**Noname manuscript No.**
(will be inserted by the editor)

# Data-Driven Stochastic Dual Dynamic Programming: Performance Guarantees and Regularization Schemes

**Hyuk Park · Zhuangzhuang Jia · Grani A. Hanasusanto**

Received: date / Accepted: date

**Abstract** We propose a data-driven scheme for multistage stochastic linear programming with Markovian random parameters by extending the stochastic dual dynamic programming (SDDP) algorithm. In our data-driven setting, only a finite number of historical trajectories are available. The proposed SDDP scheme evaluates the cost-to-go functions only at the observed sample points, where the conditional expectations are estimated empirically using kernel regression. The scheme thus avoids the construction of scenario trees, which may incur exponential time complexity during the backward induction step. However, if the training data is sparse, the resulting SDDP algorithm exhibits a high optimistic bias that gives rise to poor out-of-sample performances. To mitigate the small sample effects, we adopt ideas from the distributionally robust optimization (DRO), which replaces the empirical conditional expectation in the cost-to-go function with a worst-case conditional expectation over a polyhedral ambiguity set. We derive the theoretical out-of-sample performance guarantee of the data-driven SDDP scheme and show that the dependence of its sample complexity on the number of time stages is merely polynomial. Finally, we validate the effectiveness of the proposed algorithm and demonstrate its superiority over existing data-driven schemes through extensive numerical experiments.

**Keywords** stochastic optimization · multistage stochastic linear program · stochastic dual dynamic programming · distributionally robust optimization · Nadaraya-Watson estimator · Markov dependence

## 1 Introduction

Multistage stochastic linear programming (MSLP) is a framework for sequential decision making under uncertainty with linear objective functions and constraints. At each stage, the decision is made based on realizations of random parameters available up to that stage. This framework has been adopted to model various real-world applications, such as hydrothermal scheduling [17,45,46,60], unit commitment [4,14,29,52], portfolio optimization [11,12,16,27,42], and manufacturing and capacity planning [2,3,23,62]. Despite such exceptional modeling power, the general consensus is that multistage stochastic programs are very difficult to solve [36,59].

One popular approach to solve MSLP problems is a sampling-based Benders decomposition method called stochastic dual dynamic programming (SDDP). It was developed by Pereira and Pinto in 1991 as an effort to solve a large-scale hydrothermal scheduling problem [46] and is still considered as the state-of-the-art solution method. This algorithm iteratively approximates the expected (i.e., risk-neutral) cost-to-go functions of the dynamic programming equations by piecewise linear functions called *cuts* during the *forward* and *backward* steps. Since such cuts allow the cost-to-go function to be evaluated at well-chosen candidate solutions, the intractability arising from discretizing the state variables, known as the curse of dimensionality, can be alleviated. Asymptotic convergence of the algorithm is established in [15,25,50,55],

Hyuk Park
Department of Industrial and Enterprise Systems Engineering, University of Illinois Urbana-Champaign, Urbana, IL, 61801.
E-mail: hyukp2@illinois.edu

Zhuangzhuang Jia
Department of Industrial and Enterprise Systems Engineering, University of Illinois Urbana-Champaign, Urbana, IL, 61801.
E-mail: zj12@illinois.edu

Grani A. Hanasusanto
Department of Industrial and Enterprise Systems Engineering, University of Illinois Urbana-Champaign, Urbana, IL, 61801.
E-mail: gah@illinois.edu

while its computational complexity is recently investigated in [37,68]. The standard SDDP framework is extended to risk-averse cases [55,60], and convex [66,67] and nonconvex cost-to-go functions [1,20,69,70].

In the literature on the SDDP framework, it is commonly assumed that the underlying distribution of the uncertain data is known and the data process is stagewise independent. While these assumptions are not entirely realistic, they are crucial for several reasons. First, a known distribution assumption enables to sample scenarios in the forward step and construct a finite representation of the true problem by approximating the underlying stochastic process using a scenario tree [53]. Second, under the stagewise independent data process, we can avoid a scenario tree that grows exponentially with the number of time stages as it collapses to a recombining scenario tree, where nodes between any consecutive stages are fully connected via unconditional transition probabilities. Such a structure allows us to evaluate only one unconditional expected cost-to-go at each stage. In reality, however, there is often a correlation over time in the data process. Under stagewise dependence, the aforementioned approach is no longer valid as we have to take a new set of samples conditional on the whole history of the data process up to the current stage, which makes evaluating the conditional expected cost-to-go functions intractable [**?**,**?**,56]. To alleviate such an intractability, several papers instead assume Markov dependence (i.e., interstage dependence) with the purpose of exploiting some degree of correlation in the data process.

In a data-driven setting, where only historical data is available while the true distribution is unknown, there are two approaches to incorporate Markov dependence. One is to reformulate the random parameter as a first-order time series model with additional state variables and assume that the noise term is stagewise independent [32]. However, the time series approach requires adding auxiliary variables to state variables, which aggravates the curse of dimensionality. In addition, from a modeling perspective, this approach is limited in that uncertainty should appear only on the right-hand side of the constraints. Otherwise, the convexity of the cost-to-go is destroyed. A more general approach, known as Markov chain discretization, is to construct a recombining scenario tree and approximate the true conditional distribution using optimal quantization to reflect Markov dependence [10,13,26,38,39,48]. Unlike the time series approach, the Markov chain approach does not require the auxiliary variables and allows uncertainty in any parameters, e.g., objective function coefficients, a recourse matrix, etc, and therefore it can model a broader range of problems, such as portfolio optimization and inventory management. Löhndorf and Shapiro [38] demonstrate that the SDDP based on the Markov chain approach provides tighter lower bounds and improved policies for the hydrothermal scheduling problem compared to the time series approach. Unfortunately, despite the better empirical performance shown in the previous works, there has been no rigorous sample complexity analysis of this method, i.e., it is not clear whether the Markov chain discretization can generalize to out-of-sample data and provide asymptotically consistent policies.

Whichever approach is adopted, such data-driven schemes often suffer from overfitting issues, causing poor out-of-sample performance. The issue is aggravated particularly when the available data is limited. As a remedy for this small sample effect, distributionally robust optimization (DRO) has garnered significant attention recently. DRO relaxes the stringent assumption of known distribution by constructing an ambiguity set of plausible distributions consistent with the available information. Using DRO, an MSLP problem is formulated as a min-max problem at each stage, which yields a policy that performs best under the worst-case distribution that maximizes the expected cost-to-go. Ambiguity sets are commonly categorized into two types: moment-based [18,51,57,71] and discrepancy-based ambiguity sets [5,8,35,40,41,65]. Particularly in the distributionally robust MSLP setting, SDDP frameworks with different discrepancy-based ambiguity sets have been proposed [22,31,49,61]. Similar to this paper, Philpott et al. [49] use the modified $\chi^2$ ambiguity set [7]. They derive a closed-form solution for the inner maximization problem and use the resulting worst-case distribution to generate a cut during the backward step. They assume that the data process is stagewise independent and impose randomness only on the right-hand side. Silva et al. [61] consider a more general distributionally robust MSLP problem where the random recourse matrix has Markov dependence. They use a hidden Markov model (HMM) to capture unobservable states characterizing the distribution of the random parameters (in the recourse matrix) and then use the total variation ambiguity set to account for estimation errors of the transition probabilities among the unobservable states. In Table 1, we compare our proposed method with the existing SDDP algorithms for the distributionally robust MSLP.

This paper focuses on the incorporation of Markov dependence into risk-neutral and risk-averse MSLP problems in a data-driven setting and the sample complexity analysis of the proposed Markov discretization method. From the modeling aspect, our work is similar to [29] where the authors use the Nadaraya-Watson (NW) kernel regression estimator [43,64] for stochastic optimal control problems with endogenous state variables. However, as their stochastic dynamic programming scheme requires discretizing the endogenous state variables, it only works for low-dimensional settings. To deal with large-scale MSLP problems, we propose a data-driven SDDP framework using the NW regression under Markov dependence. Srivastava et al. [63] derive generalization bounds for the NW approximation of a static stochastic program. From a theoretical aspect, therefore, our work extends the out-of-sample guarantee derived in [63] to multistage settings.

**Table 1** Comparison of distributionally robust SDDP algorithm variants

| Algorithms | Huang et al., 2017 [31] | Philpott et al., 2018 [49] | Duque and Morton, 2020 [22] | Silva et al., 2021 [61] | This Paper |
|---|---|---|---|---|---|
| Ambiguity Set | $\infty$-norm | modified $\chi^2$ | Wasserstein | total variation | polyhedral approximation of modified $\chi^2$ |
| Random RHS | ✓ | ✓ | ✓ | ✓ | ✓ |
| Random Technology Matrix | ✓ | - | - | ✓ | ✓ |
| Random Recourse Matrix | ✓ | - | - | ✓ | ✓ |
| Random Obj. Func. Coeff. | ✓ | - | - | ✓ | ✓ |
| Markov Dependence | - | - | - | ✓ | ✓ |
| Stopping Criterion | stochastic gap | maximum iteration | maximum iteration | deterministic gap | deterministic gap |
| Out-of-Sample Guarantee | - | - | - | - | ✓ |

The main contributions of this paper can be summarized as follows:

1. We propose a data-driven SDDP framework for risk-neutral and risk-averse MSLP problems under the Markov dependence assumption. To incorporate Markov dependence, the true conditional probability is estimated using the Nadaraya-Watson kernel regression. Unlike other Markov chain discretization approaches that do not have convergence guarantees, the proposed discretization provides convergence to the true optimal policy as the number of data tends to infinity.
2. Leveraging the generalization bounds for a static stochastic program established in [63], we derive for the first time a theoretical out-of-sample guarantee for the data-driven risk-neutral MSLP problem under Markov dependence. Our result indicates that the out-of-sample suboptimality bound is at most $\widetilde{O}_p(T^{\frac{3}{2}}/N^{\frac{2}{p+4}})$,[2] where $T$ is the number of time stages. This mild (polynomial) dependence on $T$ suggests that our scheme is applicable to solve MSLP problems with a large number of stages. The result is surprising since existing analysis [59] suggests that the theoretical suboptimality bound drastically worsens as the number of stages increases.
3. Our theoretical guarantee suggests the use of a variance-based regularization scheme for improving out-of-sample performance. Unfortunately, the scheme is intractable due to nonconvexity. As a tractable alternative, we develop a conservative MSLP problem using a DRO formulation with a polyhedral outer approximation of the modified $\chi^2$ ambiguity set. We further prove that the DRO formulation asymptotically converges to the variance regularization scheme.
4. Numerical experiments in the context of portfolio optimization and wind energy commitment problems demonstrate that our data-driven schemes are superior to the stagewise independent scheme and the benchmark schemes proposed in the literature in terms of out-of-sample performance.

**Notation and terminology**. We use bold letters for vectors and regular fonts for scalars. We define $\mathbf{e}$ as the vector of all ones—its dimension will be clear from the context. The tilde symbol is used to denote random variables (e.g., $\tilde{\boldsymbol{\xi}}_t$) to differentiate from their realizations (e.g., $\boldsymbol{\xi}_t$). For any $t \in \mathbb{N}$, we define $[t]$ as the index set $\{1, \ldots, t\}$. A sequence of realizations of the random data process up to stage $t$ is denoted as $\boldsymbol{\xi}_{[t]} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_t)$. For a random variable $\tilde{Z}$, $\mathbb{E}[\tilde{Z}]$ and $\mathbb{V}[\tilde{Z}]$ denote its expectation and variance, respectively. The indicator function of a subset $\mathcal{X}$ is defined through $\mathbb{1}_{\mathcal{X}}(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in \mathcal{X}$ and $\mathbb{1}_{\mathcal{X}}(\boldsymbol{x}) = +\infty$ otherwise. The Dirac distribution, which assigns unit mass on $\boldsymbol{\xi}_t$, is denoted by $\delta_{\boldsymbol{\xi}_t}$. In asymptotic analysis, we use the standard $O$ notations for the convergence of sets of ordinary numbers and $O_p$ for the convergence of sets of random variables. In addition, $\tilde{O}_p$ is used to suppress multiplicative terms with logarithmic dependence on $n$ in the sense of convergence in probability.

## 2 Problem Statement

Consider the following risk-neutral MSLP problem under *Markov dependence*:

$$\min_{\boldsymbol{x}_1 \in \mathcal{X}_1(\boldsymbol{x}_0, \boldsymbol{\xi}_1)} \boldsymbol{c}_1^\top \boldsymbol{x}_1 + \mathbb{E}\left[ \min_{\boldsymbol{x}_2 \in \mathcal{X}_2(\boldsymbol{x}_1, \tilde{\boldsymbol{\xi}}_2)} \tilde{\boldsymbol{c}}_2^\top \boldsymbol{x}_2 + \mathbb{E}\left[ \cdots + \mathbb{E}\left[ \min_{\boldsymbol{x}_T \in \mathcal{X}_T(\boldsymbol{x}_{T-1}, \tilde{\boldsymbol{\xi}}_T)} \tilde{\boldsymbol{c}}_T^\top \boldsymbol{x}_T \,\middle|\, \tilde{\boldsymbol{\xi}}_{T-1} \right] \cdots \,\middle|\, \tilde{\boldsymbol{\xi}}_2 \right] \,\middle|\, \boldsymbol{\xi}_1 \right], \quad (2.1)$$

Here, the problem parameters are summarized by the random vectors $\tilde{\boldsymbol{\xi}}_t = (\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{b}}_t, \tilde{\boldsymbol{A}}_t, \tilde{\boldsymbol{B}}_t)$ for every $t \in [T] \setminus \{1\}$ governed by an (unknown) continuous joint distribution while the first stage parameters $\boldsymbol{\xi}_1 = (\boldsymbol{c}_1, \boldsymbol{b}_1, \boldsymbol{A}_1, \boldsymbol{B}_1)$ are deterministic with the initial state $\boldsymbol{x}_0$ is given as input. Here, the decision space is defined as the polytope $\mathcal{X}_t(\boldsymbol{x}_{t-1}, \tilde{\boldsymbol{\xi}}_t) = \{\boldsymbol{x}_t \in \mathbb{R}_+^{d_t} : \tilde{\boldsymbol{A}}_t \boldsymbol{x}_t + \tilde{\boldsymbol{B}}_t \boldsymbol{x}_{t-1} = \tilde{\boldsymbol{b}}_t\}$ and the risk measures are given by conditional expectations $\mathbb{E}[\,\cdot\,|\boldsymbol{\xi}_t]$. In this setting, the decision maker takes sequential decisions in a

[2] $N$ denotes the number of samples and $p$ is the dimension of random parameters in an MSLP problem.

nonanticipative manner as the realization of the random parameter vector $\tilde{\boldsymbol{\xi}}_t = (\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{b}}_t, \tilde{\boldsymbol{A}}_t, \tilde{\boldsymbol{B}}_t)$ is revealed at each time period. More precisely, when the decision maker takes the first-stage decision $\boldsymbol{x}_1$ from the deterministic feasible region $\mathcal{X}_1(\boldsymbol{x}_0, \boldsymbol{\xi}_1)$, the decision incurs an immediate cost $\boldsymbol{c}_1^\top \boldsymbol{x}_1$. Next, at the beginning of the second stage, the realization $\boldsymbol{\xi}_2$ of the random parameter vector $\tilde{\boldsymbol{\xi}}_2$ is revealed and the second-stage decision $\boldsymbol{x}_2$ is chosen from the feasible region $\mathcal{X}_2(\boldsymbol{x}_1, \boldsymbol{\xi}_2)$ with a cost $\boldsymbol{c}_2^\top \boldsymbol{x}_2$. The process continues until the terminal stage $T$. The goal of the problem is to optimize a sequence of functions $\{\boldsymbol{x}_t(\boldsymbol{\xi}_{[t]})\}_{t=1}^T$ called policies that minimize the expected total cost over the $T$ stages. Here, each $\boldsymbol{x}_t(\boldsymbol{\xi}_{[t]})$ is a function of the realizations $\boldsymbol{\xi}_{[t]} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_t)$ of random parameter vectors up to stage $t$. In our setting, we refer to (2.1) as the *true* problem where the random process $\tilde{\boldsymbol{\xi}}_{[T]}$ (i.e., the evolution of the random parameter vector) is governed by a continuous state Markov process. Using the Bellman's optimality principle [6], the true problem (2.1) at stage $t \in [T]$ can equivalently be expressed as the dynamic program

$$\begin{aligned} Q_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \min\ & \boldsymbol{c}_t^\top \boldsymbol{x}_t + \mathcal{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) \\ \text{s.t.}\ & \boldsymbol{x}_t \in \mathbb{R}_+^{d_t}, \\ & \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{B}_t \boldsymbol{x}_{t-1} = \boldsymbol{b}_t, \end{aligned} \tag{2.2}$$

where $\mathcal{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) = \mathbb{E}[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t]$ is the expected cost-to-go function. Here, for $t \in [T]$, the cost-to-go function $Q_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ in (2.2) represents the minimum future expected cost accrued from stage $t$ to the terminal stage $T$ given that the previous decision $\boldsymbol{x}_{t-1}$ is taken and the current stage parameter vector $\boldsymbol{\xi}_t$ is realized. For simplicity, we assume the terminal cost $\mathcal{Q}_{T+1}(\cdot) \equiv 0$. In the remainder of this paper, we focus on the dynamic programming equation since the solution scheme proposed in Section 3 is based on a cutting plane approximation of the expected cost-to-go functions in (2.2).

### 2.1 Assumptions

In the previous works studying MSLP, it is commonly assumed that i) the probability distribution of random parameters at each stage is known and ii) the data process is stagewise independent. In this paper, we depart from those assumptions with the purpose of developing a practically meaningful data-driven scheme for the MSLP problem (2.2). We make the following assumptions throughout the main paper:

(**A1**) **Markovian Random Parameters**. The first stage parameters $\boldsymbol{\xi}_1 = (\boldsymbol{c}_1, \boldsymbol{b}_1, \boldsymbol{A}_1, \boldsymbol{B}_1)$ are deterministic, whereas the subsequent stage parameters $\tilde{\boldsymbol{\xi}}_t = (\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{b}}_t, \tilde{\boldsymbol{A}}_t, \tilde{\boldsymbol{B}}_t)$ for $t \in [T] \setminus \{1\}$ are stochastic with support $\Xi_t \subset \mathbb{R}^p$. Furthermore, the stochastic process $\tilde{\boldsymbol{\xi}}_{[T]} = (\boldsymbol{\xi}_1, \ldots, \tilde{\boldsymbol{\xi}}_T)$ satisfies the Markov property. That is, the conditional distribution of $\tilde{\boldsymbol{\xi}}_{t+1}$ given the data process $\boldsymbol{\xi}_{[t]} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_t)$ up to time $t$ does not depend on $\boldsymbol{\xi}_{[t-1]} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1})$.

(**A2**) **Unknown Distribution**. The true joint distribution of $\tilde{\boldsymbol{\xi}}_{[T]} = (\boldsymbol{\xi}_1, \tilde{\boldsymbol{\xi}}_2, \ldots, \tilde{\boldsymbol{\xi}}_T)$ is unknown and only $N$ independent and identically distributed (i.i.d.) historical trajectories $\boldsymbol{\xi}_{[T]}^i = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2^i, \ldots, \boldsymbol{\xi}_T^i)$, $i \in [N]$, from the stochastic process are available. In this paper, we denote the empirical uncertainty sets by $\hat{\Xi}_1 = \{\boldsymbol{\xi}_1\}$ and $\hat{\Xi}_t = \{\boldsymbol{\xi}_t^i : i \in [N]\}$ for every $t \in [T] \setminus \{1\}$.

(**A3**) **Relatively Complete Recourse and Compactness**. For each stage $t \in [T]$, the feasible region $\mathcal{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ is nonempty and compact for any feasible $\boldsymbol{x}_{t-1}$ and any $\boldsymbol{\xi}_t \in \Xi_t$.

### 2.2 Discretized Problem

It is impossible to solve the true problem (2.2) exactly under assumptions (**A1**) and (**A2**) since the expected cost-to-go functions $\mathcal{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ for every $t \in [T-1]$ cannot be evaluated without full knowledge of the underlying (conditional) distribution. Even if we knew the true distribution, evaluating the multivariate (conditional) expectation and optimizing over the continuum of the state variables $\boldsymbol{x}_t$ make the true problem intractable to solve. To address these fundamental challenges, in this paper, we propose to estimate the true conditional probability using the historical data via the Nadaraya-Watson (NW) kernel regression [43,64] defined as follows:

$$\hat{w}_t^i(\boldsymbol{\xi}_t) = \frac{\mathcal{K}\left(\frac{\boldsymbol{\xi}_t - \boldsymbol{\xi}_t^i}{h}\right)}{\sum_{j \in [N]} \mathcal{K}\left(\frac{\boldsymbol{\xi}_t - \boldsymbol{\xi}_t^j}{h}\right)} \qquad \forall i \in [N]. \tag{2.3}$$

Here, $\mathcal{K}$ is a kernel function of choice and $h > 0$ is a smoothing parameter called bandwidth. In this paper, we use the exponential kernel function of the form $\mathcal{K}(\boldsymbol{\rho}) = \exp(-\|\boldsymbol{\rho}\|_2)$. The bandwidth parameter $h$ controls the smoothness of the estimator (2.3). A too small $h$ leads to undersmoothing, meaning that most weights are assigned to points close to $\boldsymbol{\xi}_t$. On the other hand, an extremely large $h$ reduces the weights $\hat{w}_t^i(\boldsymbol{\xi}_t)$ to

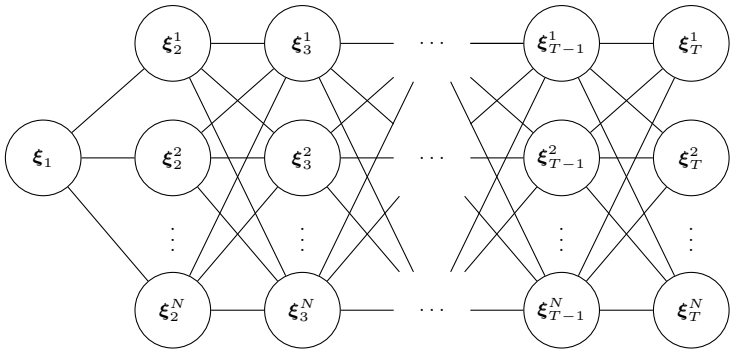

**Fig. 1** Recombining scenario tree using $N$ historical trajectories.

$1/N$, $\forall i \in [N]$. Györfi et al. [28] establish that the bandwidth scaling $h = O(1/N^{\frac{1}{p+4}})$ minimizes the expected estimation errors, which we adopt throughout the paper.

