Gradient-bounded dynamic programming for submodular and concave extensible value functions with probabilistic performance guarantees[☆],[☆☆]

Denis Lebedev *, Paul Goulart, Kostas Margellos

Department of Engineering Science, University of Oxford, Oxford, OX1 3PJ, United Kingdom

ABSTRACT

We consider stochastic dynamic programming problems with high-dimensional, discrete state-spaces and finite, discrete-time horizons that prohibit direct computation of the value function from a given Bellman equation for all states and time steps due to the "curse of dimensionality". For the case where the value function of the dynamic program is concave extensible and submodular in its state-space, we present a new algorithm that computes deterministic upper and stochastic lower bounds of the value function in the realm of dual dynamic programming. We show that the proposed algorithm terminates after a finite number of iterations. Furthermore, we derive probabilistic guarantees on the value accumulated under the associated policy for a single realisation of the dynamic program and for the expectation of this value. Finally, we demonstrate the efficacy of our approach on a high-dimensional numerical example from delivery slot pricing in attended home delivery.

© 2021 Published by Elsevier Ltd.

1. Introduction

Multi-stage optimal control problems arise in many application areas. Using dynamic programming (DP) to solve these problems is mostly restricted to low-dimensional problem instances, since high-dimensional state and action spaces prohibit direct computation of the value function of the DP due to the so-called "curse of dimensionality".

Numerous approximation approaches have been proposed in the literature. To classify our work, it is illustrative to categorise these approaches by two characteristics, namely whether they were designed to be used for problems with either continuous or
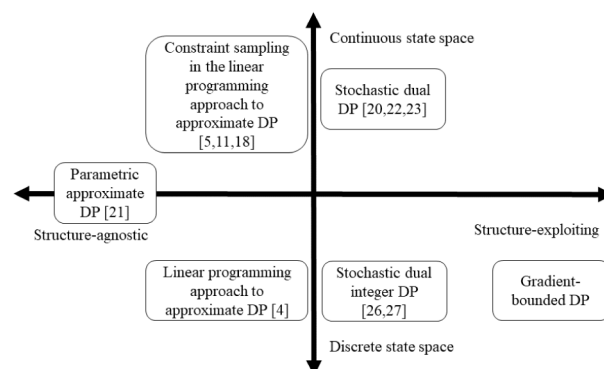

Fig. 1. Schematic grouping of selected approximate DP literature based on how general-purpose or problem-specific the approaches are and if they are limited to a discrete or continuous state space.

discrete states and whether they are rather agnostic or exploitative with regards to particular mathematical structures in the underlying problem. We schematically lay out some well-studied methods from the literature in Fig. 1.

One of the most general and widely used approaches is to model the value function of the DP as a linear combination of

---

* Corresponding author.
E-mail addresses: denis.lebedev@eng.ox.ac.uk (D. Lebedev), paul.goulart@eng.ox.ac.uk (P. Goulart), kostas.margellos@eng.ox.ac.uk (K. Margellos).

basis functions, which are sometimes also called features (Powell, 2007, Chapter 8.2). Approximation of the value function then reduces to the problem of estimating the weight coefficients of the linear combination. While this approach is very flexible, it might not be obvious which and how many basis functions should be chosen for a particular problem at hand.

A more structured way to solve such an approximate DP formulation is given by the linear programming approach to approximate DP (de Farias & Van Roy, 2003), which was originally devised for finite state and action spaces. The general idea is to change the Bellman equation of the DP by an epigraphic reformulation into a linear program, where the Bellman equation appears as a constraint for each state and action pair. Consequently, the number of constraints is often prohibitively large even for problems with finite state and action spaces, but sampling a large enough number of these constraints may be sufficient to generate a good enough solution for the particular problem at hand de Farias and Van Roy (2004). The same strategy is also exploited in Kariotoglou, Summers, Summers, Kamgarpour, and Lygeros (2013) and Mohajerin Esfahani, Sutter, Kuhn, and Lygeros (2018) to turn semi-infinite programs arising in case of infinite state and actions spaces to finite ones. Still, the quality of the approximation once again hinges on the need to choose suitable basis functions.

Certain structures of the value function can simplify the process of selecting basis functions. For example, if the value function of a DP over continuous states is convex,[1] then stochastic dual DP provides an alternative to the above-mentioned approximate DP techniques (Pereira & Pinto, 1991; Shapiro, 2011). The main objective is to under-approximate the value function by the pointwise maximum of a finite number of hyperplanes in the state space for all time steps in the DP. These hyperplanes are added iteratively by first generating an approximately optimal sample path of states forward in time and then adding a hyperplane at each time step along this sample path backward in time, which refines the value function along this sample path. Variants of this algorithm have also been developed for systems with piecewise-quadratic value functions (Warrington, Beuchat, & Lygeros, 2019).

Further developments of stochastic dual DP are concerned with discrete state systems (Zhang & Sun, 2019; Zou, Ahmed, & Sun, 2019). While exploiting convex problem structure and bi-passing the need to choose basis functions, these approaches still suffer from some limitations, such as being constrained to linear systems (Zou et al., 2019) or having to solve a non-convex optimisation problem to add hyperplanes to the representation of the approximate value function at each time step (Zhang & Sun, 2019).

In this paper, we present a variant of the stochastic dual DP algorithm, termed gradient-bounded DP, for problems with discrete states and value functions that are concave extensible and submodular. One example of a problem whose value function has these properties can be found in the so-called revenue management problem in attended home delivery (Lebedev, Goulart, & Margellos, 2019, 2021). Similar to stochastic dual dynamic (integer) programming, we represent the value function of the DP as the pointwise minimum of affine functions over states. And in contrast to the existing extensions to discrete states, our approach does not suffer from the above-mentioned limitations. We also demonstrate the effectiveness of our approximation approach on a numerical example of the revenue management problem in attended home delivery.

Furthermore, we address another problem that may be encountered by *all* approximation approaches mentioned above. In the case of finite-time multistage stochastic optimal control problems, the performance of an approximately optimal decision policy may be evaluated by simulating the DP forward in time and thus obtaining a performance metric. Since the problem is stochastic, so is this performance metric, which follows a stationary, yet unknown distribution under the approximately optimal policy (assuming that this policy is stationary, i.e. the probability of making a decision at any state–time pair is independent of the simulation run). Hence, as a second central contribution to this paper, we derive bounds on the tail and expectation of the probability distribution of a performance metric obtained from a finite number of its samples.

The paper is structured as follows: Section 2 formulates our problem of interest and the assumptions that our work builds upon. In Section 3, we present a novel algorithm to compute approximately optimal policies for value functions over discrete state-spaces under assumptions on submodularity and concave extensibility. Section 4 derives deterministic upper bounds and stochastic lower bounds to the exact value function and shows convergence of the algorithm in a finite number of iterations. In Section 5, we present an algorithm that validates the policy obtained in Section 3 by computing sample profits, their empirical mean and their standard deviation. Section 6 details our theoretical results on tail and expectation bounds of the sample profits obtained in Section 5. In Section 7, we present a numerical example on a high-dimensional problem that is unsolvable by direct computation. Finally, we conclude in Section 8 and provide directions for future research.

Notation: For any $s \in \mathbb{N}$, let 1_s be a column vector of all zeros apart from the sth entry, which equals 1. Furthermore, we define the convention that 1_0 is a vector of zeros. Let $\mathbf{1}$ denote a vector of ones. Let $\langle \cdot, \cdot \rangle$ denote the standard inner product of its arguments. Let $\lfloor \cdot \rfloor$ denote the floor function, i.e. the greatest integer less than or equal to its argument. Let $\mathbb{E}$ denote the expectation operator, let $\Pr(\cdot)$ denote the probability of its argument and let $\mathbb{1}(\cdot)$ denote the indicator function.

2. Problem statement

We consider a discrete-space, discrete-time, finite horizon DP. Define discrete states $x \in X \subset \mathbb{Z}^n$ and continuous and/or discrete decision variables $d \in D \subset \mathbb{Z}^a \times \mathbb{R}^b$. Define the set $S := \{1, 2, \ldots, n\}$. Let the transition probability between two states x and y under decision d be $P_{x,y}(d)$, where we require $P_{x,y}(d) \geq 0$ for all $(x, y, d) \in X \times X \times D$. For all $x \in X$, we impose that $\sum_{y \in Y_+(x)} P_{x,y}(d) = 1$, where $Y_+(x) := \{x + 1_s\}_{s \in S \cup \{0\}}$. This requirement implies that transitions in x are only possible in the positive direction and by at most a unit step along one dimension. Such models are typical for order-taking processes (Yang & Strauss, 2017; Yang, Strauss, Currie, & Eglese, 2016). Furthermore, we define a finite time horizon $T := \{1, 2, \ldots, \bar{t}\}$, a stage revenue function $g : \mathbb{Z}^n \times \mathbb{Z}^n \times (\mathbb{Z}^a \times \mathbb{R}^b) \to \mathbb{R}$ and a terminal cost function $C : \mathbb{Z}^n \to \mathbb{R}$ to construct the following DP:

$$V_t(x) := \max_{d \in D} \left\{ \sum_{y \in Y_+(x)} P_{x,y}(d) \left(g(x, y, d) + V_{t+1}(y) \right) \right\}$$
$$\forall (x, t) \in X \times T, \text{ where}$$
$$V_{\bar{t}+1}(x) := -C(x) \quad \forall x \in X. \tag{1}$$

It is not strictly necessary for g to be independent of t as long as the assumptions stated below can be satisfied. However, as our

---

[1] The value function needs to be convex if costs are minimised or alternatively concave if rewards are maximised.

interest lies in time-independent problems and to ease notation, we ignore time-dependency of g in this paper.[2] To represent the DP more compactly, we notice that (1) is a time-independent mapping from V_{t+1} to V_t for all $t \in T$, which makes it possible to define the so-called Bellman operator $\mathcal{T}$ through the relationship

$$V_t = \mathcal{T} V_{t+1}, \text{ for all } t \in T. \tag{2}$$

Notice that the argument of $\mathcal{T}$ is a functional. We next introduce several definitions needed to state the assumptions that we impose on the DP in (1).

Definition 1. A function $f : \mathbb{Z}^n \to \mathbb{R}$ is *submodular* if it satisfies

$$f(\max(y,z)) + f(\min(y,z)) \le f(y) + f(z) \tag{3}$$

for all $(y, z) \in \mathbb{Z}^n \times \mathbb{Z}^n$, where the maximum and minimum are taken elementwise.

