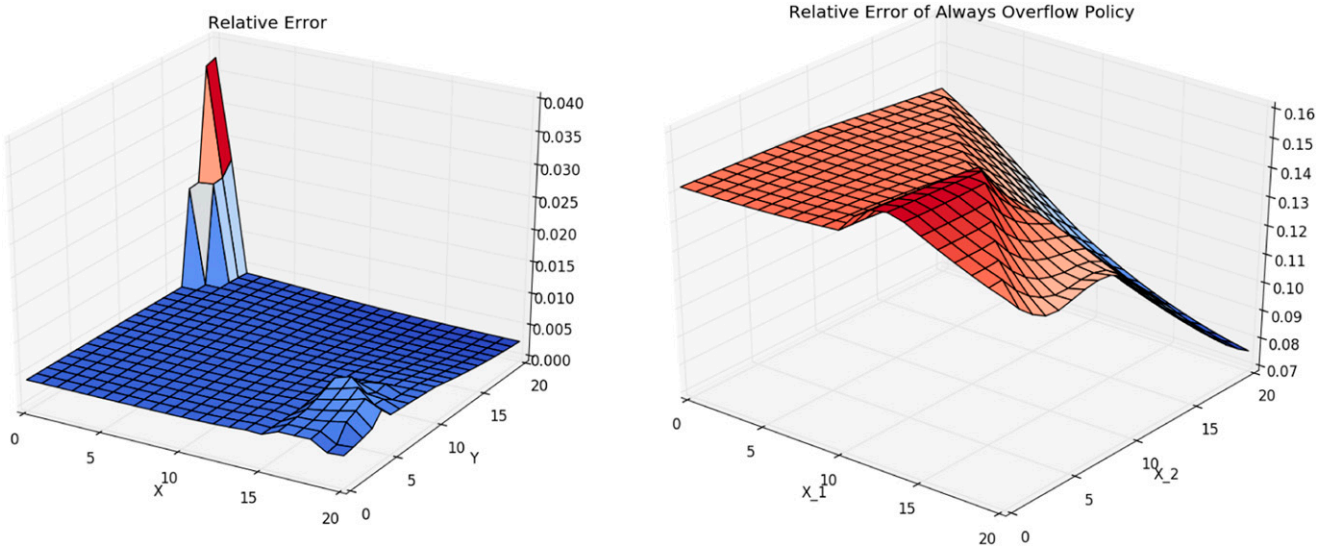
$$\begin{aligned}\big(\tilde{\sigma}_u^2\big)_{ij} &= \sum_y P^u_{0,y}\big(y_i - \epsilon\big)\big(y_j - \epsilon\big) \\ &= \mathbb{E}\big[\big(A_i(t) - \epsilon\big)\big(A_j(t) - \epsilon\big)\big], \quad \text{for } i,j \in [d].\end{aligned}$$

For all other states, $(\tilde{\mu}_u)_i(x) = (\mu_u)_i(x) + P^u_{x,0}\epsilon$ and $(\tilde{\sigma}_u^2)_{ij}(x) = (\sigma_u^2)_{ij}(x) + P^u_{x,0}\epsilon(\epsilon - x_i - x_j)$. Notice that ϵ appears only in the new corner; there are no states of the form (ϵ, x_2) or (x_1, ϵ) for $x_1, x_2 \geq 1$. Finally, the vector η is given by $(p_1, 0)$ when $x_1 = 0$ and $x_2 \geq 1$ and by $(0, p_2)$ when $x_2 = 0$ and $x_1 \geq 1$. It is $\eta(x) = (p_1, p_2)$ at $x = \tilde{0}$. We then extend to the curved boundary in a smooth way. These constructions extend to $d > 2$. □

5. Tayloring and Queues in Heavy Traffic

The impetus for the nascent literature on refined bounds—to which this paper belongs—was, to a large extent, provided by the observed accuracy of Brownian approximation–based prescriptions in queuing systems. For example, in Koçağa and Ward (2010, theorem 5.2), the limit theorem "only" guarantees that the HJB-based prescription induces an optimality gap that is $o(\sqrt{N})$ (where N is the number of servers). However, the numerical experiments (Koçağa and Ward 2010, table 1) show an extremely small optimality gap, one that does not grow with N.

