QUANTIFYING DISTRIBUTIONAL MODEL RISK VIA OPTIMAL TRANSPORT

Jose Blanchet Karthyek Murthy

Columbia University

Abstract. This paper deals with the problem of quantifying the impact of model misspecification when computing general expected values of interest. The methodology that we propose is applicable in great generality, in particular, we provide examples involving path-dependent expectations of stochastic processes. Our approach consists in computing bounds for the expectation of interest regardless of the probability measure used, as long as the measure lies within a prescribed tolerance measured in terms of a flexible class of distances from a suitable baseline model. These distances, based on optimal transportation between probability measures, include Wasserstein's distances as particular cases. The proposed methodology is well-suited for risk analysis, as we demonstrate with a number of applications. We also discuss how to estimate the tolerance region non-parametrically using Skorokhod-type embeddings in some of these applications.

Keywords. Model risk, distributional robustness, transport metric, Wasserstein distances, duality, Kullback-Liebler divergences, ruin probabilities, diffusion approximations.

MSC 60G07, 60F99, and 62P05.

1. Introduction.

One of the most ubiquitous applications of probability is the evaluation of performance or risk by means of computations such as $Ef(X) = \int f dP$ for a given probability measure P, a function f, and a random element X. In classical applications, such as in finance, insurance, or queueing analysis, X is a stochastic process and f is a path dependent functional.

The work of the modeler is to choose a probability model P which is both descriptive and, yet, tractable. However, a substantial challenge, which arises virtually always, is that the model used, P, will in general be subject to misspecification; either because of lack of data or by the choice of specific parametric distributions.

The goal of this paper is to develop a framework to assess the impact of model misspecification when computing $Ef(X)$.

The importance of constructing systematic methods that provide performance estimates which are robust to model misspecification is emphasized in the academic response to the Basel Accord 3.5, Embrechts et al. (2014): In its broadest sense, robustness has to do with (in)sensitivity to underlying model deviations and/or data changes. Furthermore, here, a whole new field of research is opening up; at the moment, it is difficult to point to the right approach.

The results of this paper allow the modeler to provide a bound for the expectation of interest regardless of the probability measure used as long as such measure remains within a prescribed tolerance of a suitable baseline model.

The bounds that we obtain are applicable in great generality, allowing us to use our results in settings which include continuous-time stochastic processes and path dependent expectations in various domains of applications, including insurance and queueing.

The authors gratefully acknowledge support from Norges Bank Investment Management and NSF grant CMMI 1436700.

Let us describe the approach that we consider more precisely. Our starting point is a technique that has been actively pursued in the literature (to be reviewed momentarily), which involves considering a family of plausible models $\mathcal{P}$, and computing distributionally robust bounds as the solution to the optimization problem

$$\sup_{P \in \mathcal{P}} \int f dP, \tag{1}$$

over all probability measures in $\mathcal{P}$. The motivation, as mentioned earlier, is to measure the highest possible risk regardless of the probability measure in $\mathcal{P}$ used. A canonical object that appears naturally in specifying the family $\mathcal{P}$ is the neighborhood $\{P : d(\mu, P) \leq \delta\}$, where μ is a chosen baseline model and δ is a non-negative tolerance level. Here, d is a metric that measures discrepancy between probability measures, and the tolerance δ interpolates between no ambiguity ($\delta = 0$) and high levels of model uncertainty (δ large). The family of plausible models, $\mathcal{P}$, can then be specified as a single neighborhood (or) a collection of such plausible neighborhoods. In this paper, we choose d in terms of a transport cost (defined precisely in Section 2), and analyze the solvability of (1) and discuss various implications.

Being a flexible class of distances that include the popular Wasserstein distances as a special case, transport costs allow easy interpretation in terms of minimum cost associated with transporting mass between probability measures, and have been widely used in probability theory and its applications (see, for example, Rachev and Rüschendorf (1998a,b); Villani (2008); Ambrosio and Caffarelli (2003) for a massive collection of classical applications and Barbour and Xia (2006); Gozlan and Leonard (2016); Nguyen (2013); Wang and Guibas (2012); Fournier and Guillin (2014); Canas and Rosasco (2012); Solomon et al. (2014); Frogner et al. (2015) for a sample of growing list of new applications).

Relative entropy (or) Kullback-Liebler (KL) divergence, despite not being a proper metric, has been the most popular choice for d, thanks to the tractability of (1) when d is chosen as relative entropy or other likelihood based discrepancy measures (see Breuer and Csiszár (2013); Lam (2013); Glasserman and Xu (2014); Atar et al. (2015) for the use in robust performance analysis, and Hansen and Sargent (2001); Iyengar (2005); Nilim and El Ghaoui (2005); Lim and Shanthikumar (2007); Jain et al. (2010); Ben-Tal et al. (2013); Wang et al. (2015); Jiang and Guan (2015); Hu and Hong (2012); Bayraksan and Love (2015) and references therein for applications to distributionally robust optimization). Our study in this paper is directly motivated to address the shortcoming that many of these earlier works acknowledge: the absolute continuity requirement of relative entropy; there can be two probability measures μ and ν that produce the same samples with high probability, despite having the relative entropy between μ and ν as infinite. Such absolute continuity requirements could be very limiting, particularly so when the models of interest are stochastic processes defined over time. For instance, Brownian motion, because of its tractability in computing path-dependent expectations, is used as approximation of piecewise linear or piecewise constant processes (such as random walk). However, the use of the relative entropy would not be appropriate to accommodate these settings because the likelihood ratio between Brownian motion and any piecewise differentiable process is not well defined.

Another instance which shows the limitations of the KL divergence arises when using an Itô diffusion as a baseline model. In such case, the region $\mathcal{P}$ contains only Itô diffusions with the same volatility parameter as the underlying baseline model, thus failing to model volatility uncertainty.

The use of a distance d based on an optimal transport cost (or) Wasserstein's distance, as we do here, also allows to incorporate the use of tractable surrogate models such as Brownian motion. For instance, consider a classical insurance risk problem in which the modeler has sufficient information on claim sizes (for example, in car insurance) to build a non-parametric reserve model. But the modeler is interested in path-dependent calculations of the reserve process (such as ruin probabilities), so she might decide to use Brownian motion based approximation as a tractable surrogate model which allows to compute $Ef(X)$ easily. A key feature of the Wasserstein's distances, is that it is computed in terms of a so-called optimal coupling. So, the modeler can use any good coupling to provide a valid bound for the tolerance parameter δ.

The use of tractable surrogate models, such as Brownian motion, diffusions, and reflected Brownian motion, etc., has enabled the analysis of otherwise intractable complex stochastic systems. As we shall illustrate, the bounds that we obtain have the benefit of being computable directly in terms of the underlying tractable surrogate model. In addition, the calibration of the tolerance parameter takes advantage of well studied coupling techniques (such as Skorokhod embedding). So, we believe that the approach we propose naturally builds on the knowledge that has been developed by the applied probability community.