The following two definitions are commonly used in discrete convex analysis:

Definition 2. Let $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Then the *concave closure* $\tilde{f} : \mathbb{R}^n \to \mathbb{R} \cup -\infty$ of a function $f : \mathbb{Z}^n \to \mathbb{R} \cup -\infty$ is defined as (Murota & Shioura, 2001, equation (2.1))

$$\tilde{f}(x) := \inf_{a,b} \left\{ \langle a, x \rangle + b \mid \langle a, y \rangle + b \ge f(y) \quad \forall y \in \mathbb{Z}^n \right\}.$$

Note that the concave closure is identical to the so-called *concave hull*.

Definition 3. A function $f : \mathbb{Z}^n \to \mathbb{R} \cup -\infty$ is *concave extensible* if and only if the evaluations of f coincide with the evaluations of its concave closure $\tilde{f}$ (Murota & Shioura, 2001, Lemma 2.3), i.e. $f(x) = \tilde{f}(x)$, for all $x \in \text{dom}(\tilde{f})$.

These definitions allow us to state the assumptions that we impose on the DP in (1):

Assumption 1. The function $-C$ is submodular and concave extensible in x.

Assumption 2. We assume that the functions D, g and $P_{x,y}$ for all $(x, y) \in X \times X$ and T are chosen such that the Bellman operator preserves concave extensibility and submodularity of any concave extensible and submodular value function, i.e. if V_{t+1} is submodular and concave extensible, then $V_t = \mathcal{T} V_{t+1}$ also has these properties for all $t \in T$.

In Lebedev, Goulart, and Margellos (2021, Theorem 2), it is shown that, under mild technical assumptions on the customer arrival rate, these assumptions are satisfied for the revenue management problem considered in Section 7.

3. Value function approximation algorithm

We first state our proposed approximation procedure in Algorithm 1 and subsequently detail the individual algorithm steps. Inspired by stochastic dual DP techniques (Shapiro, 2011), the main idea of our algorithm is to alternate between generating sample paths in "forward sweeps" and refining the value function in "backward sweeps". We term our approximation algorithm "gradient-bounded DP", since it exploits properties of the gradient of the approximate value function, namely submodularity

---

and concave extensibility, to compute an upper bound to the exact value function of the DP. The following sections describe this procedure in detail.

Algorithm 1 Gradient-bounded dynamic programming

1: Initialise parameters: $X, D, P_{x,y}, T, g, C$ and $i_{\max}$
2: Initialise $Q_t^0(x) \leftarrow \infty$, for all $(x, t) \in X \times T$
3: Initialise $Q_{\bar{t}+1}^0(x) \leftarrow -C(x)$, for all $x \in X$
4: **for** $i \in \{1, 2, \dots, i_{\max}\}$ **do**
5: $\quad x_1^i \leftarrow 0$
6: $\quad$ **for** $t \in \{1, 2, \dots, \bar{t}\}$ **do** $\qquad \triangleright$ "Forward sweep"
7: $\qquad d_t^i \leftarrow d^* \in \underset{d \in D}{\text{argmax}} \left\{ \sum_{x_{t+1}^i \in Y_+(x_t^i)} P_{x_t^i, x_{t+1}^i}(d) \right.$
$\qquad\qquad\qquad \left. \times \left(g(x_t^i, x_{t+1}^i, d) + Q_{t+1}^{i-1}(x_{t+1}^i) \right) \right\}$
8: $\qquad x_{t+1}^i \leftarrow x_t^i + \underset{x_{t+1}^i}{\text{sample}} \left\{ P_{x_t^i, x_{t+1}^i}(d_t^i) \right\}$
9: $\quad$ **end for**
10: $\quad l(i) \leftarrow \sum_{t=1}^{\bar{t}} g(x_t^i, x_{t+1}^i, d_t^i) - C(x_{\bar{t}+1})$
11: $\quad$ **for** $t \in \{\bar{t}, \bar{t}-1, \dots, 1\}$ **do** $\quad \triangleright$ "Backward sweep"
12: $\qquad Z(x_{t+1}^i) \leftarrow \{x_{t+1}^i + 1_s + 1_{s'}\}_{\substack{s \in S \cup \{0\}, \\ s' \in S \cup \{0\}}}$
13: $\qquad$ **if** Q_{t+1}^{i-1} is submodular on $Z(x_{t+1}^i)$ **then**
14: $\qquad\quad H^* \leftarrow$ unique hyperplane through $\left\{ \left(y, (\mathcal{T} Q_{t+1}^{i-1})(y) \right) \right\}_{y \in Y_+(x_{t+1}^i)}$
15: $\qquad$ **else**
16: $\qquad\quad j^* \in \underset{j \in J_{t+1}^{i-1}(x_{t+1}^i)}{\text{argmin}} \left\{ \left(\mathcal{T} H_{t+1}^{j-1} \right)(x_{t+1}^i) \right\}$
17: $\qquad\quad H^* \leftarrow \mathcal{T} H_{t+1}^{j^*-1}$
18: $\qquad$ **end if**
19: $\qquad Q_t^i \leftarrow \min\{H^*, Q_t^{i-1}\}$
20: $\qquad t \leftarrow t - 1$
21: $\quad$ **end for**
22: $\quad u(i) \leftarrow Q_1^i(0)$
23: **end for**

3.1. Initialisation

We first initialise all parameters of the DP in (1) (step 1). Denote the maximum number of iterations by $i_{\max} \in \mathbb{N}$ and let $I := \{0, 1, \dots, i_{\max}\}$. Let the value function approximation Q_t^i for all $(i, t) \in I \times T$ be the pointwise minimum of a finite number of affine functions, i.e.

$$Q_t^i(x) := \min_{j \in \{0, 1, \dots, i\}} H_t^j(x), \text{ for all } x \in X, \tag{4}$$

where $H_t^j : X \mapsto \mathbb{R}$ describes a hyperplane, i.e.

$$H_t^j(x) := \langle a_t^j, x \rangle + b_t^j, \text{ for all } x \in X, \tag{5}$$

with $a_t^j \in \mathbb{R}^n$, $b_t^j \in \mathbb{R}$ for all $(t, j) \in T \times I$. We characterise the set of supporting hyperplanes at x as

$$J_t^i(x) := \underset{j \in \{0, 1, \dots, i\}}{\text{argmin}} \left\{ \langle a_t^j, x \rangle + b_t^j \right\} \tag{6}$$

for all $(x, i, t) \in X \times I \times T$. Notice that the aforementioned functions are defined for all states $x \in X$. In practice, the number of states may be prohibitively large to compute these functions for all states, however, for our purposes, we will only ever evaluate these functions locally at certain $x \in X$, which is possible since the maximum number of hyperplanes $i_{\max}$ will be relatively moderate.

We construct Q_t^i as a successively tighter upper bound of V_t (as i increases), i.e. $V_t(x) \le Q_t^i(x) \le Q_t^{i-1}(x)$ for all $(x, i, t) \in X \times$

$(I \setminus \{0\}) \times T$. In the ith "backward sweep", H_t^i is added to Q_t^{i-1} for all $t \in T$ to form Q_t^i. To initialise Q_t^0, one could simply set Q_t^0 to be a single affine function with $a_t^0 = 0$ and $b_t^0 = \infty$, such that Q_t^0 is indeed an upper bound to V_t for all $t \in T$ (step 2). We discuss the possibility of closer initialisations in the context of our example in Section 7. We also initialise $Q_{\bar{t}+1}^i(x) := V_{\bar{t}+1}(x) = -C(x)$ for all $(x, i) \in X \times I$, which is a tight upper bound by the construction of the DP in (1) (step 3).

3.2. "Forward sweep"

Fix any iteration $i \in I \setminus \{0\}$. In each "forward sweep", we solve an approximate version of the Bellman equation in (1) forward in time, i.e. by replacing V_t with its approximation Q_t^{i-1} (step 7). Hence, we compute suboptimal d_t^i for all $t \in T$ and simulate state transitions by sampling from the transition probability distribution given the approximately optimal decisions (step 8). This defines a sample path x_t^i for all $t \in T \cup \{\bar{t} + 1\}$. At the end of each "forward sweep", we compute a stochastic lower bound on the total expected profit $V_1(0)$, which we denote by $l(i)$ for all $i \in I \setminus \{0\}$ (step 10). We show that this is indeed a stochastic lower bound in Section 6.

3.3. "Backward sweep"

Fix any iteration $i \in I$. In each "backward sweep", we first check if Q_{t+1}^{i-1} is submodular on $Z(x_{t+1}^i)$ by computing the sign of (3) for all possible pairs of points $(y, y') \in Z(x_{t+1}^i) \times Z(x_{t+1}^i)$, such that $y \neq y'$ (step 12). If the inequality in (3) holds for all these points, we locally compute the exact DP stage problem on the set $Y_+(x_{t+1}^i)$, i.e. $\{\mathcal{T}Q_{t+1}^{i-1}(y)\}_{y \in Y_+(x_{t+1}^i)}$, to construct the hyperplane through $\{(y, (\mathcal{T}Q_{t+1}^{i-1})(y))\}_{y \in Y_+(x_{t+1}^i)}$ (step 14). Then, the resulting added hyperplane is an upper bound to $V_t(x)$ for all $x \in X$, as shown in Section 4.

In the opposite case, we need to compute a submodular upper bound on Q_{t+1}^{i-1}, which is readily given by the hyperplanes from which Q_{t+1}^{i-1} is constructed. Therefore, we select the hyperplane $H_{t+1}^{j^*-1}$ that minimises the value at the evaluation point x_t^i, which therefore locally creates the tightest upper bound (step 16). It may be possible to construct other submodular upper bounds to non-submodular Q_{t+1}^{i-1}, however, steps 16 and 17 of Algorithm 1 offer a simple implementation. Finally, we update the value function approximation as the pointwise minimum of the approximation from the previous iteration and the newly constructed hyperplane (step 19). We also compute an upper bound, $u(i)$ for all $i \in I \setminus \{0\}$, on the total expected profit $V_1(0)$ (step 22). We show that this is indeed an upper bound in Section 4.

4. Approximation algorithm properties

In this section, we show our main theoretical results on bounds on the exact value function and convergence properties of Algorithm 1. Proofs not included in this section can be found in Appendix.