In our data-driven setting, the sample space of the true random process is discretized by i.i.d. historical trajectories $\{\boldsymbol{\xi}_t^i\}_{t=1}^T$ for every $i \in [N]$, giving rise to a recombining scenario tree depicted in Figure 1. Here, the true conditional distribution of $\boldsymbol{\xi}_{t+1}^i$ given $\boldsymbol{\xi}_t^j$ is approximated by the NW estimator (2.3). We refer to the MSLP problem represented by Figure 1 as the *discretized problem*. Similar to the true problem (2.2), using the dynamic programming equation, we present the discretized problem for stage $t \in [T]$ and each $\boldsymbol{\xi}_t \in \hat{\Xi}_t$ as

$$\begin{aligned} \hat{Q}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \min\ & \boldsymbol{c}_t^\top \boldsymbol{x}_t + \hat{\mathcal{Q}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) \\ \text{s.t.}\ & \boldsymbol{x}_t \in \mathbb{R}_+^{d_t}, \\ & \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{B}_t \boldsymbol{x}_{t-1} = \boldsymbol{b}_t, \end{aligned} \tag{2.4}$$

and $\hat{\mathcal{Q}}_{T+1}(\cdot) \equiv 0$. Here, the approximate expected cost-to-go function $\hat{\mathcal{Q}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ is defined as

$$\hat{\mathcal{Q}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) = \hat{\mathbb{E}}\left[\hat{Q}_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right] = \sum_{i \in [N]} \hat{w}_t^i(\boldsymbol{\xi}_t) \cdot \hat{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i). \tag{2.5}$$

Note that the approximate conditional expectation $\hat{\mathcal{Q}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ is taken with respect to $N$ data points, using the NW estimator in (2.3) and this weighted sum replaces the true expected cost-to-go function $\mathcal{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ in (2.2). This implies that the results from solving the discretized problem are not necessarily valid for the true problem. In addition, even with this drastic simplification, the discretized problem remains hard since the approximate cost-to-go functions still need to be evaluated for every $\boldsymbol{x}_t \in \mathcal{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$, which makes the discretized problem intractable. In Section 3, we discuss the stochastic dual dynamic programming algorithm that can deal with this kind of intractability.

### 2.3 Out-of-Sample Performance Guarantee

Data-driven solutions obtained from solving the discretized problem (2.4) can be suboptimal for the true problem due to the estimation errors from the NW estimator. In this paper, we are interested in understanding how well the data-driven scheme approximates the true problem and provides a reasonably good solution with a high probability.

Srivastava et al. [63] investigate the out-of-sample guarantee for the static stochastic problem with side information. In this section, we extend their result to the multi-stage setting. We state the following mild regularity conditions required for the generalization bound.

(**A4**) **Compact Uncertainty Set**. For each stage $t \in [T]$, the random parameter vector $\tilde{\boldsymbol{\xi}}_t$ is supported on a compact set $\Xi_t$.

(**A5**) **Differentiability**. For each stage $t \in [T]$, the joint density function $f(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1})$ is twice differentiable with continuous and bounded partial derivatives and the marginal density $f(\boldsymbol{\xi}_t \mid \boldsymbol{\xi}_{t-1})$ is non-zero for every $\boldsymbol{\xi}_{t-1} \in \hat{\Xi}_{t-1}$.

(**A6**) **Bandwidth**. The bandwidth parameter $h$ for the kernel function $\mathcal{K}$ is scaled such that $\lim_{N\to\infty} h_N = 0$ and $\lim_{N\to\infty} N h_N^p = \infty$.

The assumptions (**A4**)-(**A6**) are common in kernel regression estimation. The condition about the marginal density in the assumption (**A5**) ensures that the conditional probability is always positive throughout the time horizon. The condition about the bandwidth $h$ in the assumption (**A6**) guarantees that the NW

estimates asymptotically converge to the true conditional distribution. Moreover, in view of the assumption (**A4**), we can establish the following Lipschitz continuity condition for the cost-to-go fucntions.

**Lemma 1 (Lipschitz Cost-to-go Function)** *For each stage $t \in [T]$, the true cost-to-go function $Q_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ and the approximate cost-to-go function $\hat{Q}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ are $L_t$-Lipschitz continuous in $\boldsymbol{x}_{t-1}$. That is, there exists a constant $L_t > 0$ such that*

$$\begin{aligned}
\left|Q_t(\boldsymbol{x}, \boldsymbol{\xi}_t) - Q_t(\boldsymbol{x}', \boldsymbol{\xi}_t)\right| &\leq L_t \|\boldsymbol{x} - \boldsymbol{x}'\| \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}_{t-1}\left(\boldsymbol{x}_{t-2}, \boldsymbol{\xi}_{t-1}\right)\ \forall \boldsymbol{\xi}_t \in \Xi_t, \\
\left|\hat{Q}_t(\boldsymbol{x}, \boldsymbol{\xi}_t) - \hat{Q}_t(\boldsymbol{x}', \boldsymbol{\xi}_t)\right| &\leq L_t \|\boldsymbol{x} - \boldsymbol{x}'\| \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}_{t-1}\left(\boldsymbol{x}_{t-2}, \boldsymbol{\xi}_{t-1}\right)\ \forall \boldsymbol{\xi}_t \in \Xi_t.
\end{aligned}$$

The proof of Lemma 1 can be found in Appendix A.1. Based on the result, we obtain the following corollary on the Lipschitz continuity of the expected cost-to-go functions.

**Lemma 2** *At stage $t \in [T] \setminus \{1\}$, for any $\boldsymbol{\xi}_{t-1}$ we have*

$$\begin{aligned}
\left|\mathbb{E}\left[Q_t(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_t) \mid \boldsymbol{\xi}_{t-1}\right] - \mathbb{E}\left[Q_t(\boldsymbol{x}', \tilde{\boldsymbol{\xi}}_t) \mid \boldsymbol{\xi}_{t-1}\right]\right| &\leq L_t \|\boldsymbol{x} - \boldsymbol{x}'\| \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}_{t-1}\left(\boldsymbol{x}_{t-2}, \boldsymbol{\xi}_{t-1}\right), \\
\left|\hat{\mathbb{E}}\left[Q_t(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_t) \mid \boldsymbol{\xi}_{t-1}\right] - \hat{\mathbb{E}}\left[Q_t(\boldsymbol{x}', \tilde{\boldsymbol{\xi}}_t) \mid \boldsymbol{\xi}_{t-1}\right]\right| &\leq L_t \|\boldsymbol{x} - \boldsymbol{x}'\| \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}_{t-1}\left(\boldsymbol{x}_{t-2}, \boldsymbol{\xi}_{t-1}\right), \\
\left|\mathbb{E}\left[\hat{Q}_t(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_t) \,\middle|\, \boldsymbol{\xi}_{t-1}\right] - \mathbb{E}\left[\hat{Q}_t(\boldsymbol{x}', \tilde{\boldsymbol{\xi}}_t) \,\middle|\, \boldsymbol{\xi}_{t-1}\right]\right| &\leq L_t \|\boldsymbol{x} - \boldsymbol{x}'\| \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}_{t-1}\left(\boldsymbol{x}_{t-2}, \boldsymbol{\xi}_{t-1}\right), \\
\left|\hat{\mathbb{E}}\left[\hat{Q}_t(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_t) \,\middle|\, \boldsymbol{\xi}_{t-1}\right] - \hat{\mathbb{E}}\left[\hat{Q}_t(\boldsymbol{x}', \tilde{\boldsymbol{\xi}}_t) \,\middle|\, \boldsymbol{\xi}_{t-1}\right]\right| &\leq L_t \|\boldsymbol{x} - \boldsymbol{x}'\| \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}_{t-1}\left(\boldsymbol{x}_{t-2}, \boldsymbol{\xi}_{t-1}\right).
\end{aligned}$$

Lemma 2 can be verified by directly applying Lemma 1. Based on the assumption (**A4**), we also have the following lemma that plays an important role in deriving our generalization bound in Theorem 3.

**Lemma 3 (Bounded Conditional Variance)** *For each stage $t \in [T]$, there exists a constant $\sigma_{t+1}^2 > 0$ such that for any $\boldsymbol{x}_t \in \mathcal{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ and all feasible $\boldsymbol{x}_{t-1}$, and all $\boldsymbol{\xi}_t$, it holds that $\mathbb{V}[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t] \leq \sigma_{t+1}^2$.*

Under these assumptions and lemmas, we can use Theorem 1 in [63] to obtain the following generalization bound on the error of the NW estimate for any fixed state variables.

**Theorem 1 (Generalization Bound for Fixed State Variables $\boldsymbol{x}_t$)** *At stage $t \in [T]$, for any fixed $\boldsymbol{x}_t \in \mathcal{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ and any $\boldsymbol{\xi}_t \in \Xi_t$, we have*

$$\left|\mathbb{E}\left[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t\right] - \hat{\mathbb{E}}\left[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t\right]\right| \leq \sqrt{\frac{\mathbb{V}\left[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t\right]}{O(N^{\frac{4}{p+4}})(1 + o(1))g(\boldsymbol{\xi}_t)} \log\left(\frac{1}{\delta_{t+1}}\right)}, \tag{2.6}$$

*with probability at least $1 - \delta_{t+1}$. Here,*

$$g(\boldsymbol{\xi}_t) = \frac{f(\boldsymbol{\xi}_{t+1} \mid \boldsymbol{\xi}_t)}{2\int_{\mathbb{R}^p} \mathcal{K}^2(\boldsymbol{\rho}) \mathrm{d}\boldsymbol{\rho}}$$

*is the scaled marginal density of $\tilde{\boldsymbol{\xi}}_t$.*

Theorem 1 provides some insights about the error bound of the NW estimator: a small conditional variance $\mathbb{V}[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t]$ or a large scaled density $g(\boldsymbol{\xi}_t)$ can provide us a better bound. Now we extend the result to obtain a uniform generalization bound for every state variable in a continuous and bounded set under the assumptions (**A4**)-(**A6**).

**Theorem 2 (Generalization Bound for a Continuous and Bounded Feasible Region)** *At stage $t \in [T]$, for any fixed tolerance level $\eta > 0$ and $\boldsymbol{\xi}_t \in \Xi_t$, we have*

$$\begin{aligned}
&\left|\mathbb{E}\left[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t\right] - \hat{\mathbb{E}}\left[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t\right]\right| \\
&\qquad \leq \sqrt{\sigma_{t+1}^2 \frac{\log\left(\frac{O(1)(D_t/\eta)^{d_t}}{\delta_{t+1}}\right)}{O(N^{\frac{4}{p+4}})(1 + o(1))g(\boldsymbol{\xi}_t)}} + 2L_{t+1}\eta \quad \forall \boldsymbol{x}_t \in \mathcal{X}_t\left(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t\right),
\end{aligned} \tag{2.7}$$

*with probability at least $1 - \delta_{t+1}$. Here, $L_{t+1}$ is a Lipschitz constant for $Q_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1})$ defined in Lemma 1 and $D_t$ is a positive constant satisfying*

$$\sup_{\boldsymbol{x}_t, \boldsymbol{x}_t' \in \mathcal{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)} \|\boldsymbol{x}_t - \boldsymbol{x}_t'\| \leq D_t \quad \forall \boldsymbol{x}_{t-1}\ \forall \boldsymbol{\xi}_t.$$

We defer the proof of Theorem 2 to Appendix A.2. We note that Theorems 1 and 2 only provide bounds on errors purely due to the NW estimator. In this paper, we want to determine the errors between the true expected cost-to-go and the approximate expected cost-to-go that result from solving the discretized problem (2.4). Let $\boldsymbol{x}_1^\star$ and $\hat{\boldsymbol{x}}_1$ be optimal solutions to the true problem and the discretized problem, respectively. By recursively applying Theorem 2 from the terminal stage to the first stage, we derive the desired suboptimality bound in the following theorem.

**Theorem 3 (Out-of-Sample Performance Guarantee)** *For a fixed tolerance level $\eta > 0$, with probability at least $1 - \sum_{t=2}^{T} \delta_t$, we have*

$$\left(\boldsymbol{c}_1^\top \hat{\boldsymbol{x}}_1 + \mathbb{E}\left[Q_2(\hat{\boldsymbol{x}}_1, \tilde{\boldsymbol{\xi}}_2) \mid \boldsymbol{\xi}_1\right]\right) - \left(\boldsymbol{c}_1^\top \boldsymbol{x}_1^\star + \mathbb{E}\left[Q_2(\boldsymbol{x}_1^\star, \tilde{\boldsymbol{\xi}}_2) \mid \boldsymbol{\xi}_1\right]\right) \leq 2\sum_{t=2}^{T}\left(\sqrt{\sigma_t^2 \frac{\log\left(\frac{O(1)N^{t-2}\prod_{s=1}^{t-1}(D_s/\eta)^{d_s}}{\delta_t}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g_{t-1}}} + 2L_t\eta\right), \quad (2.8)$$

*where $L_t$ is a Lipschitz constant for $Q_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ defined in Lemma 1, $D_t$ is the finite diameter defined in Theorem 2, and $g_{t-1} = \min_{i\in[N]} g(\boldsymbol{\xi}_{t-1}^i)$.*

*Proof* We proceed by backward induction in the stages.
Stage $t = T$: We have $Q_T(\cdot, \boldsymbol{\xi}_T) = \hat{Q}_T(\cdot, \boldsymbol{\xi}_T)$ for every $\boldsymbol{\xi}_T$ due to the fact that $\mathcal{Q}_{T+1}(\cdot) \equiv 0$. From Theorem 2, for any fixed $\boldsymbol{x}_{T-2}$ and $\boldsymbol{\xi}_{T-1}$ we have

$$\left|\mathbb{E}\left[Q_T(\boldsymbol{x}_{T-1}, \tilde{\boldsymbol{\xi}}_T) \mid \boldsymbol{\xi}_{T-1}\right] - \hat{\mathbb{E}}\left[\hat{Q}_T(\boldsymbol{x}_{T-1}, \tilde{\boldsymbol{\xi}}_T) \,\middle|\, \boldsymbol{\xi}_{T-1}\right]\right| \leq \sqrt{\sigma_T^2 \frac{\log\left(\frac{O(1)(D_{T-1}/\eta)^{d_{T-1}}}{\delta_T}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g(\boldsymbol{\xi}_{T-1})}} + 2L_T\eta \quad \forall \boldsymbol{x}_{T-1} \in \mathcal{X}_{T-1}\left(\boldsymbol{x}_{T-2}, \boldsymbol{\xi}_{T-1}\right)$$

with probability at least $1 - \delta_T$. Taking minimization over $\mathcal{X}_{T-1}$, for any fixed $\boldsymbol{x}_{T-2}$ and $\boldsymbol{\xi}_{T-1}$ we have

$$\begin{aligned}&\left|\min_{\boldsymbol{x}_{T-1}\in\mathcal{X}_{T-1}(\boldsymbol{x}_{T-2},\boldsymbol{\xi}_{T-1})} \boldsymbol{c}_{T-1}^\top \boldsymbol{x}_{T-1} + \mathbb{E}\left[Q_T(\boldsymbol{x}_{T-1}, \tilde{\boldsymbol{\xi}}_T) \mid \boldsymbol{\xi}_{T-1}\right]\right. \\ &\qquad \left. - \min_{\boldsymbol{x}_{T-1}\in\mathcal{X}_{T-1}(\boldsymbol{x}_{T-2},\boldsymbol{\xi}_{T-1})} \boldsymbol{c}_{T-1}^\top \boldsymbol{x}_{T-1} + \hat{\mathbb{E}}\left[\hat{Q}_T(\boldsymbol{x}_{T-1}, \tilde{\boldsymbol{\xi}}_T) \,\middle|\, \boldsymbol{\xi}_{T-1}\right]\right| \\ =\ &\left|Q_{T-1}(\boldsymbol{x}_{T-2}, \boldsymbol{\xi}_{T-1}) - \hat{Q}_{T-1}(\boldsymbol{x}_{T-2}, \boldsymbol{\xi}_{T-1})\right| \leq \sqrt{\sigma_T^2 \frac{\log\left(\frac{O(1)(D_{T-1}/\eta)^{d_{T-1}}}{\delta_T}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g(\boldsymbol{\xi}_{T-1})}} + 2L_T\eta\end{aligned}$$

with probability at least $1 - \delta_T$. Applying union bound over the $N$ sample points $\boldsymbol{\xi}_{T-1}^i \in \hat{\Xi}_{T-1}$, $i \in [N]$, for any fixed $\boldsymbol{x}_{T-2}$, we have

$$\left|Q_{T-1}(\boldsymbol{x}_{T-2}, \boldsymbol{\xi}_{T-1}^i) - \hat{Q}_{T-1}(\boldsymbol{x}_{T-2}, \boldsymbol{\xi}_{T-1}^i)\right| \leq \sqrt{\sigma_T^2 \frac{\log\left(\frac{O(1)N(D_{T-1}/\eta)^{d_{T-1}}}{\delta_T}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g_{T-1}}} + 2L_T\eta \quad \forall \boldsymbol{\xi}_{T-1}^i \in \hat{\Xi}_{T-1},\ i \in [N] \quad (2.9)$$

with probability at least $1 - \delta_T$ and $g_{T-1} = \min_{i\in[N]} g(\boldsymbol{\xi}_{T-1}^i)$.
Stage $t = T-1$: From Theorem 1, for any fixed $\boldsymbol{x}_{T-2}$ and $\boldsymbol{\xi}_{T-2}$, we have

$$\left|\mathbb{E}\left[Q_{T-1}(\boldsymbol{x}_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \mid \boldsymbol{\xi}_{T-2}\right] - \hat{\mathbb{E}}\left[Q_{T-1}(\boldsymbol{x}_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \mid \boldsymbol{\xi}_{T-2}\right]\right| \leq \sqrt{\sigma_{T-1}^2 \frac{\log\left(\frac{1}{\delta_{T-1}}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g(\boldsymbol{\xi}_{T-2})}} \quad (2.10)$$

with probability at least $1 - \delta_{T-1}$. Note that $\hat{\mathbb{E}}\left[Q_{T-1}(\boldsymbol{x}_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \mid \boldsymbol{\xi}_{T-2}\right]$ in (2.10) is not the approximate expected cost-to-go defined in (2.5). Therefore, we use (2.9) to replace $Q_{T-1}(\boldsymbol{x}_{T-2}, \boldsymbol{\xi}_{T-1})$ with

$\hat{Q}_{T-1}(\boldsymbol{x}_{T-2}, \boldsymbol{\xi}_{T-1})$. Applying union bound to (2.9) and (2.10), for any fixed $\boldsymbol{x}_{T-2}$ and $\boldsymbol{\xi}_{T-2}$, we have

$$\begin{aligned}&\left|\mathbb{E}\left[Q_{T-1}(\boldsymbol{x}_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \mid \boldsymbol{\xi}_{T-2}\right] - \hat{\mathbb{E}}\left[\hat{Q}_{T-1}(\boldsymbol{x}_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \,\middle|\, \boldsymbol{\xi}_{T-2}\right]\right| \\ &\qquad \leq \sqrt{\sigma_T^2 \frac{\log\left(\frac{O(1)N(D_{T-1}/\eta)^{d_{T-1}}}{\delta_T}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g_{T-1}}} + \sqrt{\sigma_{T-1}^2 \frac{\log\left(\frac{1}{\delta_{T-1}}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g(\boldsymbol{\xi}_{T-2})}} + 2L_T\eta \end{aligned} \tag{2.11}$$

with probability at least $1-\delta_T-\delta_{T-1}$. Since the cost-to-go function $Q_{T-1}(\boldsymbol{x}_{T-2}, \boldsymbol{\xi}_{T-1})$ is $L_{T-1}$-Lipschitz continuous in $\boldsymbol{x}_{T-2}$, Lemma 2 implies that for any $\boldsymbol{x}_{T-2} \in \mathcal{X}_{T-2}(\boldsymbol{x}_{T-3}, \boldsymbol{\xi}_{T-2})$, there exists $\boldsymbol{x}'_{T-2} \in \mathcal{X}^{\eta}_{T-2}(\boldsymbol{x}_{T-3}, \boldsymbol{\xi}_{T-2})$, $\|\boldsymbol{x}_{T-2} - \boldsymbol{x}'_{T-2}\| \leq \eta$, such that:

$$\left|\mathbb{E}\left[Q_{T-1}(\boldsymbol{x}_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \mid \boldsymbol{\xi}_{T-2}\right] - \mathbb{E}\left[Q_{T-1}(\boldsymbol{x}'_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \mid \boldsymbol{\xi}_{T-2}\right]\right| \leq L_{T-1}\eta \tag{2.12}$$

Furthermore, applying union bound to (2.11) over $\boldsymbol{x}'_{T-2} \in \mathcal{X}^{\eta}_{T-2}(\boldsymbol{x}_{T-3}, \boldsymbol{\xi}_{T-2})$, we get

$$\begin{aligned}&\left|\mathbb{E}\left[Q_{T-1}(\boldsymbol{x}'_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \mid \boldsymbol{\xi}_{T-2}\right] - \hat{\mathbb{E}}\left[\hat{Q}_{T-1}(\boldsymbol{x}'_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \,\middle|\, \boldsymbol{\xi}_{T-2}\right]\right| \\ &\quad \leq \sqrt{\sigma_T^2 \frac{\log\left(\frac{O(1)N(D_{T-1}/\eta)^{d_{T-1}}(D_{T-2}/\eta)^{d_{T-2}}}{\delta_T}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g_{T-1}}} + \sqrt{\sigma_{T-1}^2 \frac{\log\left(\frac{O(1)(D_{T-2}/\eta)^{d_{T-2}}}{\delta_{T-1}}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g(\boldsymbol{\xi}_{T-2})}} + 2L_T\eta \\ &\qquad\qquad \forall \boldsymbol{x}'_{T-2} \in \mathcal{X}^{\eta}_{T-2}\left(\boldsymbol{x}_{T-3}, \boldsymbol{\xi}_{T-2}\right) \end{aligned} \tag{2.13}$$

with probability at least $1-\delta_T-\delta_{T-1}$. Using the Lipschitz continuity of $Q_{T-1}(\boldsymbol{x}_{T-2}, \boldsymbol{\xi}_{T-1})$ and $\hat{Q}_{T-1}(\boldsymbol{x}_{T-2}, \boldsymbol{\xi}_{T-1})$, we obtain

$$\begin{aligned}&\left|\mathbb{E}\left[Q_{T-1}(\boldsymbol{x}_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \mid \boldsymbol{\xi}_{T-2}\right] - \hat{\mathbb{E}}\left[\hat{Q}_{T-1}(\boldsymbol{x}_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \,\middle|\, \boldsymbol{\xi}_{T-2}\right]\right| \\ &\leq \sqrt{\sigma_T^2 \frac{\log\left(\frac{O(1)N(D_{T-1}/\eta)^{d_{T-1}}(D_{T-2}/\eta)^{d_{T-2}}}{\delta_T}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g_{T-1}}} + \sqrt{\sigma_{T-1}^2 \frac{\log\left(\frac{O(1)(D_{T-2}/\eta)^{d_{T-2}}}{\delta_{T-1}}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g(\boldsymbol{\xi}_{T-2})}} + 2L_T\eta + 2L_{T-1}\eta \\ &\qquad\qquad \forall \boldsymbol{x}_{T-2} \in \mathcal{X}_{T-2}\left(\boldsymbol{x}_{T-3}, \boldsymbol{\xi}_{T-2}\right) \end{aligned}$$

with probability at least $1-\delta_T-\delta_{T-1}$. Then minimizing over $\mathcal{X}_{T-2}$, for any fixed $\boldsymbol{\xi}_{T-2}$ and $\boldsymbol{x}_{T-3}$, we have