We summarize our main contributions in this paper below:

a) Assuming that X takes values in a Polish space, and using a wide range of optimal transport costs (that include the popular Wasserstein's distances as special cases) we arrive at a dual formulation for the optimization problem in (1), and prove strong duality – see Theorem 1.

b) Despite the infinite dimensional nature of the optimization problem in (1), we show that the dual problem admits a one dimensional reformulation which is easy to work with. We provide sufficient conditions for the existence of an optimizer P^* that attains the supremum in (1) – see Section 5. Using the result in a), for upper semicontinuous f, we show how to characterize an optimizer P^* in terms of a coupling involving μ and the chosen transport cost – see Remark 2 and Section 2.4.

c) We apply our results to various problems such as robust evaluation of ruin probabilities using Brownian motion as a tractable surrogate (see Section 3), general multidimensional first-passage time probabilities (see Section 6.1), and optimal decision making in the presence of model ambiguity (see Section 6.2).

d) We discuss a non-parametric method, based on a Skorokhod-type embedding, which allows to choose δ (these are particularly useful in tractable surrogate settings discussed earlier). The specific discussion of the Skorokhod embedding for calibration is given in Appendix A. We also discuss how δ can be chosen in a model misspecification context, see the discussion at the end of Section 6.1.

In a recent paper, Esfahani and Kuhn (2015), the authors also consider distributionally robust optimization using a very specific Wasserstein's distance. Similar formulations using Wasserstein distance based ambiguity sets have been considered in Pflug and Wozabal (2007); Wozabal (2012) and Zhao and Guan (2015) as well. The form of the optimization problem that we consider here is basically the same as that considered in Esfahani and Kuhn (2015), but there are important differences that are worth highlighting. In Esfahani and Kuhn (2015), the authors concentrate on the specific case of baseline measure being empirical distribution of samples obtained from a distribution supported in $\mathbb{R}^d$, and the function f possessing a special structure. In contrast, we allow the baseline measure μ to be supported on general Polish spaces and study a wide class of functions f (upper semicontinuous and integrable with respect to μ). The use of the Skorokhod embeddings that we consider here, for calibration of the feasible region, is also novel and not studied in the present literature.

Another recent paper, Gao and Kleywegt (2016), (which is an independent contribution made public a few days after we posted the first version of this paper in arXiv), presents a very similar version of the strong duality result shown in this paper. One difference that is immediately apparent is that the authors in Gao and Kleywegt (2016) concentrate on the case in which a specific cost function $c(\cdot,\cdot)$ used to define transport cost is of the form $c(x,y) = d(x,y)^p$, for some $p \geq 1$ and metric $d(\cdot,\cdot)$, whereas we only impose that $c(\cdot,\cdot)$ is lower semicontinuous. Allowing for general lower semicontinuous cost functions that are different from $d(x,y)^p$ is useful, as demonstrated in applications towards distributionally robust optimization and machine learning in Blanchet et al.

(2016). Another important difference is that, as far as we understand, the proof given in Gao and Kleywegt (2016) appears to implicitly assume that the space S in which the random element of interest X takes values is locally compact. For example, the proof of Lemma 2 in Gao and Kleywegt (2016) and other portions of the technical development appear to use repeatedly that a closed norm ball is compact. However, this does not hold in any infinite dimensional topological vector space[1], thus excluding important function spaces like $C[0,T]$ and $D[0,T]$ (denoting respectively the space of continuous and càdlàg functions on interval $[0,T]$), that are at the center of our applications. As mentioned earlier, our focus in this paper is on stochastic process applications, whereas the authors in Gao and Kleywegt (2016) put special emphasis on stochastic optimization in $\mathbb{R}^d$.

The rest of the paper is organized as follows: In Section 2 we shall introduce the assumptions and discuss our main result, whose proof is given in Section 4. The proof is technical due to the level of generality that is considered. More precisely, it is the fact that the cost functions used to define optimal transport costs need not be continuous, and the random elements that we consider need not take values in a locally compact space (like $\mathbb{R}^d$) which introduces technical complications, including issues like measurability of a key functional appearing in the dual formulation. In preparation to the proof, and to help the reader gain some intuition of the result, we present a one dimensional example first, in Section 3. Then, after providing the proof of our main result in Section 4 and conditions for existence of the worst-case probability distribution that attains the supremum in (1) in Section 5, we discuss additional examples in Section 6.

2. Our Main Result.

In order to state our main result, we need to introduce some notation and define the optimal transport cost between probability measures.

2.1. Notation and Definitions.

For a given Polish space S, we use $\mathcal{B}(S)$ to denote the associated Borel σ-algebra. Let us write $P(S)$ and $M(S)$, respectively, to denote the set of all probability measures and finite signed measures on $(S, \mathcal{B}(S))$. For any $\mu \in P(S)$, let $\mathcal{B}_\mu(S)$ denote the completion of $\mathcal{B}(S)$ with respect to μ; the unique extension of μ to $\mathcal{B}_\mu(S)$ that is a probability measure is also denoted by μ, and the measure μ in $\int \phi d\mu$ is to be interpreted as this extension defined on $\mathcal{B}_\mu(S)$ whenever $\phi : (S, \mathcal{B}(S)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is not measurable, but instead, $\phi : (S, \mathcal{B}_\mu(S)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is measurable. We shall use $C_b(S)$ to denote the space of bounded continuous functions from S to $\mathbb{R}$, and $\mathrm{Spt}(\mu)$ to denote the support of a probability measure μ. For any $\mu \in P(S)$ and $p \geq 1$, $L^p(d\mu)$ denotes the collection of Borel measurable functions $h : S \to \mathbb{R}$ such that $\int |h|^p d\mu < \infty$. The universal $\sigma-$algebra is defined by $\mathcal{U}(S) = \cap_{\mu \in P(S)} B_\mu(S)$. We use $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ to denote the extended real line, and $\mathfrak{m}_{\mathcal{U}}(S; \overline{\mathbb{R}})$ to denote the collection of measurable functions $\phi : (S, \mathcal{U}(S)) \to (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$. As $\mathcal{U}(S) \subseteq \mathcal{B}_\mu(S)$ for every $\mu \in P(S)$, any $\phi \in \mathfrak{m}_{\mathcal{U}}(S; \overline{\mathbb{R}})$ is also measurable when S and $\overline{\mathbb{R}}$ are equipped, respectively, with the σ-algebras $\mathcal{B}_\mu(S)$ and $\mathcal{B}(\overline{\mathbb{R}})$. Consequently, the integral $\int \phi d\mu$ is well-defined for any non-negative $\phi \in \mathfrak{m}_{\mathcal{U}}(S; \overline{\mathbb{R}})$. In addition, for any $\mu \in P(S)$, we say that μ is *concentrated* on a set $A \in B_\mu(S)$ if $\mu(A) = 1$.

