SIAM J. CONTROL OPTIM.
Vol. 61, No. 6, pp. 3559–3583

AN INTEGRATED TRANSPORTATION DISTANCE BETWEEN KERNELS AND APPROXIMATE DYNAMIC RISK EVALUATION IN MARKOV SYSTEMS*

ZHENGQI LIN† AND ANDRZEJ RUSZCZYŃSKI†

Abstract. We introduce a distance between kernels based on the Wasserstein distances between their values, study its properties, and prove that it is a metric on an appropriately defined space of kernels. We also relate it to various modes of convergence in the space of kernels. Then we consider the problem of approximating solutions to forward–backward systems, where the forward part is a Markov system described by a sequence of kernels, and the backward part calculates the values of a risk measure by operators that may be nonlinear with respect to the system's kernels. We propose recursively approximating the forward system with the use of the integrated transportation distance between kernels and we estimate the error of the risk evaluation by the errors of individual kernel approximations. We illustrate the results on stopping problems and several well-known risk measures. Then we develop a particle-based numerical procedure, in which the approximate kernels have finite support sets. Finally, we illustrate the efficacy of the approach on the financial problem of pricing an American basket option.

Key words. Wasserstein distance, dynamic risk measures, dynamic programming

MSC codes. 49M25, 60J05, 93E20

DOI. 10.1137/22M1530665

1. Introduction.

We consider a discrete-time Markov system described by the relations

$$X_{t+1} \sim Q_t(X_t), \quad t = 0, 1, \dots, T-1, \tag{1.1}$$

where $X_t \in \mathscr{X}$ represents the state at time t, $\mathscr{X}$ is a Polish space, and $Q_t : \mathscr{X} \to \mathscr{P}(\mathscr{X})$, $t = 0, 1, \dots, T-1$, are stochastic kernels (the symbol $\mathscr{P}(\mathscr{X})$ denotes the space of probability measures on $\mathscr{X}$). The initial state $X_0 = x_0$ is fixed. The model (1.1) is understood as follows: given $X_t = x$, the conditional distribution of X_{t+1} is $Q_t(x)$. The sequence of kernels Q_t, $t = 0, \dots, T$, and the distribution of the initial state λ_0 define a probability measure P on the canonical space $\mathscr{X}^{T+1}$. We also consider the filtration $\mathscr{F}_t = \mathscr{B}(\mathscr{X}^{t+1})$, $t = 0, \dots, T$.

Suppose a sequence of Borel measurable functions $c_t : \mathscr{X} \to \mathbb{R}$, $t = 0, \dots, T$, is given. Together with the dynamical system (1.1), we consider the following backward risk evaluation system:

$$\begin{aligned} v_t(x) &= c_t(x) + \sigma_t\big(x, Q_t(x), v_{t+1}(\cdot)\big), \quad x \in \mathscr{X}, \quad t = T-1, T-2, \dots, 0; \\ v_T(x) &= c_T(x), \quad x \in \mathscr{X}. \end{aligned} \tag{1.2}$$

In (1.2), the operator $\sigma_t : \mathscr{X} \times \mathscr{P}(\mathscr{X}) \times \mathscr{V} \to \mathbb{R}$, where $\mathscr{V}$ is a space of Borel measurable real functions on $\mathscr{X}$, is a *transition risk mapping.* Its first argument is the present state x. The second argument is the probability distribution $Q_t(x)$ of the state following x in the system (1.1). The last argument, the function $v_{t+1}(\cdot)$, is the next

*Received by the editors October 25, 2022; accepted for publication (in revised form) August 15, 2023; published electronically December 5, 2023.
https://doi.org/10.1137/22M1530665
Funding: This work was supported by the National Science Foundation award DMS-1907522 and by the Office of Naval Research award N00014-21-1-2161.
†Management Science and Information Systems, Rutgers University, Piscataway, NJ 08854 USA (zl458@rutgers.edu, rusz@rutgers.edu).

state's value: the risk of running the system from the next state in the time interval from $t+1$ to T. In the next section, we briefly review the background of this backward system in the dynamic risk theory and provide a more formal definition of the objects involved, but we want to stress that the evaluation (1.2) is of relevance for other problems as well.

A simple case of the transition risk mapping is the bilinear form

$$\sigma_t\big(x,\mu,v_{t+1}(\cdot)\big)=\mathbb{E}_\mu\big[v_{t+1}(\cdot)\big]. \tag{1.3}$$

In this case, the scheme (1.2) evaluates the conditional expectation of the total cost from stage t to the end of the horizon T:

$$v_t(x)=\mathbb{E}\big[c_t(X_t)+\cdots+c_T(X_T)\,\big|\,X_t=x\big],\quad x\in\mathscr{X},\quad t=0,\dots,T.$$

A more interesting application is the *optimal stopping problem,* in which $c_t(\cdot)\equiv 0$, and

$$\sigma_t\big(x,\mu,v_{t+1}(\cdot)\big)=\max\Big(r_t(x)\,;\,\mathbb{E}_\mu\big[v_{t+1}(\cdot)\big]\Big). \tag{1.4}$$

Here, $r_t:\mathscr{X}\to\mathbb{R}$, $t=0,\dots,T$, represent the rewards collected if the decision to stop at time t and state x is made. Clearly, with the mappings (1.4) used in the scheme (1.2),

$$v_t(x)=\sup_{\substack{\tau-\text{stopping time}\\ t\le\tau\le T}} r_\tau(X_\tau),\quad x\in\mathscr{X},\quad t=0,\dots,T;$$

see, *e.g.,* [10]. The most important difference between (1.3) and (1.4) is that the latter is nonlinear with respect to the probability measure μ. In the next section, we provide other examples of nonlinear transition risk mappings derived from coherent measures of risk.

One of the challenges associated with the backward system (1.2) is the numerical solution in the case when the transition risk mappings are nonlinear with respect to the probability measures involved. The objective of this paper is to present a computational method based on approximating the kernels $Q_t(\cdot)$ by simpler, easier-to-handle kernels $\widetilde{Q}_t(\cdot)$, and using them in the backward system (1.2). For this purpose, after the preliminary section, in section 3 we introduce the space of kernels under consideration and define a metric on this space. The metric generalizes the transportation (Wasserstein) metric between probability distributions. We relate it to various convergence modes in the space of kernels. In section 4 we describe an iterative scheme for building the approximate system and we estimate the error of the approximation by the distances of the kernels involved at each stage. We also illustrate the application of the theory to various specific risk evaluation systems with nonlinear transition risk mappings. Next, in section 5, we specialize our method by considering kernels supported on finite sets, and we derive tractable linear programming models for minimizing the approximation error. Finally, in section 6, we illustrate our approach on the problem of evaluating an American basket option.

The problem of approximating stochastic processes in discrete time has attracted the attention of researchers for several decades. The basic construction is that of a *scenario tree.* In [20], the construction of the tree is based on statistical parameters, such as moments and correlations. A further contribution of [22] involves copulas to capture the shape of the distributions. The use of probability metrics to reduce large scenario trees was first proposed in [19]. A concept of a distance between stochastic processes was proposed by [33], and used by [29, 26] to generate scenario trees.

The concept of nested (adapted) distance, using an extension of the Wasserstein metric for processes, was introduced in [34] and further developed in [35, 36]. Similar ideas are pursued in continuous time in [3]. Reference [4] addresses the sensitivity of the optimal value of an expected-value problem, when the probability measure perturbation is small in the nested distance. None of these contributions focuses on Markov systems and the approaches proposed do not reduce to our construction in the Markovian case.

Reference [23] considers perturbations in a transition kernel of a controlled Markov system. The distance between probability kernels defined in [23, section 3] is close to our idea, but it uses the "sup norm" over the state space, rather than the "$\mathscr{L}_p$ norm" in our case (a similar idea appeared earlier in [29] for scenario trees). This is further used to estimate the error of the value function in risk-neutral models in [46]. We discuss it in more detail in sections 3 and 4.

Finally, some recent contributions focus on mixture models, which are somehow related to our approach, but which measure the distance of mixture distributions rather than kernels. The sketched Wasserstein distance, a type of distance metric dedicated to finite mixture models, was proposed in [6]. Research on Wasserstein-based distances specifically tailored to Gaussian mixture models is reported in [7, 12, 24].

2. Preliminaries. In this section, we briefly present the mathematical foundations of the techniques discussed in the paper. In section 2.1, we summarize the relevant concepts of Markov risk evaluation, and in section 2.2 we recall the basic ideas of the transportation distance between probability measures.

2.1. Markov risk measures. A *dynamic risk measure* evaluates the sequence of random costs $Z_t = c_t(X_t)$, $t = 0, 1, 2, \dots, T$, where $c_t : \mathscr{X} \to \mathbb{R}$, $t = 0, 1, \dots, T$, are measurable functions. Because of the need to evaluate the risk of the future costs at any time period, a dynamic measure of risk is a collection of *conditional risk measures* $\rho_{t,T}(Z_t, \dots, Z_T)$, $t = 0, \dots, T$. Formally, for $t = 0, \dots, T$, we consider σ-subalgebras $\mathscr{F}_t = \mathscr{B}(\mathscr{X}^{t+1})$ and spaces $\mathscr{Z}_t$ of $\mathscr{F}_t$-measurable real random variables. A conditional risk measure is a functional $\rho_{t,T} : \mathscr{Z}_t \times \cdots \times \mathscr{Z}_T \to \mathscr{Z}_t$. We postulate three properties of each conditional risk measure:

Normalization: $\rho_{t,T}(0, \dots, 0) = 0$, $t = 0, 1, \dots, T$.
Monotonicity: For every $t = 0, \dots, T$, if $Z_s \le V_s$ for $s = t, \dots, T$, then $\rho_{t,T}(Z_t, \dots, Z_T) \le \rho_{t,T}(V_t, \dots, V_T)$.
Translation equivariance: $\rho_{t,T}(Z_t, Z_{t+1}, \dots, Z_T) = Z_t + \rho_{t,T}(0, Z_{t+1}, \dots, Z_T)$ $\forall\, t = 0, \dots, T$.

Fundamental for such a nonlinear dynamic risk evaluation is *time consistency,* discussed in various forms in [2, 8, 9, 43]. We adopt the definition and the following discussion from [41]: *A dynamic measure of risk is time consistent if for every* $t = 0, \dots, T-1$, *if* $Z_t = V_t$ *and* $\rho_{t+1,T}(Z_{t+1}, \dots, Z_T) \le \rho_{t+1,T}(V_{t+1}, \dots, V_T)$ *a.s., then*

$$\rho_{t,T}(Z_t, \dots, Z_T) \le \rho_{t,T}(V_t, \dots, V_T).$$

Such risk measures, under the conditions specified above, must have a specific recursive form [41, Thm. 1]:

$$\rho_{t,T}(Z_t, \dots, Z_T) = Z_t + \rho_t\Big(Z_{t+1} + \rho_{t+1}\big(Z_{t+2} + \cdots + \rho_{T-1}(Z_T)\cdots\big)\Big),$$

where each $\rho_t : \mathscr{Z}_{t+1} \to \mathscr{Z}_t$ is a one-step conditional risk measure. This result, generalizing the tower property of conditional expectations, is germane for our approach.

Markov risk measures evaluate the risk of future costs $Z_s = c_s(X_s)$, $s = t, \dots, T$, in a Markov system (1.1) in such a way that the risk of the future cost sequence is a function of the current state:

$$\rho_{t,T}(Z_t, \dots, Z_T) = v_t(X_t).$$

This, combined with the properties specified above, implies a very specific structure [16, 5]: transition risk mappings $\sigma_t : \mathscr{X} \times \mathscr{P}(\mathscr{X}) \times \mathscr{V} \to \mathbb{R}$, $t = 0, \dots, T-1$, exist such that the risk of each state can be evaluated by the procedure (1.2). Conversely, any collection of transition risk mappings satisfying the properties of normalization, monotonicity, and translation equivariance define via (1.2) a time-consistent Markov risk measure.

As mentioned in the introduction, the simplest transition risk mappings are the bilinear forms (1.3), which lead to the risk-neutral evaluation: the expected value of the sum of the costs. A more interesting example is the *mean–semideviation* mapping derived from the corresponding coherent risk measure [30, 31, 43]:

$$\begin{aligned}\text{(2.1)}\qquad \operatorname{msd}_p\big(x, \mu, v_{t+1}(\cdot)\big) &= \int_{\mathscr{X}} v_{t+1}(y)\, \mu(\mathrm{d}y) \\ &\quad + \varkappa(x)\bigg(\int_{\mathscr{X}} \Big[v(y) - \int_{\mathscr{X}} v(y')\, \mu(\mathrm{d}y') \Big]_+^p \mu(\mathrm{d}y) \bigg)^{1/p},\end{aligned}$$

with $p \in [1, \infty)$, and the parameter $\varkappa(x) \in [0, 1]$ controlling the degree of risk aversion.

Another example is the Average Value at Risk [39, 32, 43]:

$$\text{(2.2)}\qquad \operatorname{AVaR}_\alpha\big(x, \mu, v_{t+1}(\cdot)\big) = \inf_{\eta \in \mathbb{R}} \Big\{ \eta + \frac{1}{\alpha} \mathbb{E}_\mu \big[\max(0, v_{t+1}(\cdot) - \eta) \big] \Big\}, \quad \alpha \in (0, 1].$$

Usually, it does not occur alone, but rather in mixtures, as in *spectral* measures (see, e.g., [37, 43])

$$\text{(2.3)}\qquad \sigma_t\big(x, \mu, v_{t+1}(\cdot)\big) = \int_0^1 \operatorname{AVaR}_\alpha\big(x, \mu, v_{t+1}(\cdot)\big)\, \theta(\mathrm{d}\alpha),$$

where θ is a probability measure on $(0, 1]$.