$$\begin{aligned}&\left|\min_{\boldsymbol{x}_{T-2} \in \mathcal{X}_{T-2}\left(\boldsymbol{x}_{T-3}, \boldsymbol{\xi}_{T-2}\right)} \boldsymbol{c}_{T-2}^{\top}\boldsymbol{x}_{T-2} + \mathbb{E}\left[Q_{T-1}(\boldsymbol{x}_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \mid \boldsymbol{\xi}_{T-2}\right]\right. \\ &\qquad \left. - \min_{\boldsymbol{x}_{T-2} \in \mathcal{X}_{T-2}\left(\boldsymbol{x}_{T-3}, \boldsymbol{\xi}_{T-2}\right)} \boldsymbol{c}_{T-2}^{\top}\boldsymbol{x}_{T-2} + \hat{\mathbb{E}}\left[\hat{Q}_{T-1}(\boldsymbol{x}_{T-2}, \tilde{\boldsymbol{\xi}}_{T-1}) \,\middle|\, \boldsymbol{\xi}_{T-2}\right]\right| \\ &= \left|Q_{T-2}(\boldsymbol{x}_{T-3}, \boldsymbol{\xi}_{T-2}) - \hat{Q}_{T-2}(\boldsymbol{x}_{T-3}, \boldsymbol{\xi}_{T-2})\right| \\ &\leq \sqrt{\sigma_T^2 \frac{\log\left(\frac{O(1)N(D_{T-1}/\eta)^{d_{T-1}}(D_{T-2}/\eta)^{d_{T-2}}}{\delta_T}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g_{T-1}}} + \sqrt{\sigma_{T-1}^2 \frac{\log\left(\frac{O(1)(D_{T-2}/\eta)^{d_{T-2}}}{\delta_{T-1}}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g(\boldsymbol{\xi}_{T-2})}} + 2L_T\eta + 2L_{T-1}\eta \end{aligned}$$

with probability at least $1-\delta_T-\delta_{T-1}$. Applying union bound over the $N$ sample points $\boldsymbol{\xi}^i_{T-2} \in \hat{\Xi}_{T-2},\ i \in [N]$, for any fixed $\boldsymbol{x}_{T-3}$, we have

$$\begin{aligned}&\left|Q_{T-2}(\boldsymbol{x}_{T-3}, \boldsymbol{\xi}^i_{T-2}) - \hat{Q}_{T-2}(\boldsymbol{x}_{T-3}, \boldsymbol{\xi}^i_{T-2})\right| \\ &\leq \sqrt{\sigma_T^2 \frac{\log\left(\frac{O(1)N^2(D_{T-1}/\eta)^{d_{T-1}}(D_{T-2}/\eta)^{d_{T-2}}}{\delta_T}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g_{T-1}}} + \sqrt{\sigma_{T-1}^2 \frac{\log\left(\frac{O(1)N(D_{T-2}/\eta)^{d_{T-2}}}{\delta_{T-1}}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g_{T-2}}} + 2L_T\eta + 2L_{T-1}\eta \\ &\qquad\qquad \forall \boldsymbol{\xi}^i_{T-2} \in \hat{\Xi}_{T-2},\ i \in [N] \end{aligned}$$

with probability at least $1-\delta_T-\delta_{T-1}$ and $g_{T-2}=\min_{i\in[N]} g(\boldsymbol{\xi}_{T-2}^i)$.
For stage $t=2$, by backward induction, we obtain

$$\left|\mathbb{E}\left[Q_2(\boldsymbol{x}_1,\tilde{\boldsymbol{\xi}}_2)\mid\boldsymbol{\xi}_1\right]-\hat{\mathbb{E}}\left[\hat{Q}_2(\boldsymbol{x}_1,\tilde{\boldsymbol{\xi}}_2)\,\middle|\,\boldsymbol{\xi}_1\right]\right|\leq\underbrace{\sum_{t=2}^{T}\left(\sqrt{\sigma_t^2\frac{\log\left(\frac{O(1)N^{t-2}\prod_{s=1}^{t-1}(D_s/\eta)^{d_s}}{\delta_t}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g_{t-1}}}+2L_t\eta\right)}_{\equiv\,\epsilon(\delta_2,\delta_3,\ldots,\delta_T)}\quad\forall\boldsymbol{x}_1\in\mathcal{X}_1\left(\boldsymbol{x}_0,\boldsymbol{\xi}_1\right)\tag{2.14}$$

with probability at least $1-\sum_{t=2}^{T}\delta_t$. To ease the notation, let the error term in (2.14) be a function of $\delta_2,\delta_3\ldots,\delta_T$, namely $\epsilon(\delta_2,\delta_3,\ldots,\delta_T)$. Then, for $\boldsymbol{x}_1=\hat{\boldsymbol{x}}_1$, we have

$$\begin{aligned}\boldsymbol{c}_1^\top\hat{\boldsymbol{x}}_1+\mathbb{E}\left[Q_2(\hat{\boldsymbol{x}}_1,\tilde{\boldsymbol{\xi}}_2)\mid\boldsymbol{\xi}_1\right]&\leq\left(\boldsymbol{c}_1^\top\hat{\boldsymbol{x}}_1+\hat{\mathbb{E}}\left[\hat{Q}_2(\hat{\boldsymbol{x}}_1,\tilde{\boldsymbol{\xi}}_2)\,\middle|\,\boldsymbol{\xi}_1\right]\right)+\epsilon(\delta_2,\delta_3,\ldots,\delta_T)&\text{(2.15a)}\\&\leq\left(\boldsymbol{c}_1^\top\boldsymbol{x}_1^\star+\hat{\mathbb{E}}\left[\hat{Q}_2(\boldsymbol{x}_1^\star,\tilde{\boldsymbol{\xi}}_2)\,\middle|\,\boldsymbol{\xi}_1\right]\right)+\epsilon(\delta_2,\delta_3,\ldots,\delta_T),&\text{(2.15b)}\end{aligned}$$

where the second inequality (2.15b) holds since $\boldsymbol{x}_1^\star$ is suboptimal to the discretized problem. Similarly, for $\boldsymbol{x}_1=\boldsymbol{x}_1^\star$, we have

$$-\left(\boldsymbol{c}_1^\top\boldsymbol{x}_1^\star+\mathbb{E}\left[Q_2(\boldsymbol{x}_1^\star,\tilde{\boldsymbol{\xi}}_2)\mid\boldsymbol{\xi}_1\right]\right)+\left(\boldsymbol{c}_1^\top\boldsymbol{x}_1^\star+\hat{\mathbb{E}}\left[\hat{Q}_2(\boldsymbol{x}_1^\star,\tilde{\boldsymbol{\xi}}_2)\,\middle|\,\boldsymbol{\xi}_1\right]\right)\leq\epsilon(\delta_2,\delta_3,\ldots,\delta_T)\tag{2.16}$$

Applying union bound to (2.15b) and (2.16), we establish the desired result:

$$\left(\boldsymbol{c}_1^\top\hat{\boldsymbol{x}}_1+\mathbb{E}\left[Q_2(\hat{\boldsymbol{x}}_1,\tilde{\boldsymbol{\xi}}_2)\mid\boldsymbol{\xi}_1\right]\right)-\left(\boldsymbol{c}_1^\top\boldsymbol{x}_1^\star+\mathbb{E}\left[Q_2(\boldsymbol{x}_1^\star,\tilde{\boldsymbol{\xi}}_2)\mid\boldsymbol{\xi}_1\right]\right)\leq2\epsilon(\delta_2,\delta_3,\ldots,\delta_T)$$

with probability at least $1-\sum_{t=2}^{T}\delta_t$. Thus, the claim follows. ∎
To gain insight from Theorem 3, we derive the following simple upper bound.

**Corollary 1** *Define* $L_{\max}=\max_{t\in[T-1]}L_{t+1}$, $D_{\max}=\max_{t\in[T-1]}(D_{t+1}/\eta)^{d_{t+1}}$, $\delta_{\min}=\min_{t\in[T-1]}\delta_{t+1}$, $\sigma_{\max}=\max_{t\in[T-1]}\sigma_{t+1}^2$, *and* $g_{\min}=\min_{t\in[T-1]}g_t$. *Then, we have*

$$\begin{aligned}&\left(\boldsymbol{c}_1^\top\hat{\boldsymbol{x}}_1+\mathbb{E}\left[Q_2(\hat{\boldsymbol{x}}_1,\tilde{\boldsymbol{\xi}}_2)\mid\boldsymbol{\xi}_1\right]\right)-\left(\boldsymbol{c}_1^\top\boldsymbol{x}_1^\star+\mathbb{E}\left[Q_2(\boldsymbol{x}_1^\star,\tilde{\boldsymbol{\xi}}_2)\mid\boldsymbol{\xi}_1\right]\right)\\&\qquad\leq\frac{2T}{O(N^{\frac{2}{p+4}})}\sqrt{\frac{\sigma_{\max}^2}{(1+o(1))g_{\min}}\log\left(\frac{O(1)N^{T-2}D_{\max}^{T-1}}{\delta_{\min}}\right)}+4L_{\max}T\eta\end{aligned}\tag{2.17}$$

*with probability at least* $1-\sum_{t=2}^{T}\delta_t$.

*Proof* Since $L_{\max}=\max_{t\in[T-1]}L_{t+1}$, $D_{\max}=\max_{t\in[T]}(D_t/\eta)^{d_t}$, $\delta_{\min}=\min_{t\in[T-1]}\delta_{t+1}$, $\sigma_{\max}^2=\max_{t\in[T]}\sigma_t^2$, and $g_{\min}=\min_{t\in[T-1]}g_t$, Theorem 3 implies

$$\begin{aligned}&\left(\boldsymbol{c}_1^\top\hat{\boldsymbol{x}}_1+\mathbb{E}\left[Q_2(\hat{\boldsymbol{x}}_1,\tilde{\boldsymbol{\xi}}_2)\mid\boldsymbol{\xi}_1\right]\right)-\left(\boldsymbol{c}_1^\top\boldsymbol{x}_1^\star+\mathbb{E}\left[Q_2(\boldsymbol{x}_1^\star,\tilde{\boldsymbol{\xi}}_2)\mid\boldsymbol{\xi}_1\right]\right)\\&\leq2\sum_{t=2}^{T}\left(\sqrt{\sigma_t^2\frac{\log\left(\frac{O(1)N^{t-2}\prod_{s=1}^{t-1}(D_s/\eta)^{d_s}}{\delta_t}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g_{t-1}}}+2L_t\eta\right)\leq\frac{2T}{O(N^{\frac{2}{p+4}})}\sqrt{\sigma_{\max}^2\frac{\log\left(\frac{O(1)N^{T-2}D_{\max}^{T-1}}{\delta_{\min}}\right)}{(1+o(1))g_{\min}}}+4L_{\max}T\eta\end{aligned}$$

with probability at least $1-\sum_{t=2}^{T}\delta_t$. ∎
Corollary 1 asserts that the generalization bound in Theorem 3 is at most $\widetilde{O}_p(T^{\frac{3}{2}}/N^{\frac{2}{p+4}})$. The result is meaningful from both theoretical and practical perspectives since i) it provides the first statistical analysis of the Markov discretization method and ii) the polynomial dependence on $T$ suggests that our scheme is applicable to problems that require a large number of time stages. However, the bound also depends on the dimension $p$ of the random parameter vector $\tilde{\boldsymbol{\xi}}_t$, implying large out-of-sample errors when $p$ is large. The estimation errors from the high dimensional regression are a relatively common problem. To mitigate the issue, one can utilize a dimensionality reduction algorithm. Using the procedure, one can improve the decay rate of the errors to $\widetilde{O}_p(T^{\frac{3}{2}}/N^{\frac{2}{p'+4}})$ when the effective dimensionality $p'$ of the random parameters is much smaller than the dimensionality $p$ of the ambient space.

*Remark 1* Shapiro [54] investigates MSLP problems with stagewise independence and known uncertainty distribution. He shows that using the popular sample average approximation (SAA) method, the errors diminish slowly at the rate of $\widetilde{O}_p(T^{\frac{1}{2}}/N^{\frac{1}{2T}})$. Thus, the sample complexity of solving the true problem grows exponentially in the number of stages, which makes SAA practically inapplicable for MSLP problems with a large $T$. ∎

## 3 Data-Driven SDDP

Our solution method to solve the discretized problem in Section 2.2 is based on stochastic dual dynamic programming (SDDP) algorithm introduced by Pereira and Pinto [46], which is a sampling-based variant of the nested Benders decomposition method by Birge [9]. SDDP is an iterative cutting plane approximation algorithm [33] where each iteration consists of two steps called the *forward* and *backward* steps. In the forward step, $M$ scenarios—sample paths in the scenario tree in Figure 1 from the first stage to the terminal stage—are sampled, and a sequence of candidate solutions corresponding to each scenario is obtained under the current approximation of the expected cost-to-go function $\hat{\mathcal{Q}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ in (2.4). In the subsequent backward step, the lower and upper bounds of the expected cost-to-go functions are updated starting from the terminal stage to the first stage along the candidate solutions obtained in the forward step. As the lower and upper approximations are evaluated only at the candidate solutions, SDDP can alleviate the intractability arising from discretizing state variables.

### 3.1 Lower Bound on $\hat{\mathcal{Q}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$

At the beginning of the $k$-th iteration of the SDDP algorithm, we consider the following lower bound problem for every $t \in [T]$ and $\boldsymbol{\xi}_t \in \hat{\Xi}_t$:

$$\begin{aligned} \underline{Q}_t^{k-1}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \min\ & \boldsymbol{c}_t^\top \boldsymbol{x}_t + \underline{\mathcal{Q}}_{t+1}^{k-1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) \\ \text{s.t.}\ & \boldsymbol{x}_t \in \mathbb{R}_+^{d_t}, \quad \boldsymbol{z}_t \in \mathbb{R}_+^{d_{t-1}}, \\ & \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{B}_t \boldsymbol{z}_t = \boldsymbol{b}_t, \\ & \boldsymbol{z}_t = \boldsymbol{x}_{t-1}, \end{aligned} \tag{3.1}$$

Here, $\underline{\mathcal{Q}}_{t+1}^{k-1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) = \sum_{i\in[N]} \hat{w}_t^i(\boldsymbol{\xi}_t) \cdot \underline{Q}_{t+1}^{k-1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i)$. Here, the expected cost-to-go function $\hat{\mathcal{Q}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ in (2.4) is replaced by the lower bound approximation $\underline{\mathcal{Q}}_{t+1}^{k-1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$, which is the maximum of a collection of linear functions, known as *cuts*, derived at each iteration. Note that we add the equality constraints $\boldsymbol{z}_t = \boldsymbol{x}_{t-1}$ where the auxiliary variables $\boldsymbol{z}_t$ acting as a copy of the previous state variables $\boldsymbol{x}_{t-1}$. This has the drawback of adding one extra variable and constraint for each state variable but simplifies the implementation of the algorithm by avoiding the matrix multiplication for cut gradients [21,25].

In the forward step, a scenario[2] is sampled using the approximate conditional distributions from the NW regression. More precisely, conditional on $\boldsymbol{\xi}_1$, we sample $\hat{\boldsymbol{\xi}}_2$ from the empirical uncertainty set $\hat{\Xi}_2$ according to the probabilities $\{\hat{w}_2^i(\boldsymbol{\xi}_1)\}_{i=1}^N$. Given $\hat{\boldsymbol{\xi}}_2$, we then sample $\hat{\boldsymbol{\xi}}_3$ from $\hat{\Xi}_3$ according to the probabilities $\{\hat{w}_3^i(\hat{\boldsymbol{\xi}}_2)\}_{i=1}^N$. This process is repeated until the terminal stage to generate a scenario. For $t = 1, \ldots, T-1$, the current lower bound problem (3.1) is solved at the previous stage solution $\bar{\boldsymbol{x}}_{t-1}^k$ and the scenario $\hat{\boldsymbol{\xi}}_t$. The main output in the forward step is a sequence of optimal solutions $\bar{\boldsymbol{x}}_{[T-1]}^k$, known as *candidate solutions*. In the following backward step, for $t = T, \ldots, 2$, the lower bound problem (3.1) is solved at candidate solution $\bar{\boldsymbol{x}}_{t-1}^k$ and every sample $\boldsymbol{\xi}_t \in \hat{\Xi}_t$. Then, the objective values $\underline{V}_t^i$ and dual solutions $\boldsymbol{\pi}_t^i$, $i \in [N]$, corresponding to $\boldsymbol{z}_t = \boldsymbol{x}_{t-1}$ are used to generate a new cut for updating the $k$-th approximation $\underline{\mathcal{Q}}_t^k(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_{t-1})$ for every $\boldsymbol{\xi}_{t-1} \in \hat{\Xi}_{t-1}$, as follows:

$$\underline{\mathcal{Q}}_t^k(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_{t-1}) \leftarrow \max\left\{ \underline{\mathcal{Q}}_t^{k-1}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_{t-1}), \left(\sum_{i\in[N]} \hat{w}_{t-1}^i(\boldsymbol{\xi}_{t-1}) \cdot \boldsymbol{\pi}_t^i\right)^\top (\boldsymbol{x}_{t-1} - \bar{\boldsymbol{x}}_{t-1}^k) + \sum_{i\in[N]} \hat{w}_{t-1}^i(\boldsymbol{\xi}_{t-1}) \cdot \underline{V}_t^i \right\}.$$

Due to the convexity of $\hat{\mathcal{Q}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ in $\boldsymbol{x}_t$, which we will discuss in detail in Section 4.3, $\underline{\mathcal{Q}}_t^k(\cdot)$ is a valid lower approximation for $\hat{\mathcal{Q}}_{t+1}(\cdot)$ for every $t \in [T-1] \setminus \{1\}$ in any iteration $k$. With these updated approximations, we solve $\min_{\boldsymbol{x}_1 \in \mathcal{X}_1(\boldsymbol{x}_0, \boldsymbol{\xi}_1)} \boldsymbol{c}_1^\top \boldsymbol{x}_1 + \underline{\mathcal{Q}}_1^k(\boldsymbol{x}_0, \boldsymbol{\xi}_1)$ to obtain a deterministic lower bound on the value of optimal policies of the discretized problem.

[2] More than one scenario can be sampled in each iteration. In this case, multiple sequences of candidate solutions can be used to generate cuts in the backward step.

### 3.2 Upper Bound on $\hat{\mathcal{Q}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$

In the standard SDDP, the algorithm terminates if a gap between the *deterministic* lower bound and the *stochastic* upper bound becomes within a predetermined tolerance level; see [55] for more detail. Although a small gap may indicate that the current lower bound problem provides a good approximation of the true problem, it may not be the case if the randomness in the upper bound construction causes an early termination of the algorithm. Moreover, it fails to provide a deterministic quality of the current approximation, which would be of interest to the decision maker. We avoid such a stochastic upper bound scheme; instead, we utilize the deterministic upper bound proposed in [24,47,61].

At the $k$-th iteration, we consider the following upper bound problem for every $t \in [T]$ and $\boldsymbol{\xi}_t \in \hat{\Xi}_t$:

$$\begin{aligned} \overline{Q}_t^{k-1}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \min\ & \boldsymbol{c}_t^\top \boldsymbol{x}_t + \overline{\mathcal{Q}}_{t+1}^{k-1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) + M_t \|\boldsymbol{y}_t\|_1 \\ \text{s.t. } & \boldsymbol{x}_t \in \mathbb{R}^{d_t}_+, \quad \boldsymbol{\theta} \in \mathbb{R}^{k-1}_+, \quad \boldsymbol{y}_t \in \mathbb{R}^{d_t}, \\ & \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{B}_t \boldsymbol{x}_{t-1} = \boldsymbol{b}_t, \\ & \sum_{j \in [k-1]} \theta_j \bar{\boldsymbol{x}}_t^j + \boldsymbol{y}_t = \boldsymbol{x}_t, \\ & \mathbf{e}^\top \boldsymbol{\theta} = 1. \end{aligned} \tag{3.2}$$

Here, $\overline{\mathcal{Q}}_{t+1}^{k-1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) = \sum_{i \in [N]} \hat{w}_t^i(\boldsymbol{\xi}_t) \cdot \overline{Q}_{t+1}^{k-1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i)$. Similar to the lower bound, the upper bound problem is obtained by replacing $\hat{\mathcal{Q}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ in (2.4) with the upper bound approximation $\overline{\mathcal{Q}}_{t+1}^{k-1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$. Here, $\overline{\mathcal{Q}}_{t+1}^{k-1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ is a lower convex envelope defined on a convex combination of the previously obtained $k-1$ candidate solutions, namely, $\sum_{j \in [k-1]} \theta_j \bar{\boldsymbol{x}}_t^j$ for every $\boldsymbol{\theta} \in \mathbb{R}^{k-1}_+$ such that $\mathbf{e}^\top \boldsymbol{\theta} = 1$. Note that the auxiliary variables $\boldsymbol{y}_t$ are introduced in (3.2) to ensure the feasibility of the upper bound problem because the relatively complete recourse in the assumption (**A3**) alone does not guarantee a feasible solution $\boldsymbol{x}_t \in \mathcal{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ that also belongs to the convex combination [61]. That is, $\boldsymbol{y}_t$ becomes nonzero if there does not exists $\boldsymbol{\theta}$ that satisfies the constraints $\sum_{j \in [k-1]} \theta_j \bar{\boldsymbol{x}}_t^j = \boldsymbol{x}_t$ and $\boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{B}_t \boldsymbol{x}_{t-1} = \boldsymbol{b}_t$. Hence, a penalty proportional to a sufficiently large scalar $M_t$ is imposed whenever $\boldsymbol{y}_t$ is nonzero. This, in turn, expands the state space of $\overline{\mathcal{Q}}_{t+1}^{k-1}(\cdot, \boldsymbol{\xi}_t)$ by adding a new candidate solution $\bar{\boldsymbol{x}}_t^k$ to the current convex combination.

In the backward step, for $t = T, \ldots, 2$, the current upper bound problem (3.2) is solved at every $\boldsymbol{\xi}_t \in \hat{\Xi}_t$ and candidate solution $\bar{\boldsymbol{x}}_{t-1}^k$, and then the objective values $\overline{V}_t^i$ for every $i \in [N]$ and $\bar{\boldsymbol{x}}_{t-1}^k$ are used to update the $k$-th approximation $\overline{\mathcal{Q}}_t^k(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_{t-1})$ for every $\boldsymbol{\xi}_{t-1} \in \hat{\Xi}_{t-1}$, as follows:

$$\overline{\mathcal{Q}}_t^k(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_{t-1}) \leftarrow \text{env} \left( \min \left\{ \overline{\mathcal{Q}}_t^{k-1}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_{t-1}), \sum_{i \in [N]} \hat{w}_{t-1}^i(\boldsymbol{\xi}_{t-1}) \cdot \overline{V}_t^i + \mathbb{1}_{\{\bar{\boldsymbol{x}}_{t-1}^k\}}(\boldsymbol{x}_{t-1}) \right\} \right).$$

Here, 'env' represents the lower convex envelope where $\mathbb{1}_{\{\bar{\boldsymbol{x}}_{t-1}\}}(\cdot)$ is an indicator function for the singleton set $\{\bar{\boldsymbol{x}}_{t-1}\}$ defined as

$$\mathbb{1}_{\{\bar{\boldsymbol{x}}_{t-1}\}}(\boldsymbol{x}_{t-1}) = \begin{cases} 0 & \text{if } \boldsymbol{x}_{t-1} = \bar{\boldsymbol{x}}_{t-1} \\ +\infty & \text{otherwise.} \end{cases}$$

The convexity of $\hat{\mathcal{Q}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ in $\boldsymbol{x}_t$ guarantees that this convex envelope is a valid upper bound in any iteration $k$ as discussed in Section 4.3. As the approximations $\overline{\mathcal{Q}}_t^k(\cdot)$ for every $t \in [T] \setminus \{1\}$ are refined along the new candidate solutions in the backward step, the objective value of $\overline{Q}_1^k(\boldsymbol{x}_0, \boldsymbol{\xi}_1)$ provides a tighter deterministic upper bound on the value of optimal policies of the discretized problem. Algorithm 1 shows our data-driven SDDP (DD-SDDP) scheme where we adopt the deterministic upper bound in lieu of the standard stochastic one.