Figure 5. (Color online) Concentration of Error



Notes. $d = 2$, $N_1 = N_2 = 10$, $M = 10$, $(p_1, p_2) = (0.56, 0.56)$, $(H_1, H_2) = (1, 4)$, $(B_{12}, B_{21}) = (5, 1)$, $\alpha = 0.99$, $h = 2$, and $\lambda_i = 0.8N_i p_i$. On the left, we plot the relative error $\frac{|V_{U_h^*}(x) - V_*(x)|}{V_*(x)}$ over the entire domain $\mathbb{Z}_+^2$. The 3.7% maximal relative error reported in Table 2 is caused by only a small portion of the state space. For contrast, the plot on the right shows the distance to $V_*(x)$ of the value function under the (suboptimal) policy that overflows as many customers as possible.

Table 2. Relative Error: TAPI vs. TAPI with Exact-Improvement Step

$\lambda_i = 0.8N_ip_i$				
α	h	TAPI	+ Exact improve.	One step
0.99	1	0.0376	0.0086	0.0095
	2	0.0373	0.0081	0.0088
	4	0.0346	0.0067	0.0079
0.999	1	0.0093	0.0033	0.0051
	2	0.0082	0.0031	0.0045
	4	0.0048	0.0023	0.0032
$\lambda_i = N_ip_i$				
α	h	TAPI	+ Exact improve.	One step
0.99	1	0.0103	0.0089	0.0083
	2	0.0107	0.0069	0.0100
	4	0.0013	0.0014	0.0014
0.999	1	0.0013	0.0026	0.0030
	2	0.0012	0.0026	0.0043
	4	0.0001	0.0002	0.0002

Note. Parameters: $d = 2$, $N_1 = N_2 = 10$, $M = 10$, $(p_1, p_2) = (0.56, 0.56)$, $(H_1, H_2) = (1, 4)$, and $(B_{12}, B_{21}) = (5, 1)$.

In this short section, we wish to illustrate, informally and via the simplest of examples, the connection between Tayloring and heavy-traffic approximations.

Consider the discrete-time queue with holding cost $c(x) = x$, $P_{x,x+1} = \lambda$, and $P_{x,x-1} = \mu := 1 - \lambda > 0.5$. These parameters are fixed, so the question is one of performance approximation, but it is sufficient for illustration; Example EC.1 in the e-companion adds control.

Let us consider two avenues to approximate the infinite-horizon discounted cost.

5.1. Process Convergence

Let $\mu \downarrow \frac{1}{2}$ so that $\rho = \lambda/\mu = (1-\mu)/\mu \uparrow 1$—the queue is in *heavy traffic* (see, e.g., Whitt 2002, chapter 9). It is then a standard heavy-traffic result that extending time by $(1-\rho)^{-2}$ and shrinking space by $(1-\rho)$,

$$(1-\rho)X_{\lceil (1-\rho)^{-2}t \rceil} \approx \widehat{X}(t),$$

where $\widehat{X}(t)$ is a so-called reflected Brownian motion with drift $-\frac{1}{2}$ and diffusion coefficient $\sigma^2 \equiv 1$. This result is formalized by weak convergence arguments. With discount factor equal, for example, to 1, the infinite-horizon discounted reward of the diffusion $\widehat{V}(x) = \mathbb{E}_x[\int_0^\infty e^{-s}c(\widehat{X}(s))ds]$ solves the ordinary differential equation (ODE)

$$0 = x - \frac{1}{2}V'(x) + \frac{1}{2}V''(x) - V(x), \quad V'(0) = 0.$$

Relying on the weak convergence to $\widehat{X}$, it is then possible to show that if one takes—in the Markov chain control problem—the discount factor to be

$$\alpha_\rho = 1 - (1-\rho)^2,$$

Then, as $\rho \uparrow 1$,

$$(1-\rho)^3 V(x) = (1-\rho)^3 \mathbb{E}_x\left[\sum_{t=0}^{\infty} \alpha_\rho^t X_t\right] \approx \widehat{V}(x).$$

In words, $(1-\rho)^{-3}\widehat{V}$ approximates the value function up to an error that is small relative to $(1-\rho)^{-3}$.