Optimal transport cost. For any two probability measures μ_1 and μ_2 in $P(S)$, let $\Pi(\mu_1, \mu_2)$ denote the set of all joint distributions with μ_1 and μ_2 as respective marginals. In other words, the set $\Pi(\mu_1, \mu_2)$ represents the set of all couplings (also called transport plans) between μ_1 and μ_2. Throughout the paper, we assume that

Assumption 1 (A1). *$c : S \times S \to \mathbb{R}_+$ is a nonnegative lower semicontinuous function satisfying $c(x, y) = 0$ if and only if $x = y$.*

[1] see, for example, Example 3.100 and Theorem 5.26 in Aliprantis and Border (1999)

Then the *optimal transport cost* associated with the cost function c is defined as,

$$d_c(\mu_1, \mu_2) := \inf \left\{ \int c d\pi : \pi \in \Pi(\mu_1, \mu_2) \right\}, \quad \mu_1, \mu_2 \in P(S). \tag{2}$$

Intuitively, the quantity $c(x, y)$ specifies the cost of transporting unit mass from x in S to another element y of S. Given a lower semicontinuous cost function $c(x, y)$ and a coupling $\pi \in \Pi(\mu_1, \mu_2)$, the integral $\int c d\pi$ represents the expected cost associated with the coupling (or transport plan) π. For any non-negative lower semi-continuous cost function c, it is known that the 'optimal transport plan' that attains the infimum in the above definition exists (see Theorem 4.1 of Villani (2008)), and therefore, the optimal transport cost $d_c(\mu_1, \mu_2)$ corresponds to the lowest transport cost that is attainable among all couplings between μ_1 and μ_2.

If the cost function c is symmetric (that is, $c(y, x) = c(x, y)$ for all x, y), and it satisfies triangle inequality, one can show that the minimum transportation cost $d_c(\cdot, \cdot)$ defines a metric on the space of probability measures. For example, if S is a Polish space equipped with metric d, then taking the cost function to be $c(x, y) = d(x, y)$, renders the transport cost $d_c(\mu, \nu)$ to be simply the Wasserstein distance of first order between μ and ν. Unlike the Kullback-Liebler divergence (or) other likelihood based divergence measures, the Wasserstein distance is a proper metric on the space of probability measures. More importantly, Wasserstein distances do not restrict all the probability measures in the neighborhoods such as $\{\nu \in P(S) : d_c(\mu, \nu) \leq \delta\}$ to share the same support as that of μ (see, for example, Chapter 6 in Villani (2008) for properties of Wasserstein distances).

2.2. **Primal Problem.** Underlying our discussion we have a Polish space S, which is the space where the random elements of the given probability model $\mu \in P(S)$ takes values. Given $\delta > 0$, the objective, as mentioned in the Introduction, is to evaluate

$$\sup \left\{ \int f d\nu : d_c(\mu, \nu) \leq \delta \right\},$$

for any function f that satisfies the assumption that

Assumption 2 (A2)**.** *$f \in L^1(d\mu)$ is upper semicontinuous.*

As the integral $\int f d\nu$ may equal $\infty - \infty$ for some ν satisfying $d_c(\mu, \nu) \leq \delta$, only for the purposes of interpreting the supremum above, we let $\sup\{\infty, \infty - \infty\} = \infty$[2]. The value of this optimization problem provides a bound to the expectation of f, regardless of the probability measure used, as long as the measure lies within δ distance (measured in terms of d_c) from the baseline probability measure μ. In applied settings, μ is the probability measure chosen by the modeler as the baseline distribution, and the function f corresponds to a risk functional (or) performance measure of interest, for example, expected losses, probability of ruin, etc.

As the infimum in the definition of the optimal transport cost d_c is attained for any given non-negative lower semicontinuous cost function c (see Theorem 4.1 of Villani (2008)), we rewrite the quantity of interest as below:

$$I := \sup \left\{ \int f d\nu : \ d_c(\mu, \nu) \leq \delta \right\} = \sup \left\{ \int f(y) d\pi(x, y) : \pi \in \bigcup_{\nu \in P(S)} \Pi(\mu, \nu), \ \int c d\pi \leq \delta \right\},$$

[2]This is because, under the assumption that $f \in L^1(d\mu)$, for every probability measure ν such that $d_c(\mu, \nu) \leq \delta$ and $\int f^- d\nu = \infty$, one can identify a probability measure ν' satisfying $d_c(\mu, \nu') \leq \delta$, $\int f^+ d\nu' \geq \int f^+ d\nu$, and $\int f^- d\nu' < \infty = \int f^- d\nu$; see Corollary 3 in Appendix B for a simple construction of such a measure ν' from measure ν. Consequently, when computing the supremum of $\int f d\nu$, it is meaningful to restrict our attention to probability measures ν satisfying $\int f^- d\nu < \infty$, and interpret $\sup \left\{ \int f d\nu : d_c(\mu, \nu) \leq \delta \right\}$ as $\sup\{\int f d\nu : d_c(\mu, \nu) \leq \delta, \int f^- d\nu < \infty\}$. The notation, $\sup\{\infty, \infty - \infty\} = \infty$, simply facilitates this interpretation in this context.

which, in turn, is an optimization problem with linear objective function and linear constraints. If we let

$$I(\pi) := \int f(y)d\pi(x,y) \quad \text{ and } \quad \Phi_{\mu,\delta} := \left\{ \pi \in \bigcup_{\nu \in P(S)} \Pi(\mu,\nu) : \int c d\pi \leq \delta \right\}$$

for brevity, then

$$I = \sup\left\{ I(\pi) : \pi \in \Phi_{\mu,\delta} \right\} \tag{3}$$

is the quantity of interest.

2.2.1. *The Dual Problem and Weak Duality.* Define $\Lambda_{c,f}$ to be the collection of all pairs (λ, ϕ) such that λ is a non-negative real number, $\phi \in \mathfrak{m}_{\mathcal{U}}(S; \overline{\mathbb{R}})$ and

$$\phi(x) + \lambda c(x,y) \geq f(y), \text{ for all } x, y. \tag{4}$$

For every such $(\lambda, \phi) \in \Lambda_{c,f}$ consider

$$J(\lambda, \phi) := \lambda\delta + \int \phi d\mu.$$

As $\int f d\mu$ is finite and $\phi \geq f$ whenever ϕ satisfies (4), the integral in the definition of $J(\lambda, \phi)$ avoids ambiguities such as $\infty - \infty$ for any $(\lambda, \phi) \in \Lambda_{c,f}$. Next, for any $\pi \in \Phi_{\mu,\delta}$ and $(\lambda, \phi) \in \Lambda_{c,f}$, see that

$$\begin{aligned} J(\lambda, \phi) &= \lambda\delta + \int \phi(x) d\pi(x,y) \\ &\geq \lambda\delta + \int \big(f(y) - \lambda c(x,y) \big) d\pi(x,y) \\ &= \int f(y) d\pi(x,y) + \lambda \left(\delta - \int c(x,y) d\pi(x,y) \right) \\ &\geq \int f(y) d\pi(x,y) \\ &= I(\pi). \end{aligned}$$

Consequently, we have

$$J := \inf\big\{ J(\lambda, \phi) : (\lambda, \phi) \in \Lambda_{c,f} \big\} \geq I. \tag{5}$$

Following the tradition in optimization theory, we refer to the above infimum problem that solves for J in (5) as the dual to the problem that solves for I in (3), which we address as the primal problem. Our objective in the next section is to identify whether the primal and the dual problems have same value (that is, do we have that $I = J$?).