### 3.3 Convergence

We defer the convergence proof to Section 4.3 since the data-driven distributionally robust SDDP (DDR-SDDP) scheme introduced in the next section generalizes DD-SDDP. We emphasize that the convergence indicates the lower and upper bound problems converge to the discretized problem (2.4), not to the true problem (2.2). This implies that our DD-SDDP scheme may suffer from poor out-of-sample performance discussed in Section 2.3 if the discretized problem does not provide a good approximation of the true problem.

**Algorithm 1:** Data-Driven SDDP (DD-SDDP)

**Input:** Given observation histories $\hat{\Xi}_t = \{\boldsymbol{\xi}_t^i\}_{i=1}^N \ \forall t \in [T]$ where $\boldsymbol{\xi}_t^i = (\boldsymbol{c}_t^i, \boldsymbol{A}_t^i, \boldsymbol{B}_t^i, \boldsymbol{b}_t^i)$, tolerance level $\varepsilon$

**Initialization:** $\overline{\mathcal{Q}}_{t+1}^0(\cdot, \boldsymbol{\xi}_t^i) = +\infty,\ \underline{\mathcal{Q}}_{t+1}^0(\cdot, \boldsymbol{\xi}_t^i) = -\infty \ \forall t \in [T-1] \ \forall i \in [N]$
$UB = +\infty,\ LB = -\infty$, Iteration $k = 1$

**Output:** Optimization problems with added cuts

**while** $|UB - LB| > \varepsilon \cdot \min\{\ |UB|,\ |LB|\ \}$ **do**

  **(forward step)**

  **for** $t = 1, \ldots, T-1$ **do**

    **if** $t = 1$ **then**

      Solve $\underline{Q}_1^{k-1}(\boldsymbol{x}_0, \boldsymbol{\xi}_1)$ in (3.1) and obtain $\bar{\boldsymbol{x}}_1^k$

    **else**

      Generate a sample $\hat{\boldsymbol{\xi}}_t$ from $\hat{\Xi}_t$ with probability $\{\hat{w}_{t-1}^i(\hat{\boldsymbol{\xi}}_{t-1})\}_{i=1}^N$

      Solve $\underline{Q}_t^{k-1}(\bar{\boldsymbol{x}}_{t-1}^k, \hat{\boldsymbol{\xi}}_t)$ in (3.1) and obtain $\bar{\boldsymbol{x}}_t^k$.

  **(backward step)**

  **for** $t = T, \ldots, 2$ **do**

    **for** $i = 1, \ldots, N$ **do**

      **(LB)** Solve $\underline{Q}_t^k(\bar{\boldsymbol{x}}_{t-1}^k, \boldsymbol{\xi}_t^i)$ in (3.1). Then obtain the optimal value $\underline{V}_t^i$ and the dual solution $\boldsymbol{\pi}_t^i$ corresponding to $\boldsymbol{z}_t = \bar{\boldsymbol{x}}_{t-1}^k$

      **(UB)** Solve $\overline{Q}_t^k(\bar{\boldsymbol{x}}_{t-1}^k, \boldsymbol{\xi}_t^i)$ in (3.2) and obtain the optimal value $\overline{V}_t^i$

    **if** $t > 2$ **then**

      **for** $j = 1, \ldots, N$ **do**

        **(Update Lower Bound)**

        Compute $\underline{\mathscr{V}}_{tj}^k = \sum_{i\in[N]} \hat{w}_{t-1}^i(\boldsymbol{\xi}_{t-1}^j)\underline{V}_t^i$ and $\mathscr{G}_{tj}^k = \sum_{i\in[N]} \hat{w}_{t-1}^i(\boldsymbol{\xi}_{t-1}^j)\boldsymbol{\pi}_t^i$

        $\underline{\mathcal{Q}}_t^k(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_{t-1}^j) \leftarrow \max\{\underline{\mathcal{Q}}_t^{k-1}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_{t-1}^j),\ \mathscr{G}_{tj}^{k\top}(\boldsymbol{x}_{t-1} - \bar{\boldsymbol{x}}_{t-1}^k) + \underline{\mathscr{V}}_{tj}^k\}$

        **(Update Upper Bound)**

        Compute $\overline{\mathscr{V}}_{tj}^k = \sum_{i\in[N]} \hat{w}_{t-1}^i(\boldsymbol{\xi}_{t-1}^j)\overline{V}_t^i$

        $\overline{\mathcal{Q}}_t^k(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_{t-1}^j) \leftarrow \mathrm{env}(\min\{\overline{\mathcal{Q}}_t^{k-1}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_{t-1}^j),\ \overline{\mathscr{V}}_{tj}^k + \mathbb{1}_{\{\bar{\boldsymbol{x}}_{t-1}^k\}}(\boldsymbol{x}_{t-1})\})$

    **else**

      **(Update Lower Bound)**

      Compute $\underline{\mathscr{V}}_2^k = \sum_{i\in[N]} \underline{V}_2^i/N$ and $\mathscr{G}_2^k = \sum_{i\in[N]} \boldsymbol{\pi}_2^i/N$

      $\underline{\mathcal{Q}}_2^k(\boldsymbol{x}_1, \boldsymbol{\xi}_1) \leftarrow \max\{\underline{\mathcal{Q}}_2^{k-1}(\boldsymbol{x}_1, \boldsymbol{\xi}_1),\ \mathscr{G}_2^{k\top}(\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^k) + \underline{\mathscr{V}}_2^k\}$

      **(Update Upper Bound)**

      Compute $\overline{\mathscr{V}}_2^k = \sum_{i\in[N]} \overline{V}_2^i/N$

      $\overline{\mathcal{Q}}_2^k(\boldsymbol{x}_1, \boldsymbol{\xi}_1) \leftarrow \mathrm{env}(\min\{\overline{\mathcal{Q}}_2^{k-1}(\boldsymbol{x}_1, \boldsymbol{\xi}_1),\ \mathbb{1}_{\{\bar{\boldsymbol{x}}_1^k\}}(\boldsymbol{x}_1) + \overline{\mathscr{V}}_2^k\})$

  Solve $\underline{Q}_1^k(\boldsymbol{x}_0, \boldsymbol{\xi}_1)$ in (3.1) and obtain the optimal value $LB$

  Solve $\overline{Q}_1^k(\boldsymbol{x}_0, \boldsymbol{\xi}_1)$ in (3.2) and obtain the optimal value $UB$

  $k = k + 1$

## 4 Regularization Schemes

Based on Theorem 2, the out-of-sample performance may be poor if the constant upper bound $\sigma_{t+1}^2$ on the conditional variance in (2.7) is large. In principle, $\sigma_{t+1}^2$ can be replaced with the true conditional variance $\mathbb{V}[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t]$ to obtain a tighter generalization bound. We make use of a regularization scheme involving the conditional variance term that provides a better out-of-sample performance guarantee. The regularization scheme, however, is intractable due to nonconvexity. In this section, we propose the data-driven distributionally robust SDDP (DDR-SDDP) scheme as a tractable approximation.

### 4.1 Variance Regularization Scheme

In Theorem 2, the constant upper bound $\sigma_{t+1}^2$ on the true conditional variance $\mathbb{V}[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t]$ is used as a proxy. Since $\sigma_{t+1}^2 \geq \mathbb{V}[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t]$ for every feasible $\boldsymbol{x}_t$ and $\boldsymbol{\xi}_t \in \Xi_t$, replacing $\sigma_{t+1}^2$ with

$\mathbb{V}[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t]$ gives rise to a tighter generalization bound

$$\left|\mathbb{E}\left[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t\right] - \hat{\mathbb{E}}\left[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t\right]\right| \leq \sqrt{\frac{\mathbb{V}\left[Q_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t\right]}{g(\boldsymbol{\xi}_t)(1+o(1))N^{\frac{4}{p+4}}} \log\left(\frac{O(1)(D_t/\eta)^{d_t}}{\delta_{t+1}}\right)} + L_{t+1}\eta\left(1 + \sqrt{\frac{\log\left(\frac{O(1)(D_t/\eta)^{d_t}}{\delta_{t+1}}\right)}{g(\boldsymbol{\xi}_t)(1+o(1))N^{\frac{4}{p+4}}}}\right) \tag{4.1}$$

with probability at least $1-\delta_{t+1}$. Note that the true variance term provides a more faithful characterization of the generalization errors compared to the bound in Theorem 2. Thus, it implies that a regularization scheme involving the true conditional variance yields good out-of-sample performance, in view of the bound. Nevertheless, the true conditional variance is unknown under the assumption **(A2)**, so we utilize the empirical conditional variance given by

$$\begin{aligned}\hat{\mathbb{V}}\left[\hat{Q}_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right] &\equiv \hat{\mathbb{E}}\left[\left(\hat{Q}_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) - \hat{\mathbb{E}}\left[\hat{Q}_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right]\right)^2 \,\middle|\, \boldsymbol{\xi}_t\right] \\ &= \hat{\mathbb{E}}\left[\hat{Q}_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1})^2 \,\middle|\, \boldsymbol{\xi}_t\right] - \hat{\mathbb{E}}\left[\hat{Q}_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right]^2.\end{aligned} \tag{4.2}$$

Using this approximation, we obtain the variance regularized formulation of the data-driven dynamic program (2.4), as follows:

$$\begin{aligned}\hat{Q}_t^{\mathcal{VR}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \min\ & \boldsymbol{c}_t^\top \boldsymbol{x}_t + \hat{\mathcal{Q}}_{t+1}^{\mathcal{VR}}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) + \lambda\sqrt{\hat{\mathbb{V}}\left[\hat{Q}_{t+1}^{\mathcal{VR}}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right]} \\ \text{s.t.}\ & \boldsymbol{x}_t \in \mathbb{R}_+^{d_t}, \\ & \boldsymbol{A}_t\boldsymbol{x}_t + \boldsymbol{B}_t\boldsymbol{x}_{t-1} = \boldsymbol{b}_t,\end{aligned} \tag{4.3}$$

where $\hat{\mathcal{Q}}_{t+1}^{\mathcal{VR}}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) = \hat{\mathbb{E}}[\hat{Q}_{t+1}^{\mathcal{VR}}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t] = \sum_{i\in[N]} \hat{w}_t^i(\boldsymbol{\xi}_t) \cdot \hat{Q}_{t+1}^{\mathcal{VR}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i)$ and $\lambda \geq 0$ is a tuning parameter that controls the degree of regularization.

### 4.2 Data-Driven Distributionally Robust SDDP

Unfortunately, solving the regularization problem (4.3) is intractable since the empirical conditional variance term $\hat{\mathbb{V}}[\hat{Q}_{t+1}^{\mathcal{VR}}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t]$ in the objective function is nonconvex in $\boldsymbol{x}_t$. Instead, we apply the distributionally robust optimization (DRO) methodology to the discretized problem (2.2) as a tractable alternative to the variance regularization scheme. Specifically, we propose the following distributionally robust dynamic program for $t \in [T]$ and $\boldsymbol{\xi}_t \in \hat{\Xi}_t$,

$$\begin{aligned}\hat{Q}_t^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \min\ & \boldsymbol{c}_t^\top \boldsymbol{x}_t + \max_{\mathbb{P}_t \in \mathcal{P}_t^\lambda(\hat{\mathbb{P}}_t)} \mathbb{E}_{\mathbb{P}_t}\left[\hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right] \\ \text{s.t.}\ & \boldsymbol{x}_t \in \mathbb{R}_+^{d_t}, \\ & \boldsymbol{A}_t\boldsymbol{x}_t + \boldsymbol{B}_t\boldsymbol{x}_{t-1} = \boldsymbol{b}_t,\end{aligned} \tag{4.4}$$

where $\hat{Q}_{T+1}^{\mathcal{DRO}}(\cdot) \equiv 0$. Here, the ambiguity set $\mathcal{P}_t^\lambda(\hat{\mathbb{P}}_t)$ with radius parameter $\lambda \geq 0$ is defined as

$$\mathcal{P}_t^\lambda(\hat{\mathbb{P}}_t) = \left\{\mathbb{P}_t = \sum_{i\in[N]} w_t^i \delta_{\boldsymbol{\xi}_t^i} \ :\ \sum_{i\in[N]} \left|\frac{w_t^i - \hat{w}_t^i(\boldsymbol{\xi}_t)}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}}\right| \leq \sqrt{N}\lambda,\ \max_{i\in[N]} \left|\frac{w_t^i - \hat{w}_t^i(\boldsymbol{\xi}_t)}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}}\right| \leq \lambda,\ \mathbf{e}^\top \boldsymbol{w}_t = 1,\ \boldsymbol{w}_t \in \mathbb{R}_+^N\right\}, \tag{4.5}$$

where each candidate distribution $\mathbb{P}_t \in \mathcal{P}_t^\lambda(\hat{\mathbb{P}}_t)$ is supported on the $N$ observed data points with probabilities close to the nominal weights $\{\hat{w}_t^i(\boldsymbol{\xi}_t)\}_{i=1}^N$ up to a certain distance $\lambda$ with respect to the measure described in the ambiguity set (4.5). The distance measure of our choice is a polyhedral outer approximation to the *modified* $\chi^2$ ambiguity set; see [5,7]. More precisely, we construct the ambiguity set using 1-norm and $\infty$-norm to conservatively approximate the 2-norm in the description of the modified $\chi^2$ distance. While utilizing the polyhedral approximation is mainly motivated by the computational benefits, it also provides a connection between the variance regularization formulation (4.3) and the DRO formulation (4.4) as stated in the following proposition.

**Proposition 1** *For any $\boldsymbol{x}_t$ and $\boldsymbol{\xi}_t$, we have*

$$\hat{\mathbb{E}}\left[\hat{Q}_{t+1}^{\mathcal{VR}}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right] + \lambda\sqrt{\hat{\mathbb{V}}\left[\hat{Q}_{t+1}^{\mathcal{VR}}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right]} \leq \max_{\mathbb{P}_t \in \mathcal{P}_t^\lambda(\hat{\mathbb{P}}_t)} \mathbb{E}_{\mathbb{P}_t}\left[\hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right] + O(\lambda^2).$$

The proof of Proposition 1 can be found in Appendix B.1. Proposition 1 implies that the DRO model provides a more accurate approximation of the variance regularization scheme as we observe more sample trajectories and decrease $\lambda$ accordingly. Specifically, the suboptimality bound for the variance regularization scheme in [63] suggests to scale the regularization parameter according to $\lambda = O(1/N^{\frac{2}{p+4}})$, which converges to 0 as $N \to \infty$. In the following proposition, we present a tractable single-level reformulation for (4.4).

**Proposition 2 (A Tractable Reformulation)** *At stage $t \in [T]$, the DRO problem (4.4) can be reformulated as the linear program*

$$\begin{aligned}
\hat{Q}_t^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \min \quad & \boldsymbol{c}_t^\top \boldsymbol{x}_t + \gamma + \lambda \left( \beta + \sum_{i\in[N]} \psi_i \right) + \sum_{i\in[N]} \sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}(\mu_i - \zeta_i) \\
s.t. \quad & \boldsymbol{x}_t \in \mathbb{R}_+^{d_t}, \quad \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}_+, \quad \boldsymbol{\mu}, \boldsymbol{\zeta}, \boldsymbol{\psi} \in \mathbb{R}_+^N, \\
& \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{B}_t \boldsymbol{x}_{t-1} = \boldsymbol{b}_t, \\
& \hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i) \le \gamma + \frac{\mu_i - \zeta_i}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} \qquad \forall i \in [N], \\
& \mu_i + \zeta_i = \psi_i + \frac{\beta}{\sqrt{N}} \qquad \forall i \in [N].
\end{aligned} \tag{4.6}$$

The proof of Proposition 2 is deferred to Appendix B.2.

As a solution scheme for the DRO problem (4.6), we propose our data-driven distributionally robust SDDP (DDR-SDDP) scheme in Algorithm 2. Similiar to DD-SDDP in Section 3, we construct the lower and upper bound problems, as follows:

$$\begin{aligned}
\underline{Q}_{t,k-1}^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \min \quad & \boldsymbol{c}_t^\top \boldsymbol{x}_t + \gamma + \lambda \left( \beta + \sum_{i\in[N]} \psi_i \right) + \sum_{i\in[N]} \sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}(\mu_i - \zeta_i) \\
\text{s.t.} \quad & \boldsymbol{x}_t \in \mathbb{R}_+^{d_t}, \quad \boldsymbol{z}_t \in \mathbb{R}_+^{d_{t-1}}, \quad \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}_+, \quad \boldsymbol{\mu}, \boldsymbol{\zeta}, \boldsymbol{\psi} \in \mathbb{R}_+^N \\
& \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{B}_t \boldsymbol{z}_t = \boldsymbol{b}_t, \\
& \boldsymbol{z}_t = \boldsymbol{x}_{t-1}, \\
& \underline{Q}_{t+1,k-1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i) \le \gamma + \frac{\mu_i - \zeta_i}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} \qquad \forall i \in [N], \\
& \mu_i + \zeta_i = \psi_i + \frac{\beta}{\sqrt{N}} \qquad \forall i \in [N].
\end{aligned} \tag{4.7}$$

$$\begin{aligned}
\overline{Q}_{t,k-1}^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \min \quad & \boldsymbol{c}_t^\top \boldsymbol{x}_t + \gamma + \lambda \left( \beta + \sum_{i\in[N]} \psi_i \right) + \sum_{i\in[N]} \sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}(\mu_i - \zeta_i) \\
\text{s.t.} \quad & \boldsymbol{x}_t, \boldsymbol{y}_t^i \in \mathbb{R}_+^{d_t}, \quad \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}_+, \quad \boldsymbol{\mu}, \boldsymbol{\zeta}, \boldsymbol{\psi} \in \mathbb{R}_+^N, \quad \boldsymbol{\theta}^i \in \mathbb{R}_+^{k-1} \quad \forall i \in [N], \\
& \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{B}_t \boldsymbol{x}_{t-1} = \boldsymbol{b}_t, \\
& \overline{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i) \le \gamma + \frac{\mu_i - \zeta_i}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} \qquad \forall i \in [N], \\
& \sum_{j\in[k]} \theta_j^i \overline{Q}_{t+1,j}^{\mathcal{DRO}}(\bar{\boldsymbol{x}}_t^j, \boldsymbol{\xi}_{t+1}^i) + M_{t+1} \|\boldsymbol{y}_t^i\|_1 = \overline{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i) \qquad \forall i \in [N], \\
& \sum_{j\in[k]} \theta_j^i \bar{\boldsymbol{x}}_t^j + \boldsymbol{y}_t^i = \boldsymbol{x}_t \qquad \forall i \in [N], \\
& \mu_i + \zeta_i = \psi_i + \frac{\beta}{\sqrt{N}} \qquad \forall i \in [N], \\
& \mathbf{e}^\top \boldsymbol{\theta}^i = 1 \qquad \forall i \in [N].
\end{aligned} \tag{4.8}$$

In DDR-SDDP, we use a multicut version of SDDP, i.e., for each $\boldsymbol{\xi}_t \in \hat{\Xi}_t$, the cost-to-go function $\hat{Q}_t^{\mathcal{DRO}}(\cdot, \boldsymbol{\xi}_t)$ is approximated by $\underline{Q}_{t,k}^{\mathcal{DRO}}(\cdot, \boldsymbol{\xi}_t)$ and $\overline{Q}_{t,k}^{\mathcal{DRO}}(\cdot, \boldsymbol{\xi}_t)$. Each iteration of DDR-SDDP is tractable since the upper and lower bound problems are linear programs, which are efficiently solvable using off-the-shelf solvers.

**Algorithm 2:** Data-Driven Distributionally Robust SDDP (DDR-SDDP)

**Input:** Given observation histories $\hat{\Xi}_t = \{\boldsymbol{\xi}_t^i\}_{i=1}^N \ \forall t \in [T]$ where $\boldsymbol{\xi}_t^i = (\boldsymbol{c}_t^i, \boldsymbol{A}_t^i, \boldsymbol{B}_t^i, \boldsymbol{b}_t^i)$, tolerance level $\varepsilon$

**Initialization:** $\overline{Q}_{t+1,0}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i) = +\infty$, $\underline{Q}_{t+1,0}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i) = -\infty \ \forall t \in [T-1] \ \forall i \in [N]$
$UB = +\infty$, $LB = -\infty$, Iteration $k = 1$

**Output:** Optimization problems with added cuts

**while** $|UB - LB| > \varepsilon \cdot \min\{|UB|, |LB|\}$ **do**
  **(forward step)**
  **for** $t = 1, \ldots, T-1$ **do**
    **if** $t = 1$ **then**
      Solve $\underline{Q}_{1,k-1}^{\mathcal{DRO}}(\boldsymbol{x}_0, \boldsymbol{\xi}_1)$ in (4.7) and obtain $\bar{\boldsymbol{x}}_1^k$
    **else**
      Generate a sample $\hat{\boldsymbol{\xi}}_t$ from $\hat{\Xi}_t$ with probability $\{\hat{w}_{t-1}^i(\hat{\boldsymbol{\xi}}_{t-1})\}_{i=1}^N$
      Solve $\underline{Q}_{t,k-1}^{\mathcal{DRO}}(\bar{\boldsymbol{x}}_{t-1}^k, \hat{\boldsymbol{\xi}}_t)$ in (4.7) and obtain $\bar{\boldsymbol{x}}_t^k$.
  **(backward step)**
  **for** $t = T, \ldots, 2$ **do**
    **for** $i = 1, \ldots, N$ **do**
      **(LB)** Solve $\underline{Q}_{t,k}^{\mathcal{DRO}}(\bar{\boldsymbol{x}}_{t-1}^k, \boldsymbol{\xi}_t^i)$ in (4.7). Then obtain the optimal value $\underline{V}_t^i$ and the dual solution $\boldsymbol{\pi}_t^i$ corresponding to $\boldsymbol{z}_t = \bar{\boldsymbol{x}}_{t-1}^k$
      **(UB)** Solve $\overline{Q}_{t,k}^{\mathcal{DRO}}(\bar{\boldsymbol{x}}_{t-1}^k, \boldsymbol{\xi}_t^i)$in (4.8) and obtain the optimal value $\overline{V}_t^i$
    **for** $i = 1, \ldots, N$ **do**
      **(Update Lower Bound)**
      $\underline{Q}_{t,k}^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t^i) \leftarrow \max\{\underline{Q}_{t,k-1}^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t^i), {\boldsymbol{\pi}_t^i}^\top(\boldsymbol{x}_{t-1} - \bar{\boldsymbol{x}}_{t-1}^k) + \underline{V}_{t,k}^i\}$
      **(Update Upper Bound)**
      $\overline{Q}_{t,k}^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t^i) \leftarrow \mathrm{env}(\min\{\overline{Q}_{t,k-1}^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t^i), \ \overline{V}_{t,k}^i + \mathbb{1}_{\{\bar{\boldsymbol{x}}_{t-1}^k\}}(\boldsymbol{x}_{t-1})\})$
Solve $\underline{Q}_{1,k}^{\mathcal{DRO}}(\boldsymbol{x}_0, \boldsymbol{\xi}_1)$ in (4.7) and obtain the optimal value $LB$
Solve $\overline{Q}_{1,k}^{\mathcal{DRO}}(\boldsymbol{x}_0, \boldsymbol{\xi}_1)$ in (4.8) and obtain the optimal value $UB$
$k = k + 1$

### 4.3 Convergence

To establish the convergence of the DDR-SDDP scheme presented in Algorithm 2, we make the following assumption.