Proposition 1. *Under Assumptions 1 and 2, the approximate value function is an upper bound to the exact finite horizon value function, i.e. $Q_t^i(x) \geq V_t(x)$ for all $(x, i, t) \in X \times I \times T$.*

Corollary 2. *Under Assumptions 1 and 2, the value of $u(i)$ is an upper bound to the exact total expected profit, i.e. $u(i) \geq V_1(0)$ for all $i \in I \setminus \{0\}$.*

Proof. This result follows immediately from Proposition 1 and by observing that $u(i) = Q_1^i(0)$ for all $i \in I \setminus \{0\}$ from step 22 of Algorithm 1.

Proposition 3. *The value of $l(i)$ is a stochastic lower bound to the expected total profit, i.e. $\mathbb{E}[l(i)] \leq V_1(0)$ for all $i \in I \setminus \{0\}$.*

Proof. For any $i \in I \setminus \{0\}$, the value of $l(i)$ is obtained from suboptimal decisions d_t^i for all $t \in T$, due to the use of Q_{t+1}^{i-1} instead of the exact (yet unavailable) V_{t+1} in step 7 of Algorithm 1. It follows that d_t^i is not a maximiser of the exact DP in (1) which, by the principle of optimality, implies that the expected value accumulated under this suboptimal policy will not be greater than the value obtained under the optimal policy. Hence, $\mathbb{E}[l(i)] \leq V_1(0)$ for all $i \in I \setminus \{0\}$.

The stochastic dual DP algorithm converges asymptotically in i to the exact value function (Shapiro, 2011). We can strengthen this result for our algorithm by exploiting the fact that the set of states X is finite. Hence, the proposed algorithm converges in a finite number of steps under the following minor modification to Algorithm 1.

Algorithm 2 Resampling procedure replacing step 8 of Algorithm 1

1: $m \leftarrow \lfloor (i-1)/|X| \rfloor$

2: $x_{t+1}^i \leftarrow x_t^i + \underset{x_{t+1}^i}{\text{sample}} \left\{ P_{x_t^i, x_{t+1}^i}(d_t^i) \right\}$

3: **if** $t = \bar{t} - m$ and $x_{t+1}^i \in \left\{ x_{t+1}^j \mid m|X|+1 \leq j < i \right\}$ **then**

4: $\quad x_{t+1}^i \leftarrow$ sample (with uniform probability)

$$\text{from } X \setminus \left\{ x_{t+1}^j \mid m|X|+1 \leq j < i \right\}$$

5: **end if**

Notice that for an arbitrary (i, t), m in the if-statement in step 3 of Algorithm 2 ensures that every state $x \in X$ is sampled every $|X|$ iterations.

Proposition 4. *Under Assumptions 1 and 2, the gap $u(i) - \mathbb{E}[l(i)]$ for all $i \in I \setminus \{0\}$ converges to 0 in at most $\bar{t}|X|$ iterations of Algorithm 1, when using the resampling procedure of Algorithm 2.*

Note that it is likely to take an unacceptably large number of iterations for the algorithm to converge to the exact value function due to the large number of states $|X|$. Since the value function is computationally expensive to calculate for all states, we seek to generate closer approximations at points that are likely to be visited, i.e. points on the sample path, and to use this information to save on approximation accuracy for less likely samples.

Our ultimate objective is to solve problems with large state spaces ($|X| \approx 10^{20}$) and long time horizons ($|T| \approx 10^4$). In such scenarios, the need to resample the state as detailed in Algorithm 2 becomes negligible, because the required number of iterations to reach convergence is much larger than the maximum acceptable number of iterations. Therefore, from a practical point of view, we do not use Algorithm 2. In this case, the proposed algorithm only converges asymptotically to the exact value function instead of in a finite number of steps, just as in stochastic dual DP (Shapiro, 2011).

5. Proposed validation algorithm

As noted in the previous section, absolute convergence of the approximate value function to the exact value function cannot be achieved for industry-sized problems due to the "curse of dimensionality". The performance of the algorithm, i.e. how close the stochastic lower bound $l(i)$ is to the deterministic upper

bound $u(i)$ for any $i \in I \setminus \{0\}$, can only be validated statistically to a certain level of probabilistic confidence. To this end, we will generate a set of validation samples as described in Algorithm 3 and detailed further below.

Algorithm 3 Proposed validation algorithm

1: Compute approximation: $Q_t^{i_{\max}}$, for all $t \in T \setminus \{0\} \cup \{\bar{t} + 1\}$
2: Initialise number of validation samples $k_{\max}$
3: **for** $k \in K := \{1, 2, \ldots, k_{\max}\}$ **do**
4:      $x_1^k \leftarrow 0$
5:      **for** $t \in T$ **do**   ▷ "Forward validation sweep"
6:          $d_t^k \leftarrow d^* \in \underset{d \in D}{\operatorname{argmax}} \left\{ \sum_{x_{t+1}^k \in X} P_{x_t^k, x_{t+1}^k}(d) \right.$
                         $\left. \times \left(g(x_t^k, x_{t+1}^k, d) + Q_{t+1}^{i_{\max}}(x_{t+1}^k) \right) \right\}$
7:          $x_{t+1}^k \leftarrow x_t^k + \underset{x_{t+1}^k}{\operatorname{sample}} \left\{ P_{x_t^k, x_{t+1}^k}\left(d_t^k\right) \right\}$
8:      **end for**
9:      $l_v(k) \leftarrow \sum_{t=1}^{T} g(x_t^k, x_{t+1}^k, d_t^k) - C(x_{\bar{t}+1})$
10: **end for**
11: $\bar{l}_v \leftarrow k_{\max}^{-1} \sum_{k=1}^{k_{\max}} l_v(k)$
12: $\sigma_v \leftarrow \sqrt{(k_{\max} - 1)^{-1} \sum_{k=1}^{k_{\max}} \left(l_v(k) - \bar{l}_v\right)^2}$

We first compute the approximation obtained in Algorithm 1 (step 1). We denote the maximum number of validation samples by $k_{\max} \in \mathbb{N}$ and let $K := \{1, 2, \ldots, k_{\max}\}$ (step 2). We then compute $k_{\max}$ "forward validation sweeps", where in each of them we use our most refined estimate, $Q_{t+1}^{i_{\max}}$ as our approximate value function (steps 5–8). After each sweep $k \in K$, we compute the stochastic lower bound $l_v(k)$ on the total expected profit, similarly to $l(i)$ for all $i \in I \setminus \{0\}$ in Algorithm 1 (step 9). We then compute the sample mean profit $\bar{l}_v$ and unbiased empirical standard deviation σ_v of the set of sampled lower bounds $\{l_v(k)\}_{k \in K}$ (steps 11–12). As detailed in the next section, these quantities will be used to generate one-sided confidence intervals, quantifying the performance of the decision policy associated with the approximate value function $Q_{t+1}^{i_{\max}}$.

6. Validation algorithm properties

In this section we state the main theoretical properties of our validation procedure. The proofs can be found in Appendix. We use $\{l_v(k)\}_{k \in K}$, $\bar{l}_v$ and σ_v from Algorithm 3 to derive two different measures for the performance guarantee. The first is a probabilistic bound on the tail of the distribution of a single lower bound sample, i.e. a value for $l(k_{\max} + 1)$ that is reached or exceeded with $1 - \alpha$ confidence for a user-defined $\alpha \in (0, 1)$. As we will see later in Section 7, this bound is not necessarily indicative of the expectation of $\bar{l}_v$, since even under the profit-maximising decision policy, some variance will persist in $l(k_{\max} + 1)$ from the randomness of the state transitions. Therefore, we also derive a bound on the expectation of the empirical sample mean $\bar{l}_v$ that holds with confidence $1 - \alpha^{\mathbb{E}}$, where $\alpha^{\mathbb{E}} \in (0, 1)$ can be chosen by the user.

6.1. Tail bounds

In this section, we present two tail bounds of the distribution of $l_v(k_{\max} + 1)$. Let $[l_-, l_+]$ denote the (finite) support of the distribution of $l_v(k)$ for any $k \in K \cup \{k_{\max} + 1\}$ and let F_K denote the empirical cumulative distribution function of $\{l_v(k)\}_{k \in K}$, i.e. $F_K(l) = k_{\max}^{-1} \sum_{k \in K} \mathbb{1}(l_v(k) \geq l)$. We derive two tail bounds of the distribution of $l_v(k_{\max} + 1)$ with a given confidence level $(1 - \alpha) \in (0, 1)$, which is mildly restricted for the first bound due to the next assumption.

Assumption 3. Assume that $\alpha > \theta_C := \Pr(\sigma_v = 0)$.

The value of θ_C will often be negligibly small, since $l_v(k)$ for all $k \in K$ is highly unlikely to take identical values due to the typically high-dimensional state space and long time horizon. We show this later in Section 7.2.

Proposition 5. The inequality $\Pr(l_v(k_{\max} + 1) > l^*) \geq 1 - \alpha$ holds (i) under Assumption 3, if $\alpha \in (\theta_C, 1)$ and $l^* = l_C$, the empirical Cantelli bound given by

$$l_C := \bar{l}_v - \sigma_v \sqrt{\frac{(1 - \alpha)(k_{\max} - 1)}{(\alpha - \theta_C)k_{\max}}}, \quad or \tag{7}$$

(ii) if $\alpha \in (0, 1)$ and $l^* = l_D$, the Dvoretzky–Kiefer–Wolfowitz bound given by

$$l_D := \sup \left\{ l \in [l_-, l_+] \;\middle|\; F_K(l) \leq \alpha - \theta_D - \sqrt{\frac{\ln\left(\frac{1}{\theta_D}\right)}{2k_{\max}}} \right\}, \tag{8}$$

where $\theta_D \in (0, \alpha)$ is a user-defined parameter.

For l_D, we find the θ_D, which maximises the value of the bound, from the so-called Lambert W function.

Definition 4. Let the Lambert W function be implicitly defined as $W_i : \mathbb{R} \rightarrow \mathbb{R}$, such that $W_i(x) \exp(W_i(x)) = x$ for $i \in \{0, -1\}$, where $W_0(x) > -1$ is called the principal branch and $W_{-1}(x) \leq -1$ is called the lower branch.

Lemma 6. For any $\alpha \in (0, 1)$, the value of l_D is maximised at

$$\theta_D = \min \left\{ \alpha, \sqrt{\exp\left(W_{-1}\left(\frac{-1}{4k_{\max}}\right)\right)} \right\}. \tag{9}$$

The bounds l_C and l_D, are termed after Cantelli's inequality (Cantelli, 1928) and the Dvoretzky–Kiefer–Wolfowitz (Massart, 1990) inequality, respectively. These inequalities are critical for showing that the bounds are indeed reached or exceeded with probability $1 - \alpha$. By Proposition 5, we can always choose the tighter, i.e. greater, of the two bounds and we will see later in Section 7 that the selection of α and $k_{\max}$ influences which bound is preferred.