Summing up, the risk evaluation procedure (1.2) is not an arbitrary construction, but rather the result of assumptions of normalization, monotonicity, translation, time consistency, and the Markov property. The transition risk mappings are nonlinear operators with respect to the probability measure, and the numerical evaluation of risk is a difficult task. Structures of the form (1.2) arise also in the discretization of backward stochastic differential equations [42]. For recent applications of Markov risk measures in the control of dynamical systems, see [28, 44, 25].

2.2. The Wasserstein distance. Another essential ingredient of our construction is the Wasserstein distance between measures. As before, $\mathscr{X}$ is a Polish space, with the metric $d(\cdot, \cdot)$, and the associated Borel σ-field $\mathscr{B}(\mathscr{X})$. The symbol $\mathscr{P}(\mathscr{X})$ denotes the space of probability measures on $\mathscr{B}(\mathscr{X})$. For $p \geq 1$, we consider the space

$$\mathscr{P}_p(\mathscr{X}) := \bigg\{ \mu \in \mathscr{P}(\mathscr{X}) : \int_{\mathscr{X}} d\,(x_0, x)^p\ \mu(dx) < +\infty \bigg\},$$

where $x_0 \in \mathscr{X}$ is arbitrary. In the brief summary below, we follow [45]. The reader is referred to this monograph, as well as to [38], for an extensive exposition and historical account.

DEFINITION 2.1. *The Wasserstein distance of order $p \in [1,\infty)$ between two probability measures $\mu,\nu \in \mathscr{P}_p(\mathscr{X})$ is defined by the formula*

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathscr{X}\times\mathscr{X}} d(x,y)^p \, \pi(dx,dy) \right)^{1/p}, \tag{2.4}$$

where $\Pi(\mu,\nu)$ is the set of all probability measures in $\mathscr{P}_p(\mathscr{X}\times\mathscr{X})$ with the marginals μ and ν. The measure $\pi^* \in \Pi(\mu,\nu)$ that realizes the infimum in (2.4) *is called the* optimal coupling *or the* optimal transport plan.

For each $p \in [1,\infty)$, the function $W_p(\cdot,\cdot)$ defines a metric on $\mathscr{P}_p(\mathscr{X})$. Furthermore, for all $\mu,\nu \in \mathscr{P}_p(\mathscr{X})$ the optimal coupling realizing the infimum in (2.4) exists. From now on, the space $\mathscr{P}_p(\mathscr{X})$ will be always equipped with the distance $W_p(\cdot,\cdot)$.

Remark 2.2. Problem (2.4) has a convenient linear programming representation for discrete measures. Let μ and ν be discrete measures in $\mathscr{P}(\mathscr{X})$, supported at positions $\{x^{(i)}\}_{i=1}^N$ and $\{z^{(s)}\}_{s=1}^S$ with normalized (totaling 1) positive weight vectors w_x and w_z: $\mu = \sum_{i=1}^N w_x^{(i)} \delta_{x^{(i)}}$, $\nu = \sum_{s=1}^S w_z^{(s)} \delta_{z^{(s)}}$. For $p \geq 1$, let $D \in R_+^{N\times S}$ be the distance matrix defined as $D_{is} = d(x^{(i)}, z^{(s)})^p$. Then the pth power of the p-Wasserstein distance between the measures μ and ν is the optimal value of the following transportation problem:

$$\min_{\pi \in R_+^{N\times S}} \sum_{is} D_{is}\pi_{is} \quad \text{s.t.} \quad \pi^\top \mathbb{1}_N = w_x, \quad \pi \mathbb{1}_S = w_z. \tag{2.5}$$

Its regularized version can be efficiently solved with almost linear complexity with respect to NS; see [11, 1].

The following classical result, known as the Kantorovich–Rubinstein duality [21], provides an alternative characterization of $W_1(\cdot,\cdot)$.

THEOREM 2.3. *For any μ,ν in $\mathscr{P}_1(\mathscr{X})$,*

$$W_1(\mu,\nu) = \sup_{\|\psi\|_{Lip}\leq 1} \left\{ \int_{\mathscr{X}} \psi(x)\, \mu(dx) - \int_{\mathscr{X}} \psi(x)\, \nu(dx) \right\}, \tag{2.6}$$

where $\|\psi\|_{Lip}$ denotes the minimal Lipschitz constant of the function $\psi : \mathscr{X} \to \mathbb{R}$.

In the discrete case, it follows from the linear programming duality for problem (2.5).

We now briefly review the convergence concepts in the space $\mathscr{P}_p(\mathscr{X})$. The notation $\mu_k \rightharpoonup \mu$ means that μ_k converges weakly to μ, i.e., $\int \varphi(x)\, \mu_k(\mathrm{d}x) \to \int \varphi(x)\, \mu(\mathrm{d}x)$ for all bounded continuous functions $\varphi : \mathscr{X} \to \mathbb{R}$.

DEFINITION 2.4. *Let $(\mathscr{X},d)$ be a Polish space, and let $p \in [1,\infty)$. Let $\{\mu_k\}_{k\in N}$ be a sequence of probability measures in $\mathscr{P}_p(\mathscr{X})$, and let μ be an element of $\mathscr{P}_p(\mathscr{X})$. Then $\{\mu_k\}$ is said to converge to μ weakly in $\mathscr{P}_p(\mathscr{X})$, written $\mu_k \xrightarrow{p} \mu$, if for some (and then any) $x_0 \in \mathscr{X}$, and for all continuous functions φ with $|\varphi(x)| \leq 1 + d(x_0,x)^p$ one has*

$$\int \varphi(x)\, \mu_k(dx) \longrightarrow \int \varphi(x)\, \mu(dx). \tag{2.7}$$

The fundamental property of the Wasserstein distance $W_p(\cdot,\cdot)$ is that it metricizes the topology of weak convergence in $\mathscr{P}_p(\mathscr{X})$.

THEOREM 2.5. *Let $(\mathscr{X},d)$ be a Polish space, $p\in[1,\infty)$; then $\mu_k \xrightarrow{p} \mu$ if and only if $W_p(\mu_k,\mu)\to 0$. Furthermore, $\big(\mathscr{P}_p(\mathscr{X}),W_p\big)$ is a Polish space.*

By the triangle inequality, $W_p(\cdot,\cdot)$ is continuous on $\mathscr{P}_p(\mathscr{X})\times\mathscr{P}_p(\mathscr{X})$.

3. The integrated transportation distance between kernels. We now introduce an essential concept in our research: the integrated transportation distance between kernels.

Suppose $\mathscr{X}$ and $\mathscr{Y}$ are Polish spaces. By the measure disintegration formula, every probability measure $\mu\in\mathscr{P}(\mathscr{X}\times\mathscr{Y})$ admits a disintegration $\mu=\lambda\circledast Q$, where $\lambda\in\mathscr{P}(\mathscr{X})$ is the marginal distribution on $\mathscr{X}$, and $Q:\mathscr{X}\to\mathscr{P}(\mathscr{Y})$ is a *kernel* (a function such that for each $B\in\mathscr{B}(\mathscr{Y})$ the mapping $x\mapsto Q(B|x)$ is Borel measurable):

$$\mu(A\times B)=\int_A Q(B|x)\,\lambda(\mathrm{d}x)\quad \forall\big(A\in\mathscr{B}(\mathscr{X})\big),\,\forall\big(B\in\mathscr{B}(\mathscr{Y})\big).$$

Conversely, given a marginal $\lambda\in\mathscr{P}(\mathscr{X})$ and a kernel $Q:\mathscr{X}\to\mathscr{P}(\mathscr{Y})$, the above formula defines a probability measure $\lambda\circledast Q$ on $\mathscr{X}\times\mathscr{Y}$. Its marginal on $\mathscr{Y}$ is the *mixture distribution* $\lambda\circ Q$ given by

$$(\lambda\circ Q)(B)=\int_{\mathscr{X}} Q(B|x)\,\lambda(\mathrm{d}x)\quad\forall\,B\in\mathscr{B}(\mathscr{Y}).$$

We intend to define a distance between kernels with the use of the Wasserstein metric in the space of probability measures. To this end, we restrict the class of kernels under consideration. We use the same symbol $d(\cdot,\cdot)$ to denote the metrics on $\mathscr{X}$ and $\mathscr{Y}$; the space will be clear from the context.

DEFINITION 3.1. *The kernel space of order $p\in[1,\infty)$ is the set*

$$\mathscr{Q}_p(\mathscr{X},\mathscr{Y})=\Big\{Q:\mathscr{X}\to\mathscr{P}_p(\mathscr{Y}):\forall\big(B\in\mathscr{B}(\mathscr{Y})\big)\,Q(B|\cdot)\ \textit{is Borel measurable},\ \exists(C>0)\,\forall(x\in\mathscr{X})\int_{\mathscr{Y}} d(y,y_0)^p\,Q(dy|x)\le C\big(1+d(x,x_0)^p\big)\Big\}. \tag{3.1}$$

It is evident that the choice of the points $x_0\in\mathscr{X}$ and $y_0\in\mathscr{Y}$ is irrelevant in this definition.

DEFINITION 3.2. *The integrated transportation distance of degree p between two kernels Q and $\widetilde{Q}$ in $\mathscr{Q}_p(\mathscr{X},\mathscr{Y})$ with fixed marginal $\lambda\in\mathscr{P}_p(\mathscr{X})$ is defined as*

$$\mathscr{W}_p^{\lambda}(Q,\widetilde{Q})=\left(\int_{\mathscr{X}}\big[W_p(Q(\cdot|x),\widetilde{Q}(\cdot|x))\big]^p\,\lambda(dx)\right)^{1/p}. \tag{3.2}$$

From now on, for a fixed marginal $\lambda\in\mathscr{P}_p(\mathscr{X})$, we shall identify the kernels Q and $\widetilde{Q}$ if $W_p(Q(\cdot|x),\widetilde{Q}(\cdot|x))=0$ for λ-almost all $x\in\mathscr{X}$. Thus, we consider the space $\mathscr{Q}_p^{\lambda}(\mathscr{X},\mathscr{Y})$ of equivalence classes of $\mathscr{Q}_p(\mathscr{X},\mathscr{Y})$.

THEOREM 3.3. *For any $p\in[1,\infty)$ and any $\lambda\in\mathscr{P}_p(\mathscr{X})$, the function $\mathscr{W}_p^{\lambda}(\cdot,\cdot)$, defines a metric on the space $\mathscr{Q}_p^{\lambda}(\mathscr{X},\mathscr{Y})$.*

The proof is provided in Appendix A.

Remark 3.4. Our construction of the kernel space (3.1) and the metric (3.2) are related to the ideas used in [29] for scenario trees, and refined in [23, section 3] for Markov systems. In our notation, the authors of [23] propose the metric $\mathbb{D}_p(Q,\widetilde{Q})=\sup_{x\in\mathscr{X}}\frac{1}{\psi(x)}W_p(Q(\cdot|x),\widetilde{Q}(\cdot|x))$, with a gauge function $\psi:\mathscr{X}\to[1,\infty)$. If $\psi(\cdot)\equiv 1$,

we have $\mathscr{W}_p^{\lambda}(Q,\widetilde{Q}) \le \mathbb{D}_p(Q,\widetilde{Q})$. The uniformity (relative to the gauge function) of the approximation over all states $x \in \mathscr{X}$ is most suitable for situations when nothing is known about the distribution of x. In our approximation method in the next section, the marginal λ is not arbitrary, but it closely approximates the marginal distribution of the state in the original system. Thanks to that, the use of the metric (3.2) allows for controlling the propagation of errors in the backward system (1.2). It also eliminates the need to work with gauge functions in unbounded spaces.

For a kernel $Q \in \mathscr{Q}_p(\mathscr{X},\mathscr{Y})$, and every $\lambda \in \mathscr{P}_p(\mathscr{X})$ the measure $\lambda \circ Q$ is an element of $\mathscr{P}_p(\mathscr{Y})$, because

$$\begin{aligned}\int_{\mathscr{Y}} d(y,y_0)^p\,(\lambda\circ Q)(\mathrm{d}y) &= \int_{\mathscr{X}}\int_{\mathscr{Y}} d(y,y_0)^p\;Q(\mathrm{d}y|x)\,\lambda(\mathrm{d}x)\\ &\le C(Q)\int_{\mathscr{X}}\left(1+d(x,x_0)^p\right)\lambda(\mathrm{d}x) < \infty.\end{aligned}$$

In a similar way, the measure $\lambda \circledast Q \in \mathscr{P}_p(\mathscr{X}\times\mathscr{Y})$, because

$$\begin{aligned}&\int_{\mathscr{X}}\int_{\mathscr{Y}}\left[d(x,x_0)^p + d(y,y_0)^p\right] Q(\mathrm{d}y|x)\,\lambda(\mathrm{d}x)\\ &\quad= \int_{\mathscr{X}}\left[d(x,x_0)^p + \int_{\mathscr{Y}} d(y,y_0)^p\;Q(\mathrm{d}y|x)\right]\lambda(\mathrm{d}x)\\ &\quad\le (C(Q)+1)\int_{\mathscr{X}}\left(1+d(x,x_0)^p\right)\lambda(\mathrm{d}x) < \infty.\end{aligned}$$

The integrated transportation distance provides an upper bound on the distances between two mixture distributions and between two composition distributions.

THEOREM 3.5. *For all $\lambda \in \mathscr{P}_p(\mathscr{X})$ and all $Q,\widetilde{Q} \in \mathscr{Q}_p^{\lambda}(\mathscr{X},\mathscr{Y})$,*

$$\mathscr{W}_p^{\lambda}(Q,\widetilde{Q}) \ge W_p(\lambda\circledast Q, \lambda\circledast\widetilde{Q}) \ge W_p(\lambda\circ Q, \lambda\circ\widetilde{Q}). \tag{3.3}$$

The proof is provided in Appendix A.