(**B1**) **Basic Feasible Solution**. Candidate solution $\bar{\boldsymbol{x}}_t^k$ obtained during the forward step and dual solutions $\bar{\boldsymbol{\pi}}_{t,k}^i$ for every $i \in [N]$ obtained during the backward step are basic feasible solutions for any iteration $k$.

(**B1**) is a mild assumption since it can be easily satisfied by using the simplex method to solve the linear programs during the forward and backward steps.

As stated earlier, the DRO problem (4.4) reduces to the discretized problem (2.4) if $\lambda$ is set to 0. That is, the expected cost-to-go function $\sum_{i\in[N]} \hat{w}_t^i(\boldsymbol{\xi}_t) \cdot \hat{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i)$ in (2.4) replaces the inner maximization problem in (4.4) if the ambiguity set $\mathcal{P}_t^\lambda(\hat{\mathbb{P}}_t)$ contains only the nominal distribution $\hat{\mathbb{P}}_t$. Thus, the convergence proof presented in this section holds for DD-SDDP in Algorithm 1 as well. Before we present the convergence proof, we show that the cost-to-go function $\hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1})$ is piecewise linear convex in the state variables $\boldsymbol{x}_t$.

**Lemma 4** *At each stage $t \in [T-1]$, for any $\boldsymbol{\xi}_t \in \hat{\Xi}_t$, $\boldsymbol{\xi}_{t+1} \in \hat{\Xi}_{t+1}$, and feasible $\boldsymbol{x}_{t-1}$, the cost-to-go function $\hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1})$ is piecewise linear convex in $\boldsymbol{x}_t \in \mathcal{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ with a finite number of pieces.*

The proof of Lemma 4 can be found in Appendix B.3. Lemma 4 implies that the cost-to-go function $\hat{Q}_t^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ for every $\boldsymbol{\xi}_t \in \hat{\Xi}_t$ can be restored by a finite number of supporting hyperplanes, although the number of such hyperplanes can be prohibitively large. Furthermore, in the following lemma, we establish that the lower and upper bound problems provide valid lower and upper bounds for the discretized problem, respectively.

**Lemma 5** *For the upper bound problem in (4.8), suppose $M_{t+1} \geq L_{t+1}^{\mathcal{DRO}} \ \forall t \in [T-1]$, where $L_{t+1}^{\mathcal{DRO}}$ is a Lipschitz constant for $\hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1})$ under the 1-norm. Then, for any iteration $k$, the upper and lower bound problems provide valid bounds on the optimal value of the discretized problem, i.e.,*

$$\underline{Q}_{t+1,k}^{\mathcal{DRO}}(\cdot, \boldsymbol{\xi}_{t+1}) \leq \hat{Q}_{t+1}^{\mathcal{DRO}}(\cdot, \boldsymbol{\xi}_{t+1}) \leq \overline{Q}_{t+1,k}^{\mathcal{DRO}}(\cdot, \boldsymbol{\xi}_{t+1}) \quad \forall t \in [T-1] \ \forall \boldsymbol{\xi}_{t+1} \in \hat{\Xi}_{t+1}.$$

The proof of Lemma 5 is deferred to Appendix B.4.

An almost sure finite convergence of the standard SDDP algorithm under stagewise independence has been established in the existing works [50,55]. These papers, however, only consider the lower bound convergence. It is also worth mentioning that they have different settings on uncertain parameters. Philpott and Guan [50] consider randomness only on the right-hand side of the constraints of the linear program, whereas Shapiro [55] considers a more general MSLP problem where uncertainty can enter any parameters, e.g., objective function coefficients, a recourse matrices, etc. The following theorem presents the convergence of the data-driven distributionally robust SDDP algorithm.

**Theorem 4** *Consider the DDR-SDDP scheme described in Algorithm 2 without the termination condition in line 1. Then, the algorithm converges to an optimal solution of the following DRO problem,*

$$\min_{\boldsymbol{x}_1 \in \mathcal{X}_1(\boldsymbol{x}_0, \boldsymbol{\xi}_1)} \boldsymbol{c}_1^\top \boldsymbol{x}_1 + \max_{\mathbb{P}_1 \in \mathcal{P}_1^\lambda(\hat{\mathbb{P}}_1)} \mathbb{E}_{\mathbb{P}_1} \left[ \hat{Q}_2^{\mathcal{DRO}}(\boldsymbol{x}_1, \tilde{\boldsymbol{\xi}}_2) \,\middle|\, \boldsymbol{\xi}_1 \right]$$

*in a finite number of iterations with probability one.*

We defer convergence analysis to Appendix B.5, where we propose our upper bound convergence proof while the lower bound convergence relies on [55].

## 5 Numerical Experiments

In this section, we conduct numerical experiments to assess the performance of our proposed schemes. All optimization problems are solved using Python 3.7 with GUROBIPY 9.5.2 on on a 6-core, 2.3GHz Intel Core i7 CPU laptop with 16GB RAM.

### 5.1 Portfolio Optimization

We consider the classical multistage portfolio optimization problem where an investor aims to maximize his/her utility at the terminal stage by re-balancing the portfolio at each time stage. The portfolio can be selected from $K$ risky assets and one risk-free asset with a fixed return rate $r_f$. Initially, the investor has \$1 available in the risk-free asset. At each stage $t \in [T-1]$, the investor can either hold his/her position, buy more, or sell off part (or all) of asset $i \in [K]$ before observing the returns $\boldsymbol{\xi}_{t+1} \in \mathbb{R}^K$ of risky assets at stage $t+1$. We denote by $\boldsymbol{u}_t^+ \in \mathbb{R}_+^K$ the amount of risky assets bought and by $\boldsymbol{u}_t^- \in \mathbb{R}_+^K$ the amount of risky assets sold at stage $t$. At the end of stage $t$, the value of asset $i$, $s_{t,i}$, equals the previous value $s_{t-1,i}$, plus the realized return $\xi_{t,i} s_{t-1,i}$ during the period, plus the newly bought amount $u_{t,i}^+$, minus the newly sold amount $u_{t,i}^-$. We use $f_b$ and $f_s$ to denote per-unit transaction costs for buying and selling a unit asset, respectively.

The problem can be solved via the (true) dynamic program for $t \in [T-1]$:

$$\begin{aligned} Q_t(\boldsymbol{s}_{t-1}, \boldsymbol{\xi}_t) = \max\ & \mathbb{E}\left[ Q_{t+1}(\boldsymbol{s}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t \right] \\ \text{s.t.}\ & \boldsymbol{s}_t \in \mathbb{R}_+^{K+1}, \quad \boldsymbol{u}_t^+, \boldsymbol{u}_t^- \in \mathbb{R}_+^K, \\ & s_{t,i} = \xi_{t,i} s_{t-1,i} + u_{t,i}^+ - u_{t,i}^- \qquad \forall i \in [K], \\ & s_{t,K+1} = r_f s_{t-1,K+1} - (1+f_b)\mathbf{e}^\top \boldsymbol{u}_t^+ + (1-f_s)\mathbf{e}^\top \boldsymbol{u}_t^-. \end{aligned} \tag{5.1}$$

Here, $s_{0,i} = 0$ for every $i \in [K]$, $s_{0,K+1} = 1$, and the cost-to-go function at the terminal stage $T$ is

$$\begin{aligned} Q_T(\boldsymbol{s}_{T-1}, \boldsymbol{\xi}_T) = \max\ & U(\boldsymbol{\xi}_T^\top \boldsymbol{x}_T) \\ \text{s.t.}\ & \boldsymbol{s}_T \in \mathbb{R}_+^{K+1}, \\ & s_{T,i} = \xi_{T,i} s_{T-1,i} \qquad \forall i \in [K], \\ & s_{T,K+1} = r_f s_{T-1,K+1}. \end{aligned} \tag{5.2}$$

Here, the function $U : \mathbb{R}_{++} \to \mathbb{R}_+$ is a linear approximation to the log utility function. Note that no immediate cost is imposed at stage $t \in [T-1]$ since our goal is to maximize the utility of cumulative wealth at the terminal stage $T$.

We derive a single-level DRO reformulation for the true problem based on (4.6) and compare our data-driven SDDP scheme (DD-SDDP) in Algorithm 1 and the distributionally robust version (DDR-SDDP) in Algorithm 2 against the stagewise independent scheme (Independent), the equally weighted portfolio (Equal), and the hidden Markov model based SDDP scheme (HMM-SDDP) proposed by Silva et al. [61]. The stagewise independent scheme is an SDDP scheme that ignores correlation over time in the random

return process. Thus, we replace the NW estimates for the conditional distribution with discrete uniform distribution. The equally weighted portfolio is to allocate an equal amount of the current wealth in all assets at each stage, hence known as the $1/n$ portfolio strategy. DeMiguel et al. [19] show that this seemingly naïve strategy is optimal under certain conditions. The hidden Markov model based SDDP scheme assumes three unobservable states governing random returns $\boldsymbol{\xi}_t$, namely *bull*, *regular*, and *bear* states and apply DRO to account for estimation errors of the transition probabilities among those states. Similar to our schemes, the resulting MSLP problem is then solved by SDDP.

We test the schemes with the historical weekly returns of the following data sets from December 2003 to January 2023: the 10 Industry Portfolios and 12 Industry Portfolios from the Fama-French online data library[3], which include US stock portfolios categorized by industries; and the iShares Exchange-Trades Funds (iShares) data set downloaded from *yfinance*[9], which includes the following nine funds: EWG (Germany), EWH (Hong Kong), EWI (Italy), EWK (Belgium), EWL (Switzerland), EWN (Netherlands), EWP (Spain), EWQ (France), and EWU (United Kingdom).

We obtain 120 historical trajectories of the weekly closing prices from each data set by setting the time horizon to $T = 8$ weeks. The first 50 trajectories are used as training data for our models, while the remaining 70 trajectories are used for the out-of-sample tests. In our experiment, we utilize a five-fold cross-validation procedure to determine the radius of the ambiguity set $\lambda$ for DDR-SDDP and HMM-SDDP. In each trial, we split the training data into five equal-sized subsets where four of the five subsets are put together to train the model. The resulting policy is then tested on the remaining set for $\lambda$ in $[10^{-3}, 10^{1}]/N^{2/(p+4)}$ on a logarithm search grid with 10 equidistant points. This process is repeated five times for different partitions of the data to choose $\lambda$ that performs best overall.

First, we show the convergence of our data-driven schemes. Figure 2 depicts the evolution of the upper and lower bounds using the 10 Industry Portfolios training data set. We observe that both DD-SDDP and DDR-SDDP are able to close the optimality gap to less than 3% after 800 iterations, and the gap for DDR-SDDP is about three times smaller than one for DD-SDDP.

Table 2 reports the out-of-sample performance for five different schemes. We fixed the number of iterations $k = 200$ for every SDDP scheme. It took about 30 minutes to solve an MSLP problem instance in this setup. Along with other statistics, we include the Sharpe ratio to measure the performance of a portfolio relative to a risk-free asset, after adjusting for its risk. Here, the Sharpe ratio is defined as

$$\text{Sharpe ratio} = \frac{\text{mean}\{R_i\}_{i=1}^{70} - R_f}{\text{std}\{R_i\}_{i=1}^{70}},$$

where $R_i$ is the return of the portfolio along test trajectory $i$ over 8 weeks and $R_f$ is the return of the risk-free asset over 8 weeks. The results indicate that the DDR-SDDP scheme performs favorably relative to its competitors: it achieves the largest utility, the largest mean return, and the largest Sharpe ratio over all data sets. In addition, compared to DD-SDDP, DDR-SDDP provides a less risky policy in terms of standard deviation for every data set, illustrating the connection with the variance regularization scheme discussed in Section 4. Meanwhile, we observe that DD-SDDP also performs significantly better than the stagewise independent scheme, showing the benefit of incorporating stagewise dependence into the model.

**Table 2** Out-of-sample statistics of different schemes

| Data set | Model | Utility | Mean return | Std. dev. | Sharpe ratio |
|---|---|---|---|---|---|
| 10 Industry Fama-French | DDR-SDDP | **0.03190** | **0.03411** | 0.06348 | **0.47656** |
| | DD-SDDP | 0.02654 | 0.02904 | 0.06957 | 0.36176 |
| | Independent | 0.01582 | 0.01762 | 0.06228 | 0.22088 |
| | HMM-SDDP | 0.02428 | 0.02540 | 0.04697 | 0.45878 |
| | Equal | -0.00215 | -0.00197 | **0.02743** | -0.03518 |
| 12 Industry Fama-French | DDR-SDDP | **0.02486** | **0.02669** | 0.05974 | **0.38238** |
| | DD-SDDP | 0.02348 | 0.02591 | 0.06899 | 0.31962 |
| | Independent | 0.02014 | 0.02203 | 0.06316 | 0.28780 |
| | HMM-SDDP | 0.01968 | 0.02086 | 0.04970 | 0.34211 |
| | Equal | -0.00152 | -0.00136 | **0.02715** | -0.03166 |
| iShares | DDR-SDDP | **0.01380** | **0.01425** | 0.03605 | **0.28860** |
| | DD-SDDP | 0.00734 | 0.00945 | 0.06855 | 0.08160 |
| | Independent | 0.00343 | 0.00534 | 0.06511 | 0.02283 |
| | HMM-SDDP | 0.00568 | 0.00609 | 0.03433 | 0.06526 |
| | Equal | -0.05000 | -0.04865 | **0.02729** | -0.31754 |

[3] `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

[9] `https://pypi.org/project/yfinance/`

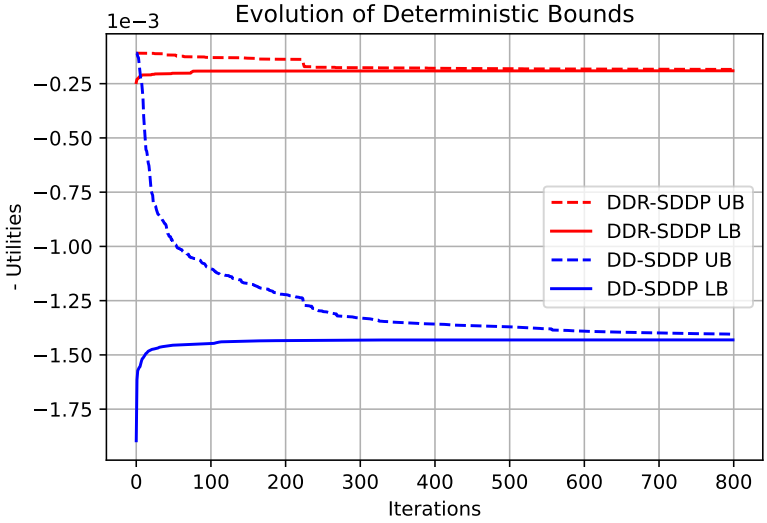


**Fig. 2** Evolution of both upper and lower bounds for DDR-SDDP and DD-SDDP over 800 iterations using the 10 Industry Portfolios data set.

### 5.2 Day-Ahead Wind Energy Commitment

In this experiment, we explore the wind energy commitment problem on the day-ahead market considered in [30] and [34]. At the beginning of day $t$, the day-ahead electricity prices $\boldsymbol{p}_t \in \mathbb{R}^{24}_+$ for the next 24 hours are known to a wind energy producer. The producer determines how much wind energy to commit for each of the next 24 hours without knowing the actual hourly amounts of wind energy $\boldsymbol{w}_{t+1} \in \mathbb{R}^{24}_+$. That is, the energy commitment levels $\boldsymbol{u}_t \in \mathbb{R}^{24}_+$ are determined after the observation of the day-ahead prices $\boldsymbol{p}_t$ but before the actual amounts of wind energy $\boldsymbol{w}_{t+1}$ are realized. On day $t+1$, the actual production can be used to satisfy the scheduled commitment $\boldsymbol{u}_t$ or charge three storage devices indexed by $l \in \{1,2,3\}$. It can also be dumped if all three storage devices are fully charged. The hourly commitment $u_{t,h}$ can be satisfied directly from the newly generated wind energy $w_{t+1,h}$ or by discharging the storage devices. Otherwise, there is a penalty of twice the respective day-ahead price $p_{t,h}$ for the unsatisfied commitment. Each of the devices has a different capacity $\bar{s}^l$, hourly leakage $\lambda^l$, charging efficiency $\lambda^l_c$ and discharging efficiency $\lambda^l_d$. We denote by $s^l_t \in \mathbb{R}^{24}_+$ the hourly levels of storage $l$ over the next 24 hours. The state variables $\boldsymbol{s}_t = (s^1_{t,24}, s^2_{t,24}, s^3_{t,24}) \in \mathbb{R}^3_+$ represent the storage levels at the end of day $t$ while the random parameters include the day-ahead prices $\boldsymbol{p}_t$ and the wind energy production levels $\boldsymbol{w}_t$ during the day, i.e., $\boldsymbol{\xi}_t = (\boldsymbol{p}_t, \boldsymbol{w}_t)$. The wind producer's objective is to maximize the expected profit over the time horizon of $T = 7$ days. The optimal bidding and storage strategy for the wind energy producer is obtained by solving the (true) dynamic program for $t \in [T]$

$$
\begin{aligned}
Q_t(\boldsymbol{s}_t, \boldsymbol{\xi}_t) = \max\ & \boldsymbol{p}_t^\top \boldsymbol{u}_t - 2\boldsymbol{p}_t^\top \mathbb{E}\left[\boldsymbol{e}^u_{t+1} \mid \boldsymbol{\xi}_t\right] + \mathbb{E}[Q_{t+1}(\boldsymbol{s}_{t+1}, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t] \\
\text{s.t. } & \boldsymbol{u}_t, \boldsymbol{e}^{\{s,u,d\}}_{t+1} \in \mathbb{R}^{24}_+, \quad \boldsymbol{e}^{\{+,-\},l}_{t+1}, \boldsymbol{s}^l_{t+1} \in \mathbb{R}^{24}_+ && \forall l \in [3], \\
& w_{t+1,h} = e^s_{t+1,h} + e^{+,1}_{t+1,h} + e^{+,2}_{t+1,h} + e^{+,3}_{t+1,h} + e^d_{t+1,h} && \forall h \in [24], \\
& u_{t,h} = e^s_{t+1,h} + e^{-,1}_{t+1,h} + e^{-,2}_{t+1,h} + e^{-,3}_{t+1,h} + e^u_{t+1,h} && \forall h \in [24], \\
& s^l_{t+1,h} = \lambda^l s^l_{t+1,h-1} + \lambda^l_c e^{+,l}_{t+1,h} - \frac{1}{\lambda^l_d} e^{-,l}_{t+1,h} && \forall h \in [24]\ \forall l \in [3], \\
& s^l_{t+1,h} \le \bar{s}^l && \forall h \in [24]\ \forall l \in [3],
\end{aligned}
\tag{5.3}
$$

where $Q_{T+1}(\cdot) \equiv 0$. Here, $\boldsymbol{e}^s_{t+1}$ and $\boldsymbol{e}^u_{t+1}$ represent the amounts of satisfied and unsatisfied energy commitments, respectively, while $\boldsymbol{e}^d_{t+1}$ represents the amounts of dumped wind energy. In addition, $\boldsymbol{e}^{+,l}_{t+1}$ represents the amounts of wind energy used to charge storage $l$ and $\boldsymbol{e}^{-,l}_{t+1}$ the amounts of energy discharged from storage $l$ to meet the commitments.

We obtain the hourly day-ahead prices in the PJM market and the hourly wind energy from 2002 to 2011 at the following locations: **Ohio** (41.8125N, 81.5625W) and **North Carolina** (33.9375N, 77.9375W). By setting $T = 7$ days, we obtain 520 historical trajectories for each location. As the wind energy and the day-ahead prices show clear seasonality patterns, we divide the trajectories into four parts according to different seasons and evaluate the out-of-sample performance for each of them separately. We perform principal component analysis on the high dimensional data $\boldsymbol{\xi}_t = (\boldsymbol{p}_t, \boldsymbol{w}_t) \in \mathbb{R}^{48}_+$ to obtain a 6-dimensional

subspace that explains more than 90% of the variability of the historical observations, which mitigates large out-of-sample errors as discussed in Section 2.3.

Table 3 presents the out-of-sample performance for DDR-SDDP, DD-SDDP, and the stagewise independent scheme (Independent). Similar to the previous example, our data-driven schemes outperform the stagewise independent scheme in all criteria. We observe that DDR-SDDP wins in all the categories. Particularly, it is more robust in terms of the 10th percentile compared to other schemes while the stagewise independent scheme has a significant risk of incurring a loss (negative profits).

**Table 3** Out-of-sample statistics of profit (in $100,000)

| Data Set | Model | Mean | Variance | 10th pct. |
|---|---|---|---|---|
| Ohio | DDR-SDDP | **7.12** | **18.61** | **2.72** |
| | DD-SDDP | 6.71 | 22.57 | 2.07 |
| | Independent | 5.61 | 21.58 | 0.78 |
| North Carolina | DDR-SDDP | **8.62** | **42.82** | **1.47** |
| | DD-SDDP | 8.27 | 53.47 | 1.07 |
| | Independent | 7.55 | 57.12 | -0.30 |

## 6 Conclusion

In this paper, we introduced a data-driven SDDP (DD-SDDP) scheme for solving a risk-neutral multistage stochastic linear programming (MSLP) problem under an unknown underlying Markov process for random parameters. We utilized the NW kernel regression to estimate the true conditional distribution and established for the first time the theoretical out-of-sample guarantee for the data-driven MSLP problem. This theoretical result inspired us to develop a data-driven distributionally robust SDDP (DDR-SDDP) scheme as a tractable regularization scheme. The numerical experiments demonstrated that our robust scheme outperformed all other benchmarks in real-world applications. In the future, it would be interesting to extend our scheme to solve general convex and nonconvex problems as well as problems with discrete decisions.