6.2. Expectation bounds

Similarly to the tail bounds, we now state our theoretical results on two bounds on the expectation of $\bar{l}_v$, denoted by $\mathbb{E}\bar{l}_v$.

Proposition 7. Fix any significance level $\alpha^{\mathbb{E}} \in (0, 1)$. Then $\Pr(\mathbb{E}\bar{l}_v > l^*) \geq 1 - \alpha^{\mathbb{E}}$, for all $l^* \in \{l_B^{\mathbb{E}}, l_D^{\mathbb{E}}\}$, where:
(i) $l_B^{\mathbb{E}}$ is the empirical Bernstein bound given by

$$l_B^{\mathbb{E}} := \bar{l}_v - \sqrt{\frac{2\sigma_v^2 \ln(2/\alpha^{\mathbb{E}})}{k_{\max}}} - \frac{7(l_+ - l_-)\ln(2/\alpha^{\mathbb{E}})}{3(k_{\max} - 1)} \quad and \tag{10}$$

(ii) $l_D^{\mathbb{E}}$ is the expectation Dvoretzky–Kiefer–Wolfowitz bound given by

$$l_D^{\mathbb{E}} := \int_{\max\{0, l_-\}}^{\max\{0, l_+\}} 1 - \min\left\{1, F_K(l) + \sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}}\right\} dl$$

$$- \int_{\min\{0, l_-\}}^{\min\{0, l_+\}} \min\left\{1, F_K(l) + \sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}}\right\} dl$$

$$+ \max\{0, l_-\} - \min\{0, l_+\}. \tag{11}$$

The bounds $l_B^{\mathbb{E}}$ and $l_D^{\mathbb{E}}$ are termed after the empirical Bernstein Maurer and Pontil (2009) and Dvoretzky–Kiefer–Wolfowitz (Massart, 1990) inequalities, respectively. The proof of Proposition 7(i) is given in Maurer and Pontil (2009). It can be shown that $l_D^{\mathbb{E}}$ is at least as tight as Hoeffding's concentration bound (Hoeffding, 1963), given by

$$l_H^{\mathbb{E}} := \bar{l}_v - (l_+ - l_-)\sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}}. \tag{12}$$

In fact, under an additional technical assumption, we show that the expectation Dvoretzky–Kiefer–Wolfowitz bound is strictly better than Hoeffding's bound.

Assumption 4. We assume that α and $k_{\max}$ are chosen to satisfy $\sqrt{\ln(1/\alpha^{\mathbb{E}})/(2k_{\max})} > k_{\max}^{-1}$.

Assumption 4 is very mild, since even for only a single observation $k_{\max} = 1$, the critical value of $\alpha^{\mathbb{E}}$ would be $e^{-2} \approx 13.5\%$, which is much larger than typical significance levels, e.g. 5% or 1%. Taking any smaller value of $\alpha^{\mathbb{E}}$ than the critical value will ensure that Assumption 4 is always satisfied. Furthermore, for $k_{\max} > 1$, the constraint on $\alpha^{\mathbb{E}}$ is even less restrictive.

Proposition 8. *Under Assumption 4, the expectation Dvoretzky–Kiefer–Wolfowitz bound is strictly tighter than Hoeffding's concentration bound, i.e. $l_D^{\mathbb{E}} > l_H^{\mathbb{E}}$ for all $\alpha^{\mathbb{E}} \in (0, 1)$.*

Finally, we note that other bounds have also been proposed in the literature, e.g. Shapiro (2011) assumes that the distribution of $\bar{l}_v$ is Gaussian and determines confidence intervals based on the corresponding standard score, i.e. a Gaussian lower bound on the expectation of $\bar{l}_v$ would be

$$l_G^{\mathbb{E}} := \bar{l}_v - z(\alpha^{\mathbb{E}})\frac{\sigma_v}{\sqrt{k_{\max}}}, \tag{13}$$

where $z(\alpha^{\mathbb{E}})$ is the standard score of the Gaussian distribution (in fact, Student's t-distribution, especially for small sample sizes $k_{\max}$, since the true variance of the underlying distribution of $\bar{l}_v$ is approximated by σ_v^2). We compare this with our proposed bounds in Section 7.

7. Numerical example

We demonstrate our algorithm on an example of the so-called revenue management problem in attended home delivery. The objective is to price delivery time windows, called "slots", dynamically over a finite time horizon to control the customer purchasing process to maximise profits while ensuring that all orders can still be fulfilled.

In this problem, S is the set of delivery slots and the components of x are the number of orders placed in every delivery slot. The feasible set of states X is defined by the maximum state vector $\bar{x}$, i.e. $X := \{x \in \mathbb{Z}^n \mid 0 \le x \le \bar{x}\}$. The set of delivery slot price vectors is $D := \{d \in \mathbb{R}^n \mid d_s \in [\underline{d}, \bar{d}], s = 1, 2, \ldots, n\}$. Customer choice follows a multinomial logit model (Dong, Kouvelis, & Tian, 2009):

$$P_{x,x}(d) := (1 - \lambda) + \frac{\lambda}{\sum_{k \in S} \exp(\beta_c + \beta_k + \beta_d d_k) + 1},$$

$$P_{x,x+1_s}(d) := \frac{\lambda \exp(\beta_c + \beta_s + \beta_d d_s)}{\sum_{k \in S} \exp(\beta_c + \beta_k + \beta_d d_k) + 1} \tag{14}$$

for all $(x, d, s) \in X \times D \times S$, where $\lambda \in (0, 1)$ is the probability that a customer arrives on the booking website, $\beta_c \in \mathbb{R}$ denotes a constant offset, $\beta_s \in \mathbb{R}$ represents a measure of the popularity for all delivery slots $s \in S$ and $\beta_d < 0$ is a parameter for the price sensitivity. More details on the estimation of these parameters

Table 1
Numerical example parameters.

S	$\{1, 2, \ldots, 17\}$
$\bar{x}$	$[6, 6, \ldots, 6]$
λ	0.8
$[\underline{d}, \bar{d}]$	$[£0, £10]$
r	£20
$\bar{t}$	200
$C(x)$	$£10 \times \langle \mathbf{1}, x \rangle$ if $x \in X$ and ∞ otherwise

can be found in Yang et al. (2016). The average revenue of an order is r and the length of the time horizon, representing the booking period, is $\bar{t}$. The cost function C represents the delivery cost for all lists of orders $x \in X$ accumulated at the end of the booking period. The challenge is to price the slots dynamically to maximise profits, which corresponds to solving a DP of the form of (1), where $g(x, y, d) := r + d_s$ if $y = x + 1_s$ for all $s \in S$ and otherwise, $g(x, y, d) := 0$, i.e. the stage revenue is the average revenue plus delivery price for slot s if slot s is chosen and otherwise, it is zero. The DP in our numerical example takes the parameters in Table 1, adapted from a real-world, multi-subarea case study by Yang and Strauss (2017) to a single delivery subarea scenario. Furthermore, we also adopt the customer choice parameters $(\beta_c, \beta_d, \{\beta_s\}_{s \in S})$ from that paper.

We have chosen $C(0) = 0$, i.e. we ignore fixed costs, which have no effect on the pricing policy. Notice that for direct value function computation we would have to evaluate (1) for all $(x, t) \in X \times T$, i.e. $(6 + 1)^{17} \times 200 \approx 4.7 \times 10^{16}$ evaluations in our example. This is prohibitively large for any available computer capabilities. Hence, we use an approximate algorithm.

For this type of DP, Lebedev, Goulart, and Margellos (2021, Theorem 2) showed that the Bellman operator preserves strict submodularity, i.e. the condition in (3) holds with strict inequality, if a small enough $\lambda > 0$ is chosen. We set $\lambda = 0.8$ as shown in Table 1. This setup is similar to the numerical example in our preliminary work in Lebedev, Goulart, and Margellos (2020). However, we changed the expected number of customers arriving on the booking website to $\lambda \times \bar{t} = 0.8 \times 200 = 160$. This is an interesting variation for several reasons:
(1) Increasing the number of customers arriving on the booking website increases the need to actively control the sales process much earlier in the booking period.
(2) Increasing λ while reducing $\bar{t}$ in the model speeds up computation time, since it depends linearly on the number of time steps. At the same time, we do not observe any decrease in profit generation performance in comparison with smaller values of λ.

Since it is more difficult to maximise profits in this scenario, the need to select an appropriate initialisation for the approximate value function also gains importance. To illustrate this, we compare two initialisation strategies.
(1) A trivial way to initialise the value function is to set Q_t^0 to a large constant for all $t \in T$, i.e. a number that exceeds the maximum attainable profit. We choose $Q_t^0(x) = 10^6$ for all $(x, t) \in X \times T$ in our example.
(2) An alternative to this is to initialise Q_t^0 for all $t \in T$ using the fixed point of DP, V^*, which is a known upper bound to the exact value function at any $(x, t) \in X \times T$, i.e. $V^*(x) \ge V_t(x)$. This is always the case, since $\mathcal{T}$ in (2) is a monotone operator (Bertsekas, 2012, Chapter 3). In Lebedev et al. (2019), it is shown that the fixed point is given analytically as

$$V^*(x) := (\bar{d} + r)\langle \mathbf{1}, \bar{x} - x \rangle - C(\bar{x}), \text{ for all } x \in X. \tag{15}$$

Hence, we use this result to set $Q_t^0(x) = V^*(x)$ for all $(x, t) \in X \times T$.

Note that the fixed point in (15) is an affine function, so the initialiser has low complexity, i.e. only one affine function describes Q_t^0. There is also an intuitive interpretation for the
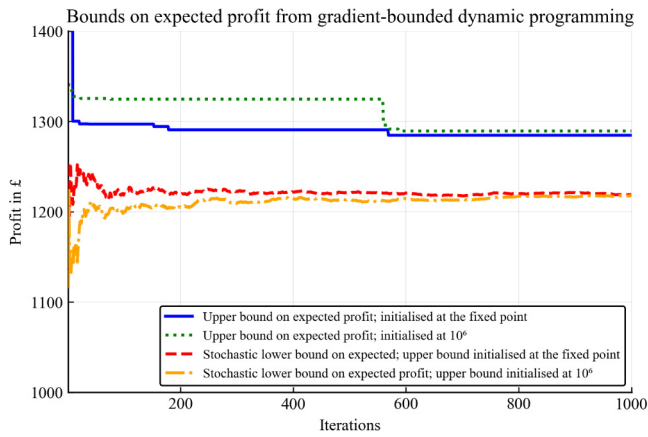
Fig. 2. Deterministic upper and stochastic lower bounds with initialisation of the value function at the fixed point and at an arbitrary large constant, 10^6 in this case.