5.2. Tayloring

The TCP for given λ and μ is given by

$$0 = x + \alpha\left((\lambda - \mu)V'(x) + \frac{1}{2}V''(x)\right) - (1-\alpha)V(x)$$

and $V'(0) = 0$. With $\lambda < \mu$, this ODE has the solution

$$\widehat{V}^\alpha(x) = -\alpha\frac{\mu - \lambda}{(1-\alpha)^2} + \frac{x}{1-\alpha} + c_1 e^{\gamma_- x},$$

where $\gamma_- = -\sqrt{(\mu-\lambda)^2 + 2\frac{1}{\alpha}(1-\alpha)} + (\mu - \lambda) < 0$ and $c_1 = -\frac{1}{(1-\alpha)\gamma_-}$.

Straightforward differentiation gives

$$\left|D^3\widehat{V}(x)\right| = \frac{\gamma_-^2}{1-\alpha}e^{\gamma_- x} \le \frac{\gamma_-^2}{1-\alpha},$$

so

$$\mathbb{E}_x\left[\sum_{t=0}^{\infty}\alpha^t\left|D^3\widehat{V}^\alpha\right|^*_{X_t \pm 1}\right] \le \frac{1}{1-\alpha}\sup_{x\ge 0}\left|D^3\widehat{V}^\alpha(x)\right| \le \frac{\gamma_-^2}{(1-\alpha)^2} \le \frac{\Gamma}{(\mu-\lambda)^2} \text{ as } \alpha \uparrow 1,$$

where Γ does not depend on α, λ, μ. If μ and λ are chosen so that $\rho = \lambda/\mu$ is away from 1, then the error bound remains bounded as $\alpha \uparrow 1$ while both the approximate value $\widehat{V}^\alpha$ and the true value grow like $1/(1-\alpha)$ as $\alpha \uparrow 1$.[6]

Let $\mu \downarrow \frac{1}{2}$ and $\lambda = 1 - \mu$ to place the queue in heavy traffic as before. Taking $\alpha_\rho = 1 - (1-\rho)^2$, we have

$$\left|V^{\alpha_\rho}(x) - \widehat{V}^{\alpha_\rho}(x)\right| \le \frac{\Gamma}{(1-\rho)^2},$$

consistent with the $o(1/(1-\rho)^3))$ accuracy of the Brownian approximation derived through process convergence.

Because $V^{\alpha_\rho}(x) \ge \frac{1}{(1-\rho)^3}$ for $x \ge \frac{1}{1-\rho}$, it follows that

$$\left|V^{\alpha_\rho}(x) - \widehat{V}^{\alpha_\rho}(x)\right| \le \Gamma(1-\rho)V^{\alpha_\rho}(x), \text{for all } x \ge \frac{1}{1-\rho}.$$

This establishes the asymptotic correctness of a Brownian approximation by means of Tayloring rather than by those of weak convergence. In contrast to Brownian approximations, *Tayloring is a purely analytical device.*

Finally, this is an opportunity to revisit the contribution of corners to the bound in Theorem 2. In a

queuing network in which all stations operate at utilization of $1-\sqrt{1-\alpha}=1-\rho$ as before, the fraction of time spent near corners at which two or more stations are idle is of the order of (or smaller than) $(1-\alpha)=(1-\rho)^2$.

6. Final Comments

In this paper, we have introduced Tayloring as a rigorous framework for value function approximation. Applied to a controlled chain in discrete time and space, we derive bounds grounded in PDE theory and propose a solution algorithm with performance guarantees. This paper is a first and by no means last step. Much remains open in terms of the scope—continuous time and space, finite and long-run average problems—and various algorithmic aspects. Here is a short informal discussion of these directions.