**Funding** This research was supported by the National Science Foundation grants no. 1752125 and 2153606.

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

1. Ahmed, S., Cabral, F.G., Freitas Paulo da Costa, B.: Stochastic lipschitz dynamic programming. Mathematical Programming **191**(2), 755–793 (2022)
2. Ahmed, S., King, A.J., Parija, G.: A multi-stage stochastic integer programming approach for capacity expansion under uncertainty. Journal of Global Optimization **26**(1), 3–24 (2003)
3. Ahmed, S., Sahinidis, N.V.: An approximation scheme for stochastic integer programs arising in capacity expansion. Operations Research **51**(3), 461–471 (2003)
4. Barth, R., Brand, H., Meibom, P., Weber, C.: A stochastic unit-commitment model for the evaluation of the impacts of integration of large amounts of intermittent wind power. In: International Conference on Probabilistic Methods Applied to Power Systems, pp. 1–8. IEEE (2006)
5. Bayraksan, G., Love, D.K.: Data-driven stochastic programming using phi-divergences. In: The Operations Research Revolution, pp. 1–19. INFORMS (2015)
6. Bellman, R.: Dynamic Programming. Princeton University Press (1957)
7. Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., Rennen, G.: Robust solutions of optimization problems affected by uncertain probabilities. Management Science **59**(2), 341–357 (2013)
8. Bertsimas, D., Gupta, V., Kallus, N.: Data-driven robust optimization. Mathematical Programming **167**(2), 235–292 (2018)
9. Birge, J.R.: Decomposition and partitioning methods for multistage stochastic linear programs. Operations Research **33**(5), 989–1007 (1985)
10. Bonnans, J.F., Cen, Z., Christel, T.: Energy contracts management by stochastic programming techniques. Annals of Operations Research **200**(1), 199–222 (2012)
11. Bradley, S.P., Crane, D.B.: A dynamic model for bond portfolio management. Management Science **19**(2), 139–151 (1972)
12. Carino, D.R., Kent, T., Myers, D.H., Stacy, C., Sylvanus, M., Turner, A.L., Watanabe, K., Ziemba, W.T.: The russell-yasuda kasai model: An asset/liability model for a Japanese insurance company using multistage stochastic programming. Interfaces **24**(1), 29–49 (1994)
13. Castro, M.P., Bodur, M., Song, Y.: Markov chain-based policies for multi-stage stochastic integer linear programming with an application to disaster relief logistics. arXiv preprint arXiv:2207.14779 (2022)

14. Cerisola, S., Baíllo, Á., Fernández-López, J.M., Ramos, A., Gollmer, R.: Stochastic power generation unit commitment in electricity markets: A novel formulation and a comparison of solution methods. Operations Research **57**(1), 32–46 (2009)
15. Chen, Z.L., Powell, W.B.: Convergent cutting-plane and partial-sampling algorithm for multistage stochastic linear programs with recourse. Journal of Optimization Theory and Applications **102**(3), 497–524 (1999)
16. Dantzig, G.B., Infanger, G.: Multi-stage stochastic linear programs for portfolio optimization. Annals of Operations Research **45**(1), 59–76 (1993)
17. De Matos, V.L., Philpott, A.B., Finardi, E.C.: Improving the performance of stochastic dual dynamic programming. Journal of Computational and Applied Mathematics **290**, 196–208 (2015)
18. Delage, E., Ye, Y.: Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations Research **58**(3), 595–612 (2010)
19. DeMiguel, V., Garlappi, L., Uppal, R.: Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? The review of Financial studies **22**(5), 1915–1953 (2009)
20. Downward, A., Dowson, O., Baucke, R.: Stochastic dual dynamic programming with stagewise-dependent objective uncertainty. Operations Research Letters **48**(1), 33–39 (2020)
21. Dowson, O.: Applying stochastic optimisation to the new zealand dairy industry. Ph.D. thesis, University of Auckland (2018)
22. Duque, D., Morton, D.P.: Distributionally robust stochastic dual dynamic programming. SIAM Journal on Optimization **30**(4), 2841–2865 (2020)
23. Escudero, L.F., Kamesam, P.V., King, A.J., Wets, R.J.: Production planning via scenario modelling. Annals of Operations Research **43**(6), 309–335 (1993)
24. Georghiou, A., Tsoukalas, A., Wiesemann, W.: Robust dual dynamic programming. Operations Research **67**(3), 813–830 (2019)
25. Girardeau, P., Leclere, V., Philpott, A.B.: On the convergence of decomposition methods for multistage stochastic convex programs. Mathematics of Operations Research **40**(1), 130–145 (2015)
26. Gjelsvik, A., Belsnes, M.M., Haugstad, A.: An algorithm for stochastic medium-term hydrothermal scheduling under spot price uncertainty. In: Proceedings of 13th Power Systems Computation Conference (1999)
27. Golub, B., Holmer, M., McKendall, R., Pohlman, L., Zenios, S.A.: A stochastic programming model for money management. European Journal of Operational Research **85**(2), 282–296 (1995)
28. Györfi, L., Kohler, M., Krzyzak, A., Walk, H., et al.: A Distribution-Free Theory of Nonparametric Regression, vol. 1. Springer (2002)
29. Hanasusanto, G.A., Kuhn, D.: Robust data-driven dynamic programming. In Advances in Neural Information Processing Systems **26** (2013)
30. Hannah, L., Dunson, D.B.: Approximate dynamic programming for storage problems. In: International Conference on Machine Learning (2011)
31. Huang, J., Zhou, K., Guan, Y.: A study of distributionally robust multistage stochastic optimization. arXiv preprint arXiv:1708.07930 (2017)
32. Infanger, G., Morton, D.P.: Cut sharing for multistage stochastic linear programs with interstage dependency. Mathematical Programming **75**(2), 241–256 (1996)
33. Kelley Jr, J.E.: The cutting-plane method for solving convex programs. Journal of the society for Industrial and Applied Mathematics **8**(4), 703–712 (1960)
34. Kim, J.H., Powell, W.B.: Optimal energy commitments with storage and intermittent supply. Operations Research **59**(6), 1347–1360 (2011)
35. Klabjan, D., Simchi-Levi, D., Song, M.: Robust stochastic lot-sizing by means of histograms. Production and Operations Management **22**(3), 691–710 (2013)
36. Kuhn, D., Wiesemann, W., Georghiou, A.: Primal and dual linear decision rules in stochastic and robust optimization. Mathematical Programming **130**(1), 177–209 (2011)
37. Lan, G.: Complexity of stochastic dual dynamic programming. Mathematical Programming **191**(2), 717–754 (2022)
38. Löhndorf, N., Shapiro, A.: Modeling time-dependent randomness in stochastic dual dynamic programming. European Journal of Operational Research **273**(2), 650–661 (2019)
39. Löhndorf, N., Wozabal, D., Minner, S.: Optimizing trading decisions for hydro storage systems using approximate dual dynamic programming. Operations Research **61**(4), 810–823 (2013)
40. Love, D., Bayraksan, G.: Phi-divergence constrained ambiguous stochastic programs for data-driven optimization. Technical report, Department of Integrated Systems Engineering, The Ohio State University, Columbus, Ohio (2015)
41. Mehrotra, S., Zhang, H.: Models and algorithms for distributionally robust least squares problems. Mathematical Programming **146**(1), 123–141 (2014)
42. Mulvey, J.M., Vladimirou, H.: Stochastic network programming for financial planning problems. Management Science **38**(11), 1642–1664 (1992)
43. Nadaraya, E.A.: On estimating regression. Theory of Probability & Its Applications **9**(1), 141–142 (1964)
44. Namkoong, H., Duchi, J.C.: Variance-based regularization with convex objectives. Advances in neural information processing systems **30** (2017)
45. Pereira, M.V., Pinto, L.M.: Stochastic optimization of a multireservoir hydroelectric system: A decomposition approach. Water Resources Research **21**(6), 779–792 (1985)
46. Pereira, M.V., Pinto, L.M.: Multi-stage stochastic optimization applied to energy planning. Mathematical Programming **52**(1), 359–375 (1991)
47. Philpott, A., de Matos, V., Finardi, E.: On solving multistage stochastic programs with coherent risk measures. Operations Research **61**(4), 957–970 (2013)
48. Philpott, A.B., De Matos, V.L.: Dynamic sampling algorithms for multi-stage stochastic programs with risk aversion. European Journal of Operational Research **218**(2), 470–483 (2012)
49. Philpott, A.B., De Matos, V.L., Kapelevich, L.: Distributionally robust SDDP. Computational Management Science **15**(3), 431–454 (2018)
50. Philpott, A.B., Guan, Z.: On the convergence of stochastic dual dynamic programming and related methods. Operations Research Letters **36**(4), 450–455 (2008)
51. Scarf, H.: A min-max solution of an inventory problem. Studies in the Mathematical Theory of Inventory and Production (1958)

52. Sen, S., Yu, L., Genc, T.: A stochastic programming approach to power portfolio optimization. Operations Research **54**(1), 55–72 (2006)
53. Shapiro, A.: Inference of statistical bounds for multistage stochastic programming problems. Mathematical Methods of Operations Research **58**(1), 57–68 (2003)
54. Shapiro, A.: On complexity of multistage stochastic programs. Operations Research Letters **34**(1), 1–8 (2006)
55. Shapiro, A.: Analysis of stochastic dual dynamic programming method. European Journal of Operational Research **209**(1), 63–72 (2011)
56. Shapiro, A.: Tutorial on risk neutral, distributionally robust and risk averse multistage stochastic programming. European Journal of Operational Research **288**(1), 1–13 (2021)
57. Shapiro, A., Ahmed, S.: On a class of minimax stochastic programs. SIAM Journal on Optimization **14**(4), 1237–1249 (2004)
58. Shapiro, A., Dentcheva, D., Ruszczynski, A.: Lectures on Stochastic Programming: Modeling and Theory. SIAM (2021)
59. Shapiro, A., Nemirovski, A.: On complexity of stochastic programming problems. In: V. Jeyakumar, A. Rubinov (eds.) Continuous Optimization: Current Trends and Modern Applications, pp. 111–146. Springer (2005)
60. Shapiro, A., Tekaya, W., da Costa, J.P., Soares, M.P.: Risk neutral and risk averse stochastic dual dynamic programming method. European Journal of Operational Research **224**(2), 375–391 (2013)
61. Silva, T., Valladão, D., Homem-de Mello, T.: A data-driven approach for a class of stochastic dynamic optimization problems. Computational Optimization and Applications **80**(3), 687–729 (2021)
62. Singh, K.J., Philpott, A.B., Wood, R.K.: Dantzig-wolfe decomposition for solving multistage stochastic capacity-planning problems. Operations Research **57**(5), 1271–1286 (2009)
63. Srivastava, P.R., Wang, Y., Hanasusanto, G.A., Ho, C.P.: On data-driven prescriptive analytics with side information: A regularized Nadaraya-Watson approach. arXiv preprint arXiv:2110.04855 (2021)
64. Watson, G.S.: Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A pp. 359–372 (1964)
65. Wiesemann, W., Kuhn, D., Sim, M.: Distributionally robust convex optimization. Operations Research **62**(6), 1358–1376 (2014)
66. Zhang, S., Sun, X.A.: On distributionally robust multistage convex optimization: new algorithms and complexity analysis. arXiv preprint arXiv:2010.06759 (2020)
67. Zhang, S., Sun, X.A.: On distributionally robust multistage convex optimization: Data-driven models and performance. arXiv preprint arXiv:2210.08433 (2022)
68. Zhang, S., Sun, X.A.: Stochastic dual dynamic programming for multistage stochastic mixed-integer nonlinear optimization. Mathematical Programming pp. 1–51 (2022)
69. Zou, J., Ahmed, S., Sun, X.A.: Multistage stochastic unit commitment using stochastic dual dynamic integer programming. IEEE transactions on Power Systems **34**(3), 1814–1823 (2018)
70. Zou, J., Ahmed, S., Sun, X.A.: Stochastic dual dynamic integer programming. Mathematical Programming **175**(1), 461–502 (2019)
71. Zymler, S., Kuhn, D., Rustem, B.: Distributionally robust joint chance constraints with second-order moment information. Mathematical Programming **137**(1), 167–198 (2013)

## A Proofs in Section 2

### A.1 Proof of Lemma 1

The proof of Lemma 1 relies on the Hoffman's Lemma shown below.

**Lemma 6 (Hoffman Lemma (Theorem 9.14 in [58]))** *Let $\mathcal{M}(\boldsymbol{u}) = \{\boldsymbol{x} \in \mathbb{R}^n_+ \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{u}\}$ be a nonempty polyhedron parameterized by the right-hand side vector $\boldsymbol{u} \in \mathbb{R}^m$. Consider $\boldsymbol{u}' \in \mathbb{R}^m$ such that $\mathcal{M}(\boldsymbol{u}') \neq \emptyset$. Then, there exists a positive constant $r = \max_{\boldsymbol{\lambda}_1 \in S_0} \|\boldsymbol{\lambda}_1\|_1$, such that for any $\boldsymbol{x} \in \mathcal{M}(\boldsymbol{u})$,*

$$dist\ (\boldsymbol{x}, \mathcal{M}(\boldsymbol{u}')) \leq r\|\boldsymbol{u} - \boldsymbol{u}'\|.$$

*Here, $S_0$ is a bounded polyhedral set satisfying $S = S_0 + C$, where $S = \{(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \ : \ \|\boldsymbol{A}^\top \boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2\|_1 \leq 1\}$ is a polyhedron set that depends only on $\boldsymbol{A}$, and $C$ is a polyhedral cone.*

*Proof (Proof of Lemma 1)* We proceed by backward induction from the terminal stage. At stage $T$, we have $Q_T(\boldsymbol{x}_{T-1}, \boldsymbol{\xi}_T) = \hat{Q}_T(\boldsymbol{x}_{T-1}, \boldsymbol{\xi}_T) = \min\left\{\boldsymbol{c}_T^\top \boldsymbol{x}_T \ : \ \boldsymbol{A}_T\boldsymbol{x}_T + \boldsymbol{B}_T\boldsymbol{x}_{T-1} = \boldsymbol{b}_T,\ \boldsymbol{x}_T \in \mathbb{R}^{d_T}_+\right\}$. Hence, it is sufficient to show the result for $Q_T(\boldsymbol{x}_{T-1}, \boldsymbol{\xi}_T)$. For any $\boldsymbol{\xi}_T \in \Xi_T$, let $\boldsymbol{x}_{T-1}$ and $\boldsymbol{x}'_{T-1}$ be two points in the domain of $Q_T(\boldsymbol{x}_{T-1}, \boldsymbol{\xi}_T)$. Fix any feasible point $\boldsymbol{x}_T \in \mathcal{X}_T(\boldsymbol{x}_{T-1}, \boldsymbol{\xi}_T)$. Applying Lemma 6 with right-hand side vectors $\boldsymbol{u}_T = \boldsymbol{b}_T - \boldsymbol{B}_T\boldsymbol{x}_{T-1}$ and $\boldsymbol{u}'_T = \boldsymbol{b}_T - \boldsymbol{B}_T\boldsymbol{x}'_{T-1}$, we have that there exists a feasible point $\boldsymbol{x}'_T \in \mathcal{X}_T(\boldsymbol{x}'_{T-1}, \boldsymbol{\xi}_T)$ such that

$$\|\boldsymbol{x}_T - \boldsymbol{x}'_T\| \leq r_T\|\boldsymbol{u}_T - \boldsymbol{u}'_T\| \leq r_T\|(\boldsymbol{b}_T - \boldsymbol{B}_T\boldsymbol{x}_{T-1}) - (\boldsymbol{b}_T - \boldsymbol{B}_T\boldsymbol{x}'_{T-1})\| \leq r_T\|\boldsymbol{B}_T\| \cdot \|\boldsymbol{x}_{T-1} - \boldsymbol{x}'_{T-1}\|,$$

where $r_T > 0$ is a constant that depends only on $\boldsymbol{A}_T$. Therefore, the Lipschitz continuity of $\boldsymbol{c}_T^\top \boldsymbol{x}_T$ implies

$$Q_T(\boldsymbol{x}'_{T-1}, \boldsymbol{\xi}_T) \leq \boldsymbol{c}_T^\top \boldsymbol{x}'_T \leq \boldsymbol{c}_T^\top \boldsymbol{x}_T + r_T\|\boldsymbol{c}_T\| \cdot \|\boldsymbol{B}_T\| \cdot \|\boldsymbol{x}_{T-1} - \boldsymbol{x}'_{T-1}\|.$$

Taking minimization over $\boldsymbol{x}_T \in \mathcal{X}_T(\boldsymbol{x}_{T-1}, \boldsymbol{\xi}_T)$, we have

$$Q_T(\boldsymbol{x}'_{T-1}, \boldsymbol{\xi}_T) \leq Q_T(\boldsymbol{x}_{T-1}, \boldsymbol{\xi}_T) + r_T\|\boldsymbol{c}_T\| \cdot \|\boldsymbol{B}_T\| \cdot \|\boldsymbol{x}_{T-1} - \boldsymbol{x}'_{T-1}\|.$$

By symmetry, we have

$$\left|Q_T(\boldsymbol{x}_{T-1}, \boldsymbol{\xi}_T) - Q_T(\boldsymbol{x}'_{T-1}, \boldsymbol{\xi}_T)\right| \leq r_T\|\boldsymbol{c}_T\| \cdot \|\boldsymbol{B}_T\| \cdot \|\boldsymbol{x}_{T-1} - \boldsymbol{x}'_{T-1}\|.$$

By the compactness of the uncertainty set in the assumption (A4), we know there exists $\bar{r}_T, \bar{\boldsymbol{c}}_T$, and $\bar{\boldsymbol{B}}_T$ such that the right-hand side is upper bounded by $\bar{r}_T\|\bar{\boldsymbol{c}}_T\| \cdot \|\bar{\boldsymbol{B}}_T\| \cdot \|\boldsymbol{x}_{T-1} - \boldsymbol{x}'_{T-1}\|$. Setting $L_T = \bar{r}_T\|\bar{\boldsymbol{c}}_T\| \cdot \|\bar{\boldsymbol{B}}_T\|$, we conclude that

$$\left|Q_T(\boldsymbol{x}_{T-1}, \boldsymbol{\xi}_T) - Q_T(\boldsymbol{x}'_{T-1}, \boldsymbol{\xi}_T)\right| \leq L_T\|\boldsymbol{x}_{T-1} - \boldsymbol{x}'_{T-1}\|, \quad \forall \boldsymbol{x}_{T-1},\ \boldsymbol{x}'_{T-1} \in \mathcal{X}_{T-1}\left(\boldsymbol{x}_{T-2}, \boldsymbol{\xi}_{T-1}\right),\ \forall \boldsymbol{\xi}_T \in \Xi_T.$$

Hence the desired result holds for stage $T$. By induction, suppose the statement in the lemma holds for stage $t+1 \leq T$. That is, there exists a constant $L_{t+1} > 0$ such that

$$\begin{aligned}
&\left|Q_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}) - Q_{t+1}(\boldsymbol{x}'_t, \boldsymbol{\xi}_{t+1})\right| \leq L_{t+1}\|\boldsymbol{x}_t - \boldsymbol{x}'_t\|, \quad \forall \boldsymbol{x}_t,\ \boldsymbol{x}'_t \in \mathcal{X}_t\left(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t\right)\ \forall \boldsymbol{\xi}_{t+1} \in \Xi_{t+1}, \text{ and} \\
&\left|\hat{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}) - \hat{Q}_{t+1}(\boldsymbol{x}'_t, \boldsymbol{\xi}_{t+1})\right| \leq L_{t+1}\|\boldsymbol{x}_t - \boldsymbol{x}'_t\|, \quad \forall \boldsymbol{x}_t,\ \boldsymbol{x}'_t \in \mathcal{X}_t\left(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t\right)\ \forall \boldsymbol{\xi}_{t+1} \in \Xi_{t+1}.
\end{aligned}$$

We show the Lipschitz continuity of $Q_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$, the Lipschitz continuity of $\hat{Q}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ can be proved in a similar way. For any feasible points $\boldsymbol{x}_t,\ \boldsymbol{x}'_t \in \mathcal{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ and any $\boldsymbol{\xi}_t \in \Xi_t$, the following chain of inequalities holds for the true expected cost-to-go function.

$$\begin{aligned}
\left|\mathcal{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) - \mathcal{Q}_{t+1}(\boldsymbol{x}'_t, \boldsymbol{\xi}_t)\right| &= \left| \int_{\boldsymbol{\xi}_{t+1} \in \Xi_{t+1}} Q_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}) f\left(\boldsymbol{\xi}_{t+1} \mid \boldsymbol{\xi}_t\right) \mathrm{d}\boldsymbol{\xi}_{t+1} \right. \\
&\qquad \left. - \int_{\boldsymbol{\xi}_{t+1} \in \Xi_{t+1}} Q_{t+1}(\boldsymbol{x}'_t, \boldsymbol{\xi}_{t+1}) f\left(\boldsymbol{\xi}_{t+1} \mid \boldsymbol{\xi}_t\right) \mathrm{d}\boldsymbol{\xi}_{t+1} \right| \\
&\leq \int_{\boldsymbol{\xi}_{t+1} \in \Xi_{t+1}} \left|Q_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}) - Q_{t+1}(\boldsymbol{x}'_t, \boldsymbol{\xi}_{t+1})\right| f\left(\boldsymbol{\xi}_{t+1} \mid \boldsymbol{\xi}_t\right) \mathrm{d}\boldsymbol{\xi}_{t+1} \\
&\leq L_{t+1}\|\boldsymbol{x}_t - \boldsymbol{x}'_t\|.
\end{aligned}$$

Therefore, $\boldsymbol{c}_t^\top \boldsymbol{x}_t + \mathcal{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ is a Lipschitz function with constant $L_{t+1} + \|\boldsymbol{c}_t\|$. For any $\boldsymbol{\xi}_t \in \Xi_t$, let $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}'_{t-1}$ be two points in the domain of $Q_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$. Fix any feasible point $\boldsymbol{x}_t \in \mathcal{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$. Applying Lemma 6 with right-hand side vectors $\boldsymbol{u}_t = \boldsymbol{b}_t - \boldsymbol{B}_t\boldsymbol{x}_{t-1}$ and $\boldsymbol{u}'_t = \boldsymbol{b}_t - \boldsymbol{B}_t\boldsymbol{x}'_{t-1}$, we have that there exists a feasible point $\boldsymbol{x}'_t \in \mathcal{X}_t(\boldsymbol{x}'_{t-1}, \boldsymbol{\xi}_t)$ such that

$$\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| \leq r_t\|\boldsymbol{u}_t - \boldsymbol{u}'_t\| \leq r_t\|(\boldsymbol{b}_t - \boldsymbol{B}_t\boldsymbol{x}_{t-1}) - (\boldsymbol{b}_t - \boldsymbol{B}_t\boldsymbol{x}'_{t-1})\| \leq r_t\|\boldsymbol{B}_t\| \cdot \|\boldsymbol{x}_{t-1} - \boldsymbol{x}'_{t-1}\|$$

where $r_t > 0$ is a constant that depends only on $\boldsymbol{A}_t$. Therefore, the Lipschitz continuity of $\boldsymbol{c}_t^\top \boldsymbol{x}_t + \mathcal{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ implies

$$\begin{aligned}
Q_t(\boldsymbol{x}'_{t-1}, \boldsymbol{\xi}_t) &\leq \boldsymbol{c}_t^\top \boldsymbol{x}'_t + \mathcal{Q}_{t+1}(\boldsymbol{x}'_t, \boldsymbol{\xi}_t) \\
&\leq \boldsymbol{c}_t^\top \boldsymbol{x}_t + \mathcal{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) + r_t(L_{t+1} + \|\boldsymbol{c}_t\|) \cdot \|\boldsymbol{B}_t\| \cdot \|\boldsymbol{x}_{t-1} - \boldsymbol{x}'_{t-1}\|.
\end{aligned}$$

Taking minimization over $\boldsymbol{x}_t \in \mathcal{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$, we have

$$Q_t(\boldsymbol{x}'_{t-1}, \boldsymbol{\xi}_t) \leq Q_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) + r_t(L_{t+1} + \|\boldsymbol{c}_t\|) \cdot \|\boldsymbol{B}_t\| \cdot \|\boldsymbol{x}_{t-1} - \boldsymbol{x}'_{t-1}\|.$$

By symmetry, we have

$$\left|Q_t(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)-Q_t(\boldsymbol{x}'_{t-1},\boldsymbol{\xi}_t)\right| \le r_t(L_{t+1}+\|\boldsymbol{c}_t\|)\cdot\|\boldsymbol{B}_t\|\cdot\|\boldsymbol{x}_{t-1}-\boldsymbol{x}'_{t-1}\|.$$

By the compactness of the uncertainty set in the assumption (A4), we know there exists $\bar{r}_t, \bar{\boldsymbol{c}}_t$, and $\bar{\boldsymbol{B}}_t$ such that the right-hand side is upper bounded by $\bar{r}_t(L_{t+1}+\|\bar{\boldsymbol{c}}_t\|)\cdot\|\bar{\boldsymbol{B}}_t\|\cdot\|\boldsymbol{x}_{t-1}-\boldsymbol{x}'_{t-1}\|$ . Setting $L_t=\bar{r}_t(L_{t+1}+\|\bar{\boldsymbol{c}}_t\|)\cdot\|\bar{\boldsymbol{B}}_t\|$, we conclude that

$$\left|Q_t(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)-Q_t(\boldsymbol{x}'_{t-1},\boldsymbol{\xi}_t)\right| \le L_t\|\boldsymbol{x}_{t-1}-\boldsymbol{x}'_{t-1}\|, \quad \forall \boldsymbol{x}_{t-1},\ \boldsymbol{x}'_{t-1}\in\mathcal{X}_{t-1}\left(\boldsymbol{x}_{t-2},\boldsymbol{\xi}_{t-1}\right),\ \forall \boldsymbol{\xi}_t\in\Xi_T.$$

Therefore, the result holds for $Q_t(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)$. We omit the proof for $\hat{Q}_t(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)$ since it is similar to $Q_t(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)$. Hence, the desired result holds for stage $t$ and this completes the induction. ∎

### A.2 Proof of Theorem 2

*Proof* Based on the assumptions (A3) and (A4), there exists a positive constant $D_t$ satisfying

$$\sup_{\boldsymbol{x}_t,\boldsymbol{x}'_t\in\mathcal{X}_t(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)}\|\boldsymbol{x}_t-\boldsymbol{x}'_t\|\le D_t, \quad \forall \boldsymbol{x}_{t-1},\ \forall \boldsymbol{\xi}_t,$$

i.e., $\mathcal{X}_t\left(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t\right)\subseteq\mathbb{R}^{d_t}$ has a finite diameter less than or equal to $D_t$ for any $\boldsymbol{x}_{t-1}$ and $\boldsymbol{\xi}_t$.