The inequalities in Theorem 3.5 may be strict, as illustrated in the example of $\mathscr{X}=\{0,\varepsilon\}$ with $\varepsilon\in(0,1)$, $\mathscr{Y}=\{0,1\}$, $\lambda=(1/2,1/2)$, $Q(\cdot|x)=\delta_{\{\mathrm{sign}(x)\}}$, and $\widetilde{Q}(\cdot|x)=\delta_{\{1-\mathrm{sign}(x)\}}$, in which $\mathscr{W}_p^{\lambda}(Q,\widetilde{Q})=1$, $W_p(\lambda\circ Q,\lambda\circ\widetilde{Q})=0$, and $W_p(\lambda\circledast Q,\lambda\circledast\widetilde{Q})=\varepsilon$.

We can define a topology of weak convergence in the space $\mathscr{Q}_p^{\lambda}(\mathscr{X},\mathscr{Y})$.

DEFINITION 3.6. *The sequence of kernels $\{Q_k\}$ converges weakly to Q in $\mathscr{Q}_p^{\lambda}(\mathscr{X},\mathscr{Y})$, where $\lambda\in\mathscr{P}_p(\mathscr{X})$ if for every continuous function $f:\mathscr{X}\times\mathscr{Y}\to\mathbb{R}$ such that $|f(x,y)|\le 1+d\,(y_0,y)^p\;\forall\,(x\in\mathscr{X},\,y\in\mathscr{Y})$,*

$$\int_{\mathscr{X}}\int_{\mathscr{Y}} f(x,y)\,Q_k(dy|x)\,\lambda(dx) \longrightarrow \int_{\mathscr{X}}\int_{\mathscr{Y}} f(x,y)\,Q(dy|x)\,\lambda(dx).$$

This entails that $\lambda\circledast Q_k \rightharpoonup \lambda\circledast Q$, and, due to Definition 2.4(i), $\lambda\circ Q^k \xrightarrow{p} \lambda\circ Q$. The latter property is essential to our approximation scheme, because it allows us to derive the convergence of integrals or other functionals of the mixture distributions in the space $\mathscr{P}_p(\mathscr{X})$. It also implies that $\mathscr{W}_p^{\lambda}(Q_k,\delta_{\{y_0\}}) \to \mathscr{W}_p^{\lambda}(Q,\delta_{\{y_0\}})$ (see (A.1) in Appendix A).

The distance $\mathscr{W}_p^{\lambda}(\cdot,\cdot)$ metrizes the topology of weak convergence in $\mathscr{Q}_p^{\lambda}(\mathscr{X},\mathscr{Y})$.

THEOREM 3.7. *Let $\mathscr{X}$ and $\mathscr{Y}$ be Polish spaces, $p \in [1,\infty)$, and $\lambda \in \mathscr{P}_P(\mathscr{X})$. Then the following statements are equivalent:* (i) $Q_k \to Q$ *weakly in* $\mathscr{Q}_p^\lambda(\mathscr{X},\mathscr{Y})$; (ii) $\mathscr{W}_p^\lambda(Q_k, Q) \to 0$.

The proof is provided in Appendix A.

By the triangle inequality, we obtain the following corollary.

COROLLARY 3.8. *The functional $\mathscr{W}_p^\lambda(\cdot,\cdot)$ is continuous with respect to the weak convergence in the space $\mathscr{Q}_p^\lambda(\mathscr{X},\mathscr{Y}) \times \mathscr{Q}_p^\lambda(\mathscr{X},\mathscr{Y})$.*

We can also establish an extension of the Kantorovich–Rubinstein duality.

THEOREM 3.9. *For all $Q, \widetilde{Q} \in \mathscr{Q}_1^\lambda(\mathscr{X},\mathscr{Y})$ we have*

(3.4)
$$\mathscr{W}_1^\lambda(Q,\widetilde{Q}) = \sup_{f(\cdot,\cdot)\in F} \left\{ \int_{\mathscr{X}\times\mathscr{Y}} f(x,y)\,(\lambda \circledast Q)(dx\ dy) - \int_{\mathscr{X}\times\mathscr{Y}} f(x,y)(\lambda \circledast \widetilde{Q})(dx\ dy) \right\},$$

where F is the set of measurable functions on $\mathscr{X}\times\mathscr{Y}$ such that $\|f(x,\cdot)\|_{Lip} \le 1$ for λ-almost all $x \in \mathscr{X}$. With no loss of generality, we may also assume that $f(\cdot, y_0) \equiv 0$, for all $f \in F$.

The proof is provided in Appendix A.

4. Approximate risk evaluation in Markov systems. Our objective in this section is to propose and analyze a method for approximating forward–backward Markov systems which are described by (1.1)–(1.2), with the use of the integrated transportation distance as the criterion for constructing the approximation and a measure of its accuracy. Throughout this section, the parameter $p \in [1,\infty)$ is fixed.

The method proceeds in stages for $t = 0, 1, \dots, T$. At each stage t, for all $\tau = 0, \dots, t-1$, we already have approximate transition kernels $\widetilde{Q}_\tau : \mathscr{X} \to \mathscr{P}_p(\mathscr{X})$, $\tau = 0, \dots, t-1$. These kernels define the approximate marginal distribution

$$\widetilde{\lambda}_t = \lambda_0 \circ \widetilde{Q}_0 \circ \widetilde{Q}_1 \circ \cdots \circ \widetilde{Q}_{t-1} = \widetilde{\lambda}_{t-1} \circ \widetilde{Q}_{t-1}. \tag{4.1}$$

We also have the subspaces $\mathscr{X}_\tau \subset \mathscr{X}$ as $\mathscr{X}_\tau = \operatorname{supp}(\widetilde{\lambda}_\tau)$, $\tau = 0, 1, \dots, t$. For $t = 0$, $\widetilde{\lambda}_0 = \lambda_0$, and $\mathscr{X}_0 = \operatorname{supp}\lambda_0$.

At the stage t, we construct a kernel $\widetilde{Q}_t : \mathscr{X}_t \to \mathscr{P}_p(\mathscr{X})$ such that

$$\mathscr{W}_p^{\widetilde{\lambda}_t}(Q_t, \widetilde{Q}_t) \le \Delta_t. \tag{4.2}$$

If $t < T-1$, we increase t by one, and continue; otherwise, we stop. Observe that the approximate marginal distribution $\widetilde{\lambda}_t$ is well-defined at each step of this abstract scheme.

We then solve the approximate version of the risk evaluation algorithm (1.2), with the true kernels Q_t replaced by the approximate kernels $\widetilde{Q}_t$, $t = 0, \dots, T-1$:

$$\widetilde{v}_t(x) = c_t(x) + \sigma_t\big(x, \widetilde{Q}_t(x), \widetilde{v}_{t+1}(\cdot)\big), \quad x \in \mathscr{X}_t, \quad t = 0, 1, \dots, T-1; \tag{4.3}$$

we assume that $\widetilde{v}_T(\cdot) \equiv v_T(\cdot) \equiv c_T(\cdot)$.

Our plan is to estimate the error of this evaluation in terms of the kernel errors Δ_t. For this purpose, we make the following general assumptions:

(A1) For every $t = 0, 1, \dots, T-1$ and for every $x \in \mathscr{X}_t$, the operator $\sigma_t(x, \cdot, v_{t+1})$ is Lipschitz continuous with respect to the metric $W_p(\cdot,\cdot)$ with the constant L_t:

$$\begin{aligned} &\big|\sigma_t\big(x,\mu,v_{t+1}(\cdot)\big) - \sigma_t\big(x,\nu,v_{t+1}(\cdot)\big)\big| \\ &\quad \le L_t\, W_p(\mu,\nu) \quad \forall\, \mu,\nu \in \mathscr{P}_p(\mathscr{X}). \end{aligned}$$

(A2) For every $x \in \mathscr{X}_t$ and for every $t = 0, 1, \dots, T-1$, the operator $\sigma_t(x, \widetilde{Q}_t(x), \cdot)$ is Lipschitz continuous with respect to the norm in the space $\mathscr{L}_p(\mathscr{X}, \mathscr{B}(\mathscr{X}), \widetilde{Q}_t(x))$ with the constant K_t:

$$\begin{aligned}&\left|\sigma_t\big(x, \widetilde{Q}_t(x), v(\cdot)\big) - \sigma_t\big(x, \widetilde{Q}_t(x), w(\cdot)\big)\right| \le K_t \, \|v - w\|_p \\ &\qquad \forall\, v, w \in \mathscr{L}_p(\mathscr{X}, \mathscr{B}(\mathscr{X}), \widetilde{Q}_t(x)).\end{aligned}$$

These are fairly schematic conditions, but they are exactly what we need for the analysis below. After the theorem, we discuss several important cases, in which these conditions are satisfied.

THEOREM 4.1. *If assumptions* (A1)–(A2) *are satisfied, then for all* $t = 0, \dots, T-1$ *we have*

$$\left(\int_{\mathscr{X}} \left|\widetilde{v}_t(x) - v_t(x)\right|^p \widetilde{\lambda}_t(dx)\right)^{1/p} \le \sum_{\tau=t}^{T-1} L_\tau \left(\prod_{j=t}^{\tau-1} K_j\right) \Delta_\tau. \tag{4.4}$$

In particular, for $t = 0$,

$$\left|\widetilde{v}_0(x_0) - v_0(x_0)\right| \le \sum_{\tau=0}^{T-1} L_\tau \left(\prod_{j=0}^{\tau-1} K_j\right) \Delta_\tau. \tag{4.5}$$

Proof. First, we prove by induction backward in time that for all $t = 0, 1, \dots, T-1$ and all $x \in \mathscr{X}_t$ we have

$$\left|\widetilde{v}_t(x) - v_t(x)\right| \le \sum_{\tau=t}^{T-1} L_\tau \left(\prod_{j=t}^{\tau-1} K_j\right) \mathscr{W}_p^{\delta_x \circ \widetilde{Q}_t \circ \cdots \circ \widetilde{Q}_{\tau-1}}(\widetilde{Q}_\tau, Q_\tau). \tag{4.6}$$

At the time $t = T-1$, assumption (A1) yields the inequality

$$\begin{aligned}\left|\widetilde{v}_{T-1}(x) - v_{T-1}(x)\right| &\le \left|\sigma_{T-1}\big(x, \widetilde{Q}_{T-1}(x), v_T(\cdot)\big) - \sigma_{T-1}\big(x, Q_{T-1}(x), v_T(\cdot)\big)\right| \\ &\le L_{T-1} \, W_p(\widetilde{Q}_{T-1}(x), Q_{T-1}(x)) = L_{T-1} \, \mathscr{W}_p^{\delta_x}(\widetilde{Q}_{T-1}, Q_{T-1}),\end{aligned}$$

which is the same as (4.6) for $T-1$. Supposing (4.6) is true for t, we verify it for $t-1$. Using assumptions (A1) and (A2) we obtain

$$\begin{aligned}&\left|\widetilde{v}_{t-1}(x) - v_{t-1}(x)\right| \\ &\quad \le \left|\sigma_{t-1}\big(x, \widetilde{Q}_{t-1}(x), v_t(\cdot)\big) - \sigma_{t-1}\big(x, Q_{t-1}(x), v_t(\cdot)\big)\right| \\ &\qquad + \left|\sigma_{t-1}\big(x, \widetilde{Q}_{t-1}(x), \widetilde{v}_t(\cdot)\big) - \sigma_{t-1}\big(x, \widetilde{Q}_{t-1}(x), v_t(\cdot)\big)\right| \\ &\quad \le L_{t-1} \, W_p(\widetilde{Q}_{t-1}(x), Q_{t-1}(x)) + K_{t-1}\left(\int_{\mathscr{X}} \left|\widetilde{v}_t(y) - v_t(y)\right|^p \widetilde{Q}_{t-1}(\mathrm{d}y|x)\right)^{1/p}.\end{aligned}$$

The substitution of (4.6) and the application of the Minkowski inequality yield

$$\begin{aligned}&\left|\widetilde{v}_{t-1}(x) - v_{t-1}(x)\right| \le L_{t-1} \, \mathscr{W}_p^{\delta_x}(\widetilde{Q}_{t-1}, Q_{t-1}) \\ &\quad + K_{t-1} \sum_{\tau=t}^{T-1} L_\tau \left(\prod_{j=t}^{\tau-1} K_j\right) \left(\int_{\mathscr{X}} \left[\mathscr{W}_p^{\delta_y \circ \widetilde{Q}_t \circ \cdots \circ \widetilde{Q}_{\tau-1}}(\widetilde{Q}_\tau, Q_\tau)\right]^p \widetilde{Q}_{t-1}(\mathrm{d}y|x)\right)^{1/p}.\end{aligned}$$

Observing that

(4.7)
$$\int_{\mathscr{X}} \big[\mathscr{W}_p^{\delta_y\circ\widetilde{Q}_t\circ\cdots\circ\widetilde{Q}_{\tau-1}}(\widetilde{Q}_\tau, Q_\tau)\big]^p\, \widetilde{Q}_{t-1}(\mathrm{d}y|x) = \big[\mathscr{W}_p^{\delta_x\circ\widetilde{Q}_{t-1}\circ\widetilde{Q}_t\circ\cdots\circ\widetilde{Q}_{\tau-1}}(\widetilde{Q}_\tau, Q_\tau)\big]^p,$$

we can write the preceding displayed inequality as

$$\begin{aligned}\big|\widetilde{v}_{t-1}(x) - v_{t-1}(x)\big| \le{}& L_{t-1}\,\mathscr{W}_p^{\delta_x}(\widetilde{Q}_{t-1}, Q_{t-1})\\ &+ K_{t-1}\sum_{\tau=t}^{T-1} L_\tau\left(\prod_{j=t}^{\tau-1} K_j\right)\mathscr{W}_p^{\delta_x\circ\widetilde{Q}_{t-1}\circ\widetilde{Q}_t\circ\cdots\circ\widetilde{Q}_{\tau-1}}(\widetilde{Q}_\tau, Q_\tau),\end{aligned}$$

which is the same as (4.6) for $t-1$. By induction, (4.6) is true for all t.