2.3. **Main Result: Strong Duality Holds.** Recall that the feasible sets for the primal and dual problems, respectively, are:

$$\Phi_{\mu,\delta} := \left\{ \pi \in \bigcup_{\nu \in P(S)} \Pi(\mu,\nu) : \int c d\pi \leq \delta \right\} \text{ and} \tag{6a}$$

$$\Lambda_{c,f} := \big\{ (\lambda, \phi) : \ \lambda \geq 0, \ \phi \in \mathfrak{m}_{\mathcal{U}}(S; \overline{\mathbb{R}}), \ \phi(x) + \lambda c(x,y) \geq f(y) \text{ for all } x, y \in S \big\}. \tag{6b}$$

The corresponding primal and dual problems are

$$I := \sup\left\{ \int f(y) d\pi(x,y) : \pi \in \Phi_{\mu,\delta} \right\} \text{ and } J := \inf\left\{ \lambda\delta + \int \phi d\mu : (\lambda, \phi) \in \Lambda_{c,f} \right\}$$

For brevity, we have identified the primal and dual objective functions as $I(\pi)$ and $J(\lambda, \phi)$ respectively. As the identified primal and dual problems are infinite dimensional, it is not immediate whether they have same value (that is, is $I = J$?). The objective of the following theorem is to verify that, for a broad class of performance measures f, indeed I equals J.

Theorem 1. *Under the Assumptions (A1) and (A2),*
(a) $I = J$. In other words,

$$\sup \left\{ I(\pi) : \pi \in \Phi_{\mu,\delta} \right\} = \inf \left\{ J(\lambda, \phi) : (\lambda, \phi) \in \Lambda_{c,f} \right\}.$$

(b) For any $\lambda \geq 0$, define $\phi_{\lambda} : S \to \mathbb{R} \cup \{\infty\}$ as follows:

$$\phi_{\lambda}(x) := \sup_{y \in S} \left\{ f(y) - \lambda c(x, y) \right\}.$$

There exists a dual optimizer of the form $(\lambda, \phi_{\lambda})$, for some $\lambda \geq 0$. In addition, any feasible $\pi^* \in \Phi_{\mu,\delta}$ and $(\lambda^*, \phi_{\lambda^*}) \in \Lambda_{c,f}$ are primal and dual optimizers, satisfying $I(\pi^*) = J(\lambda^*, \phi_{\lambda^*})$, if and only if

$$f(y) - \lambda^* c(x, y) = \sup_{z \in S} \left\{ f(z) - \lambda^* c(x, z) \right\}, \quad \pi^* a.s., \text{ and} \tag{8a}$$

$$\lambda^* \left(\int c(x, y) d\pi^*(x, y) - \delta \right) = 0. \tag{8b}$$

As the measurability of the function ϕ_{λ} is not immediate, we establish that $\phi_{\lambda} \in \mathfrak{m}_{\mathcal{U}}(S; \overline{\mathbb{R}})$ in Section 4, where the proof of Theorem 1 is also presented. Sufficient conditions for the existence of a primal optimizer $\pi^* \in \Phi_{\mu,\delta}$, satisfying $I(\pi^*) = I$, are presented in Section 5. For now, we are content discussing the insights that can be obtained from Theorem 1.

Remark 1 (on the value of the dual problem)**.** First, we point out the following useful characterization of the optimal value, I, as a consequence of Theorem 1:

$$I = \inf_{\lambda \geq 0} \left\{ \lambda\delta + E_{\mu} \left[\sup_{y \in S} \left\{ f(y) - \lambda c(X, y) \right\} \right] \right\}, \tag{9}$$

where the right hand side[3] is simply a univariate reformulation of the dual problem. This characterization of I follows from the strong duality in Theorem 1 and the observation that $J(\lambda, \phi_{\lambda}) \leq J(\lambda, \phi)$, for every $(\lambda, \phi) \in \Lambda_{c,f}$. The significance of this observation lies in the fact that the only probability measure involved in the right-hand side of (9) is the baseline measure μ, which is completely characterized, and is usually chosen in a way that it is easy to work with (or) draw samples from. In effect, the result indicates that the infinite dimensional optimization problem in (3) is easily solved by working on the univariate reformulation in the right hand side of (9).

Remark 2 (on the structure of primal optimal transport plan[4])**.** It is evident from the characterisation of an optimal measure π^* in Theorem 1(b) that if π^* exists, it is concentrated on $\left\{ (x, y) \in S \times S : y \in \arg\max_{z \in S} \{ f(z) - \lambda^* c(x, z) \} \right\}$. Thus, a worst-case joint probability measure π^* gets identified with a transport plan that transports mass from x to the optimizer(s) of the local optimization problem $\sup_{z \in S} \left\{ f(z) - \lambda^* c(x, z) \right\}$. According to the complementary slackness conditions (8a) and (8b), such a transport plan π^* satisfies one of the following two cases:
CASE 1: $\lambda^* > 0$: the transport plan π^* necessarily costs $\int c d\pi^* = \delta$,
CASE 2: $\lambda^* = 0$: the transport plan π^* satisfies $\int c d\pi^* \leq \delta$ (follows from primal feasibility of π^*) and $f(y)$ equals the constant $\sup_{z \in S} f(z)$, π^* almost surely (follows from (8a) assuming that $\arg\max_{z \in S} f(z)$ exists). Recall from (9) that $I = \lambda^* \delta + \sup_{z \in S} \{ f(z) - \lambda^* c(x, z) \} = \sup_{z \in S} f(z)$, which is in agreement with the structure described for the primal optimizer here when $\lambda^* = 0$. The interpretation is that the budget for quantifying ambiguity, δ, is sufficiently large to move all the probability mass to $\arg\max_{z \in S} f(z)$, thus making I as large as possible.

[3]As $\phi_{\lambda}(x) \geq f(x)$, there is no ambiguity, such as the form $\infty - \infty$, in the definition of integral $\int \phi_{\lambda}(x) d\mu(x)$

[4]In the literature of optimal transportation of probability measures, it is common to refer joint probability measures alternatively as transport plans, and the integral $\int c d\pi$ as the cost of transport plan π. We follow this convention to allow easy interpretations.