Next, we define a finite set of points $\mathcal{X}_t^\eta(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)\subseteq\mathcal{X}_t(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)$. For any $\boldsymbol{x}_t\in\mathcal{X}_t(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)$, there exists $\boldsymbol{x}'_t\in\mathcal{X}_t^\eta(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)$ such that $\|\boldsymbol{x}_t-\boldsymbol{x}'_t\|\le\eta$, i.e., for a fixed tolerance level $\eta$, we have the cardinality $\left|\mathcal{X}_t^\eta\left(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t\right)\right| = O(1)(D_t/\eta)^{d_t}$. Since the cost-to-go function $Q_{t+1}(\boldsymbol{x}_t,\boldsymbol{\xi}_{t+1})$ is $L_{t+1}$-Lipschitz continuous in $\boldsymbol{x}_t$, Lemma 2 implies that for any $\boldsymbol{x}_t\in\mathcal{X}_t(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)$, there exists $\boldsymbol{x}'_t\in\mathcal{X}_t^\eta(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)$, $\|\boldsymbol{x}_t-\boldsymbol{x}'_t\|\le\eta$, such that:

$$\left|\mathbb{E}\left[Q_{t+1}(\boldsymbol{x}_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right]-\mathbb{E}\left[Q_{t+1}(\boldsymbol{x}'_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right]\right|\le L_{t+1}\eta. \tag{A.1}$$

Furthermore, from Theorem 1, we have that for a fixed $\boldsymbol{x}'_t\in\mathcal{X}_t^\eta(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)$,

$$\left|\mathbb{E}\left[Q_{t+1}(\boldsymbol{x}'_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right]-\hat{\mathbb{E}}\left[Q_{t+1}(\boldsymbol{x}'_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right]\right|\le\sqrt{\frac{\mathbb{V}\left[Q_{t+1}(\boldsymbol{x}'_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right]}{O(N^{\frac{4}{p+4}})(1+o(1))g(\boldsymbol{\xi}_t)}\log\left(\frac{1}{\delta_{t+1}}\right)} \tag{A.2}$$

with probability at least $1-\delta_{t+1}$. Applying union bound, we get

$$\begin{aligned}\left|\mathbb{E}\left[Q_{t+1}(\boldsymbol{x}'_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right]-\hat{\mathbb{E}}\left[Q_{t+1}(\boldsymbol{x}'_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right]\right| &\le\sqrt{\frac{\mathbb{V}\left[Q_{t+1}(\boldsymbol{x}'_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right]}{O(N^{\frac{4}{p+4}})(1+o(1))g(\boldsymbol{\xi}_t)}\log\left(\frac{\left|\mathcal{X}_t^\eta\left(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t\right)\right|}{\delta_{t+1}}\right)}\\ &\le\sqrt{\sigma_{t+1}^2\frac{\log\left(\frac{O(1)(D_t/\eta)^{d_t}}{\delta_{t+1}}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g(\boldsymbol{\xi}_t)}} \qquad \forall \boldsymbol{x}'_t\in\mathcal{X}_t^\eta\left(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t\right)\end{aligned} \tag{A.3}$$

with probability at least $1-\delta_{t+1}$. Using the Lipschitz continuity of $Q_{t+1}(\boldsymbol{x}_t,\boldsymbol{\xi}_{t+1})$, from Lemma 2, we get

$$\left|\mathbb{E}\left[Q_{t+1}(\boldsymbol{x}_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right]-\hat{\mathbb{E}}\left[Q_{t+1}(\boldsymbol{x}_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right]\right|\le\sqrt{\sigma_{t+1}^2\frac{\log\left(\frac{O(1)(D_t/\eta)^{d_t}}{\delta_{t+1}}\right)}{O(N^{\frac{4}{p+4}})(1+o(1))g(\boldsymbol{\xi}_t)}}+2L_{t+1}\eta \qquad \forall \boldsymbol{x}_t\in\mathcal{X}_t\left(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t\right)$$

with probability at least $1-\delta_{t+1}$. This completes the proof. ∎

## B Proofs in Section 3

### B.1 Proof of Proposition 1

*Proof* The proof of this proposition follows from the approach discussed in [44]. To simplify the notation, we define a random variable $\tilde{z}=\hat{Q}_{t+1}(\boldsymbol{x}_t,\tilde{\boldsymbol{\xi}}_{t+1})$ and a vector $\boldsymbol{z}\in\mathbb{R}^N$ where $z_i=\hat{Q}_{t+1}(\boldsymbol{x}_t,\boldsymbol{\xi}_{t+1}^i)$. We denote

$$\begin{aligned}\overline{z}&=\hat{\mathbb{E}}\left[\hat{Q}_{t+1}(\boldsymbol{x}_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right]=\sum_{i\in[N]}\hat{w}_t^i(\boldsymbol{\xi}_t)\cdot\hat{Q}_{t+1}(\boldsymbol{x}_t,\boldsymbol{\xi}_{t+1}^i)\\ s&=\hat{\mathbb{V}}\left[\hat{Q}_{t+1}(\boldsymbol{x}_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right]=\sum_{i\in[N]}\hat{w}_t^i(\boldsymbol{\xi}_t)\cdot\left[\hat{Q}_{t+1}(\boldsymbol{x}_t,\boldsymbol{\xi}_{t+1}^i)-\overline{z}\right]^2.\end{aligned}$$

The DRO problem $\max_{\mathbb{P}_t \in \mathcal{P}_t^\lambda(\hat{\mathbb{P}}_t)} \mathbb{E}_{\mathbb{P}_t}[\hat{Q}_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \mid \boldsymbol{\xi}_t]$ can be equivalently written as

$$\max_{\boldsymbol{w}_t} \left\{ \boldsymbol{w}_t^\top \boldsymbol{z} \ : \ \sum_{i\in[N]} \left| \frac{w_t^i - \hat{w}_t^i(\boldsymbol{\xi}_t)}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} \right| \le \sqrt{N}\lambda, \ \max_{i\in[N]} \left| \frac{w_t^i - \hat{w}_t^i(\boldsymbol{\xi}_t)}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} \right| \le \lambda, \ \mathbf{e}^\top \boldsymbol{w}_t = 1, \ \boldsymbol{w}_t \in \mathbb{R}_+^N \right\},$$

which is lowered bounded by

$$\max_{\boldsymbol{w}_t} \left\{ \boldsymbol{w}_t^\top \boldsymbol{z} \ : \ \sum_{i\in[N]} \frac{[w_t^i - \hat{w}_t^i(\boldsymbol{\xi}_t)]^2}{\hat{w}_t^i(\boldsymbol{\xi}_t)} \le \lambda^2, \ \mathbf{e}^\top \boldsymbol{w}_t = 1, \ \boldsymbol{w}_t \in \mathbb{R}_+^N \right\}.$$

By change of variable $\boldsymbol{u}_t = \boldsymbol{w}_t - \hat{\boldsymbol{w}}_t(\boldsymbol{\xi}_t)$, the above problem is equivalent to

$$\max_{\boldsymbol{u}_t} \left\{ \overline{z} + \boldsymbol{u}_t^\top (\boldsymbol{z} - \overline{z} \cdot \mathbf{e}) \ : \ \|\boldsymbol{u}_t\|_W \le \lambda, \ \mathbf{e}^\top \boldsymbol{u}_t = 0, \ \boldsymbol{u}_t + \hat{\boldsymbol{w}}_t(\boldsymbol{\xi}_t) \ge \mathbf{0} \right\},$$

where $\|\boldsymbol{u}_t\|_W := \sqrt{\sum_{i\in[N]} \frac{1}{\hat{w}_t^i(\boldsymbol{\xi}_t)} (u_t^i)^2}$ is defined to be a weighted norm. We further define its dual norm $\|\boldsymbol{u}_t\|_{W^{-1}} := \sqrt{\sum_{i\in[N]} \hat{w}_t^i(\boldsymbol{\xi}_t) \cdot (u_t^i)^2}$, and the upper bound of the above optimization problem is

$$\overline{z} + \boldsymbol{u}_t^\top (\boldsymbol{z} - \overline{z} \cdot \mathbf{e}) \le \overline{z} + \|\boldsymbol{u}_t\|_W \ \|\boldsymbol{z} - \overline{z} \cdot \mathbf{e}\|_{W^{-1}} \le \overline{z} + \lambda \ \|\boldsymbol{z} - \overline{z} \cdot \mathbf{e}\|_{W^{-1}} = \overline{z} + \lambda\sqrt{s},$$

where the last equality holds because

$$\|\boldsymbol{z} - \overline{z} \cdot \mathbf{e}\|_{W^{-1}} = \sqrt{\sum_{n\in[N]} \hat{w}_t^i(\boldsymbol{\xi}_t) \cdot (z_i - \overline{z})^2} = \sqrt{\hat{\mathbb{V}}\left[\hat{Q}_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right]}.$$

The above upper bound can be achieved by selecting

$$u_t^i = \frac{\lambda \hat{w}_t^i(\boldsymbol{\xi}_t) \cdot (z_i - \overline{z})}{\sqrt{s}}.$$

The above choice of $\boldsymbol{u}_t$ satisfies the constraints $\|\boldsymbol{u}_t\|_W^2 \le \lambda$ and $\boldsymbol{u}_t^\top \mathbf{e} = 0$. Therefore, such $\boldsymbol{u}_t$ is feasible as long as

$$u_t^i = \frac{\lambda \hat{w}_t^i(\boldsymbol{\xi}_t) \cdot (z_i - \overline{z})}{\sqrt{s}} \ge -\hat{w}_t^i(\boldsymbol{\xi}_t) \iff \frac{\lambda(z_i - \overline{z})}{\sqrt{s}} \ge -1.$$

By Lipschitz continuity of the cost-to-go function and compactness of the feasible region according to the assumption **(A3)**, $\hat{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i)$ is bounded. Denote $\overline{U}_{t+1}$ to be the upper bound, that is, $\hat{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i) \le \overline{U}_{t+1}, \forall \boldsymbol{\xi}_{t+1}^i \in \hat{\Xi}_{t+1}$. Hence, $|z_i - \overline{z}| = |\hat{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i) - \sum_{n\in[N]} \hat{w}_t^i \cdot \hat{Q}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i)| \le \overline{U}_{t+1}$, then a sufficient condition of the above is

$$\frac{\lambda^2 \overline{U}_{t+1}^2}{s} \le 1 \iff s \ge \lambda^2 \overline{U}_{t+1}^2 \iff \lambda\sqrt{s} \ge \lambda^2 \overline{U}_{t+1}.$$

if $s - \lambda^2 \overline{U}_{t+1}^2 \ge 0$, $\boldsymbol{u}_t$ is a feasible solution. On the other hand, $\boldsymbol{u}_t = \mathbf{0}$ is another feasible solution for this problem. Thus

$$\begin{aligned} \overline{z} + \left( \lambda \sqrt{\hat{\mathbb{V}}\left[\hat{Q}_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right]} - \lambda^2 \overline{U}_{t+1} \right) &\le \max_{\boldsymbol{w}_t} \left\{ \boldsymbol{w}_t^\top \boldsymbol{z} \ : \ \sum_{i\in[N]} \frac{(w_t^i - \hat{w}_t^i(\boldsymbol{\xi}_t))^2}{\hat{w}_t^i(\boldsymbol{\xi}_t)} \le \lambda^2, \ \mathbf{e}^\top \boldsymbol{w}_t = 1, \ \boldsymbol{w}_t \in \mathbb{R}_+^N \right\} \\ &\le \max_{\mathbb{P}_t \in \mathcal{P}_t^\lambda(\hat{\mathbb{P}}_t)} \mathbb{E}_{\mathbb{P}_t}\left[\hat{Q}_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right]. \end{aligned}$$

Rearranging the terms, we complete the proof. ∎

## B.2 Proof of Proposition 2

*Proof* We consider the inner maximization problem, which given a feasible solution $\boldsymbol{x}_t \in \mathcal{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$, yields the distribution with the worst-case expected loss as given below

$$\max_{\mathbb{P}_t \in \mathcal{P}_t^\lambda(\hat{\mathbb{P}}_t)} \hat{\mathcal{Q}}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_t).$$

To simplify the notation, we define the vector $\boldsymbol{z} \in \mathbb{R}^i$ whose $n$-th component $z_i = \hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i)$ denotes the loss function evaluated for the $n$-th data point. We then have the following linear programming formulation for the worst-case expected

loss:

$$\begin{array}{lll} \max & \sum\limits_{n\in[N]} z_i \hat{w}_t^i(\boldsymbol{\xi}_t) & \\ \text{s.t.} & \boldsymbol{w}_t \in \mathbb{R}_+^N, \quad \boldsymbol{f} \in \mathbb{R}^N, & \\ & \mathbf{e}^\top \boldsymbol{w}_t = 1 & : (\gamma), \\ & \dfrac{1}{\sqrt{N}} \sum\limits_{n\in[N]} f_i \le \lambda & : (\beta), \\ & \dfrac{w_t^i - \hat{w}_t^i(\boldsymbol{\xi}_t)}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} \le f_i \qquad \forall i \in [N] & : (\boldsymbol{\mu}), \\ & \dfrac{-w_t^i + \hat{w}_t^i(\boldsymbol{\xi}_t)}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} \le f_i \qquad \forall i \in [N] & : (\boldsymbol{\zeta}), \\ & f_i \le \lambda \qquad \forall i \in [N] & : (\boldsymbol{\psi}). \end{array} \tag{B.1}$$

Strong linear programming duality holds because the ambiguity set (4.5) is always nonempty with the nominal distribution as the trivial solution. Therefore, the dual problem can be written as:

$$\begin{array}{lll} \min & \gamma + \lambda \left( \beta + \sum\limits_{n\in[N]} \psi_i \right) + \sum\limits_{n\in[N]} \sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}(\mu_i - \zeta_i) & \\ \text{s.t.} & \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}_+, \quad \boldsymbol{\mu}, \boldsymbol{\zeta}, \boldsymbol{\psi} \in \mathbb{R}_+^N, & \\ & z_i \le \gamma + \dfrac{\mu_i - \zeta_i}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} & \forall i \in [N], \\ & \mu_i + \zeta_i = \psi_i + \dfrac{\beta}{\sqrt{N}} & \forall i \in [N]. \end{array} \tag{B.2}$$

Combining with the outer minimization problem, we have the following desired result:

$$\begin{array}{lll} \min & \boldsymbol{c}_t^\top \boldsymbol{x}_t + \gamma + \lambda \left( \beta + \sum\limits_{n\in[N]} \psi_i \right) + \sum\limits_{n\in[N]} \sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}(\mu_i - \zeta_i) & \\ \text{s.t.} & \boldsymbol{x}_t \in \mathbb{R}_+^{d_t}, \quad \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}_+, \quad \boldsymbol{\mu}, \boldsymbol{\zeta}, \boldsymbol{\psi} \in \mathbb{R}_+^N, & \\ & \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{B}_t \boldsymbol{x}_{t-1} = \boldsymbol{b}_t, & \\ & \hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i) \le \gamma + \dfrac{\mu_i - \zeta_i}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} & \forall i \in [N], \\ & \mu_i + \zeta_i = \psi_i + \dfrac{\beta}{\sqrt{N}} & \forall i \in [N]. \end{array} \tag{B.3}$$

■

## B.3 Proof of Lemma 4

*Proof* We proceed by backward induction in the stages.
Stage $t = T$: For any $\boldsymbol{\xi}_T \in \hat{\Xi}_T$, using the assumption **(A3)** and strong duality, we have

$$\begin{aligned} \hat{Q}_T^{\mathcal{DRO}}(\boldsymbol{x}_{T-1}, \boldsymbol{\xi}_T) &= \min_{\boldsymbol{x}_T} \left\{ \boldsymbol{c}_T^\top \boldsymbol{x}_T \ : \ \boldsymbol{A}_T \boldsymbol{x}_T + \boldsymbol{B}_T \boldsymbol{x}_{T-1} = \boldsymbol{b}_T, \ \boldsymbol{x}_T \in \mathbb{R}_+^{d_T} \right\} \\ &= \max_{\boldsymbol{\pi}_T} \left\{ \boldsymbol{\pi}_T^\top (\boldsymbol{b}_T - \boldsymbol{B}_T \boldsymbol{x}_{T-1}) \ : \ \boldsymbol{A}_T^\top \boldsymbol{\pi}_T \le \boldsymbol{c}_T, \ \boldsymbol{\pi}_T \in \mathbb{R}^{d_T} \right\} \\ &= \max \left\{ \boldsymbol{\pi}_{T,j}^\top (\boldsymbol{b}_T - \boldsymbol{B}_T \boldsymbol{x}_{T-1}) \ : \ j \in [\mathcal{S}_T] \right\} \end{aligned}$$

where $\boldsymbol{\pi}_{T,j}$, $j \in [\mathcal{S}_T]$ are the extreme points of the dual problem. Note that the number of dual extreme points in $\mathcal{S}_T$ is finite. Therefore the claim holds for the stage $T$.
Suppose $\hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1})$ is piecewise linear convex with finite pieces in $\boldsymbol{x}_t$ for any $\boldsymbol{\xi}_{t+1} \in \hat{\Xi}_{t+1}$. For a specific $\boldsymbol{\xi}_{t+1}^i \in \hat{\Xi}_{t+1}$, $i \in [N]$, let $\mathcal{PL}_{t+1}^i$ denote the set of all finite pieces of $\hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i)$, i.e.,

$$\mathcal{PL}_{t+1}^i = \max \left\{ \boldsymbol{\mathscr{G}}_{t,j}^{i\top} \boldsymbol{x}_t + \mathscr{V}_{t,j}^i \ : \ j \in [\mathcal{S}_{t+1}^i] \right\}$$

where $\mathcal{S}_{t+1}^i < \infty$ is the total number of piecewise linear functions that describe $\hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i)$. Now for any $\boldsymbol{\xi}_t \in \hat{\boldsymbol{\Xi}}_t$, we can rewrite (4.6) as

$$\begin{aligned}
\hat{Q}_t^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \min\ & \boldsymbol{c}_t^\top \boldsymbol{x}_t + \gamma + \lambda \left( \beta + \sum_{n\in[N]} \psi_i \right) + \sum_{n\in[N]} \sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}(\mu_i - \zeta_i) \\
\text{s.t. } & \boldsymbol{x}_t \in \mathbb{R}_+^{d_t}, \quad \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}_+, \quad \boldsymbol{\mu}, \boldsymbol{\zeta}, \boldsymbol{\psi} \in \mathbb{R}_+^N, \\
& \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{B}_t \boldsymbol{x}_{t-1} = \boldsymbol{b}_t, && : (\boldsymbol{\pi}_t), \\
& {\boldsymbol{g}_{t,j}^i}^\top \boldsymbol{x}_t + \mathscr{V}_{t,j}^i \le \gamma + \frac{\mu_i - \zeta_i}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} \qquad \forall i \in [N],\ \forall j \in [\mathcal{S}_{t+1}^i] && : (y_{t,j}^i), \\
& \mu_i + \zeta_i = \psi_i + \frac{\beta}{\sqrt{N}} \qquad \forall i \in [N] && : (r_t^i).
\end{aligned} \tag{B.4}$$

Under assumption (**A3**), problem (B.4) has a finite optimal solution. The corresponding dual problem is

$$\begin{aligned}
\hat{Q}_t^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \max\ & \boldsymbol{\pi}_t^\top (\boldsymbol{b}_t - \boldsymbol{B}_t \boldsymbol{x}_{t-1}) + \sum_{i\in[N]} \sum_{j\in[\mathcal{S}_{t+1}^i]} y_{t,j}^i \mathscr{V}_{t,j}^i \\
\text{s.t. } & \boldsymbol{\pi}_t \in \mathbb{R}^M, \quad \boldsymbol{r}_t \in \mathbb{R}^N, \quad y_{t,j}^i \in \mathbb{R}_+ \ \forall i \in [N], \forall j \in [\mathcal{S}_{t+1}^i], \\
& \boldsymbol{A}_t^\top \boldsymbol{\pi}_t \le \boldsymbol{c}_t + \sum_{i\in[N]} \sum_{j\in[\mathcal{S}_{t+1}^i]} y_{t,j}^i \boldsymbol{g}_{t,j}^i, \\
& \sum_{i\in[N]} \sum_{j\in[\mathcal{S}_{t+1}^i]} y_{t,j}^i = 1, \\
& \frac{\sum_{i\in[N]} r_t^i}{\sqrt{N}} \le \lambda, \\
& \frac{\sum_{j\in[\mathcal{S}_{t+1}^i]} y_{t,j}^i}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} \le \sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)} + r_t^i, \\
& \sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)} \le \frac{\sum_{j\in[\mathcal{S}_{t+1}^i]} y_{t,j}^i}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} + r_t^i, \\
& r_t^i \le \lambda.
\end{aligned} \tag{B.5}$$

Note that here $\{\sum_{j\in[\mathcal{S}_{t+1}^i]} y_{t,j}^i,\ i \in [N]\}$ is actually the worst-case distribution. Using a finite number, $\mathcal{S}_t$, of extreme points of the feasible region in (B.5), we can write the above problem, as follows:

$$\hat{Q}_t^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \max \left\{ \boldsymbol{\pi}_{t,l}^\top (\boldsymbol{b}_t - \boldsymbol{B}_t \boldsymbol{x}_{t-1}) + \sum_{i\in[N]} \sum_{j\in[\mathcal{S}_{t+1}^i]} \left(y_{t,j}^i\right)_l \mathscr{V}_{t,j}^i \ :\ l \in [\mathcal{S}_t] \right\} \tag{B.6}$$

which implies $\hat{Q}_t^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ is piecewise linear convex with finite pieces in $\boldsymbol{x}_{t-1}$ for any $\boldsymbol{\xi}_t \in \hat{\Xi}_t$. This completes the proof. ∎

## B.4 Proof of Lemma 5

*Proof* We proceed by backward induction in the stages. Before we start, we present the lower bound problem at the $k$-th iteration:

$$\begin{aligned}
\underline{Q}_{t,k}^{\mathcal{DRO}}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) = \min\ & \boldsymbol{c}_t^\top \boldsymbol{x}_t + \gamma + \lambda \left( \beta + \sum_{n\in[N]} \psi_i \right) + \sum_{n\in[N]} \sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}(\mu_i - \zeta_i) \\
\text{s.t. } & \boldsymbol{x}_t \in \mathbb{R}_+^{d_t}, \quad \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}_+, \quad \boldsymbol{\mu}, \boldsymbol{\zeta}, \boldsymbol{\psi} \in \mathbb{R}_+^N \\
& \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{B}_t \boldsymbol{x}_{t-1} = \boldsymbol{b}_t, && : (\boldsymbol{\pi}_t), \\
& {\boldsymbol{g}_{t,j}^i}^\top \boldsymbol{x}_t + \mathscr{V}_{t,j}^i \le \gamma + \frac{\mu_i - \zeta_i}{\sqrt{\hat{w}_t^i(\boldsymbol{\xi}_t)}} \qquad \forall i \in [N], \forall j \in [\mathcal{S}_{t+1}^{i,k}] && : (y_{t,j}^i), \\
& \mu_i + \zeta_i = \psi_i + \frac{\beta}{\sqrt{N}} \qquad \forall i \in [N] && : (r_t^i).
\end{aligned} \tag{B.7}$$

Here, we use $\mathcal{S}_{t+1}^{i,k}$ to denote the total number of cuts generated in the first $k$ iterations for $\hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i)$. Note in (B.4), we use $\mathcal{S}_{t+1}^i$ to the total number of piecewise linear functions that describe $\hat{Q}_{t+1}^{\mathcal{DRO}}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}^i)$. Now we proceed to the main proof.