The formula (4.4) follows now by integrating the right-hand side of (4.6) and using the identity

$$\int_{\mathscr{X}} \big[\mathscr{W}_p^{\delta_x\circ\widetilde{Q}_t\circ\cdots\circ\widetilde{Q}_{\tau-1}}(\widetilde{Q}_\tau, Q_\tau)\big]^p\, \widetilde{\lambda}_t(\mathrm{d}x) = \big[\mathscr{W}_p^{\widetilde{\lambda}_\tau}(\widetilde{Q}_\tau, Q_\tau)\big]^p, \quad \tau = t,\dots,T-1. \tag{4.8}$$

The formula (4.5) is a special case of (4.4) resulting from $\lambda_0 = \widetilde{\lambda}_0 = \delta_{x_0}$. ∎

Remark 4.2. At each time t, the ingredients of the formula (4.4): $\widetilde{\lambda}_t$ and Δ_t, are known. The identities (4.7) and (4.8) explain the use of the marginal $\tilde{\lambda}$ in (4.2), and the mechanism of the error control. Compared to [46, Thm. 6.2], which deals with the expected value problem in the backward system, the error estimate (4.5) is linear in the Δ_τ's, $\tau = 1,\dots,T-1$.

Assumptions (A1) and (A2) can be verified in several relevant special cases.

Example 4.1. Consider the transition risk mappings of the following form:

$$\begin{aligned}&\sigma(x,\mu,v)\\ &\quad= \mathbb{E}_\mu\Big[f_1\Big(x, \mathbb{E}_\mu\big[f_2\big(x, \mathbb{E}_\mu[\cdots f_k(x, \mathbb{E}_\mu[f_{k+1}(x, v(\cdot))], v(\cdot))]\cdots, v(\cdot)\big)\big], v(\cdot)\Big)\Big],\end{aligned} \tag{4.9}$$

where $v:\mathscr{X}\to\mathbb{R}$, $\mathbb{E}_\mu\big[f(v(\cdot))\big] = \int_{\mathscr{X}} f(v(y))\,\mu(\mathrm{d}y)$, and $f_j:\mathscr{X}\times\mathbb{R}^{m_j}\times\mathbb{R}\to\mathbb{R}^{m_{j-1}}$, $j=1,\dots,k$, with $m_0 = 1$ and $f_{k+1}:\mathscr{X}\times\mathbb{R}\to\mathbb{R}^{m_k}$. This is a fairly general class, considered in [13], which covers several risk measures, such as the mean–semideviation measure (2.1). Indeed, if $p=1$, we can write (2.1) in the form (4.9), with $k=1$ and $f_1(x,\eta,v(\cdot)) = \eta + \varkappa(x)[v(\cdot)-\eta]_+$, $f_2(x,v(\cdot)) = v(\cdot)$.

The model (4.9) also covers the mapping (1.4) in the stopping problem. In this case, $k=1$ again, and $f_1(x,\eta,v(\cdot)) = \max(r(x);\eta)$, $f_2(x,v(\cdot)) = v(\cdot)$.

Suppose the functions $f_j(x,\cdot,\cdot)$, $j=1,\dots,k$, and $f_{k+1}(x,\cdot)$ are Lipschitz continuous (it is true in both special cases mentioned above). Furthermore, let the function $v(\cdot)$ be Lipschitz continuous as well. Then, by virtue of the Kantorovich–Rubinstein duality, the functional $g_{k+1}(\mu) = \mathbb{E}_\mu\big[f_{k+1}(x,v(\cdot))\big]$ is Lipschitz continuous in the space $\mathscr{P}_1(\mathscr{X})$. In a similar way, the mapping $g_k(\mu) = \mathbb{E}_\mu\big[f_k(x, \mathbb{E}_\mu[f_{k+1}(x,v(\cdot))], v(\cdot))\big] = \mathbb{E}_\mu\big[f_k(x, g_{k+1}(\mu), v(\cdot))\big]$ is Lipschitz continuous as well. Proceeding in this way, we conclude that assumption (A1) is satisfied with $p=1$, as long as the optimal value functions $v_t(\cdot)$, $t=1,\dots,T$, are Lipschitz continuous.

Consider assumption (A2). With a fixed measure μ (corresponding to $\widetilde{Q}_t(x)$ in (A2)), we observe that the functional $\varphi_{k+1}(v) = \mathbb{E}_\mu\big[f_{k+1}(x,v(\cdot))\big]$ is Lipschitz continuous in the space $\mathscr{L}_1(\mathscr{X},\mathscr{B}(\mathscr{X}),\mu)$. This, in turn, implies that the functional

$$\varphi_k(v) = \mathbb{E}_\mu\big[f_k(x, \mathbb{E}_\mu[f_{k+1}(x,v(\cdot))], v(\cdot))\big] = \mathbb{E}_\mu\big[f_k(x, \varphi_{k+1}(v), v(\cdot))\big]$$

is Lipschitz continuous in $\mathscr{L}_1(\mathscr{X},\mathscr{B}(\mathscr{X}),\mu)$. Proceeding in a similar way, we conclude that the assumption (A2) is satisfied as well.

Example 4.2. Consider the transition risk mapping (2.2) derived from the Average Value at Risk. If $v(\cdot)$ is Lipschitz continuous, then the mapping $\mu \mapsto \mathrm{AVaR}_\alpha(x,\mu,v(\cdot))$ is Lipschitz continuous on the space $\mathscr{P}_1(\mathscr{X})$. Thus, assumption (A1) is satisfied with $p=1$. Furthermore, for a fixed μ, the mapping $v(\cdot)\mapsto \mathbb{E}_\mu\big[\max(0,v(y)-\eta)\big]$ is Lipschitz continuous (with the modulus 1) in the space $\mathscr{L}_1(\mathscr{X},\mathscr{B}(\mathscr{X}),\mu)$. Indeed, suppose η_v achieves the infimum in (2.2). Then

$$\begin{aligned}&\mathrm{AVaR}_\alpha(x,\mu,w(\cdot)) - \mathrm{AVaR}_\alpha(x,\mu,v(\cdot))\\ &\quad\le \frac{1}{\alpha}\mathbb{E}_\mu\big[\max(0,w(y)-\eta_v)\big] - \frac{1}{\alpha}\mathbb{E}_\mu\big[\max(0,v(y)-\eta_v)\big] \le \frac{1}{\alpha}\|w-v\|_1.\end{aligned}$$

Reversing the roles of v and w we obtain the Lipschitz continuity of (2.2) on the space $\mathscr{L}_1(\mathscr{X},\mathscr{B}(\mathscr{X}),\mu)$. If the infimum is not achieved, which may happen for $\alpha=1$, then $\mathrm{AVaR}_1(x,\mu,v(\cdot)) = \mathbb{E}_\mu[v(\cdot)]$ and the Lipschitz continuity is evident. Therefore, assumption (A2) is satisfied.

The last example allows for deriving the Lipschitz continuity in the space $\mathscr{P}_1(\mathscr{X})$ of a broad class of coherent risk mappings in the spectral form (2.3), or, more generally, enjoying the Kusuoka representation [27]. We refer the reader to [14, Thm. 6.5] for the details.

Example 4.3. Consider now the mean–semideviation mapping (2.1) for $p>1$. By [14, Lem. 6.6], if $v(\cdot)$ is Lipschitz continuous, then the functional $\mu\mapsto \mathrm{msd}_p(x,\mu,v)$ is Lipschitz continuous on the space $\mathscr{P}_p(\mathscr{X})$. Thus, assumption (A1) is satisfied.

Furthermore, for a fixed μ, the continuity of the mapping $v\mapsto \mathrm{msd}_p(x,\mu,v)$ on the space $\mathscr{L}_p(\mathscr{X},\mathscr{B}(\mathscr{X}),\mu)$ is evident, because it is a sum of a linear mapping and the norm. Thus, (A2) holds true as well.

It follows from the above examples that the assumptions (A1) and (A2) are indeed satisfied for a wide range of transition risk mappings. The Lipschitz continuity of the value functions $v_t(\cdot)$, $t=2,\dots,T$, is crucial in this context.

This can be guaranteed by a simple induction argument. Suppose each function $c_t(\cdot)$ and operator $(x,\mu)\mapsto \sigma_t(x,\mu,v_{t+1}(\cdot))$ are Lipschitz continuous in $\mathscr{X}$ and $\mathscr{X}\times\mathscr{Q}^{\lambda_t}(\mathscr{X},\mathscr{X})$, respectively, provided the function $v_{t+1}(\cdot)$ is Lipschitz continuous. Moreover, let the kernels $Q_t:\mathscr{X}\to\mathscr{P}_p(\mathscr{X})$, $t=1,\dots,T-1$, be Lipschitz continuous as well: a constant L_Q exists, such that

$$W_p(Q_t(x),Q_t(x')) \le L_{Q_t}\, d(x,x') \quad \forall\, x,x'\in\mathscr{X}\,. \tag{4.10}$$

Then the function $v_t(\cdot)$ in (1.2) is a composition of Lipschitz continuous mappings, and it is thus Lipschitz continuous. By induction, all value functions are Lipschitz continuous. Their Lipschitz constants, though, may grow exponentially with the horizon $T-t$ if $L_Q>1$. The constant L_Q is known as the *ergodicity coefficient*; see [40] and the references therein.

We can also study the accuracy of the marginal distributions $\widetilde{\lambda}_t$, $t=1,\dots,T$. First, we establish a useful continuity result.

LEMMA 4.3. *If a kernel $Q:\mathscr{X}\to\mathscr{P}_p(\mathscr{X})$ is Lipschitz continuous, then the mapping $\mu\mapsto\mu\circ Q$ is Lipschitz continuous on $\mathscr{P}_p(\mathscr{X})$ with the same modulus.*

Proof. If $\lambda(\mathrm{d}y\,\mathrm{d}y'|x,x')$ is the optimal transport plan from $Q(x)$ to $Q(x')$, then

$$W_p(Q(x),Q(x'))^p = \int_{\mathscr{X}\times\mathscr{X}} d(y,y')^p\,\lambda(\mathrm{d}y\,\mathrm{d}y'|x,x') \le L_Q^p d(x,x')^p,$$

where L_Q is the Lipschitz constant of Q. Suppose $\pi(\mathrm{d}x\,\mathrm{d}x')$ is the optimal coupling of μ and ν. Consider the transport plan $\Pi = \pi\circ\lambda$, with π considered as a marginal on $\mathscr{X}\times\mathscr{X}$, and λ as a kernel from $\mathscr{X}\times\mathscr{X}$ to $\mathscr{P}(\mathscr{X}\times\mathscr{X})$. We have

$$\begin{aligned}\Pi(A\times\mathscr{X}) &= \int_{\mathscr{X}\times\mathscr{X}}\int_{\mathscr{X}} \lambda(A,\mathrm{d}y'|x,x')\,\pi(\mathrm{d}x\,\mathrm{d}x')\\ &= \int_{\mathscr{X}\times\mathscr{X}} Q(A|x)\,\pi(\mathrm{d}x\,\mathrm{d}x') = \int_{\mathscr{X}} Q(A|x)\,\mu(\mathrm{d}x') = [\mu\circ Q](A).\end{aligned}$$

In a similar way, $\Pi(\mathscr{X}\times B)[\nu\circ Q](B)$, and thus Π is a feasible transport plan from $\mu\circ Q$ to $\nu\circ Q$. Therefore,

$$\begin{aligned}W_p\big(\mu\circ Q,\nu\circ Q\big) &\le \int_{\mathscr{X}\times\mathscr{X}} d(y,y')^p\,\Pi(\mathrm{d}y\,\mathrm{d}y')\\ &= \int_{\mathscr{X}\times\mathscr{X}}\int_{\mathscr{X}\times\mathscr{X}} d(y,y')^p\,\lambda(\mathrm{d}y\,\mathrm{d}y'|x,x')\,\pi(\mathrm{d}x\,\mathrm{d}x')\\ &\le L_Q^p\int_{\mathscr{X}\times\mathscr{X}} d(x,x')^p\,\pi(\mathrm{d}x\,\mathrm{d}x') = L_Q^p W_p(\mu,\nu)^p.\end{aligned}$$

It follows that L_Q is the Lipschitz constant of the mapping $\mu\mapsto\mu\circ Q$. ∎

We can now easily estimate the errors of the marginal distributions.

THEOREM 4.4. *If the kernels $Q_t : \mathscr{X}\to\mathscr{P}_p(\mathscr{X})$ are Lipschitz continuous with constants L_{Q_t}, then*

$$W_p(\widetilde{\lambda}_t,\lambda_t) = \sum_{\tau=1}^{t-1}\Delta_\tau\prod_{i=\tau+1}^{t-1} L_{Q_i},\quad t=1,\dots,T. \tag{4.11}$$

Proof. The estimate (4.11) is true for $t=1$. Supposing it is valid for $t-1$, we verify it for t:

$$\begin{aligned}W_p(\widetilde{\lambda}_t,\lambda_t) &= W_p(\widetilde{\lambda}_{t-1}\circ\widetilde{Q}_{t-1},\lambda_{t-1}\circ Q_{t-1})\\ &\le W_p(\widetilde{\lambda}_{t-1}\circ\widetilde{Q}_{t-1},\widetilde{\lambda}_{t-1}\circ Q_{t-1}) + W_p(\widetilde{\lambda}_{t-1}\circ Q_{t-1},\lambda_{t-1}\circ Q_{t-1})\\ &\le \mathscr{W}_p^{\widetilde{\lambda}_{t-1}}\big(\widetilde{Q}_{t-1},Q_{t-1}\big) + W_p(\widetilde{\lambda}_{t-1}\circ Q_{t-1},\lambda_{t-1}\circ Q_{t-1})\\ &\le \Delta_{t-1} + L_{Q_{t-1}}W_p(\widetilde{\lambda}_{t-1},\lambda_{t-1}).\end{aligned}$$

The substitution of (4.11) for $t-1$ yields the same estimate for t. By induction, it is true for all t. ∎

5. Kernel approximation by particles. In the general method discussed in the previous section, we iteratively constructed approximate kernels $\widetilde{Q}_t$, proceeding from $t=0$ to $t=T-1$, and we used their error estimates (4.2) to estimate the error of the risk evaluation.