Remark 3 (On the uniqueness of a primal optimal transport plan)**.** Suppose that there exists a primal optimizer $\pi^* \in \Phi_{\mu,\delta}$ and a dual optimizer $(\lambda^*, \phi_{\lambda^*}) \in \Lambda_{c,f}$ satisfying $I(\pi^*) = J(\lambda^*, \phi_{\lambda^*}) < \infty$. In addition, suppose that for $\mu-$almost every $x \in S$, there is a unique $y \in S$ that attains the supremum in $\sup_{y\in S}\{f(y) - \lambda^* c(x,y)\}$. Then the primal optimizer π^* is unique because of the following reasoning: For every $x \in S$, if we let $T(x)$ denote the unique maximizer, $\arg\max\{f(y) - \lambda^* c(x,y)\}$, then it follows from Proposition 7.50(b) of Bertsekas and Shreve (1978) that the map $T : (S, \mathcal{U}(S)) \to (S, \mathcal{B}(S))$ is measurable. If (X, Y) is a pair jointly distributed according to π^*, then due to complementarity slackness condition (8a), we must have that $Y = T(X)$, almost surely. As any primal optimizer must necessarily assign entire probability mass to the set $\{(x,y) : y = T(x)\}$ (due to complementarity slackness condition (8a)), the primal optimizer is unique.

Remark 4 (on $\varepsilon-$optimal transport plans)**.** Given $\varepsilon > 0$ and a dual optimal pair $(\lambda^*, \phi_{\lambda^*})$, any $\pi_\varepsilon \in \Phi_{\mu,\delta}$ is $\varepsilon-$primal optimal if and only if

$$\int \big(\phi_{\lambda^*}(x) - \big(f(y) - \lambda^* c(x,y)\big)\big)\, d\pi_\varepsilon(x,y) \quad + \quad \lambda^* \left(\delta - \int c\, d\pi_\varepsilon\right) \;\leq\; \varepsilon, \tag{10}$$

where both the summands in the above expression are necessarily nonnegative. This follows trivially by substituting the following expressions for I and $I(\pi_\varepsilon)$ in $I - I(\pi_\varepsilon) \leq \varepsilon$.

$$I = J(\lambda^*, \phi^*_\lambda) = \lambda^*\delta + \int \phi_{\lambda^*}(x) d\pi_\varepsilon(x,y) \text{ and}$$
$$I(\pi_\varepsilon) = \int f(y) d\pi_\varepsilon(x,y) = \int \big(f(y) - \lambda^* c(x,y)\big) d\pi_\varepsilon(x,y) + \lambda^* \int c(x,y) d\pi_\varepsilon(x,y).$$

As both the summands in (10) are nonnegative, for any $\pi_\varepsilon \in \Phi_{\mu,\delta}$ such that $I(\pi_\varepsilon) = \pi_\varepsilon(S \times A) \geq I - \varepsilon$, we have,

$$\int \big(\phi_{\lambda^*}(x) - f(y) - \lambda^* c(x,y)\big) d\pi_\varepsilon(x,y) \;\leq\; \varepsilon \quad \text{and} \quad \left(\delta - \frac{\varepsilon}{\lambda^*}\right)^+ \leq \int c\, d\pi_\varepsilon \leq \delta, \tag{11}$$

where $a^+ := \max\{a, 0\}$ for any $a \in \mathbb{R}$.

Remark 5. If $\sup_{y\in S} f(y)/(1 + c(x,y)) = \infty$ (for example, if $f(y)$ grows to ∞ at a rate faster than the rate at which the transport cost function $c(x,y)$ grows) for every x in a set $A \subseteq S$ such that $\mu(A) > 0$, then $I = J = \infty$. This is because, for every $\lambda \geq 0$ and $x \in A$, $\phi_\lambda(x) = \sup_{y\in S}\{f(y) - \lambda c(x,y)\} = \infty$, and consequently, $\lambda\delta + \int \phi_\lambda d\mu = \infty$ for very $\lambda \geq 0$; therefore $I = J = \infty$ as a consequence of Theorem 1. Requiring the objective f to not grow faster than the transport cost c may be useful from a modeling viewpoint as it offers guidance in understanding choices of transport cost functions that necessarily yield ∞ as the robust estimate $\sup\{\int f d\nu : d_c(\mu,\nu) \leq \delta\}$.

We next discuss an important special case of Theorem 1, namely, computing worst case probabilities. The applications of this special case, in the context of ruin probabilities, is presented in Section 3 and Section 6.1. Other applications of Theorem 1, broadly in the context of distributionally robust optimization, are available in Esfahani and Kuhn (2015); Zhao and Guan (2015), Gao and Kleywegt (2016) and Blanchet et al. (2016), and as well in Example 4 in Section 6.2 of this paper.

2.4. Application of Theorem 1 for computing worst-case probabilities.

Suppose that we are interested in computing,

$$I = \sup\{P(A) : d_c(\mu, P) \leq \delta\}, \tag{12}$$

where A is a nonempty closed subset of the Polish space S. Since A is closed, the indicator function $f(x) = \mathbf{1}_A(x)$ is upper semicontinuous and therefore we can apply Theorem 1 to address (12). In order to apply Theorem 1, we first observe that

$$\sup_{y\in S}\{\mathbf{1}_A(y) - \lambda c(x,y)\} = \big(1 - \lambda c(x,A)\big)^+,$$

where $c(x,A) := \inf\{c(x,y) : y \in A\}$ is the lowest cost possible in transporting unit mass from x to some y in the set A, and a^+ denotes the positive part of the real number a. Consequently, Theorem 1 guarantees that the quantity of interest, I, in (12), can be computed by simply solving,

$$I = \inf_{\lambda \geq 0} \left\{ \lambda\delta + E_\mu \left[\left(1 - \lambda c(X,A)\right)^+ \right] \right\}. \tag{13}$$

If the infimum in the above expression for I is attained at $\lambda^* = 0$, then merely by substituting $\lambda = 0$ in $\lambda\delta + E_\mu[(1 - \lambda c(X,A))^+]$, we obtain $I = 1$. In addition, due to complementary slackness conditions (8a) and (8b), we have $\pi^*(\mathbf{1}_A(y) = 1) = 1$, and $\int c d\pi^* \leq \delta$, for any optimal transport plan $\pi^* \in \Phi_{\mu,\delta}$ satisfying $I = I(\pi^*)$. Here, as $f(x) = \mathbf{1}_A(x)$, we have

$$I(\pi) = \int \mathbf{1}_A(y) d\pi(x,y) = \pi(S \times A), \qquad \pi \in \Phi_{\mu,\delta}.$$

Next, as we turn our attention towards the structure of an optimal transport plan when $\lambda^* > 0$, let us assume, for ease of discussion, that $Proj_A(x) := \{y \in A : c(x,y) = c(x,A)\} \neq \varnothing$ for every $x \in S$. Unless indicated otherwise, let us assume that $\lambda^* > 0$ in the rest of this discussion. For every $x \in S$ and $\lambda^* > 0$, observe that

$$\arg\max_{y \in S} \left\{ \mathbf{1}_A(y) - \lambda^* c(x,y) \right\} = \begin{cases} Proj_A(x) & \text{if } 0 \leq c(x,A) < \frac{1}{\lambda^*}, \\ Proj_A(x) \cup \{x\} & \text{if } c(x,A) = \frac{1}{\lambda^*}, \\ \{x\} & \text{otherwise.} \end{cases}$$