Fig. 3. "Violin" plots of sample profits obtained by 100 validation samples from Algorithm 3. Left halves show samples where the approximate value function was initialised at the fixed point, right halves show samples with trivial initialisation at a large constant, 10^6 in this case. Horizontal lines indicate sample median.

simplicity of the fixed point: The fixed point corresponds to the expected profit obtained in an artificial scenario with infinite booking horizon. It happens that the uniform gradient $-(\bar{d} + r)$ of the fixed point implies that the optimal delivery charge is the maximum admissible price $\bar{d}$ for all feasible orders. The intuition behind this is that (assuming non-zero choice probability for this set of prices) all orders will sell out in a finite number of time steps, such that one should always charge the maximum admissible price to maximise profits in the infinite horizon case. This is a hypothetical scenario, yet, it provides the means to identify state–time pairs for which some of the slots should be priced at the maximum admissible delivery charge.

7.1. Computation of approximate value function

We run $i_{\max} = 1000$ iterations of Algorithm 1 for both initial approximate value functions. Computation of our Julia (Bezanson, Edelman, Karpinski, & Shah, 2017) code takes 9 min, 49 s on an i7-8565U CPU at 1.80 GHz processor base frequency and with 16 GB RAM. In each iteration $i \in \{1, 2, \ldots i_{\max}\}$, we compute the deterministic upper bound on the expected profit $u(i)$ (Corollary 2) and the stochastic lower bound $l(i)$ (Proposition 3). We show the behaviour of these bounds, for both initialisations, over all iterations in Fig. 2.

A gap between upper and lower bounds of approximately 5% remains even if the algorithm is run for 10,000 iterations. This indicates that the numerical problem instance actually violates Assumption 2, i.e. the exact value function of the problem is not concave extensible, such that there remains a gap between the concave extensible upper bound and the exact value function. However, as numerical evidence suggests the upper and lower bounds established in Corollary 2 and Proposition 3 remain valid. However, the fact that the gap is non-zero makes our upper and lower bounds non-trivial thus facilitating the subsequent numerical investigation. We discuss on choices of λ that result in a zero gap and allow numerical satisfaction of Assumption 2 in Section 7.3. Notice also that the fixed point initialisation outperforms the trivial initialisation in several ways:
(1) The stochastic lower bound based on initialising the upper bound at the fixed point is greater (tighter) than the other stochastic lower bound, especially in the first 200 iterations, while they approach each other over iterations and become very similar after 800 iterations.
(2) The deterministic upper bound based on the fixed point initialisation is substantially lower (tighter) than for the other

initialiser for the first 560 iterations, while still being slightly tighter for larger iteration indices.

The relative advantage of the fixed point initialisation strategy can also be seen in Fig. 3, where we have generated "violin" plots from 100 validation samples of Algorithm 3 for both initialisers.

Especially the first iteration samples have higher value when initialised at the fixed point and not at an arbitrary large constant. However, the effect decreases for larger iteration numbers. Since in either case a substantial variation in sample profits remains and since, upper and lower bounds for either case do not converge in the number of iterations performed, we want to compute confidence bounds on the tail and the expectation of the distribution of the stochastic lower bound in the next step.

Another important consideration when computing these bounds is the number of samples drawn. We would like to note that the computation time for the above validation samples depends on the iteration number, since it corresponds to the number of hyperplanes that the value function is comprised of. As we need to find the minimum over this number of hyperplanes for every approximate value function evaluation, the computation time also grows approximately linearly in the number iterations, e.g. computing 100 validation samples takes 0.05 s per iteration count, i.e. approximately 50 s for $i = 1000$. Overall, validation (Algorithm 3) tends to be much faster than approximation (Algorithm 1). However, in time-critical applications, it might be prohibitive to choose overly large validation sample sizes.

7.2. Computation of bounds

We first compute the tail bounds on the value of profit obtained by a single sample under the approximate policy after 1000 iterations of Algorithm 1. To this end, we compute the empirical Cantelli bound l_C from (7) and the Dvoretzky–Kiefer–Wolfowitz tail bound l_D from (8). Due to the 17-dimensional state-space, $\theta_C \approx 0$. To see this, upper bound the probability of the most likely event at every stage, namely no order being placed by $P_{x,x}(d) \leq P_{x,x}(\mathbf{1}\bar{d}) \approx 0.6732$. Due to time-independence of the transition probabilities, we can exponentiate this number by the number of time steps in the DP to obtain the probability of 0 orders at the end of the booking period. This needs to happen for all $k_{\max}$ (independent) validation samples, hence we again exponentiate this number by $k_{\max}$, which we assume is at least 10. This gives us the probability of all validation samples having 0 orders. This is the most likely, but only one of $|X|$ states,
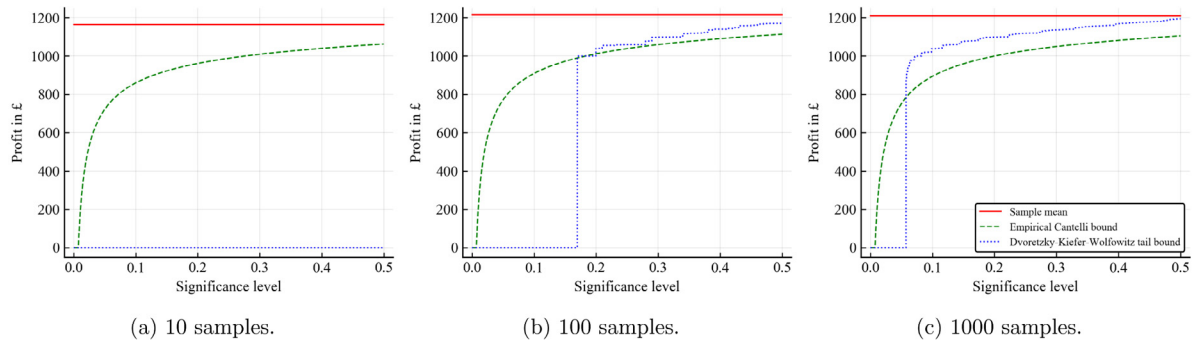
(a) 10 samples.

(b) 100 samples.

(c) 1000 samples.

Fig. 4. Probabilistic bounds — empirical Cantelli bound (green dashed line, Proposition 5(i)) and Dvoretzky–Kiefer–Wolfowitz tail bound (blue dotted line, Proposition 5(ii)) – on the profit obtained from a single validation sample as functions of various significance levels for 10 (a), 100 (b) and 1000 (c) validation samples.
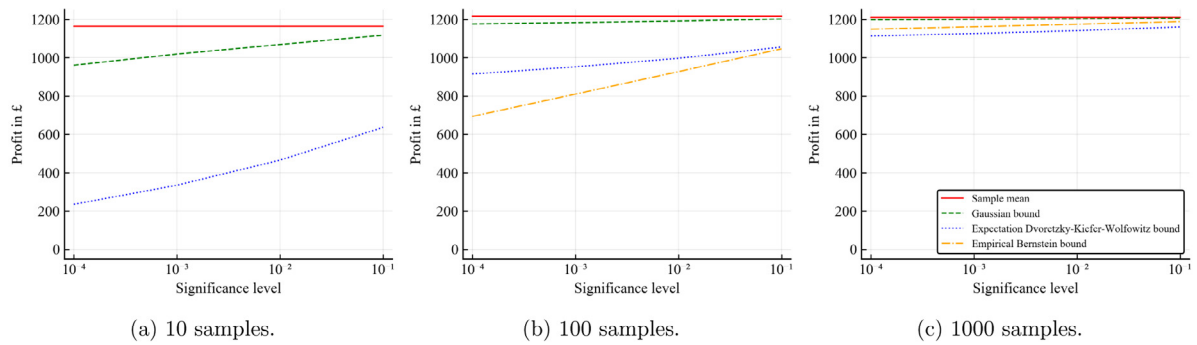

(a) 10 samples.

(b) 100 samples.

(c) 1000 samples.

Fig. 5. Probabilistic bounds — Gaussian bound (green dashed line, (13)), expectation Dvoretzky–Kiefer–Wolfowitz bound (blue dotted line, Proposition 7(ii)) and empirical Bernstein bound (orange dash-dotted line, Proposition 7(i)) – on the expectation of the profit obtained as functions of various significance levels and for 10 (a), 100 (b) and 1000 (c) validation samples.

so we multiply this number by $|X|$ to obtain $\Pr(\sigma_v = 0) \leq |X|P_{x,x}(\mathbf{1}\bar{d})^{\bar{t}k_{\max}} \approx 7^{17} \times 0.6732^{200 \times 10} \approx 0$.

As we see in Fig. 4, the tail bounds do not converge to the sample average, since there is an inherent variance in the customer choice model. This can be seen by inspecting the high variation of the sample profits in Fig. 3 for all iteration steps. In Fig. 4, notice that the choice of optimal bound changes with sample size: In our example, the empirical Cantelli bound is preferred for all significance levels when $k_{\max} = 10$, whereas the Dvoretzky–Kiefer–Wolfowitz tail bound is preferred for significance levels $\alpha > 20\%$ when $k_{\max} = 100$ and for significance levels $\alpha > 5.74\%$, where we have chosen the optimal parameter θ_D from Lemma 6.

To get more meaningful measures of the convergence of Algorithm 1, we now compute the bounds on the expectation of the profit obtained after 1000 iterations of Algorithm 1. To this end, we compute the empirical Bernstein bound $l_B^{\mathbb{E}}$ from (10) and the expectation Dvoretzky–Kiefer–Wolfowitz bound $l_D^{\mathbb{E}}$ from (11). We also compute the Gaussian bound $l_G^{\mathbb{E}}$ from (13), following Shapiro (2011). This final bound relies on the additional assumption that the exact distribution of the mean sample profit $\bar{l}_v$ is Gaussian. This is only asymptotically true by a Central Limit Theorem (see Hajek, 2015, Proposition 2.16). As seen in Fig. 5, the Gaussian bound is always the most optimistic, however not accompanied by theoretical guarantees. Moreover, for large validation samples sizes $k_{\max}$, the gap with the empirical Bernstein bound and with the expectation Dvoretzky–Kiefer–Wolfowitz bound is small. Therefore, we suggest to avoid using the Gaussian bound if probabilistic guarantees are sought for problems, such as our example, where the validation sample values are not normally distributed. Out of the empirical Bernstein bound and the expectation Dvoretzky–Kiefer–Wolfowitz bound, the earlier only tends to perform better for large sample sizes ($k_{\max} = 1000$). Note

that we omit the empirical Bernstein bound in Fig. 5(a) since its negative values are not meaningful.