Stage $t = T$: Since the cost-to-go function $\hat{Q}_{T+1}^{\mathcal{DRO}}(\cdot) \equiv 0$, there are no cuts. For any feasible $\boldsymbol{x}_{T-1}$ and $\boldsymbol{\xi}_T \in \hat{\Xi}_T$, let $\hat{\boldsymbol{\pi}}_T$ be a dual feasible extreme point (here $\hat{\boldsymbol{\pi}}_T$ is the only dual variable) to problem (B.7), we have

$$\begin{aligned}
{\boldsymbol{g}_{T-1}}^\top \boldsymbol{x}_{T-1} + \mathscr{V}_{T-1} &= \hat{\boldsymbol{\pi}}_T^\top (\boldsymbol{b}_T - \boldsymbol{B}_T \boldsymbol{x}_{T-1}) \\
&\le \max \left\{ \boldsymbol{\pi}_T^\top (\boldsymbol{b}_T - \boldsymbol{B}_T \boldsymbol{x}_{T-1}) \ :\ \boldsymbol{A}_T^\top \boldsymbol{\pi}_T \le \boldsymbol{c}_T,\ \boldsymbol{\pi}_T \in \mathbb{R}^{d_T} \right\} \\
&= \min \left\{ \boldsymbol{c}_T^\top \boldsymbol{x}_T \ :\ \boldsymbol{A}_T \boldsymbol{x}_T + \boldsymbol{B}_T \boldsymbol{x}_{T-1} = \boldsymbol{b}_T,\ \boldsymbol{x}_T \in \mathbb{R}_+^{d_T} \right\} \\
&= \hat{Q}_T^{\mathcal{DRO}}(\boldsymbol{x}_{T-1}, \boldsymbol{\xi}_T)
\end{aligned}$$

where $\boldsymbol{\mathscr{g}}_{T-1}$ is the slope and $\mathscr{v}_{T-1}$ is the intercept. The second equality holds by strong duality. This proves the validity of the cut in stage $T-1$. Hence,

$$\underline{Q}^{\mathcal{DRO}}_{T-1,k}(\boldsymbol{x}_{T-2},\boldsymbol{\xi}_{T-1}) \leq \hat{Q}^{\mathcal{DRO}}_{T-1}\left(\boldsymbol{x}_{T-2},\boldsymbol{\xi}_{T-1}\right). \tag{B.8}$$

For the upper bound problem, first we have $\overline{Q}^{\mathcal{DRO}}_{T}(\cdot,\boldsymbol{\xi}_T) = \hat{Q}^{\mathcal{DRO}}_{T}(\cdot,\boldsymbol{\xi}_T)$ for $\boldsymbol{\xi}_T \in \hat{\Xi}_T$ due to the fact that there are no cost-to-go functions at stage $T$. By the convexity of $\hat{Q}^{\mathcal{DRO}}_{T}(\cdot,\boldsymbol{\xi}_T)$, we have

$$\hat{Q}^{\mathcal{DRO}}_{T}(\boldsymbol{x}_{T-1},\boldsymbol{\xi}^i_T) - \mathscr{F}_T^\top \boldsymbol{y}^i_{T-1} \leq \hat{Q}^{\mathcal{DRO}}_{T}\left(\sum_{j\in[k]}\theta^i_j \bar{\boldsymbol{x}}^j_{T-1},\boldsymbol{\xi}^i_T\right) \leq \sum_{j\in[k]}\theta^i_j \hat{Q}^{\mathcal{DRO}}_{T}(\bar{\boldsymbol{x}}^j_{T-1},\boldsymbol{\xi}^i_T), \tag{B.9}$$

where $\mathscr{F}_T$ is any subgradient of $\hat{Q}^{\mathcal{DRO}}_{T}(\boldsymbol{x}_{T-1},\boldsymbol{\xi}^i_T)$ at $\boldsymbol{x}_{T-1}$. Therefore,

$$\begin{aligned}
\sum_{j\in[k]}\theta^i_j \overline{Q}^{\mathcal{DRO}}_{T,j}(\bar{\boldsymbol{x}}^j_{T-1},\boldsymbol{\xi}^i_T) + M_T\|\boldsymbol{y}^i_{T-1}\|_1 &= \sum_{j\in[k]}\theta^i_j \hat{Q}^{\mathcal{DRO}}_{T}(\bar{\boldsymbol{x}}^j_{T-1},\boldsymbol{\xi}^i_T) + M_T\|\boldsymbol{y}^i_{T-1}\|_1 \\
&\geq \sum_{j\in[k]}\theta^i_j \hat{Q}^{\mathcal{DRO}}_{T}(\bar{\boldsymbol{x}}^j_{T-1},\boldsymbol{\xi}^i_T) + \|\mathscr{F}_T^\top\|_1 \cdot \|\boldsymbol{y}^i_{T-1}\|_1 \\
&\geq \sum_{j\in[k]}\theta^i_j \hat{Q}^{\mathcal{DRO}}_{T}(\bar{\boldsymbol{x}}^j_{T-1},\boldsymbol{\xi}^i_T) + \|\mathscr{F}_T^\top\|_2 \cdot \|\boldsymbol{y}^i_{T-1}\|_2 \\
&\geq \hat{Q}^{\mathcal{DRO}}_{T}(\sum_{j\in[k]}\theta^i_j \bar{\boldsymbol{x}}^j_{T-1},\boldsymbol{\xi}^i_T) + \mathscr{F}_T^\top \boldsymbol{y}^i_{T-1} \\
&\geq \hat{Q}^{\mathcal{DRO}}_{T}(\boldsymbol{x}_{T-1},\boldsymbol{\xi}^i_T).
\end{aligned}$$

The first inequality holds because $M_T \geq L^{\mathcal{DRO}}_T$ and thus any subgradient of $\hat{Q}^{\mathcal{DRO}}_{T}(\boldsymbol{x}_{T-1},\boldsymbol{\xi}^i_T)$ under 1-norm is bounded by $M_T$. The second inequality comes from the fact that $\|\boldsymbol{y}\|_2 \leq \|\boldsymbol{y}\|_1$ for any $\boldsymbol{y}$. The convexity and the Cauchy-Schwarz inequality imply the third inequality, whereas the last inequality comes from (B.9). Hence,

$$\hat{Q}^{\mathcal{DRO}}_{T-1}\left(\boldsymbol{x}_{T-2},\boldsymbol{\xi}_{T-1}\right) \leq \overline{Q}^{\mathcal{DRO}}_{T-1,k}(\boldsymbol{x}_{T-2},\boldsymbol{\xi}_{T-1}).$$

Combining this with (B.8), we have

$$\underline{Q}^{\mathcal{DRO}}_{T-1,k}(\boldsymbol{x}_{T-2},\boldsymbol{\xi}_{T-1}) \leq \hat{Q}^{\mathcal{DRO}}_{T-1}\left(\boldsymbol{x}_{T-2},\boldsymbol{\xi}_{T-1}\right) \leq \overline{Q}^{\mathcal{DRO}}_{T-1,k}(\boldsymbol{x}_{T-2},\boldsymbol{\xi}_{T-1}).$$

Therefore the claim holds for stage $T-1$.
Suppose the result holds for $t+1 \leq T-1$,

$$\underline{Q}^{\mathcal{DRO}}_{t+1,k}(\boldsymbol{x}_{t},\boldsymbol{\xi}_{t+1}) \leq \hat{Q}^{\mathcal{DRO}}_{t+1}\left(\boldsymbol{x}_{t},\boldsymbol{\xi}_{t+1}\right) \leq \overline{Q}^{\mathcal{DRO}}_{t+1,k}(\boldsymbol{x}_{t},\boldsymbol{\xi}_{t+1}), \quad \forall \boldsymbol{\xi}_{t+1} \in \hat{\Xi}_{t+1}.$$

This implies the cuts generated for the lower bound problem are valid for $t+1$, i.e.,

$$\max\left\{{\boldsymbol{\mathscr{g}}^i_{t,j}}^\top \boldsymbol{x}_t + \mathscr{v}^i_{t,j} \ : \ j \in [\mathcal{S}^{i,k}_{t+1}]\right\} \leq \hat{Q}^{\mathcal{DRO}}_{t+1}(\boldsymbol{x}_t,\boldsymbol{\xi}^i_{t+1}), \quad \forall \boldsymbol{\xi}^i_{t+1} \in \hat{\Xi}_{t+1}.$$

Hence the cuts in the feasible region of (B.7) are valid. Let $\mathcal{D}_t$ denote the feasible region of $\hat{Q}^{\mathcal{DRO}}_{t}(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)$ in (B.4), and $\mathcal{D}^k_t$ denote the feasible region of (B.7). Note the only difference between (B.4) and (B.7) is the set of cuts: in (B.4) $\mathcal{S}^i_{t+1}$ represent the total number of cuts for $\hat{Q}^{\mathcal{DRO}}_{t+1}(\boldsymbol{x}_t,\boldsymbol{\xi}^i_{t+1})$, while in (B.7) $\mathcal{S}^{i,k}_{t+1}$ represent the total number of cuts generated in the first $k$ iterations for $\hat{Q}^{\mathcal{DRO}}_{t+1}(\boldsymbol{x}_t,\boldsymbol{\xi}^i_{t+1})$. Let $(\hat{\boldsymbol{\pi}}_t,\hat{\boldsymbol{y}})$ be a dual feasible extreme point to problem (B.7). Using similar result as (B.6), we have

$$\begin{aligned}
{\boldsymbol{\mathscr{g}}_{t-1}}^\top \boldsymbol{x}_{t-1} + \mathscr{v}_{t-1} &= \hat{\boldsymbol{\pi}}_t^\top(\boldsymbol{b}_t - \boldsymbol{B}_t\boldsymbol{x}_{t-1}) + \sum_{i\in[N]}\sum_{j\in[\mathcal{S}^{i,k}_{t+1}]} \hat{y}^i_{t,j}\mathscr{v}^i_{t,j} \\
&\leq \max\left\{\boldsymbol{\pi}_t^\top(\boldsymbol{b}_t - \boldsymbol{B}_t\boldsymbol{x}_{t-1}) + \sum_{i\in[N]}\sum_{j\in[\mathcal{S}^{i,k}_{t+1}]} y^i_{t,j}\mathscr{v}^i_{t,j} \ : \ (\boldsymbol{\pi}_t,\boldsymbol{y}) \in \mathcal{D}^k_t\right\} \\
&\leq \max\left\{\boldsymbol{\pi}_t^\top(\boldsymbol{b}_t - \boldsymbol{B}_t\boldsymbol{x}_{t-1}) + \sum_{i\in[N]}\sum_{j\in[\mathcal{S}^{i}_{t+1}]} y^i_{t,j}\mathscr{v}^i_{t,j} \ : \ (\boldsymbol{\pi}_t,\boldsymbol{y}) \in \mathcal{D}_t\right\} \\
&= \hat{Q}^{\mathcal{DRO}}_{t}(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t).
\end{aligned}$$

The first inequality holds because $(\hat{\boldsymbol{\pi}}_t,\hat{\boldsymbol{y}}) \in \mathcal{D}^k_t$ whereas the second inequality holds because $\mathcal{D}_t$ includes all the cuts that define $\hat{Q}^{\mathcal{DRO}}_{t}(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)$ while the cuts indexed by $\mathcal{S}^{i,k}_{t+1}$ in $\mathcal{D}^k_t$ by assumption only includes parts of these valid cuts. The last equality comes from strong duality. This proves the validity of the cut in stage $t$. Hence,

$$\underline{Q}^{\mathcal{DRO}}_{t,k}(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t) \leq \hat{Q}^{\mathcal{DRO}}_{t}(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t). \tag{B.10}$$

For the upper bound problem, by hypothesis we have $\hat{Q}^{\mathcal{DRO}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}^i_{t+1}) \leq \overline{Q}^{\mathcal{DRO}}_{t+1,k}(\boldsymbol{x}_t, \boldsymbol{\xi}^i_{t+1})$, $\forall \boldsymbol{\xi}^i_{t+1} \in \hat{\Xi}_{t+1}$. Now replace $\overline{Q}^{\mathcal{DRO}}_{t+1,k}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1})$ with $\hat{Q}^{\mathcal{DRO}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1})$ in the constraints of (4.8) and call the new problem $\overline{Q}^{\mathcal{DRO}_L}_{t,k}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$, we have

$$\overline{Q}^{\mathcal{DRO}_L}_{t,k}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) \leq \overline{Q}^{\mathcal{DRO}}_{t,k}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$$

Therefore, it is sufficient to show $\overline{Q}^{\mathcal{DRO}_L}_{t,k}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ is an upper bound for $\hat{Q}^{\mathcal{DRO}}_{t,k}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$. Denote $\mathscr{F}_t$ to be any subgradient of $\hat{Q}^{\mathcal{DRO}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}^i_{t+1})$ at $\boldsymbol{x}_t$. We have

$$\begin{aligned}
\hat{Q}^{\mathcal{DRO}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}^i_{t+1}) &\leq \sum_{j\in[k]} \theta^i_j \hat{Q}^{\mathcal{DRO}}_{t+1}(\bar{\boldsymbol{x}}^j_t, \boldsymbol{\xi}^i_{t+1}) + \mathscr{F}_t^\top \boldsymbol{y}^i_t \\
&\leq \sum_{j\in[k]} \theta^i_j \hat{Q}^{\mathcal{DRO}}_{t+1}(\bar{\boldsymbol{x}}^j_t, \boldsymbol{\xi}^i_{t+1}) + \|\mathscr{F}_t^\top\|_2 \cdot \|\boldsymbol{y}^i_t\|_2 \\
&\leq \sum_{j\in[k]} \theta^i_j \hat{Q}^{\mathcal{DRO}}_{t+1}(\bar{\boldsymbol{x}}^j_t, \boldsymbol{\xi}^i_{t+1}) + \|\mathscr{F}_t^\top\|_1 \cdot \|\boldsymbol{y}^i_t\|_1 \\
&\leq \sum_{j\in[k]} \theta^i_j \hat{Q}^{\mathcal{DRO}}_{t+1}(\bar{\boldsymbol{x}}^j_t, \boldsymbol{\xi}^i_{t+1}) + M_t \|\boldsymbol{y}^i_t\|_1 \\
&= \overline{Q}^{\mathcal{DRO}_L}_{t+1,k}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1}).
\end{aligned}$$

The first inequality comes from the convexity of $\hat{Q}^{\mathcal{DRO}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1})$, whereas the second inequality comes from the Cauchy-Schwarz inequality. The third inequality holds due to the fact that $\|\boldsymbol{y}\|_2 \leq \|\boldsymbol{y}\|_1$ for any $\boldsymbol{y}$. The last inequality holds because $M_t \geq L^{\mathcal{DRO}}_t$ and thus any subgradient of $\hat{Q}^{\mathcal{DRO}}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}^i_t)$ under 1-norm is bounded by $M_t$. Hence,

$$\hat{Q}^{\mathcal{DRO}}_{t,k}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) \leq \overline{Q}^{\mathcal{DRO}_L}_{t,k}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) \leq \overline{Q}^{\mathcal{DRO}}_{t,k}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t).$$

Combining this with (B.10), we have

$$\underline{Q}^{\mathcal{DRO}}_{t,k}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) \leq \hat{Q}^{\mathcal{DRO}}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t) \leq \overline{Q}^{\mathcal{DRO}}_{t,k}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t),$$

which completes the proof. ∎

## B.5 Proof of Theorem 4

*Proof* We first discuss the lower bound convergence, using the proof of the standard SDDP algorithm presented by [55].

As discussed in Lemma 4, $\hat{Q}^{\mathcal{DRO}}_t(\cdot, \boldsymbol{\xi}^i_t)$ $\forall i \in [N]$ are piecewise linear convex with a finite number of pieces under the assumption (**B1**) since there exist only a finite number of basic feasible solutions of the dual problem of $\hat{Q}^{\mathcal{DRO}}_t(\cdot, \boldsymbol{\xi}^i_t)$ $\forall i \in [N]$. In addition, the total number of possible scenarios is finite (i.e., $\prod_{t=2}^T |\hat{\Xi}_t| = N^{T-1}$) due to the discretization, and there is a nonzero probability for any possible scenario to occur in the forward step since scenarios are generated by Monte Carlo sampling. Hence, any possible scenario occurs infinitely many times in the forward step unless the algorithm terminates.

As shown in [55], the lower bound convergence holds if policy $\bar{\boldsymbol{x}}_t(\boldsymbol{\xi}_{[t]})$ for the current lower bound problem satisfies the following Bellman's optimality condition,

$$\bar{\boldsymbol{x}}_t(\boldsymbol{\xi}_{[t]}) \in \operatorname*{argmin}_{\boldsymbol{x}_t \in \mathcal{X}_t(\bar{\boldsymbol{x}}_{t-1}(\boldsymbol{\xi}_{[t-1]}), \boldsymbol{\xi}_t)} \boldsymbol{c}_t^\top \boldsymbol{x}_t + \max_{\mathbb{P}_t \in \mathcal{P}^\lambda_t(\hat{\mathbb{P}}_t)} \mathbb{E}_{\mathbb{P}_t}\left[\hat{Q}^{\mathcal{DRO}}_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) \,\middle|\, \boldsymbol{\xi}_t\right] \tag{B.11}$$

for every stage and possible scenario. Note that the DRO formulation (B.11) is equivalent to the single-level DRO reformulation (4.6) (here, we use (B.11) to save space). Let $\bar{\boldsymbol{x}}^k_t(\boldsymbol{\xi}_{[t]})$ be a policy obtained by lower bound approximation $\underline{Q}^{\mathcal{DRO}}_{t,k}(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}^i_t)$ $\forall t = 2, \ldots, T$ and $\forall i \in [N]$ at iteration $k$. Suppose that (B.11) does not hold for some (or all) stage $t \in \{2, \ldots, T\}$ and some (or all) possible scenario, i.e., a policy for the current lower bound approximation is not optimal for the DRO problem. Let stage $t'$ be the largest stage $t$ such that the function $\bar{\boldsymbol{x}}^k_{t'}(\boldsymbol{\xi}_{[t']})$ does not satisfy (B.11), i.e., a candidate solution $\bar{\boldsymbol{x}}^k_{t'} = \bar{\boldsymbol{x}}^k_{t'}(\boldsymbol{\xi}_{[t']})$ is suboptimal for the current scenario. Then, at some iteration $k' > k$, we add a new cut corresponding to $\bar{\boldsymbol{x}}^k_{t'}$, updating the current approximation $\underline{Q}^{\mathcal{DRO}}_{t',k'}(\cdot, \cdot)$. Similarly, a new cut is added until (B.11) holds for stage $t'$. Such cut generations continue for some stage $t < t'$ until (B.11) holds. After a sufficiently large number of iterations, (B.11) holds for every stage and possible scenario. This completes the proof for the lower bound convergence.

Now we discuss the upper bound convergence. Let $\underline{k}^*$ be the iteration after which the lower bound convergence holds. The lower bound convergence under the assumption (**B1**) implies that there only exist a finite number of optimal policies after the convergence, i.e., iteration $k \geq \underline{k}^*$. In other words, for each scenario, we have only a finite number of sequences of candidate solutions $\bar{\boldsymbol{x}}_{[T]}$ at which the upper bound approximation is evaluated (recall that candidate solutions are obtained by solving the lower bound problem in Algorithm 2). Hence, we will show that the gap between the lower and upper approximation is closed at those finite number of candidate solutions.

Without loss of generality, let us assume that there is only one optimal policy for each scenario. Let $\mathcal{S}$ be a set of all possible scenarios and $\bar{\boldsymbol{x}}^s_{[T]}$ be the optimal sequence of solutions for scenario $s \in \mathcal{S}$ after the lower bound convergence. Here, we show

$$\overline{Q}^{\mathcal{DRO}}_{t,k}(\bar{\boldsymbol{x}}^s_{t-1}, \boldsymbol{\xi}^i_t) = \underline{Q}^{\mathcal{DRO}}_{t,k}(\bar{\boldsymbol{x}}^s_{t-1}, \boldsymbol{\xi}^i_t)\ \forall i \in [N],\ s \in \mathcal{S} \text{ for some iteration } k \geq \underline{k}^* \tag{B.12}$$

for every stage.

Let $\overline{k}_t$ be the iteration for which (B.12) holds at stage t. At the terminal stage $T$, (B.12) holds after the upper bound approximation at stage $T$ is evaluated at all possible candidate solutions $\bar{\boldsymbol{x}}^s_T$ $\forall s \in \mathcal{S}$ since $\overline{Q}^{\mathcal{DRO}}_{T+1,k}(\cdot, \cdot) = \underline{Q}^{\mathcal{DRO}}_{T+1,k}(\cdot, \cdot) \equiv 0$. Proceeding by induction, for $t = T-1, \ldots, 2$, there exists iteration $k \geq \overline{k}_{t+1}$ such that (B.12) holds at stage $t$ because $\overline{Q}^{\mathcal{DRO}}_{t+1,k}(\bar{\boldsymbol{x}}^s_t, \boldsymbol{\xi}^i_{t+1}) = \underline{Q}^{\mathcal{DRO}}_{t+1,k}(\bar{\boldsymbol{x}}^s_t, \boldsymbol{\xi}^i_{t+1})$ $\forall i \in [N],\ s \in \mathcal{S}$. This completes the upper bound convergence proof.

∎