Now we aim at an implementable method to realize this general scheme. The most important assumption is that the spaces $\mathscr{X}_t$, $t=0,1,\dots,T$, be *finite.* We assume that we start from $\mathscr{X}_0=\{x_0\}$. At each stage t, we aim to construct a finite

set $\mathscr{X}_{t+1} \subset \mathscr{X}$ of cardinality M_{t+1} and a kernel $\widetilde{Q}_t : \mathscr{X}_t \to \mathscr{P}(\mathscr{X}_{t+1})$ by solving the following problem:

$$\min_{\mathscr{X}_{t+1}, \widetilde{Q}_t} \mathscr{W}_p^{\widetilde{\lambda}_t}(Q_t, \widetilde{Q}_t) \quad \text{s.t.} \quad \operatorname{supp}(\widetilde{\lambda}_t \circ \widetilde{Q}_t) = \mathscr{X}_{t+1} \quad \text{and} \quad \left|\mathscr{X}_{t+1}\right| \le M_{t+1}. \tag{5.1}$$

After (approximately) solving this problem, we increase t by one and continue. Evidently, the objective function of this problem is motivated by its direct effect on the error estimates in Theorems 4.1 and 4.4.

Let us focus on effective and scalable ways for constructing an approximate solution to problem (5.1). We represent the (unknown) support of $\widetilde{\lambda}_t \circ \widetilde{Q}_t$ by $\mathscr{X}_{t+1} = \left\{z_{t+1}^j\right\}_{j=1,\dots,M_{t+1}}$ and the (unknown) transition probabilities by $\widetilde{Q}(z_{t+1}^j | z_t^s)$, $s = 1, \dots, M_n$, $j = 1, \dots, M_{n+1}$. With the use of Definition 3.2, problem (5.1) can be equivalently rewritten as

$$\begin{aligned} \min_{\mathscr{X}_{t+1}, \widetilde{Q}_t} \quad & \sum_{s=1}^{M_n} \widetilde{\lambda}_t^s W_p\big(Q_t(\cdot|z_t^s), \widetilde{Q}_t(\cdot|z_t^s)\big)^p \\ \text{s.t.} \quad & \operatorname{supp}\big(\widetilde{Q}_t(\cdot|z_t^s)\big) \subset \mathscr{X}_{t+1}, \quad s = 1, \dots, M_n, \\ & \left|\mathscr{X}_{t+1}\right| \le M_{t+1}. \end{aligned} \tag{5.2}$$

Let π_t^s be a transportation plan from $Q_t(\cdot|z_t^s)$ to $\widetilde{Q}_t(\cdot|z_t^s)$. Then it follows from the definition of the Wasserstein distance that $W_p\big(Q_t(\cdot|z_t^s), \widetilde{Q}_t(\cdot|z_t^s)\big)^p$ is the optimal value of the problem

$$\begin{aligned} \min_{\pi_t^s \ge 0} \quad & \sum_{j=1}^{M_{t+1}} \int_{\mathscr{X}} \|x - z_{t+1}^j\|^p \, \pi_t^{sj}(\mathrm{d}x) \\ \text{s.t.} \quad & \int_{\mathscr{X}} \pi_t^{sj}(\mathrm{d}x) = \widetilde{Q}_t(z_{t+1}^j | z_t^s), \quad j = 1, \dots, M_{t+1}, \\ & \sum_{j=1}^{M_{t+1}} \pi_t^{sj}(A) = Q_t(A|z_t^s) \quad \forall\, A \in \mathscr{B}(\mathscr{X}). \end{aligned} \tag{5.3}$$

The integration of problems (5.2)–(5.3) leads to a very difficult nonconvex infinite-dimensional problem which can be only solved in very special cases. To develop a tractable approach in large-scale applications, we restrict the supports of the kernels under consideration to finite sets. We may remark that the optimal quantization of probability distributions with the use of the Wasserstein metric was systematically studied in [18]. Our problem is slightly different because we want to obtain a "quantization" of kernels.

In our particle approach, for $t = 0, 1, \dots, T-1$, each distribution $Q_t(\cdot|z_t^s)$ is represented by finitely many points (particles) $\left\{x_{t+1}^{s,i}\right\}_{i \in \mathscr{I}_{t+1}^s}$, drawn independently from $Q_t(\cdot|z_t^s)$. If the state space $\mathscr{X}$ is finite-dimensional, the expected error of this approximation is well-investigated in [15, 17], as a function of the sample size $\left|\mathscr{I}_{t+1}^s\right|$, the dimension of the state space, and the moments of the distribution (see formula (5.8) below). From this point, we consider the error of this large-size discrete approximation as fixed, and we focus on constructing smaller support with as small an error to the particle distributions as possible. To this end, we introduce the sets $\mathscr{Z}_{t+1} = \left\{\zeta_{t+1}^k\right\}_{k=1,\dots,K_{t+1}}$, which are preselected potential locations of the next-stage representative states z_{t+1}^j, $j = 1, \dots, M_{t+1}$. In the simplest case, we may consider the

union of the sets of particles, $\{x_{t+1}^{s,i},\ i \in \mathscr{I}_{t+1}^s,\ s = 1, \ldots, M_t\}$ as the potential locations, but often computational efficiency requires that $K_{t+1} \ll \sum_{s=1}^{M_t} |\mathscr{I}_{t+1}^s|$. There are several heuristic ways to choose the set $\mathscr{Z}_{t+1}$ of potential points. For instance, they may be sampled independently along with successors at the particle generation step, or they may be sampled from a different distribution. In any case, we still have $M_{t+1} \ll K_{t+1}$, which makes the problem of finding the best representative points nontrivial.

Suppose temporarily the next-stage representative points $\{z_{t+1}^j\}_{j=1,\ldots,M_{t+1}}$ have been found. Then the particle version of problem (5.3) (for a fixed s) takes on the form

$$\begin{aligned} \min_{\pi_t^s \geq 0} \quad & \sum_{j=1}^{M_{t+1}} \sum_{i \in \mathscr{I}_{t+1}^s} \|x_{t+1}^{s,i} - z_{t+1}^j\|^p \, \pi_t^{s,i,j} \\ \text{s.t.} \quad & \sum_{j=1}^{M_{t+1}} \pi_t^{s,i,j} = \frac{1}{|\mathscr{I}_{t+1}^s|}, \quad i \in \mathscr{I}_{t+1}^s. \end{aligned} \tag{5.4}$$

It has a straightforward solution: find for each particle i the closest representative point, $j^*(i) = \operatorname{argmin}_{j=1,\ldots,M_{t+1}} \|x_{t+1}^{s,i} - z_{t+1}^j\|$, and set $\pi_t^{s,i,j^*(k)} = \frac{1}{|\mathscr{I}_{t+1}^s|}$; for other j, we set it to 0. The implied approximate kernel is

$$\widetilde{Q}_t(z_{t+1}^j | z_t^s) = \sum_{i \in \mathscr{I}_{t+1}^s} \pi_t^{s,i,j}, \quad s = 1, \ldots, M_t, \quad j = 1, \ldots, M_{t+1}, \tag{5.5}$$

which simply counts the particles from $\mathscr{I}_{t+1}^s$ which were assigned to z_{t+1}^j.

These considerations allow us to integrate problems (5.4) into (5.2). We introduce the binary variables

$$\gamma_k = \begin{cases} 1 & \text{if the point } \zeta_{t+1}^k \text{ has been selected to } \mathscr{X}_{t+1}, \\ 0 & \text{otherwise,} \end{cases} \quad k = 1, \ldots, K_{t+1},$$

and we rescale the transportation plans:

$$\beta^{s,i,k} = |\mathscr{I}_{t+1}^s| \pi_t^{s,i,k}, \quad s = 1, \ldots, M_t,\ i \in \mathscr{I}_{t+1}^s,\ k = 1, \ldots, K_{t+1}.$$

We obtain from (5.2) the following linear mixed-integer optimization problem:

$$\begin{aligned} \min_{\gamma,\beta} \quad & \sum_{s=1}^{M_n} \frac{\widetilde{\lambda}_t^s}{|\mathscr{I}_{t+1}^s|} \sum_{k=1}^{K_{t+1}} \sum_{i \in \mathscr{I}_{t+1}^s} \|x_{t+1}^{s,i} - \zeta_{t+1}^k\|^p \, \beta^{s,i,k} \\ \text{s.t.} \quad & \beta^{s,i,k} \leq \gamma_k, \quad s = 1, \ldots, M_t, \quad i \in \mathscr{I}_{t+1}^s, \quad k = 1, \ldots, K_{t+1}, \\ & \sum_{k=1}^{K_{t+1}} \beta^{s,i,k} = 1, \quad s = 1, \ldots, M_t, \quad i \in \mathscr{I}_{t+1}^s, \\ & \sum_{k=1}^{K_{t+1}} \gamma_k \leq M_{t+1}, \\ & \beta^{s,i,k} \in [0,1],\ \gamma_k \in \{0,1\}, \quad s = 1, \ldots, M_t,\ i \in \mathscr{I}_{t+1}^s,\ k = 1, \ldots, K_{t+1}. \end{aligned} \tag{5.6}$$

The complicating element is that the γ_k's are binary variables. However, we may solve the relaxation of (5.6) in which we require only that $\gamma_k \in [0,1]$, $k = 1, \ldots, K_{t+1}$,

while still bounding their sum by M_{t+1}. After that, we may randomly assign to fractional γ_k's values 0 or 1, by using independent Bernoulli random variables with parameters γ_k, and then resolve (5.6) with respect to the β variables only. This can be accomplished by assigning each point $x_{t+1}^{s,i}$ to the closest ζ_{t+1}^k having $\gamma_k = 1$. The implied approximate kernel is given by (5.5):

$$\widetilde{Q}_t(\zeta_{t+1}^k | z_t^s) = \frac{1}{|\mathscr{I}_{t+1}^s|} \sum_{i \in \mathscr{I}_{t+1}^s} \beta^{s,i,k}, \quad s = 1, \dots, M_t, \quad k = 1, \dots, K_{t+1}. \tag{5.7}$$

By construction, these probabilities can be positive only when $\gamma_k = 1$.

Finally, $\widetilde{\lambda}_{t+1} = \widetilde{\lambda}_t \circ \widetilde{Q}_t$, and the iteration continues until $t = T$.

At each stage t, the estimate of the error Δ_t in (4.2) can be computed: it is the sum of the pth root of the objective value of (5.6) and the particle distribution error. Denoting by $\widehat{Q}_t$ the approximate kernel defined by all the particles sampled, due to Theorem 3.3, we have

$$\mathscr{W}_p^{\tilde{\lambda}_t}(\widetilde{Q}_t, Q_t) \le \mathscr{W}_p^{\tilde{\lambda}_t}(\widetilde{Q}_t, \widehat{Q}_t) + \mathscr{W}_p^{\tilde{\lambda}_t}(\widehat{Q}_t, Q_t).$$

To recall a bound on the expected value of the second term, we assume that the state space is finite-dimensional, and that for each point z_t^s the measure $Q_t(\cdot | z_t^s)$ has a finite moment m_u for some $u > p$. The following inequality due to [15, 17] is true for all $N = |\mathscr{I}_{t+1}^s|$:

$$\mathbb{E}\big[W_p\big(\widehat{Q}_t(\cdot | z_t^s), Q_t(\cdot | z_t^s)\big)\big] \le C m_u^{p/u} \times \begin{cases} N^{-1/2} + N^{-(u-p)/u} & \text{if } p > n/2 \text{ and } u \ne 2p, \\ N^{-1/2} \ln(1+N) + N^{-(u-p)/u} & \text{if } p = n/2 \text{ and } u \ne 2p, \\ N^{-p/n} + N^{-(u-p)/u} & \text{if } p < n/2 \text{ and } u \ne \frac{n}{n-p}, \end{cases} \tag{5.8}$$

where $n = \dim(\mathscr{X})$, and C is a constant depending only on p, u, and n. If the number N of particles sampled from each $Q_t(\cdot | z_t^s)$ is the same for all $s = 1, \dots, M_t$, the expected distance $\mathbb{E}\big[\mathscr{W}_p^{\tilde{\lambda}_t}(\widehat{Q}_t, Q_t)\big]$ is bounded by the expression (5.8) as well.

Our procedure adds to this error a fully controllable part $\mathscr{W}_p^{\tilde{\lambda}_t}(\widetilde{Q}_t, \widehat{Q}_t)$ by constructing a set of representative points z_{t+1}^j, $j = 1, \dots, M_{t+1}$, each of which may serve as a "descendant" of multiple points z_t^s. Our experience shows that for large M_t the total number of these points, M_{t+1}, is comparable to M_t, and thus much smaller than the number of particles NM_t. As a result, the total number of representative points, while still exponential in the dimension of the state space, grows only linearly with the number of time steps. We elaborate on it in the next section.