Finally, we would like to comment on the relative computational cost of Algorithms 1 and 3. It is efficient to evaluate the forward part of Algorithm 1 compared to its backward part. However, as the iteration count increases, hyperplanes are added to the approximate value function, making it more expensive to evaluate. For example, to evaluate the value function at some state–time pair (x, t) after i iterations, one needs to compare the values of i hyperplanes, since the value function is the pointwise minimum of these i hyperplanes. Since Algorithm 3 uses *only* the most refined value function, i.e. the one with the most hyperplanes, value function evaluations take relatively long. Thus, a complex approximation will also take long to be validated. If time is limited, this could mean that $i_{\max}$ needs to be lowered to allow sufficient time for large enough number of validation samples to be generated and the bounds to be computed. In our example, it took 13 min, 10 s to compute 100 validation samples with Algorithm 3 after 1000 iterations of Algorithm 1. Recalling that running Algorithm 1 took 9 min, 49 s, there could be a trade-off between accuracy of the approximation and quality of the validation bound if time were more limited, e.g. in another application.

7.3. Algorithm convergence analysis

As noted in Section 7.1, there is a persistent gap of about 5% between the upper and stochastic lower bound for the particular problem instance. We conjecture that this is due to Assumption 2 not holding for these parameters. In particular, the customer arrival rate may be prohibitively high for the Bellman operator to produce concave extensible value functions in the exact
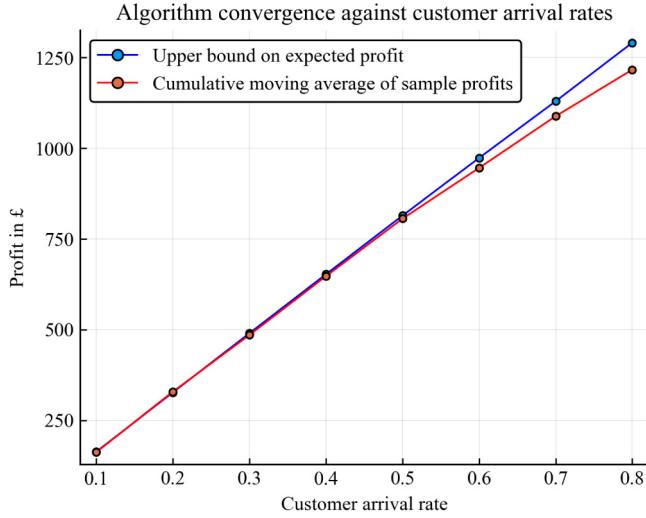
Fig. 6. Deterministic upper and stochastic lower bounds for different customer arrival rates.

problem. Hence, there may always be a gap between the exact value function and the approximate value function, since the latter is defined as the pointwise minimum of a finite number of hyperplanes and is thus concave. As shown in Lebedev, Goulart, and Margellos (2021, Theorem 2), there exists a small enough customer arrival rate λ, for which Assumption 2 holds. We investigate this theorem and also demonstrate the convergence result from this paper (Proposition 4) by repeating the experiment from Section 7.1 with reduced values of λ, in steps of 0.1 down to a minimum value of 0.1. The resulting upper and stochastic lower bounds after 1000 iterations and using the fixed point initialisation strategy are shown in Fig. 6.

Notice that both bounds monotonically increase with the customer arrival rate λ, which is to be expected, since we keep the number of time steps $\bar{t}$ constant across all experiments. Therefore, both the expected number of customer arrivals $\lambda\bar{t}$ and thus the expected profit monotonically increase with λ. The gap between the bounds is no greater than 1.0% for all customer arrival rates $\lambda \leq 0.5$, indicating that a policy has been found that produces near-optimal profits for these cases. In the other cases, the gap increases with λ, such that the associated performance guarantees are increasingly loose. Tightening these bounds further is a topic of current research.

8. Conclusions and future work

In this paper, we addressed two problems: First, we presented a new algorithm, termed gradient-bounded dynamic programming, for approximately solving high-dimensional multi-stage optimisation problems characterised by dynamic programming formulations with submodular, concave extensible value functions over discrete states. We accompanied the algorithm with finite convergence guarantees as well as deterministic upper and stochastic lower bounds to the exact value function. In future work, these bounds may be used to compare the profit generation efficiency with other approximate dynamic programming algorithms, which may not provide an upper bound to the exact value function. A comparative study of gradient-bounded dynamic programming and other approximate dynamic programming approaches can be found in Lebedev, Margellos, and Goulart (2021). One possible direction for future numerical studies would be the analysis of gradient-bounded dynamic programming in other application areas.

Second, we derived bounds on the tail and expectation of the (unknown) distribution of samples of the value obtained under an approximately optimal decision policy. These bounds may be used to validate the performance of approximately optimal decision policies also in other multistage stochastic optimisation problems without additional assumptions on this distribution other than its finite support. Hence, these bounds can be used to obtain probabilistic performance certificates for a wide range of multi-stage optimisation problems. Finally, we demonstrated our results in an example of the revenue management problem in attended home delivery.

Appendix

A.1. Proof of Proposition 1

We show this result by induction on t. In the base case (the terminal condition), $Q^i_{\bar{t}+1}(x) := V_{\bar{t}+1}(x) = -C(x)$ for all $(x, i) \in X \times I$, which satisfies the proposition trivially by Assumption 1. Assume for an induction hypothesis that $Q^{i-1}_{t+1}(x) \geq V_{t+1}(x)$ for some $(i, t) \in I \setminus \{0\} \times T$ and for all $x \in X$. Fix any x in X and distinguish the two cases of the if-statement in step 12 of Algorithm 1.

Case I: Suppose that Q^{i-1}_{t+1} is submodular on $Z(x^i_{t+1})$. Then H^* is the unique hyperplane through the set $\{(y, (\mathcal{T}Q^i_{t+1})(y))\}_{y \in Y_+(x^i_{t+1})}$. By (4), Q^i_{t+1} is concave extensible since it is the pointwise minimum of a finite number of hyperplanes. Hence, we invoke Assumption 2 to conclude that $\mathcal{T}Q^{i-1}_{t+1}$ is concave extensible and submodular. As shown by Lebedev, Goulart, and Margellos (2021, Appendix B4), this implies that H^* is a separating hyperplane, i.e. $H^*(x) \geq \mathcal{T}Q^{i-1}_{t+1}(x)$ for all $x \in X$. Define d^V to be the maximiser of (1) and define d^Q to be the maximiser of (1) with $V_{t+1}(y)$ replaced by $Q^{i-1}_{t+1}(y)$. We now show that the Bellman operator of the DP preserves the inequality $Q^{i-1}_{t+1}(x) \geq V_{t+1}(x)$, i.e. $\mathcal{T}Q^{i-1}_{t+1}(x) \geq \mathcal{T}V_{t+1}(x)$. To this end, fix $x \in X$ and consider

$$
\begin{aligned}
(\mathcal{T}Q^{i-1}_{t+1})(x) &= g(x, d^Q) + \sum_{y \in Y_+(x)} P_{x,y}(d^Q) Q^{i-1}_{t+1}(y) \\
&\geq g(x, d^V) + \sum_{y \in Y_+(x)} P_{x,y}(d^V) Q^{i-1}_{t+1}(y) \\
&\geq g(x, d^V) + \sum_{y \in Y_+(x)} P_{x,y}(d^V) V_{t+1}(y) \\
&= (\mathcal{T}V_{t+1})(x), \quad\quad\quad\quad (A.1)
\end{aligned}
$$

where the first inequality follows from the supoptimality of d^V for $(\mathcal{T}Q^{i-1}_{t+1})(x)$ and the second inequality follows from the induction hypothesis.

Case II: Now consider the case when Q^{i-1}_{t+1} is not submodular on $Z(x^i_{t+1})$. Then $H^* \in \left\{ \mathcal{T}H^{i-1}_{t+1} \mid j \in J^{i-1}_{t+1} \right\}$. Furthermore, by (4) and the induction hypothesis,

$$H^{j-1}_{t+1}(x) \geq Q^{i-1}_{t+1}(x) \geq V_{t+1}(x), \text{ for all } (x, j) \in X \times J^{i-1}_{t+1}. \quad (A.2)$$

We now show that all possible realisations of H^* constitute upper bounds on $\mathcal{T}V_{t+1}$. To this end, fix any $(x, j) \in X \times J^{i-1}_{t+1}$. Define d^H to be the maximiser of (1) with $V_{t+1}(y)$ replaced by $H^{j-1}_{t+1}(y)$. We can

show that the Bellman operator of the DP preserves the inequality $Q_{t+1}^{i-1}(x) \geq V_{t+1}(x)$ using a similar argument as before:

$$(\mathcal{T}H_{t+1}^{j-1})(x) \geq (\mathcal{T}V_{t+1})(x), \tag{A.3}$$

which follows from the suboptimality of d^V (see Case I) for $(\mathcal{T}H_{t+1}^{j-1})(x)$ and the fact that $H_{t+1}^{j-1}(x) \geq V_{t+1}(x)$ (see (A.2)). Therefore, we conclude that $H^*(x) \geq \mathcal{T}V_t(x)$ for all $x \in X$ in the second case as well.

Since both cases lead to an upper bound, i.e. $H^*(x) \geq \mathcal{T}V_{t+1}(x)$ for all $x \in X$, we infer that

$$Q_t^i(x) = \min\left\{H^*(x), Q_{t+1}^{i-1}(x)\right\} \geq \mathcal{T}V_{t+1}(x) \tag{A.4}$$

for all $x \in X$. This concludes our induction argument and shows that $Q_t^i(x) \geq V_t(x)$ for all $(x, i, t) \in X \times I \times T$.