6.1. Continuous Time

Consider the $M/M/1$ queue, which is the continuous-time version of the discrete-time queue in Example 2. This queue has Poisson arrivals with rate λ and a single server with service times that are exponential with (controlled) parameter $u(x)$. Given holding and service rate cost $r_u(x)$, consider the problem

$$\min_U \mathbb{E}_x\left[\int_0^\infty e^{-(1-\alpha)t} r(X(t),U(t))dt\right].$$

The Bellman equation is given by

$$0=\min_{u\geq 0}\{r_u(x)+\lambda(V(x+1)-V(x)) + u\mathbb{1}\{x>0\}(V(x-1)-V(x))-(1-\alpha)V(x)\}.$$

Second-order Tayloring leads to the TCP

$$0=\min_{u\geq 0}\left\{r_u(x)+(\lambda-u)V'(x)+\frac{1}{2}(\lambda+u)V''(x) -(1-\alpha)V(x)\right\},\quad x>0,$$

and we use the boundary condition $V'(0)=0$. Generally, for a continuous-time chain on $\mathbb{Z}_+^d$ with transition-rate matrix $q_u(x,y)$ at state x, we have

$$(\mu_u)_i(x)=\sum_y q_u(x,y)(y_i-x_i) \text{ and}$$
$$(\sigma_u)_{ij}(x)=\sum_y q_u(x,y)(y_i-x_i)(y_j-x_j),\quad \text{for } i,j\in[d].$$

It is reasonable to conjecture that versions of Theorems 1 and 2 can be derived for the continuous-time case *with bounded rates* because uniformization provides an immediate mapping between the continuous and discrete-time models and hence between their TCPs. In a variety of practical models, the transition rates are unbounded as in the case of queues with abandonment (see, e.g., Weerasinghe and Mandelbaum 2013). Queueing models such as this one may serve as test cases for the extension to continuous time.

6.2. Continuous State Space

It is reasonable to conjecture that Theorems 1 and 2 still hold with modifications to μ_u and σ_u: for all $x\in\mathbb{R}_+^d$,

$$(\mu_u)_i(x)=\mathbb{E}_x^u[(X_1)_i-x_i] =\int_{\mathbb{R}_+} P^u(x,dy)(y_i-x_i),\quad i\in[d], \tag{34}$$

$$(\sigma_u^2)_{ij}(x)=\mathbb{E}_x^u\left[((X_1)_i-x_i)\left((X_1)_j-x_j\right)\right] =\int_{\mathbb{R}_+^2} P^u(x,dy)(y_i-x_i)(y_j-x_j),\quad i,j\in[d]. \tag{35}$$

Indeed, the continuous state space may seem initially to simplify things insofar as there is no need to extend μ_u and σ_u or to smooth the state space. This simplification, however, implies also losing the freedom, for example, to extend μ and σ^2 in a Lipschitz continuous way or smooth the corners of the state space. It is these freedoms that facilitate, in this paper, the application of the PDE theory of classical solutions in smooth domains.

6.3. Average Reward

Consider the average reward problem

$$V_*(x)=\max_U \liminf_{n\to\infty}\frac{1}{n+1}\mathbb{E}_x\left[\sum_{t=0}^n r(X_t,U(X_t))\right].$$

Consider the equation

$$\beta+h(x)=\max_{u\in\mathcal{U}(x)}\{r_u(x)+P^u h(x)\},$$

with $P^u h(x):=\sum_y P^u_{x,y}h(y)$. Under suitable conditions (e.g., Bertsekas 2007, chapter 4), if a constant β (together with a function h) solves this equation, then $V_*(x)\equiv\beta$. That is, β is the optimal long-run average.

Second-order Tayloring then gives rise here to the equation

$$\beta=\max_{u\in\mathcal{U}(x)}\{r_u(x)+\mathcal{L}_u h(x)\}.$$

Huang and Gurvich (2018) follow the Tayloring path in identifying nearly optimal policies for an $M/G/1$ service-rate control problem under a long-run average criterion. There the volume of arrivals λ serves as a natural scale parameter against which optimality can be measured. That is, because the value function scales with λ, one can express near optimality. It is meaningful to write optimality gap $=o(V_*^\lambda)$.