6. Numerical illustration. Consider n stocks $\{S_t^{(i)}\}$, $i = 1, \dots, n$, in an arbitrage-free and complete market, following (under the risk-neutral probability measure $\mathbb{Q}$) the equations

$$\mathrm{d}S_t^{(i)} = r S_t^{(i)} \, \mathrm{d}t + \sigma^{(i)} S_t^{(i)} \, \mathrm{d}W_t^{\mathbb{Q}}, \quad i = 1, \dots, n, \quad t \in [0, T]. \tag{6.1}$$

Here, $\{W_t^{\mathbb{Q}}\}$ is an n-dimensional Brownian motion under $\mathbb{Q}$, r is the risk-free interest rate, and $\sigma^{(i)}$ is the n-dimensional (row) vector of volatility coefficients of stock i. We assume that the coefficients r and σ are constant, but our methodology is applicable to problems with varying coefficients as well.

An option is one of the most common financial derivatives that give buyers the right, but not the obligation, to buy or sell an underlying asset at an agreed-upon

price during a certain period of time. The American option is the type of option that can be exercised anytime, prior to the maturity time T. If exercised at time t, the option pays $\Phi(S_t)$ for some known function $\Phi : \mathbb{R}^n \to [0, +\infty)$. The price of the American option is given by the optimal value of the stopping problem:

$$V_t(x) = \sup_{\substack{\tau-\text{stopping time} \\ t \le \tau \le T}} E^{\mathbb{Q}}\left[e^{-r(\tau-t)}\Phi\left(S_\tau\right) \middle| S_t = x\right], \quad x \in R^n, \tag{6.2}$$

In our example, $\Phi(S_t) = \max\left(0, K - \sum_{i=1}^n w_i S_t^{(i)}\right)$ is the value of the basket put option, with the basket weights w_i, $i = 1, \ldots, n$.

To develop a numerical scheme for approximating the option value, we first partition the time interval $[0, T]$ into short intervals of length $\Delta t = T/N$: $\Gamma_N = \{t_i = i\Delta t : i = 0, 1, \ldots, N\}$.

With the exercise times restricted to Γ_N, we approximate the option value by

$$V_t^{(N)}(x) = \sup_{\substack{\tau-\text{stopping time} \\ \tau \in \Gamma_N}} E^{\mathbb{Q}}\left[e^{-r(\tau-t)}\Phi\left(S_\tau\right) \middle| S_t = x\right], \quad t \in \Gamma_N, \quad x \in R^n. \tag{6.3}$$

We view $V_t^{(N)}(x)$ as an approximation to the actual American option price when N increases to infinity. It satisfies the following dynamic programming equations:

$$V_{t_N}^{(N)}(x) = \Phi(x), \quad x \in \mathbb{R}^n,$$
$$V_{t_i}^{(N)}(x) = \max\left\{\Phi(x), E^{\mathbb{Q}}\left[e^{-r\Delta t} V_{t_{i+1}}^{(N)}\left(S_{t_{i+1}}\right) \middle| S_{t_i} = x\right]\right\}, \quad i = 0, 1, \ldots, N-1,$$

which is a special case of (1.4). We apply two methods to simulate the movements of stocks and compare the values of the approximation of the American basket option. The first method is the grid point selection method based on the integrated transportation distance. For every time step t_i, we select the representative point(s) $z_i^j, j = 1, \ldots, M_i$, to represent the state space at time t_i, as outlined in section 5. We compare the above method with the binomial tree method, a lattice method based on the random walk approximation to the Brownian motion. Between the start and expiration dates, each grid point in a lattice represents the state of the system at a given time step. Starting from the grid points at the final time step, the prices at the preceding grid points are computed in a backward direction. Since every node of the lattice has 2^n descendants, the number of lattice points in the binomial tree method grows exponentially, as the number of time steps increases. In the grid point selection method, the total number of representative points grows approximately at a linear rate with respect to the total number of time steps N.

In the initial experiment, both methods are applied to evaluate the American basket put option with $n = 2$ and the payoff function for the American basket put is $\Phi_p(S_t) = \max(K - \sum_{i=1}^n w_i S_t^{(i)}, 0)$, where w_i is the percentage of stock i held in the portfolio and K is the strike price. The values of the parameters are $S_0 = [10, 10]$, $r = 0.03$, $K = 10$, $w = (0.5, 0.5)$, and $T = 1$. The volatility coefficients were $\sigma = \begin{bmatrix} 0.5 & -0.2; & -0.2 & 0.5 \end{bmatrix}$.

Table 1 compares the approximated option prices using the grid point selection method and the binomial tree method. Figure 1 summarizes the convergence of the American basket put option as the number of time steps increases. Moreover, the upper bound of the error in estimating value function is determined by the integrated transportation distance at every time stage. For the grid point selection method, we have computed the the integrated transportation distances for the first few time stages. $\Delta_0 = 0.239$, $\Delta_1 = 0.211$, $\Delta_2 = 0.192$, $\Delta_3 = 0.190$, and $\Delta_4 = 0.181$.

TABLE 1

Convergence of the American basket put option prices with respect to the number of time discretization steps.

N	Grid	Binomial
1	0.832	0.824
2	0.869	1.009
5	0.880	0.896
10	0.880	0.873
25	0.884	0.887
50	0.887	0.889

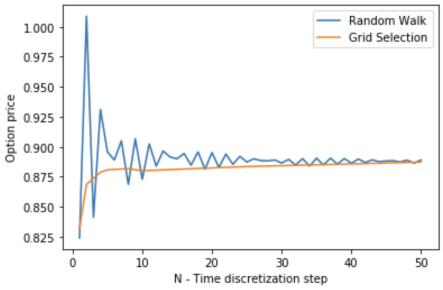


FIG. 1. *The approximate value of the American basket put option as a function of the number of time steps.*

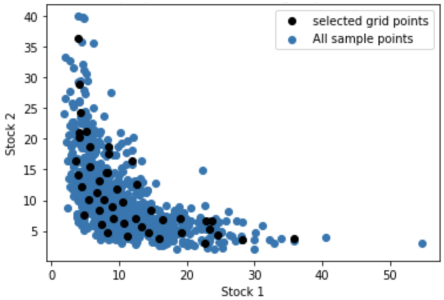
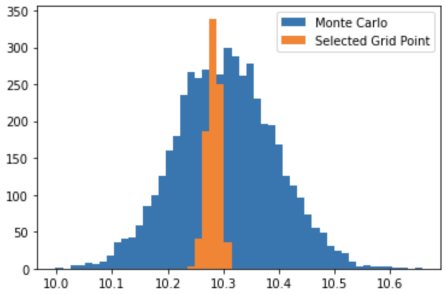
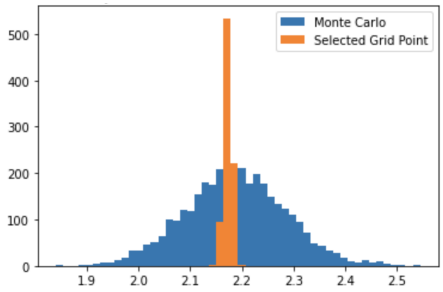


FIG. 2. *All sample points (blue, $N = 1000$) versus selected grid points (black, $M = 46$). (Figure in color online.)*

In order to demonstrate the stability of the approximate prices using our grid point selection method, we will also apply this method on risk measures at $T = 1$. A practically relevant law-invariant coherent measure of risk is the mean–semideviation of order $p \geq 1$, defined in (2.1). Figure 2 illustrates an example of selecting grid points from the simulated stock prices at $T = 1$. In the grid selection method, we set the number of grid points to be around 400 selected out of 1000 randomly sampled points. We repeated the experiment over 900 times and recorded the mean and semideviation estimates. In the Monte Carlo experiment, we sampled 1000 points and evaluated plug-in estimates of the mean and the semideviation; this experiment was repeated 5000 times. In Figure 3a, we plot the histograms of the estimated expected values, and in Figure 3b, the histograms of the estimated semideviations. It is obvious that the approximated values from the grid selection method are more stable than those from the Monte Carlo simulation.



(a) Histogram of the estimated expected value of the stock price.



(b) Histogram of the estimated semideviation of the stock price.

FIG. 3. *Monte Carlo simulation versus the grid selection method.*

TABLE 2

Convergence of the estimates of the American put price with respect to the number of time discretization steps N.

N	Put - grid	Put - binomial	M-grid	M-binomial
1	1.168	1.179	30343	33
2	1.188	1.223	38740	276
3	1.207	1.239	50891	1300
4	1.213	1.240	56970	4425
5	1.231	1.241	74044	12201
6	1.231	1.242	81022	29008
7	1.240	1.242	94592	61776
8	1.239	1.244	97639	120825
9	1.250	1.244	127981	220825
10	1.254	1.244	136378	381876
11	1.258	1.245	148528	630708
12	1.259	1.246	154607	1002001

In our more challenging experiment, we estimated the American put option value for a five-dimensional stock basket. The values of the parameters are $S_0 = [10, 10, 10, 10, 10]$, $r = 0.03$, $K = 10$, $w = (0.2, 0.2, 0.2, 0.2, 0, 2)$, $T = 1$, and

$$\sigma = \begin{bmatrix} 0.5 & 0.2 & 0.3 & -0.2 & 0.15 \\ 0.2 & 0.5 & -0.15 & 0.3 & 0.12 \\ 0.3 & -0.15 & 0.75 & -0.1 & 0.1 \\ -0.2 & 0.03 & -0.1 & 0.3 & 0.05 \\ 0.15 & 0.12 & 0.1 & 0.05 & 0.4 \end{bmatrix}.$$

Table 2 displays the convergence of the American put option prices as we increase the number N of time discretization points, using the grid selection method and the binomial tree method. M refers to the total number of grid points used. As shown in the table, the binomial tree method cannot go beyond $N = 12$ because the total number of grid points, M increases exponentially with N. The grid point selection method achieves similar results to that of the binomial tree method while requiring only linear growth of the total number of representative points with the number of stages.

Appendix A. Proofs of the statements in section 3.

Proof of Theorem 3.3. It is obvious that $\mathscr{W}_p^\lambda(Q,\widetilde{Q}) \geq 0$ for any $Q,\widetilde{Q} \in \mathscr{Q}_p^\lambda(\mathscr{X},\mathscr{Y})$ and $\mathscr{W}_p^\lambda(Q,\widetilde{Q}) = 0$ if and only if $Q = \widetilde{Q}$ λ-a.s.. We next verify the triangle inequality. For all $Q,Q',\widetilde{Q} \in \mathscr{Q}_p^\lambda(\mathscr{X},\mathscr{Y})$, by the triangle inequality for $W_p(\cdot,\cdot)$ and then by the Minkowski inequality, we obtain

$$\begin{aligned}\mathscr{W}_p^\lambda(Q,\widetilde{Q}) &\leq \left(\int_{\mathscr{X}} \left[W_p(Q(\cdot|x),Q'(\cdot|x)) + W_p(Q'(\cdot|x),\widetilde{Q}(\cdot|x))\right]^p \lambda(\mathrm{d}x)\right)^{1/p}\\ &\leq \left(\int_{\mathscr{X}} \left[W_p(Q(\cdot|x),Q'(\cdot|x))\right]^p \lambda(\mathrm{d}x)\right)^{1/p}\\ &\quad + \left(\int_{\mathscr{X}} \left[W_p(Q'(\cdot|x),\widetilde{Q}(\cdot|x))\right]^p \lambda(\mathrm{d}x)\right)^{1/p}\\ &= \mathscr{W}_p^\lambda(Q,Q') + \mathscr{W}_p^\lambda(Q',\widetilde{Q}).\end{aligned}$$

Furthermore, setting $Q'(\cdot|x) = \delta_{\{y_0\}}(\cdot)$ and using (3.1), we get

$$\begin{aligned}\left[\mathscr{W}_p^\lambda(Q,\delta_{\{y_0\}})\right]^p &= \int_{\mathscr{X}} \left[W_p(Q(\cdot|x),\delta_{\{y_0\}})\right]^p \lambda(\mathrm{d}x)\\ &= \int_{\mathscr{X}}\int_{\mathscr{Y}} d(y,y_0)^p\, Q(\mathrm{d}y|x)\,\lambda(\mathrm{d}x)\\ &\leq C(Q)\int_{\mathscr{X}} \left(1 + d(x,x_0)^p\right)\lambda(\mathrm{d}x) < \infty,\end{aligned} \tag{A.1}$$

which proves the finiteness of $\mathscr{W}_p^\lambda(Q,\widetilde{Q})$ if $\lambda \in \mathscr{P}_p(\mathscr{X})$. ∎

Proof of Theorem 3.5. From (3.2) we obtain

$$\left[\mathscr{W}_p^\lambda(Q,\widetilde{Q})\right]^p = \int_{\mathscr{X}}\int_{\mathscr{Y}\times\mathscr{Y}} d(y,y')^p\,\pi^*(\mathrm{d}y,\mathrm{d}y'|x)\,\lambda(\mathrm{d}x),$$

where $\pi^*(\cdot,\cdot|x)$ is the optimal transportation plan between $Q(\cdot|x)$ and $\widetilde{Q}(\cdot|x)$. By the measurable selection theorem, the mapping $x \mapsto \pi^*(\cdot,\cdot|x)$ may be viewed as a kernel from $\mathscr{X}$ to $\mathscr{P}(\mathscr{Y}\times\mathscr{Y})$.