A.2. Proof of Proposition 4

We will show the proposition by induction on t. Consider the base case, when $Q_{\bar{t}+1}^0(x) = V_{\bar{t}+1}(x)$ for all $x \in X$. Then notice that in the "backward sweep", the proposed algorithm computes the Bellman equation from $\bar{t} + 1 \to \bar{t}$ exactly for every $x \in X$. This is because $Q_{\bar{t}+1}^0$ is submodular by Assumption 1 and hence, the if-statement in step 12 of Algorithm 1 is true. Using Algorithm 2, $x_{\bar{t}+1}^i$ is resampled if for the time step transition $\bar{t} + 1 \to \bar{t}$, the algorithm has not visited this state in iteration $m, m+1, \ldots, m+ |X|$, where $m = 0$. Therefore, the value function is computed exactly at all $x \in X$ for the time step transition $\bar{t} + 1 \to \bar{t}$ after at most $|X|$ iterations of the proposed algorithm, i.e. $Q_{\bar{t}}^{\hat{i}}(x) = V_{\bar{t}}(x)$ for all $x \in X$, where $\hat{i} \leq |X|$.

Now suppose by means of an induction hypothesis that for some $(t, i) \in T \times I$, $Q_{t+1}^i(x) = V_{t+1}(x)$ for all $x \in X$. Then by Assumptions 1 and 2, V_{t+1} is submodular and hence, Q_{t+1}^i is also submodular. By a similar argument to the base case, the proposed algorithm computes the exact value function for the time step transition $t + 1 \to t$ in another $\hat{i} \leq |X|$ iterations. Notice that for an arbitrary (i, t), m in the if-statement in step 3 of Algorithm 2 ensures that resampling only occurs if states have been visited that are relevant for this particular time step t.

Hence, we conclude that for every time step transition, the proposed algorithm needs at most $|X|$ iterations to compute the exact value function for any one time step $t \in T$, which gives at most $\bar{t}|X|$ iterations for the total time horizon. Hence, after any $i \geq \bar{t}|X|$ iterations, $Q_t^i(x) = V_t(x)$ for all $(x, t) \in X \times T$. Therefore, both $\mathbb{E}[l(i)] = V_1(0)$ and $u(i) = Q_1^i(0) = V_1(0)$, which finally implies that $\mathbb{E}[l(i)] = u(i)$ for all $i \geq \bar{t}|X|$ iterations.

A.3. Proof of Proposition 5(i)

The proof is a finite sample adaptation of the one-sided Chebyshev's inequality, i.e. Cantelli's inequality (Ghosh, 2002, Theorem 1). We distinguish the following two cases:

Case I: Suppose that $\sigma_v \neq 0$. Fix any $k \in K$ and consider the conditional probability that $l_C := l_v(k_{max} + 1)$ is no greater than $\bar{l}_v - m\hat{\sigma}$ for some $m > 0$:

$$\begin{aligned} &\Pr(l_C \leq \bar{l}_v - m\sigma_v | \sigma_v \neq 0) \\ &= \Pr(m\sigma_v \leq \bar{l}_v - l_C | \sigma_v \neq 0) \\ &= \frac{1}{k_{max}} \sum_{k \in K} \Pr(m\sigma_v \leq \bar{l}_v - l_v(k) | \sigma_v \neq 0) \\ &= \frac{1}{k_{max}} \mathbb{E}\left(\sum_{k \in K} \mathbb{1}(m\sigma_v \leq \bar{l}_v - l_v(k)) \middle| \sigma_v \neq 0 \right), \end{aligned} \tag{A.5}$$

where the second last equality follows from the observation that $l_v(k)$ for all $k \in K \cup \{k_{max} + 1\}$ are independently and identically distributed. Next, we want to upper bound the indicator function in (A.5) by a quadratic function. A suitable expression is given for any $c > 0$ by

$$\begin{aligned} &\Pr(l_C \leq \bar{l}_v - m\sigma_v | \sigma_v \neq 0) \\ &\leq \frac{1}{k_{max}} \mathbb{E}\left(\sum_{k \in K} \frac{(\bar{l}_v - l_v(k) + c\sigma_v)^2}{(m\sigma_v + c\sigma_v)^2} \middle| \sigma_v \neq 0 \right). \end{aligned} \tag{A.6}$$

Note that each element in the summation is always non-negative and no smaller than one if $m\sigma_v \leq \bar{l} - l_v(k)$ and hence is an upper bound to (A.5). We simplify this expression using the definitions of $\bar{l}_v$ and σ_v from Algorithm 3 in Section 5 as

$$\begin{aligned} &\frac{1}{k_{max}} \mathbb{E}\left(\sum_{k \in K} \frac{(\bar{l}_v - l_v(k) + c\sigma_v)^2}{(m\sigma_v + c\sigma_v)^2} \middle| \sigma_v \neq 0 \right) \\ &= \frac{1}{k_{max}} \mathbb{E}\left(\frac{(k_{max} - 1)\sigma_v^2 + k_{max}c^2\sigma_v^2}{(m + c)^2 \sigma_v^2} \middle| \sigma_v \neq 0 \right) \\ &= \frac{1}{k_{max}} \mathbb{E}\left(\frac{k_{max} - 1 + k_{max}c^2}{(m + c)^2} \right) \\ &= \frac{k_{max} - 1 + k_{max}c^2}{k_{max}(m + c)^2}, \end{aligned} \tag{A.7}$$

where σ_v cancels, since $\sigma_v \neq 0$, and the expectation operator drops, since its argument is a constant. We minimise (A.7) by considering its first order condition, i.e.

$$\begin{aligned} 0 &= \frac{\partial}{\partial c} \frac{k_{max} - 1 + k_{max}c^2}{k_{max}(m + c)^2} \\ &= (k_{max}(m + c)^2)^{-2} \left\{ 2k_{max}^2 c(m + c)^2 \right. \\ &\quad \left. - (k_{max} - 1 + k_{max}c^2)2k_{max}(m + c) \right\} \\ \Rightarrow c &= \frac{k_{max} - 1}{k_{max}m}. \end{aligned} \tag{A.8}$$

The second-order condition shows that c minimises (A.7). Substituting c into (A.7) and simplifying gives:

$$\Pr(l_C \leq \bar{l}_v - m\sigma_v | \sigma_v \neq 0) \leq \frac{k_{max} - 1}{k_{max}m^2 + k_{max} - 1}. \tag{A.9}$$

Case II: Suppose that $\sigma_v = 0$. We repeat the derivation of Case I with $\Pr(l_C \leq \bar{l}_v - m\sigma_v | \sigma_v \neq 0)$ replaced by $\Pr(l_C \leq \bar{l}_v - m\sigma_v | \sigma_v = 0)$ until (A.5), where we note that due to $\sigma_v = 0$, we have $l_v(k) = l_v(k')$ for all $(k, k') \in K \times K$ and hence, $\Pr(l_C \leq \bar{l}_v - m\sigma_v | \sigma_v = 0) = 1$.

Recalling that $\theta_C := \Pr(\sigma_v = 0)$ and taking both cases together, we obtain by the total probability theorem that

$$\Pr(l_C \leq \bar{l}_v - m\sigma_v) \leq \frac{(1 - \theta_C)(k_{max} - 1)}{k_{max}m^2 + k_{max} - 1} + \theta_C. \tag{A.10}$$

We want this probability to be at most the significance level α. Hence, we solve this expression for m, yielding

$$\begin{aligned} &\alpha \geq (1 - \theta_C)\frac{k_{max} - 1}{k_{max}m^2 + k_{max} - 1} + \theta_C \\ &\iff m \geq \sqrt{\frac{(1 - \alpha)(k_{max} - 1)}{(\alpha - \theta_C)k_{max}}}, \end{aligned} \tag{A.11}$$

which is real-valued, since $\alpha > \theta_C$ by Assumption 3. Substituting for m on the left-hand side of (A.10) gives the desired property:

$$\Pr\left(l_C \leq \bar{l} - \sigma_v \sqrt{\frac{(1 - \alpha)(k_{max} - 1)}{(\alpha - \theta_C)k_{max}}} \right) \leq \alpha. \tag{A.12}$$

A.4. Proof of Proposition 5(ii)

Fix any $\alpha \in (0, 1)$, any $\theta_D \in (0, \alpha)$ and compute

$$l_D = \sup \left\{ l \in [l_-, l_+] \,\middle|\, F_K(l) \leq \alpha - \theta_D - \sqrt{\frac{\ln(1/\theta_D)}{2k_{\max}}} \right\}. \tag{A.13}$$

We will now show that $l_v(k_{\max} + 1) > l_D$ with probability at least $1 - \alpha$. By the total probability theorem, we write

$$\Pr(l_v(k_{\max} + 1) > l_D) = \Pr(B|E)\Pr(E) \\ + \Pr(B|E^c)\Pr(E^c), \tag{A.14}$$

where B denotes the random event that $l_v(k_{\max} + 1) > l_D$, i.e. $\Pr(B) = 1 - F(l_D)$, and E denotes the random event that

$$F(l_D) \leq F_K(l_D) + \sqrt{\frac{\ln(1/\theta_D)}{2k_{\max}}}, \tag{A.15}$$

which according to the Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990) has probability $\Pr(E) \geq 1 - \theta_D$. E^c denotes the complementary event of E. Notice that $\Pr(E^c)$ and $\Pr(B|E^c)$ are non-negative and hence, we can create the following lower bound:

$$\Pr(l_v(k_{\max} + 1) > l_D) \geq \Pr(B|E)\Pr(E). \tag{A.16}$$

Since we condition on E, we can lower bound $\Pr(B|E) \geq 1 - \left(F_K(l_D) + \sqrt{\ln(1/\theta_D)/(2k_{\max})}\right)$ according to (A.15), which yields

$$\Pr(l_v(k_{\max} + 1) > l_D) \\ \geq \left[1 - \left(F_K(l_D) + \sqrt{\frac{\ln(1/\theta_D)}{2k_{\max}}} \right) \right] (1 - \theta_D) \\ > 1 - \left(F_K(l_D) + \sqrt{\frac{\ln(1/\theta_D)}{2k_{\max}}} \right) - \theta_D \\ \geq 1 - (\alpha - \theta_D) - \theta_D = 1 - \alpha, \tag{A.17}$$

where the final inequality follows from the choice of l_D in (A.13), which thus concludes our proof.

A.5. Proof of Lemma 6

Consider the first-order optimality condition, i.e.

$$0 = \frac{\partial}{\partial \theta_D} \left\{ \alpha - \theta_D - \sqrt{\frac{\ln(\frac{1}{\theta_D})}{2k_{\max}}} \right\} \\ \Rightarrow 0 = -1 + \frac{1}{2\theta_D \sqrt{-2k_{\max}\ln(\theta_D)}}. \tag{A.18}$$

Since $\theta_D > 0$, we can simplify to arrive at

$$\theta_D^2 \ln(\theta_D^2) = \frac{-1}{4k_{\max}} \\ \Rightarrow \theta_D^2 = \exp\left(W_i\left(\frac{-1}{4k_{\max}} \right) \right), \tag{A.19}$$

where $i \in \{0, -1\}$ and W_i is the Lambert W function (see Definition 4). The second-order conditions show that $i = -1$ gives a local maximum. Hence, we take the square root and note that θ_D is bounded from above by α to arrive at the desired result.