Then, Part (b) of Theorem 1 allows us to conclude that, an optimal transport plan $\pi^* \in \Phi_{\mu,\delta}$ satisfying $\pi^*(S \times A) = I$, if it exists, is concentrated on

$$S^* = \left\{ (x,y) : c(x,A) \leq \frac{1}{\lambda^*},\ y \in Proj_A(x) \right\} \cup \left\{ (x,x) : c(x,A) \geq \frac{1}{\lambda^*} \right\}.$$

Next, let $\Pi_{S^*}(\mu)$ denote the set of probability measures π with μ as marginal for the first component, and satisfying $\pi(S^*) = 1$. In other words,

$$\Pi_{S^*}(\mu) := \left\{ \pi \in P(S \times S) :\ \pi(A \times S) = \mu(A) \text{ for all } A \in \mathcal{B}(S),\ \pi(S^*) = 1 \right\}.$$

Recall that $c(x,x) = 0$ for any $x \in S$, and $c(x,y) = c(x,A)$ for any $y \in Proj_A(x)$. Therefore, for a pair (X,Y) distributed jointly according to some $\pi \in \Pi_{S^*}(\mu)$, it follows from the definition of the collection $\Pi_{S^*}(\mu)$ that

$$E_\pi \left[c(X,Y) \mid X \right] = \begin{cases} E_\pi \left[c(X,A) \mid X \right] = c(X,A) & \text{if } c(X,A) < 1/\lambda^*, \\ E_\pi \left[c(X,X) \mid X \right] = 0 & \text{if } c(X,A) > 1/\lambda^*, \\ E_\pi \left[I\left(Y \in Proj_A(X)\right) \mid X \right] c(X,A) & \text{if } c(X,A) = 1/\lambda^*, \end{cases}$$

almost surely. Then, as $c(\cdot,\cdot)$ is non-negative, $E_\pi\left[c(X,Y)\right] = E_\pi\left[E_\pi\left[c(X,Y) \mid X\right]\right]$ satisfies,

$$E_\pi \left[c(X,A); c(X,A) < \frac{1}{\lambda^*} \right] \leq E_\pi \left[c(X,Y) \right] \leq E_\pi \left[c(X,A); c(X,A) < \frac{1}{\lambda^*} \right] + E_\pi \left[c(X,A); c(X,A) = \frac{1}{\lambda^*} \right].$$

Since the marginal distribution of X is μ (refer the definiton of $\Pi_{S^*}(\mu)$ above), it follows that $\int c d\pi = E_\pi[c(X,Y)]$ satisfies $\underline{c} \leq \int c d\pi \leq \bar{c}$, where

$$\underline{c} := \int_{\{x : c(x,A) < \frac{1}{\lambda^*}\}} c(x,A) d\mu(x) \quad \text{and} \quad \bar{c} := \int_{\{x : c(x,A) \leq \frac{1}{\lambda^*}\}} c(x,A) d\mu(x). \tag{14}$$

Further, as any optimal measure $\pi^* \in \Phi_{\mu,\delta}$ satisfying $\pi^*(S \times A) = I$ is a member of $\Pi_{S^*}(\mu)$, it follows from the complementary slackness condition (8b) that $\int c d\pi^*$ has to equal δ whenever $\lambda^* > 0$, and consequently, $\underline{c} \leq \delta \leq \bar{c}$, whenever a primal optimizer $\pi^* \in \Phi_{\mu,\delta}$ satisfying $\pi^*(S \times A) = I$ exists. The observation that $\underline{c} \leq \delta \leq \bar{c}$ holds regardless of whether a primal optimizer exists or not, and this is the content of Lemma 2 below, whose proof is provided towards the end of this section in Subsection 2.4.1.

Lemma 2. *Suppose that Assumption (A1) is in force, and A is a nonempty closed subset of the Polish space S. In addition, suppose that $\lambda^* \in (0,\infty)$ attains the infimum in (13). Then, $\underline{c} \leq \delta \leq \overline{c}$. On the other hand, if the infimum in (13) is attained at $\lambda^* = 0$, then $\delta \geq \overline{c} = \underline{c}$.*

If $\lambda^* > 0$ and $\underline{c} = \overline{c} = \delta$, any coupling in $\Pi_{S^*}(\mu)$ is primal optimal (because it satisfies the complementary slackness conditions (8a) and (8b) in addition to the primal feasibility condition that $\pi^* \in \Phi_{\mu,\delta}$). In particular, one can describe a convenient optimal coupling $\pi^* \in \Pi_{S^*}(\mu)$ satisfying $I(\pi^*) := \pi^*(S \times A) = I$ as follows: First, sample X with distribution μ, and let $\xi : S \to A$ be any universally measurable map such that $\xi(x) \in Proj_A(x)$, μ-almost surely. We then write,

$$Y^* = \xi(X) \cdot I\left(c(X,A) \leq \frac{1}{\lambda^*}\right) + X \cdot I\left(c(X,A) > \frac{1}{\lambda^*}\right),$$

to obtain (X, Y^*) distributed according to π^*. Such a universally measurable map $\xi(\cdot)$ always exists, assuming that $c(\cdot,\cdot)$ is lower semicontinuous and that $Proj_A(x)$ is not empty[5] for μ-almost every x (see, for example, Proposition 7.50(b) of Bertsekas and Shreve (1978)). In this case,

$$I = \pi^*(S \times A) = \pi^*\left\{(x,y) \in S \times S : c(x,A) \leq \frac{1}{\lambda^*}\right\} = \mu\left\{x \in S : c(x,A) \leq \frac{1}{\lambda^*}\right\}.$$

The second equality follows from the observation that for the described coupling (X, Y^*) distributed according to π^*, we have, $Y^* \in A$ if and only if $c(X,A) \leq 1/\lambda^*$, almost surely.

The reformulation that $I = \mu(x : c(x,A) \leq 1/\lambda^*)$ is extremely useful, as it re-expresses the worst-case probability of interest in terms of the probability of a suitably inflated neighborhood, $\{x \in S : c(x,A) \leq 1/\lambda^*\}$, evaluated under the reference measure μ, which is often tractable (see Section 3 and 6.1 for some applications). However, as the reasoning that led to this reformulation relies on the existence of a primal optimal transport plan, we use $\varepsilon-$optimal transport plans in Theorem 3 below to arrive at the same conclusion. The proof of Theorem 3 is presented towards the end of this section in Subsection 2.4.1.