Now, we construct from π^* a transportation plan $\Pi^* \in \mathscr{P}\big((\mathscr{X}\times\mathscr{Y})\times(\mathscr{X}\times\mathscr{Y})\big)$: for all $A_X, B_X \in \mathscr{B}(\mathscr{X})$ and all $A_Y, B_Y \in \mathscr{B}(\mathscr{Y})$ we set

$$\Pi^*\big((A_X\times A_Y)\times(B_X\times B_Y)\big) = \int_{A_X\cap B_X} \pi^*(A_Y\times B_Y|x)\,\lambda(\mathrm{d}x). \tag{A.2}$$

Setting $B_X = \mathscr{X}$ and $B_Y = \mathscr{Y}$ we obtain the marginal of Π^*:

$$\begin{aligned}\Pi^*\big((A_X\times A_Y)\times(\mathscr{X}\times\mathscr{Y})\big) &= \int_{A_X} \pi^*(A_Y\times\mathscr{Y}|x)\,\lambda(\mathrm{d}x)\\ &= \int_{A_X} Q(A_Y|x)\,\lambda(\mathrm{d}x) = (\lambda\circledast Q)(A_X\times A_Y).\end{aligned}$$

The second marginal is verified in an analogous way and thus Π^* moves $\lambda\circledast Q$ to $\lambda\circledast\widetilde{Q}$. Then, by virtue of (A.2),

$$\begin{aligned}W_p(\lambda\circledast Q,\lambda\circledast\widetilde{Q})^p &\leq \int_{(\mathscr{X}\times\mathscr{Y})\times(\mathscr{X}\times\mathscr{Y})} d\big((x,y),(x',y')\big)^p\,\Pi^*(\mathrm{d}x\,\mathrm{d}y,\mathrm{d}x'\,\mathrm{d}y')\\ &= \int_{\mathscr{Y}\times\mathscr{Y}} d(y,y')^p \int_{\mathscr{X}} \pi^*(\mathrm{d}y,\mathrm{d}y'|x)\,\lambda(\mathrm{d}x) = \left[\mathscr{W}_p^\lambda(Q,\widetilde{Q})\right]^p,\end{aligned}$$

which verifies the left inequality in (3.3).

Next, for the optimal transportation plan $\widehat{\Pi} \in \mathscr{P}\big((\mathscr{X} \times \mathscr{Y}) \times (\mathscr{X} \times \mathscr{Y})\big)$, with marginals $\lambda \circledast Q$ and $\lambda \circledast \widetilde{Q}$, we construct a transportation plan $\hat{\pi} \in \mathscr{P}(\mathscr{Y} \times \mathscr{Y})$ as

$$\hat{\pi}(A_Y \times B_Y) = \widehat{\Pi}\big((\mathscr{X} \times A_Y) \times (\mathscr{X} \times B_Y)\big) \quad \forall\, A_Y, B_Y \in \mathscr{B}(\mathscr{Y}).$$

Then

$$\hat{\pi}(A_Y \times \mathscr{Y}) = \widehat{\Pi}\big((\mathscr{X} \times A_Y) \times (\mathscr{X} \times \mathscr{Y})\big) = [\lambda \circledast Q](\mathscr{X} \times A_Y) = [\lambda \circ Q](A_Y).$$

The second marginal is verified analogously and thus $\hat{\pi}$ moves $\lambda \circ Q$ to $\lambda \circ \widetilde{Q}$. Therefore,

$$\begin{aligned} W_p(\lambda \circ Q, \lambda \circ \widetilde{Q})^p &\le \int_{\mathscr{Y} \times \mathscr{Y}} d(y, y')^p \, \hat{\pi}(\mathrm{d}y, \mathrm{d}y') \\ &= \int_{(\mathscr{X} \times \mathscr{Y}) \times (\mathscr{X} \times \mathscr{Y})} d(y, , y')^p \, \widehat{\Pi}(\mathrm{d}x \, \mathrm{d}y, \mathrm{d}x' \, \mathrm{d}y') \\ &\le \int_{(\mathscr{X} \times \mathscr{Y}) \times (\mathscr{X} \times \mathscr{Y})} d\big((x, y), (x', y')\big)^p \, \widehat{\Pi}(\mathrm{d}x \, \mathrm{d}y, \mathrm{d}x' \, \mathrm{d}y') \\ &= W_p(\lambda \circledast Q, \lambda \circledast \widetilde{Q})^p. \end{aligned}$$

which is the right inequality in (3.3). ∎

Proof of Theorem 3.7. The implication (ii)⇒(i) follows from Theorem 3.5, because the first inequality in (3.3) yields $W_p(\lambda \circledast Q_k, \lambda \circledast Q) \to 0$, and thus $\lambda \circledast Q_k \xrightarrow{p} \lambda \circledast Q$, by virtue of [45, Thm. 6.9]. The latter convergence implies that Definition 3.6 is satisfied.

To prove the implication (i)⇒(ii), we adopt some ideas of the proof of [45, Thm. 6.9]. From (3.2) we obtain

$$\big[\mathscr{W}_p^{\lambda}(Q_k, Q)\big]^p = \int_{\mathscr{X}} \int_{\mathscr{Y} \times \mathscr{Y}} d(y, y')^p \, \pi_k(\mathrm{d}y, \mathrm{d}y' | x) \, \lambda(\mathrm{d}x),$$

where $\pi_k(\cdot, \cdot | x)$ is the optimal transport plan between $Q_k(\cdot | x)$ and $Q(\cdot | x)$. By the measurable selection theorem, the mapping $x \mapsto \pi_k(\cdot, \cdot | x)$ may be viewed as a kernel from $\mathscr{X}$ to $\mathscr{P}(\mathscr{Y} \times \mathscr{Y})$. Since Definition 3.6 implies that $\lambda \circledast Q_k \rightharpoonup \lambda \circledast Q$, it follows that $Q_k(\cdot | x) \rightharpoonup Q(\cdot | x)$ for λ-almost all $x \in \mathscr{X}$. For every such x, by virtue of the Prohorov theorem, the sequence $\{Q_k(\cdot | x)\}$ is tight, and thus the sequence $\{\pi_k(\cdot, \cdot | x)\}$ is tight as well [45, Lem. 4.4]. By passing to a subsequence, if necessary, we conclude that the sequence $\{\pi_k(\cdot, \cdot | x)\}$ is weakly convergent to some limit $\{\pi^*(\cdot, \cdot | x)\}$. The limit must be the optimal transport from $Q(\cdot | x)$ to itself: $Q(\cdot | x) \circ \mathbb{I}$, where $\mathbb{I}$ is the identity kernel $y \mapsto \delta_y$. It follows that the limit does not depend on the subsequence; the entire sequence $\{\pi_k(\cdot, \cdot | x)\}$ is weakly convergent to $\pi^*(\cdot, \cdot | x)$ for λ-almost all x.

For any $R > 0$, we have a simple upper bound:

$$\begin{aligned} \big[\mathscr{W}_p^{\lambda}(Q_k, Q)\big]^p \le &\int_{\mathscr{X}} \int_{\mathscr{Y} \times \mathscr{Y}} \big[d(y, y') \wedge R\big]^p \, \pi_k(\mathrm{d}y, \mathrm{d}y' | x) \, \lambda(\mathrm{d}x) \\ &+ \int_{\mathscr{X}} \int_{\mathscr{Y} \times \mathscr{Y}} \big[d(y, y')^p - R^p\big]_+ \, \pi_k(\mathrm{d}y, \mathrm{d}y' | x) \, \lambda(\mathrm{d}x). \end{aligned}$$

Using the inequality

$$\big[d(y, y')^p - R^p\big]_+ \le 2^p d(y, y_0)^p \mathbb{1}_{\{d(y, y_0) \ge R/2\}} + 2^p d(y_0, y')^p \mathbb{1}_{\{d(y_0, y') \ge R/2\}},$$

we can continue the upper bound as follows:

$$\begin{aligned}\big[\mathscr{W}_p^\lambda(Q_k,Q)\big]^p &\le \int_{\mathscr{X}}\int_{\mathscr{Y}\times\mathscr{Y}} \big[d(y,y')\wedge R\big]^p\,\pi_k(\mathrm{d}y,\mathrm{d}y'|x)\,\lambda(\mathrm{d}x)\\ &\quad + 2^p \int_{\{d(y,y_0)\ge R/2\}} d(y,y_0)^p\,\pi_k(\mathrm{d}y,\mathrm{d}y'|x)\,\lambda(\mathrm{d}x)\\ &\quad + 2^p \int_{\{d(y_0,y')\ge R/2\}} d(y_0,y')^p\,\pi_k(\mathrm{d}y,\mathrm{d}y'|x)\,\lambda(\mathrm{d}x)\\ &= \int_{\mathscr{X}}\int_{\mathscr{Y}\times\mathscr{Y}} \big[d(y,y')\wedge R\big]^p\,\pi_k(\mathrm{d}y,\mathrm{d}y'|x)\,\lambda(\mathrm{d}x)\\ &\quad + 2^p \int_{\{d(y,y_0)\ge R/2\}} d(y,y_0)^p\,Q_k(\mathrm{d}y|x)\,\lambda(\mathrm{d}x)\\ &\quad + 2^p \int_{\{d(y_0,y')\ge R/2\}} d(y_0,y')^p\,Q(\mathrm{d}y'|x)\,\lambda(\mathrm{d}x).\end{aligned}$$

As the sequence $\{\pi_k(\cdot,\cdot|x)\}$ converges weakly to $\pi^*(\cdot,\cdot|x)$ for λ-almost all x, the first term on the right-hand side converges to 0 for every $R>0$. Furthermore, by Definition 2.4(ii), since $\lambda\circ Q_k \xrightarrow{p} \lambda\circ Q$,

$$\lim_{R\to\infty}\limsup_{k\to\infty}\int_{\{d(y,y_0)\ge R/2\}} d(y,y_0)^p\,Q_k(\mathrm{d}y|x)\,\lambda(\mathrm{d}x)=0.$$

The same is true for the third term. Putting these estimates together, we conclude that $\lim_{k\to\infty}\mathscr{W}_p^\lambda(Q_k,Q)=0$. ∎

Proof of Theorem 3.9. Theorem 2.3 implies that for all $f\in F$,

$$\begin{aligned}\mathscr{W}_1^\lambda(Q,\widetilde{Q}) &= \int_{\mathscr{X}} W_1(Q(\cdot|x),\widetilde{Q}(\cdot|x))\,\lambda(\mathrm{d}x)\\ &\ge \int_{\mathscr{X}}\left\{\int_{\mathscr{Y}} f(x,y)\,Q(\mathrm{d}y|x)-\int_{\mathscr{Y}} f(x,y)\,\widetilde{Q}(\mathrm{d}y|x)\right\}\lambda(\mathrm{d}x)\\ &= \int_{\mathscr{X}\times\mathscr{Y}} f(x,y)\,(\lambda\circledast Q)(\mathrm{d}x\;\mathrm{d}y)-\int_{\mathscr{X}\times\mathscr{Y}} f(x,y)\,(\lambda\circledast\widetilde{Q})(\mathrm{d}x\;\mathrm{d}y).\end{aligned}$$

This verifies the inequality "$\ge$" in (3.4). To verify the reverse inequality, let $\varepsilon>0$ and define the multifunction $F_\varepsilon:\mathscr{X}\rightrightarrows \mathrm{Lip}(\mathscr{Y},\mathbb{R})$ as follows:

$$\begin{aligned}F_\varepsilon(x)=\Bigg\{&\psi\in\mathrm{Lip}(\mathscr{Y},\mathbb{R}):\|\psi\|_{\mathrm{Lip}}\le 1,\\ &\int_{\mathscr{Y}}\psi(y)\,Q(\mathrm{d}y|x)-\int_{\mathscr{Y}}\psi(y)\,\widetilde{Q}(\mathrm{d}y|x)\ge W_1(Q(\cdot|x),\widetilde{Q}(\cdot|x))-\varepsilon\Bigg\},\quad x\in\mathscr{X}.\end{aligned}$$

It is measurable and, owing to Theorem 2.3, has nonempty closed values. Therefore, by the measurable selection theorem, a selector $\Psi_\epsilon:\mathscr{X}\to\mathrm{Lip}(\mathscr{Y},\mathbb{R})$ exists, such that $\Psi_\varepsilon(x)\in F_\varepsilon(x)$ for all $x\in\mathscr{X}$. Define $f_\varepsilon(x,y)=\big[\Psi_\varepsilon(x)\big](y)$, $x\in\mathscr{X}$, $y\in\mathscr{Y}$. By construction, $f_\epsilon\in F$ and

$$\begin{aligned}\mathscr{W}_1^\lambda(Q,\widetilde{Q}) &\le \int_{\mathscr{X}}\left\{\int_{\mathscr{Y}} f_\varepsilon(x,y)\,Q(\mathrm{d}y|x)-\int_{\mathscr{Y}} f_\varepsilon(x,y)\,\widetilde{Q}(\mathrm{d}y|x)+\varepsilon\right\}\lambda(\mathrm{d}x)\\ &\le \sup_{f(\cdot,\cdot)\in F}\left\{\int_{\mathscr{X}\times\mathscr{Y}} f(x,y)\,(\lambda\circledast Q)(\mathrm{d}x\;\mathrm{d}y)-\int_{\mathscr{X}\times\mathscr{Y}} f(x,y)(\lambda\circledast\widetilde{Q})(\mathrm{d}x\;\mathrm{d}y)\right\}+\varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, the inequality "$\leq$" (and then the equality) in (3.4) is true. As subtracting $f(\cdot, y_0)$ from $f(\cdot,\cdot)$ does not affect the right-hand side of (3.4), we may restrict F to contain only the functions whose value at y_0 is 0. ∎

Appendix B. Comparison of kernel distances on Gaussian mixture models. In this section, we consider Gaussian mixture models with varying dimensions and numbers of centers, each having a different weight (marginal probability). We denote by $\mathscr{X}_0$ the set of the centers, and by λ_0 the marginal distribution.