A.6. Proof of Proposition 7(ii)

We start by writing the expectation of the random variable l in terms of its exact yet unknown cumulative distribution function

F (see Evans & Rosenthal, 2004, Definition 3.7.1), i.e.

$$\mathbb{E}l := \int_{l=0}^{\infty} 1 - F(l)\mathrm{d}l - \int_{l=-\infty}^{0} F(l)\mathrm{d}l. \\ = \int_{l=0}^{\infty} 1 - F(l)\mathrm{d}l + \int_{l=-\infty}^{0} -F(l)\mathrm{d}l \\ = \int_{l=\max\{0,l_-\}}^{\max\{0,l_+\}} 1 - F(l)\mathrm{d}l + \max\{0, l_-\} \\ + \int_{l=\min\{0,l_-\}}^{\min\{0,l_+\}} -F(l)\mathrm{d}l - \min\{0, l_+\}, \tag{A.20}$$

where we changed the limits of integration, since F has the finite support $[l_-, l_+]$. The max- and min-operators in the integration limits ensure that equality holds independent of the sign of either l_- or l_+. By the Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990), we can write with confidence $1 - \alpha^{\mathbb{E}}$ that

$$F(l) \leq \min \left\{ 1, F_K(l) + \sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}} \right\} \\ \Longleftrightarrow -F(l) \geq -\min \left\{ 1, F_K(l) + \sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}} \right\} \\ \Longleftrightarrow 1 - F(l) \geq 1 - \min \left\{ 1, F_K(l) + \sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}} \right\}. \tag{A.21}$$

Using these expressions for $-F(l)$ and $1 - F(l)$, we lower bound (A.20) by

$$\mathbb{E}l \geq \int_{l=\max\{0,l_-\}}^{\max\{0,l_+\}} 1 - \min \left\{ 1, F_K(l) + \sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}} \right\} \mathrm{d}l \\ - \int_{l=\min\{0,l_-\}}^{\min\{0,l_+\}} \min \left\{ 1, F_K(l) + \sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}} \right\} \mathrm{d}l \\ + \max\{0, l_-\} - \min\{0, l_+\}. \tag{A.22}$$

Finally, $\mathbb{E}\bar{l}_v = \mathbb{E}l$, since $\mathbb{E}$ is a linear operator and $l_v(k)$ for all $k \in K$ are independent, thus concluding the proof.

A.7. Proof of Proposition 8

We can express the empirical mean in Hoeffding's bound (Hoeffding, 1963) as an integral over the empirical cumulative distribution function:

$$l_H^{\mathbb{E}} := \bar{l}_v - (l_+ - l_-)\sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}} \\ = \int_{l=0}^{\infty} 1 - F_K(l)\mathrm{d}l + \int_{l=-\infty}^{0} -F_K(l)\mathrm{d}l \\ - (l_+ - l_-)\sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}} \\ = \int_{l=\max\{0,l_-\}}^{\max\{0,l_+\}} 1 - \left(F_K(l) + \sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}} \right) \mathrm{d}l \\ + \int_{l=\min\{0,l_-\}}^{\min\{0,l_+\}} -\left(F_K(l) + \sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}} \right) \mathrm{d}l \\ + \max\{0, l_-\} - \min\{0, l_+\}$$

$$< \int_{l=\max\{0,l_-\}}^{\max\{0,l_+\}} 1 - \min\left\{1, F_K(l) + \sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}}\right\} \mathrm{d}l$$

$$+ \int_{l=\min\{0,l_-\}}^{\min\{0,l_+\}} - \min\left\{1, F_K(l) + \sqrt{\frac{\ln(1/\alpha^{\mathbb{E}})}{2k_{\max}}}\right\} \mathrm{d}l$$

$$+ \max\{0, l_-\} - \min\{0, l_+\}$$

$$= l_{\mathrm{D}}^{\mathbb{E}}, \tag{A.23}$$

where the third equality follows from the fact that the (finite) support of $F_K(l)$ is $[l_-, l_+]$ and the last inequality is strict by Assumption 4, since F_K is a stair function with step height $1/k_{\max}$ and $\sqrt{\ln(1/\alpha^{\mathbb{E}})/(2k_{\max})} > 1/k_{\max}$ implies that $F_K(l^*) + \sqrt{\ln(1/\alpha^{\mathbb{E}})/(2k_{\max})} > 1$ for some $l^* < l_+$.

References

Bertsekas, D. P. (2012). *Dynamic programming and optimal control, Vol. II* (4th ed.). Athena Scientific.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1), 65–98.

Cantelli, F. P. (1928). Sui confini della probabilita. In *Atti del Congresso Internazional del Matematici*: vol. 6. (pp. 47–59). Bologna.

Dong, L., Kouvelis, P., & Tian, Z. (2009). Dynamic pricing and inventory control of substitute products. *Manufacturing & Service Operations Management*, 11(2), 317–339.

Evans, M. J., & Rosenthal, J. S. (2004). *Probability and statistics: The science of uncertainty*. Macmillan.

de Farias, D. P., & Van Roy, B. (2003). The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6), 850–865.

de Farias, D. P., & Van Roy, B. (2004). On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, 29(3), 462–478.

Ghosh, B. K. (2002). Probability inequalities related to Markov's theorem. *The American Statistician*, 56(3), 186–190.

Hajek, B. (2015). *Random processes for engineers*. Cambridge University Press.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301), 13–30.

Kariotoglou, N., Summers, S., Summers, T., Kamgarpour, M., & Lygeros, J. (2013). Approximate dynamic programming for stochastic reachability. In *2013 European control conference* (pp. 584–589). http://dx.doi.org/10.23919/ECC.2013.6669603.

Lebedev, D., Goulart, P., & Margellos, K. (2019). A concave value function extension for the dynamic programming approach to revenue management in attended home delivery. In *2019 18th European control conference* (pp. 999–1004). http://dx.doi.org/10.23919/ECC.2019.8795697.

Lebedev, D., Goulart, P., & Margellos, K. (2020). Gradient-bounded dynamic programming with submodular and concave extensible value functions. In *21st IFAC world congress 2020*. Available at: https://arxiv.org/pdf/2005.11213.pdf.

Lebedev, D., Goulart, P., & Margellos, K. (2021). A dynamic programming framework for optimal delivery time slot pricing. *European Journal of Operational Research*, 292(2), 456–468.

Lebedev, D., Margellos, K., & Goulart, P. (2021). Convexity and feedback in approximate dynamic programming for delivery time slot pricing. *IEEE Transactions on Control Systems Technology*.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3), 1269–1283.

Maurer, A., & Pontil, M. (2009). *Empirical Bernstein bounds and sample variance penalization*: Technical report, https://arxiv.org/abs/0907.3740.

Mohajerin Esfahani, P., Sutter, T., Kuhn, D., & Lygeros, J. (2018). From infinite to finite programs: Explicit error bounds with applications to approximate dynamic programming. *SIAM Journal on Optimization*, 28(3), 1968–1998.

Murota, K., & Shioura, A. (2001). Relationship of M-/L-convex functions with discrete convex functions by Miller and Favati–Tardella. *Discrete Applied Mathematics*, 115(1), 151–176.

Pereira, M. V. F., & Pinto, L. M. V. G. (1991). Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, 52(1), 359–375.

Powell, W. B. (2007). Approximate dynamic programming: Solving the curses of dimensionality. In *Wiley series in probability and statistics*, New York, NY, USA: Wiley-Interscience.

Shapiro, A. (2011). Analysis of stochastic dual dynamic programming method. *European Journal of Operational Research*, 209(1), 63–72.

Warrington, J., Beuchat, P. N., & Lygeros, J. (2019). Generalized dual dynamic programming for infinite horizon problems in continuous state and action spaces. *IEEE Transactions on Automatic Control*, 64(12), 5012–5023.

Yang, X., & Strauss, A. K. (2017). An approximate dynamic programming approach to attended home delivery management. *European Journal of Operational Research*, 263(3), 935–945.

Yang, X., Strauss, A. K., Currie, C. S. M., & Eglese, R. (2016). Choice-based demand management and vehicle routing in e-fulfillment. *Transportation Science*, 50(2), 473–488.

Zhang, S., & Sun, X. A. (2019). *Stochastic dual dynamic programming for multistage stochastic mixed-integer nonlinear optimization*: Technical report.

Zou, J., Ahmed, S., & Sun, X. A. (2019). Stochastic dual dynamic integer programming. *Mathematical Programming*, 175(1), 461–502.

Denis Lebedev received the M.Eng. degree in Engineering, Economics and Management from the University of Oxford, U.K. in 2017 and the D.Phil. degree in Engineering Science from the University of Oxford, U.K. in 2021. His main research interests include approximate dynamic programming techniques for revenue management with focus on applications in e-commerce.

Paul Goulart received the B.Sc. and M.Sc. degrees in aeronautics and astronautics from the Massachusetts Institute of Technology, Cambridge, MA, USA, and the Ph.D. degree in control engineering from the University of Cambridge, Cambridge, U.K., in 2007. He was selected as a Gates Scholar at the University of Cambridge. From 2007 to 2011, he was a Lecturer in control systems in the Department of Aeronautics, Imperial College London, London, U.K., and from 2011 to 2014 as a Senior Researcher in the Automatic Control Laboratory, ETH Zurich, Zurich, Switzerland. He is currently an Associate Professor in the Department of Engineering Science and a Tutorial Fellow at St. Edmund Hall, University of Oxford, Oxford, U.K. His research interests include model predictive control, robust optimisation, and control of fluid flows.

Kostas Margellos received the Diploma in electrical engineering from the University of Patras, Greece, in 2008, and the Ph.D. in control engineering from ETH Zurich, Switzerland, in 2012. He spent 2013, 2014 and 2015 as a post-doctoral researcher at ETH Zurich, UC Berkeley and Politecnico di Milano, respectively. In 2016 he joined the Control Group, Department of Engineering Science, University of Oxford, where he is currently an Associate Professor. He is also a Fellow of Reuben College and a Lecturer at Worcester College. His research interests include optimisation and control of complex uncertain systems, with applications to generation and load side control for power networks.