In venturing outside of queues, the long-run average cost criterion imposes a challenge: what is a

natural notion of near optimality? This is in contrast to the discounted case, in which the discount factor and the initial condition supported generalizable notions of scaling.

6.4. Finite-Horizon Problems

Consider a discrete time and space dynamic program on a finite horizon of length T. Let $V_t(x)$ be the value with t steps to go and starting at state x. The Bellman equation is given by

$$V_t(x) = \max_u \left\{ r_u(x) + \sum_y P^{u,t}_{x,y} V_{t-1}(y) \right\},$$

which we can rewrite as

$$0 = \max_u \left\{ r_u(x) + \sum_y P^{u,t}_{x,y}\big(V_t(x) - V_t(y)\big) + V_{t-1}(x) - V_t(x) - \sum_y P^{u,t}_{x,y}\big[V_{t-1}(x) - V_{t-1}(y) + V_t(x) - V_t(y)\big] \right\}.$$

Let

$$\mu_{u,t}(x) = \sum_y P^{u,t}_{x,y}(y-x), \quad \text{and } \sigma^2_{u,t}(x) = \sum_y P^{u,t}_{x,y}(y-x)^2.$$

Taking a second-order expansion in x and a first-order expansion in t, we arrive at the equation

$$0 = \max_u \left\{ r_u(x) + \mu_{u,t}(x)\frac{\partial}{\partial x} V_t(x) + \frac{1}{2}\sigma^2_{u,t}(x)\frac{\partial^2}{\partial x^2} V_t(x) - \frac{\partial}{\partial t} V_t(x) \right\}.$$

We drop any consideration of boundary condition from this informal outline. The approximation errors should depend on the second derivative in t, the third derivative in the state x, and the cross-derivative in x and t that arises from the term $\sum_y P^{u,t}_{x,y}[V_{t-1}(x) - V_{t-1}(y) + V_t(x) - V_t(y)]$. The connection between the original chain and the Taylored equation seems a straightforward extension of what we have done in this paper, yet it remains for future work to discover how all other ingredients, because they rely on PDE bounds, can (if at all) be combined to produce similar optimality-gap bounds.

6.5. State-Space Collapse (SSC)

A key benefit of asymptotic analysis in controlled queues is so-called state-space collapse—the reduction of problem dimensionality through the convergence of parts of the state space to degenerate points. Roughly speaking, SSC is rooted in the fact that controls can instantaneously direct more "power" to certain queues, a power that is greater in order of magnitude than the natural scale of the workload. Under SSC, some states in the original state space "disappear" in the asymptotic limit and become identical with a metastate that attracts them. For example, in a two-class, single-server queue with the longest-queue-first policy, all queue–state pairs (q_1, q_2) with the same sum $q_1 + q_2$ are "quickly" attracted to the metastate that captures this total queue length (that is equally split between the individual queues). In a setting absent of scaling (except the discount factor), it is not clear how to restore such effects.

It is reasonable to conjecture that—for a queuing network that exhibits SSC under heavy-traffic analysis—all states that are quickly attracted to a common metastate in the asymptotic limit will have a similar value-function value. It is therefore plausible that a flexible enough aggregation procedure (such as the one discussed in Remark 8) will capture this effect. This is yet to be explored.

6.6. TAPI Computation with Nonuniform Grids

In our computational examples, we do not rationalize the choice of the coarseness level h, and once h is chosen, we use it uniformly in the state space. Our bounds, however, might suggest a direction for improvement.