Theorem 3. *Suppose that Assumption (A1) is in force, and A is a nonempty closed subset of the Polish space S. In addition, suppose that $\lambda^* \in [0,\infty)$ attains the infimum in (13), and the quantities $\underline{c}, \overline{c}$, defined in (14) in terms of λ^*, are such that $\underline{c} = \overline{c}$. Then,*

$$\sup\left\{P(A) : \ d_c(\mu, P) \leq \delta\right\} = \mu\left\{x : c(x,A) \leq \frac{1}{\lambda^*}\right\}. \tag{15}$$

Recall from Lemma 2 that λ^*, if it is strictly positive, is such that $\delta \in (\underline{c}, \overline{c})$. To arrive at the characterization (15), we had assumed that the integral

$$h(u) := \int_{\{x:c(x,A)\leq u\}} c(x,A)d\mu(x) \tag{16}$$

is continuous at $u = 1/\lambda^*$, and consequently $\underline{c} = \overline{c}$. However, if this is not the case, for example, because μ is atomic, then $\underline{c} < \delta \leq \overline{c}$. In this scenario, one can identify an optimal coupling by randomizing between the extreme cases $\underline{c}$ and $\overline{c}$. Remark 6 below provides a characterization of the primal optimizer when $\underline{c} < \delta \leq \overline{c}$.

Remark 6. *Suppose that λ^* is such that $\underline{c} < \delta \leq \overline{c}$, and that all the assumptions in Theorem 3, except the condition that $\underline{c} = \overline{c} = \delta$, are satisfied. Let Z be an independent Bernoulli random variable with success probability, $p := (\delta - \underline{c})/(\overline{c} - \underline{c})$. As before, sample X with distribution μ and let $\xi : S \to A$ be any universally measurable map such that $\xi(x) \in Proj_A(x)$, μ-almost surely. Then,*

$$Y^* := \xi(X) \cdot I\left(c(X,A) < \frac{1}{\lambda^*}\right) + X \cdot I\left(c(X,A) > \frac{1}{\lambda^*}\right) + (Z \cdot \xi(X) + (1-Z) \cdot X) \cdot I\left(c(X,A) = \frac{1}{\lambda^*}\right)$$

[5]The assumption that $Proj_A(x) \neq \varnothing$ holds, for example, when A is compact and nonempty (as c is lower semi-continuous), or when $c(x,\cdot)$ has compact sub-level sets for each x (recall that A is a closed set in the discussion)

is such that $\Pr((X, Y^*) \in S^*) = 1$ *(thus satisfying complementary slackness condition* (8a)*), and*

$$E\big[c(X, Y^*)\big] = \int_{\{c(x,A)<\frac{1}{\lambda^*}\}} c(x, A) d\mu(x) \; + \; \Pr(Z = 1) \int_{\{c(x,A)=\frac{1}{\lambda^*}\}} c(x, A) d\mu(x) \; + \; 0 = \underline{c} + p\big(\overline{c} - \underline{c}\big) = \delta.$$

This verifies the complementary condition (8b) *as well, and hence the marginal distribution of* Y^* *in the coupling* (X, Y^*) *attains the supremum in* $\sup\{P(A) : d_c(\mu, P) \leq \delta\}$.

All the examples considered in this paper have the function $h(u)$ as continuous, and hence we have the characterization (15) which offers an useful reformulation for computing worst-case probabilities. For instance, let us take the cost function c as a distance metric d (as in Wasserstein distances), and let the baseline measure μ and the set A be such that $h(\cdot)$, defined in (16), is continuous. Then (15) provides an interpretation that the worst-case probability $P(A)$ in the neighborhood of μ is just the same as $\mu(A_{1/\lambda^*})$, where $A_\varepsilon := \big\{y \in S : d(x, y) \leq \varepsilon \text{ for some } x \in A\big\}$ denotes the ε-neighborhood of the set A. Thus, the problem of finding a probability measure with worst-case probability $P(A)$ in the neighborhood of measure μ amounts to simply searching for a suitably inflated neighborhood of the set A itself. For example, if $S = \mathbb{R},\ c(x, y) = |x - y|$, then

$$\sup\big\{P[a, \infty) : d_c(\mu, P) \leq \delta\big\} = \mu\left[a - \frac{1}{\lambda^*}, \infty\right),$$

for any $\mu \in \mathcal{B}(\mathbb{R})$, where λ^* is the solution to the univariate optimization problem in (13). Alternatively, due to Lemma 2, one can characterize $1/\lambda^*$ as $h^{-1}(\delta) := \inf\{u \geq 0 : h(u) \geq \delta\}$, where $h(u)$ is the monotonically increasing right-continuous function (with left limits) defined in (16). If $h(u)$ does not admit a closed-form expression, one approach is to obtain samples of X under the reference measure μ, and either solve the sampled version of (13), or compute a Monte Carlo approximation of the integral $h(u)$ to identify the level $1/\lambda^*$ as $h^{-1}(\delta) = \inf\{u : h(u) \geq \delta\}$.

2.4.1. *Proofs of Lemma 2 and Theorem 3.* We conclude this section with proofs for Lemma 2 and Theorem 3. Given $\lambda^* \geq 0$ and $n > 1$, define $C_n := C_n^{(1)} \cup C_n^{(2)}$, where

$$C_n^{(1)} := \left\{(x, y) \in S \times S : c(x, A) \leq \frac{1}{\lambda^*}\left(1 + \frac{1}{n}\right),\ y \in A,\ c(x, y) < c(x, A) + \frac{1}{\lambda^* n}\right\} \text{ and}$$

$$C_n^{(2)} := \left\{(x, y) \in S \times S : c(x, A) > \frac{1}{\lambda^*}\left(1 - \frac{1}{n}\right),\ y \notin A,\ c(x, y) < \frac{1}{\lambda^* n}\right\}.$$

In addition, define

$$D_n^{(1)} := \{(x, y) \in C_n : c(x, A) \leq (1 - 1/n)/\lambda^*\}, \quad D_n^{(2)} := \{(x, y) \in C_n : c(x, A) > (1 + 1/n)/\lambda^*\},$$

and $D_n^{(3)} := C_n \setminus \left(D_n^{(1)} \cup D_n^{(2)}\right)$. The above definitions yield, $C_n^{(1)} = D_n^{(1)} = S \times A$, and $C_n^{(2)} = D_n^{(2)} = D_n^{(3)} = \varnothing$, when $\lambda^* = 0$.