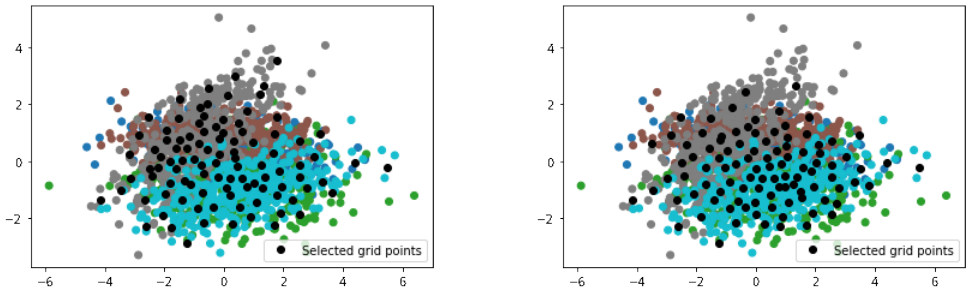
In each example, we select grid points from the same set of sample points. The point selection process employs two metrics: $\mathbb{D}_1(Q, \widetilde{Q}) = \sup_{x \in \mathscr{X}_0} W_1(Q(\cdot|x), \widetilde{Q}(\cdot|x))$, and the integrated transportation distance, $\mathscr{W}_1^{\lambda_0}(Q, \widetilde{Q}) = \sum_{x \in \mathscr{X}_0} \lambda_0(x) W_1(Q(\cdot|x), \widetilde{Q}(\cdot|x))$. The number of points to be selected by both methods is the same.

Table 3 presents the dimensions of the mixture model (dim), the number of centers (center), the number of particles sampled from each center (particles), the number of selected particles (selected), the solution times for both methods (in seconds), and the corresponding Wasserstein distance W_1 of the selected points to the particle distribution. For the sake of simplicity, we refer to the supremum distance as "sup" and the integrated transportation distance as "ITD" in the table header. The selection algorithm utilizing the integrated transportation distance consistently achieves a lower W_1 distance and faster execution time in all examples.

In Figures 4–6, the subfigures (a) and (b) illustrate the sample points and the grid points z^k (represented by black dots) selected using the supremum distance and the integrated transportation distance, respectively, for the three two-dimensional examples. The sample points x^{si} are depicted in different colors to represent the various Gaussian distributions.

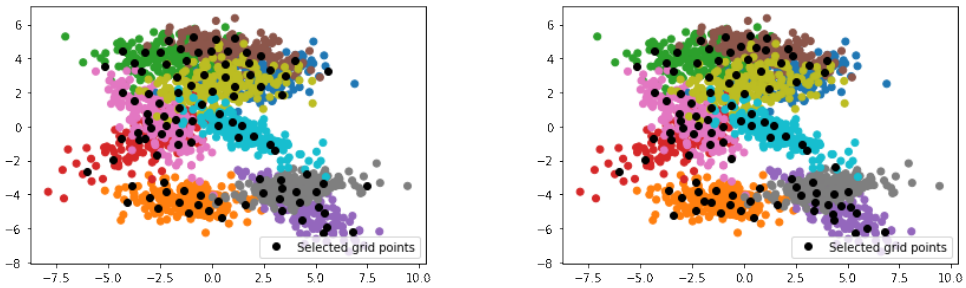
TABLE 3
Comparison of the supremum distance and the integrated transportation distance.

dim	Centers	Particles	Selected	sup (s)	ITD (s)	sup W_1	ITD W_1
2	5	400	100	1329.27	1320.15	0.288	0.268
2	10	200	100	1426.99	1296.43	0.466	0.457
2	16	160	128	1365.93	812.32	0.645	0.604
3	3	500	375	1296.96	530.36	0.913	0.901
3	5	400	500	1931.75	1253.03	0.953	0.784
5	3	800	600	1683.79	1235.43	1.963	1.812



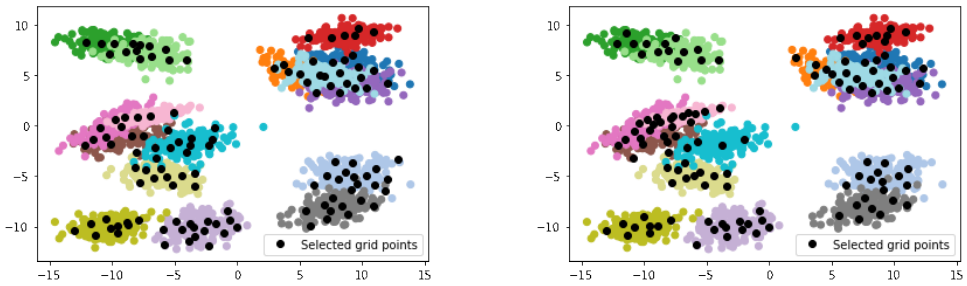
(a) The supremum distance selection. (b) The ITD selection.

FIG. 4. *Gaussian Mixture model with* five *centers and samples of* 400 *drawn from each center;* $dim(\beta) = 1000000$, $dim(\gamma) = 500$, *and* 100 *selected representative points.*



(a) The supremum distance selection. (b) The ITD selection.

FIG. 5. *Gaussian Mixture model with* 10 *centers and samples of* 200 *drawn from each center;* $dim(\beta) = 1000000$, $dim(\gamma) = 500$, *and* 100 *selected representative points.*



(a) The supremum distance selection. (b) The ITD selection.

FIG. 6. *Gaussian Mixture model with* 16 *centers and samples of* 100 *drawn from each center;* $dim(\beta) = 1638400$, $dim(\gamma) = 640$, *and* 128 *selected representative points.*

In all experiments, the integrated transportation distance model was solved faster and resulted in a more accurate representation of the mixture distribution. In experiments with problems of higher dimension these differences were dramatic.

All numerical results were obtained using Python (version 3.7) on a Macintosh HD laptop with a 2.9 GHz CPU and 16GB memory. The data are available in the working paper version.

REFERENCES

[1] J. ALTSCHULER, J. NILES-WEED, AND P. RIGOLLET, *Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration*, in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., Curran Associates, Red Hook, NY, 2017, pp. 1961–1971.

[2] P. ARTZNER, F. DELBAEN, J.-M. EBER, D. HEATH, AND H. KU, *Coherent multiperiod risk adjusted values and Bellman's principle*, Ann. Oper. Res., 152 (2007), pp. 5–22.

[3] J. BACKHOFF-VERAGUAS, D. BARTL, M. BEIGLBÖCK, AND M. EDER, *Adapted Wasserstein distances and stability in mathematical finance*, Finance Stoch., 24 (2020), pp. 601–632.

[4] D. BARTL AND J. WIESEL, *Sensitivity of Multi-Period Optimization Problems in Adapted Wasserstein Distance*, preprint, https://arxiv.org/abs/2208.05656, 2022.

[5] N. BÄUERLE AND A. GLAUNER, *Markov decision processes with recursive risk measures*, European J. Oper. Res., 296 (2022), pp. 953–966.

[6] X. BING, F. BUNEA, AND J. NILES-WEED, *The Sketched Wasserstein Distance for Mixture Distributions*, preprint, https://arxiv.org/abs/2206.12768, 2022.

[7] Y. Chen, J. Ye, and J. Li, *Aggregated Wasserstein distance and state registration for hidden Markov models*, IEEE Trans. Pattern Anal. Mach. Intell., 42 (2020), pp. 2133–2147.
[8] P. Cheridito, F. Delbaen, and M. Kupper, *Dynamic monetary risk measures for bounded discrete-time processes*, Electron. J. Probab., 11 (2006), pp. 57–106.
[9] P. Cheridito and M. Kupper, *Composition of time-consistent dynamic monetary risk measures in discrete time*, Int. J. Theor. Appl. Finance, 14 (2011), pp. 137–162.
[10] Y. Chow, H. Robbins, and D. Siegmund, *Great Expectations: The Theory of Optimal Stopping*, Houghton Mifflin, Boston, 1971.
[11] M. Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, in Advances in Neural Information Processing Systems 26, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds., Curran Associates, Red Hook, NY, 2013, pp. 2292–2300.
[12] J. Delon and A. Desolneux, *A Wasserstein-type distance in the space of Gaussian mixture models*, SIAM J. Imaging Sci., 13 (2020), pp. 936–970, https://doi.org/10.1137/19M1301047.
[13] D. Dentcheva, S. Penev, and A. Ruszczyński, *Statistical estimation of composite risk functionals and risk optimization problems*, Ann. Inst. Statist. Math., 69 (2017), pp. 737–760.
[14] D. Dentcheva and A. Ruszczyński, *Mini-batch risk forms*, SIAM J. Optim., 33 (2023), pp. 615–637, https://doi.org/10.1137/22M1503774.
[15] S. Dereich, M. Scheutzow, and R. Schottstedt, *Constructive quantization: Approximation by empirical measures*, Ann. Inst. Henri Poincare Probab. Stat., 49 (2013), pp. 1183–1203.
[16] J. Fan and A. Ruszczyński, *Process-based risk measures and risk-averse control of discrete-time systems*, Math. Program., 191 (2022), pp. 113–140.
[17] N. Fournier and A. Guillin, *On the rate of convergence in Wasserstein distance of the empirical measure*, Probab. Theory Relat. Fields, 162 (2015), pp. 707–738.
[18] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, Springer, 2007.
[19] H. Heitsch and W. Römisch, *Scenario tree modeling for multistage stochastic programs*, Math. Program., 118 (2009), pp. 371–406.
[20] K. Høyland and S. W. Wallace, *Generating scenario trees for multistage decision problems*, Manag. Sci., 47 (2001), pp. 295–307.
[21] L. V. Kantorovich and S. G. Rubinshtein, *On a space of totally additive functions*, Vestnik St. Petersburg Univ. Math., 13 (1958), pp. 52–59.
[22] M. Kaut and S. W. Wallace, *Shape-based scenario generation using copulas*, Comput. Manag. Sci., 8 (2011), pp. 181–199.
[23] P. Kern, A. Simroth, and H. Zähle, *First-order sensitivity of the optimal value in a markov decision model with respect to deviations in the transition probability function*, Math. Methods Oper. Res., 92 (2020), pp. 165–197.
[24] S. Kolouri, G. K. Rohde, and H. Hoffmann, *Sliced Wasserstein distance for learning Gaussian mixture models*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3427–3436.
[25] U. Köse and A. Ruszczyński, *Risk-averse learning by temporal difference methods with Markov risk measures*, J. Mach. Learn. Res., 22 (2021), 38.
[26] R. M. Kovacevic and A. Pichler, *Tree approximation for discrete time stochastic processes: A process distance approach*, Ann. Oper. Res., 235 (2015), pp. 395–421.
[27] S. Kusuoka, *On law-invariant coherent risk measures*, in Advances in Mathematical Economics 3, S. Kusuoka and T. Maruyama, eds., Springer, Tokyo, 2001, pp. 83–95.
[28] A. Majumdar and M. Pavone, *How should a robot assess risk? Towards an axiomatic theory of risk in robotics*, in Robotics Research, Springer, 2020, pp. 75–84.
[29] R. Mirkov and G. C. Pflug, *Tree approximations of dynamic stochastic programs*, SIAM J. Optim., 18 (2007), pp. 1082–1105, https://doi.org/10.1137/060658552.
[30] W. Ogryczak and A. Ruszczyński, *From stochastic dominance to mean-risk models: Semideviations as risk measures*, European J. Oper. Res., 116 (1999), pp. 33–50.
[31] W. Ogryczak and A. Ruszczyński, *On consistency of stochastic dominance and mean–semideviation models*, Math. Program., 89 (2001), pp. 217–232.
[32] W. Ogryczak and A. Ruszczyński, *Dual stochastic dominance and related mean-risk models*, SIAM J. Optim., 13 (2002), pp. 60–78, https://doi.org/10.1137/S1052623400375075.
[33] G. C. Pflug, *Scenario tree generation for multiperiod financial optimization by optimal discretization*, Math. Program., 89 (2001), pp. 251–271.
[34] G. C. Pflug, *Version-independence and nested distributions in multistage stochastic optimization*, SIAM J. Optim., 20 (2009), pp. 1406–1420, https://doi.org/10.1137/080718401.
[35] G. C. Pflug and A. Pichler, *A distance for multistage stochastic optimization models*, SIAM J. Optim., 22 (2012), pp. 1–23, https://doi.org/10.1137/110825054.

[36] G. C. Pflug and A. Pichler, *Dynamic generation of scenario trees*, Comput. Optim. Appl., 62 (2015), pp. 641–668.
[37] G. C. Pflug and W. Römisch, *Modeling, Measuring and Managing Risk*, World Scientific, 2007.
[38] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems: Volume* I*: Theory*, Springer, 1998.
[39] R. T. Rockafellar and S. Uryasev, *Optimization of conditional value-at-risk*, J. Risk, 2 (2000), pp. 21–42.
[40] D. Rudolf and N. Schweizer, *Perturbation theory for Markov chains via Wasserstein distance*, Bernoulli, 24 (2018), pp. 2610–2639.
[41] A. Ruszczyński, *Risk-averse dynamic programming for Markov decision processes*, Math. Program., 125 (2010), pp. 235–261.
[42] A. Ruszczyński and J. Yao, *A dual method for evaluation of dynamic risk in diffusion processes*, ESAIM Control Optim. Calc. Var., 26 (2020), 96.
[43] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, SIAM, Philadelphia, 2021, https://doi.org/10.1137/1.9781611976595.
[44] P. Sopasakis, D. Herceg, A. Bemporad, and P. Patrinos, *Risk-averse model predictive control*, Automatica, 100 (2019), pp. 281–288.
[45] C. Villani, *Optimal Transport: Old and New*, Springer, 2009.
[46] H. Zähle, *A concept of copula robustness and its applications in quantitative risk management*, Finance Stoch., 26 (2022), pp. 825–875.