Finite difference methods for PDEs use finer grids in regions in which large gradients are expected and coarser grids in which the function is relatively "flat." This brings computational efficiency at little cost to accuracy. A similar logic applies to TAPI, as is nicely captured in the bound (30), which we rewrite here as

$$\mathbb{E}^{\widehat{U}_*}_x\left[\sum_{t=0}^{\infty} \bar{\alpha}^t_h h^{2+\beta}\left[D^2\widehat{V}_*\right]^*_{\beta, X^h_t \pm h}\right] + \mathbb{E}^{U^h_*}_x\left[\sum_{t=0}^{\infty} \bar{\alpha}^t_h h^{2+\beta}\left[D^2\widehat{V}_*\right]^*_{\beta, X^h_t \pm h}\right]. \tag{36}$$

The error depends on the interaction of the step size h with the supremum of $[D^2\widehat{V}_*]$ over a neighborhood of x. It makes sense to use large "boxes" in which $[D^2\widehat{V}_*]$ is relatively flat and vice versa. Ad hoc knowledge of the problem can help here. In the inventory problem of Section 4, for example, $\widehat{V}$ is approximately linear far from the origin, suggesting that one could use a large h in that part of the state space. We can use such understanding to build a (computationally beneficial) TCP-equivalent chain with nonuniform spacing over the state space.

Acknowledgments

The authors are grateful to the members of the review team for their constructive comments.

Endnotes

[1] This stream of the literature is in its infancy relative to the well-developed literature on convergence-based asymptotic optimality.

A key benefit of asymptotic analysis in controlled queues is so-called state-space collapse—the reduction of problem dimensionality through the convergence of parts of the state space to degenerate points. A framework to incorporate such dimensionality reduction into a Stein-type analysis is still absent (see Section 6).

[2] The question of how to truncate the state space of an MDP and how the truncated values converge to the true one is of general interest in MDP (see, e.g., Altman 1993 and the references therein). It is natural to choose M large enough so that the value functions remain relatively unchanged as one increases M further.

[3] Moving from OD boundary to FOT boundary does not change the conclusions for this example. The solution to the FOT-boundary TCP has the form $\widehat{V}_U(x) = g^\alpha(x)$, where $|D^3 g^\alpha(x)| \le \Gamma(1-\alpha)$ and $|D^2 g^\alpha(0)| \le \Gamma\sqrt{1-\alpha}$.

[4] It is here where we use the assumption that the action set is of the form ${}^0\mathcal{U}(x) = \{u \in \mathbb{R}^{d_a} : Au \le b(x)\} \cap \mathbb{D}$.

[5] $\widehat{U}_*(x_n) \to \widehat{U}_*(x)$ for all $x \in \partial\mathbb{R}^d_+$ and sequences $\{x_n\}$ with $x_n \in \mathbb{R}^d_{++}$ and $x_n \to x$.

[6] Rewrite $\widehat{V}^\alpha(x) = c_1 - \alpha\frac{\mu-\lambda}{(1-\alpha)^2} + \frac{x}{1-\alpha} + c_1(e^{\gamma_- x} - 1) = c_1 - \alpha\frac{\mu-\lambda}{(1-\alpha)^2} + \frac{1}{(1-\alpha)} \cdot \left(x - \frac{1}{\gamma_-}(e^{\gamma_- x} - 1)\right)$. It can be easily shown that $c_1 - \alpha\frac{\mu-\lambda}{(1-\alpha)^2} \approx \frac{1}{2(1-\alpha)(\mu-\lambda)}$ as $\alpha \uparrow 1$, so, because $\frac{1}{\gamma_-}(e^{\gamma_- x} - 1) \ge x$, $\widehat{V}^\alpha(x) \approx \frac{1}{2(1-\alpha)(\mu-\lambda)} + \frac{1}{1-\alpha} \cdot \left(x - \frac{1}{\gamma_-}(e^{\gamma_- x} - 1)\right) \ge \frac{1}{2(1-\alpha)(\mu-\lambda)}$.

References

Altman E (1993) Asymptotic properties of constrained Markov decision processes. *Zeitschrift für Oper. Res.* 37(2):151–170.

Ata B, Gurvich I (2012) On optimality gaps in the Halfin–Whitt regime. *Ann. Appl. Probab.* 22(1):407–455.

Bertsekas DP (2007) *Approximate Dynamic Programming*, Dynamic Programming and Optimal Control, vol. 2, 3rd ed. (Athena Scientific, Belmont, MA).

Bertsekas DP (2011) Approximate policy iteration: A survey and some new methods. *J. Control Theory Appl.* 9(3):310–335.

Bertsekas DP (2019) Feature-based aggregation and deep reinforcement learning: A survey and some new implementations. *IEEE/CAA J. Automatica Sinica* 6(1):1–31.

Bertsekas DP, Tsitsiklis JN (1996) *Neuro-Dynamic Programming* (Athena Scientific, Belmont, MA).

Borkar V, Budhiraja A (2004) Ergodic control for constrained diffusions: Characterization using HJB equations. *SIAM J. Control Optim.* 43(4):1467–1492.

Braverman A, Dai JG (2017) Stein's method for steady-state diffusion approximations of $M/Ph/n + M$ systems. *Ann. Appl. Probab.* 27(1):550–581.

Chen W, Huang D, Kulkarni AA, Unnikrishnan J, Zhu Q, Mehta P, Meyn S, Wierman A (2009) Approximate dynamic programming using fluid and diffusion approximations with applications to power management. *Proc. 48th IEEE Conf. Decision Control* (IEEE, Piscataway, NJ), 3575–3580.

Dai JG, Shi P (2017) A two-time-scale approach to time-varying queues in hospital inpatient flow management. *Oper. Res.* 65(2):514–536.

Dupuis PG, Ishii H (1990) On oblique derivative problems for fully nonlinear second-order elliptic partial differential equations on nonsmooth domains. *Nonlinear Anal.: Theory Methods Appl.* 15(12):1123–1138.

Dupuis PG, James MR (1998) Rates of convergence for approximation schemes in optimal control. *SIAM J. Control Optim.* 36(2):719–741.

Gilbarg D, Trudinger NS (2001) *Elliptic Partial Differential Equations of Second Order* (Springer-Verlag, New York).

Gurvich I (2014) Diffusion models and steady-state approximations for exponentially ergodic Markovian queues. *Ann. Appl. Probab.* 24(6):2527–2559.

Harrison JM (2013) *Brownian Motion and Stochastic Flow Systems* (Cambridge University Press, New York).

Huang J, Gurvich I (2018) Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue. *Oper. Res.* 66(4):1168–1188.

Koçağa YL, Ward AR (2010) Admission control for a multi-server queue with abandonment. *Queueing Systems* 65(3):275–323.

Kushner H, Dupuis PG (2013) *Numerical Methods for Stochastic Control Problems in Continuous Time*, vol. 24 (Springer-Verlag, New York).

Larsson S, Thomée V (2008) *Partial Differential Equations with Numerical Methods*, vol. 45 (Springer-Verlag, Berlin, Heidelberg).

Lieberman GM (2013) *Oblique Derivative Problems for Elliptic Equations* (World Scientific, Hackensack, NJ).

McShane EJ (1934) Extension of range of functions. *Bull. Amer. Math. Soc.* 40(12):837–842.

Moallemi C, Kumar S, Van Roy B (2008) Approximate and data-driven dynamic programming for queueing networks. Working paper, Stanford University, CA.

Powell WB (2007) *Approximate Dynamic Programming: Solving the Curses of Dimensionality* (John Wiley & Sons, Hoboken, NJ).

Weerasinghe A, Mandelbaum A (2013) Abandonment vs. blocking in many-server queues: Asymptotic optimality in the QED regime. *Queueing Systems* 75(2):279–337.

Whitt W (2002) *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues* (Springer-Verlag, New York).

Zhang BZ, Gurvich I (2018) Aggregation via local moment matching. Working paper, Cornell University, Ithaca, NY.

Anton Braverman is an assistant professor of the Operations Group at Kellogg School of Management, Northwestern University. His research is focused on stochastic modeling and applied probability. Some application domains of interest include ride-sharing services and healthcare operations.

Itai Gurvich is a professor at Cornell's School of Operations Research and Information Engineering and at Cornell Tech. His research focuses on the performance analysis and optimization of processing networks.

Junfei Huang is an associate professor in the Department of Decision Sciences and Managerial Economics at the Chinese University of Hong Kong. His research interests are in asymptotic analysis and optimal control of queuing systems and their applications in manufacturing and services.