Lemma 4. *Suppose that Assumption (A1) is in force, and A is a nonempty closed subset of the Polish space S. In addition, suppose that $\lambda^* \in [0, \infty)$ attains the infimum in* (13)*. Then, there exists a collection of probability measures $\{\pi_n : n > 1\} \subseteq \Phi_{\mu,\delta}$ such that $\pi_n(C_n) \geq 1 - 1/n$, $\int_{(S\times S)\setminus C_n} c d\pi_n = 0$, and $I(\pi_n) := \pi_n(S \times A) \geq I - 2/n$.*

Proof. Proof of Lemma 4. Since $I := \sup\{\pi(S \times A) : \pi \in \Phi_{\mu,\delta}\}$, we consider a collection $\{\tilde{\pi}_n : n > 1\} \subseteq \Phi_{\mu,\delta}$ such that $I(\tilde{\pi}_n) \geq I - 1/n^2$. For every $n > 1$, if we let

$$B_n := \left\{(x, y) \in S \times S : \mathbf{1}_A(y) - \lambda^* c(x, y) > \big(1 - \lambda^* c(x, A)\big)^+ - 1/n\right\},$$

as a consequence of Markov's inequality and the characterization (11) in Remark 4, we obtain,

$$\tilde{\pi}_n(B_n) \geq 1 - \frac{1}{n} \int \big(\phi_{\lambda^*}(x) - \mathbf{1}_A(y) - \lambda^* c(x, y)\big) d\tilde{\pi}_n(x, y) \geq 1 - \frac{1/n^2}{1/n} = 1 - \frac{1}{n}.$$

For every $n > 1$, given a pair $(\tilde{X}_n, \tilde{Y}_n)$ jointly distributed with law $\tilde{\pi}_n$, we define a new jointly distributed pair (X_n, Y_n) defined as follows,

$$(X_n, Y_n) := \begin{cases} (\tilde{X}_n, \tilde{Y})_n & \text{if } (X_n, Y_n) \in B_n, \\ (\tilde{X}_n, \tilde{X}_n) & \text{otherwise.} \end{cases}$$

Our objective now is to show that the collection $\{\pi_n : n > 1\}$, where $\pi_n := \mathrm{Law}(X_n, Y_n)$, satisfies the desired properties. We begin by verifying that $\pi_n \in \Phi_{\mu,\delta}$ here: As $\tilde{\pi}_n \in \Phi_{\mu,\delta}$ and $\tilde{X}_n := X_n$, it is immediate that $\mathrm{Law}(X_n) = \mu$. In addition, as $\int c d\pi_n = \int_{B_n} c d\tilde{\pi}_n + 0$, because $c(x,x) = 0$, it follows from the non-negativity of $c(\cdot,\cdot)$ and $\int c d\tilde{\pi}_n \leq \delta$ that $\pi_n \in \Phi_{\mu,\delta}$. Next, as $B_n \subseteq C_n$ for every $n > 1$, we have that $\pi_n(C_n) \geq \pi_n(B_n) \geq 1 - 1/n$, and $\int_{(S\times S)\setminus C_n} c(x,y) d\pi_n(x,y) = \int_{(S\times S)\setminus C_n} c(x,x) d\pi_n(x,y) = 0$. Finally, for every $n \geq 1$, $I(\pi_n) \geq I - 2/n$ is also immediate once we observe that

$$\begin{aligned} I(\pi_n) = \pi_n(S \times A) &\geq \pi_n((S \times A) \cap B_n) = \tilde{\pi}_n((S \times A) \cap B_n) \\ &\geq \tilde{\pi}_n(S \times A) - \tilde{\pi}_n((S \times S) \setminus B_n) \geq I - \frac{1}{n^2} - \frac{1}{n}, \end{aligned}$$

thus verifying all the desired poperties of the collection $\{\pi_n : n > 1\}$. □

Proof. Proof of Lemma 2. Consider a collection $\{\pi_n : n > 1\} \subseteq \Phi_{\mu,\delta}$ such that $\pi_n(C_n) \geq 1 - 1/n$, $\int_{(S\times S)\setminus C_n} c d\pi_n = 0$, and $I(\pi_n) := \pi_n(S \times A) \geq I - 2/n$. Such a collection exists because of Lemma 4. We first observe that $\int c d\pi_n = \int_{C_n} c d\pi_n$, because $\int_{(S\times S)\setminus C_n} c d\pi_n = 0$. Next, recalling the definitions of subsets $D_n^{(i)}, i = 1,2,3$ introduced before stating Lemma 4, we use $D_n^{(i)} \subseteq C_n^{(i)}$, $i = 1, 2$, to observe that,

$$\begin{aligned} c(x,A) \leq c(x,y) &< c(x,A) + 1/(n\lambda^*) && \text{if } (x,y) \in D_n^{(1)}, \\ 0 \leq c(x,y) &< 1/(n\lambda^*) && \text{if } (x,y) \in D_n^{(2)}, \text{ and} \\ 0 \leq c(x,y) &< c(x,A) + 1/(n\lambda^*) && \text{if } (x,y) \in D_n^{(3)}. \end{aligned}$$

Case 1: $\lambda^* > 0$. Let us first restrict ourself to the case where $\lambda^* > 0$. Then $\int c d\pi_n = \int_{C_n} c d\pi_n$ can be bounded from above and below as follows:

$$\int_{D_n^{(1)}} c(x,A) d\pi_n(x,y) \leq \int c d\pi_n \leq \int_{D_n^{(1)} \cup D_n^{(3)}} \left(c(x,A) + \frac{1}{n\lambda^*} \right) d\pi_n(x,y) + \frac{\pi_n(D_n^{(2)})}{n\lambda^*}.$$

Next, as $\lambda^* > 0$ and $I(\pi_n) = \pi_n(S \times A) \geq I - 2/n$, we use the second part of (11) to reason that

$$\int_{D_n^{(1)}} c(x,A) d\pi_n(x,y) \leq \int c d\pi_n \leq \delta, \text{ and} \tag{17}$$

$$\delta - \frac{1}{2n\lambda^*} \leq \int c d\pi_n \leq \int_{D_n^{(1)} \cup D_n^{(3)}} \left(c(x,A) + \frac{1}{n\lambda^*} \right) d\pi_n(x,y) + \frac{\pi_n(D_n^{(2)})}{n\lambda^*}, \tag{18}$$

for every $n > 1$. The next few steps are dedicated towards re-expressing the integrals in the left hand side of (17) and right hand side of (18) in terms of $\underline{c}$ and $\bar{c}$ to obtain,

$$\overline{\lim_{n\to\infty}} \int_{D_n^{(1)} \cup D_n^{(3)}} \left(c(x,A) + \frac{1}{n\lambda^*} \right) d\pi_n(x,y) \leq \bar{c} \quad \text{and} \quad \underline{c} \leq \underline{\lim_{n\to\infty}} \int_{D_n^{(1)}} c(x,A) d\pi_n(x,y). \tag{19}$$

As every $(x,y) \in D_n^{(1)} \cup D_n^{(3)}$ satisfies $c(x,A) \leq (1 + 1/n)/\lambda^*$, we have

$$\int_{D_n^{(1)} \cup D_n^{(3)}} \left(c(x,A) + \frac{1}{n\lambda^*} \right) d\pi_n(x,y) \leq \int_{\{x : c(x,A) \leq (1+1/n)/\lambda^*\}} c(x,A) d\pi_n(x,y) = \int_{\{x : c(x,A) \leq (1+1/n)/\lambda^*\}} c(x,A) d\mu(x),$$

thus establishing the first inequality in (19) as a consequence of bounded convergence theorem. Next, as $\{(x,y) \notin D_n^{(1)} : c(x,A) < 1/\lambda^*\}$ is contained in the union of $(S \times S) \setminus C_n$ and $\{(x,y) \in C_n : (1-1/n)/\lambda^* < c(x,A) < 1/\lambda^*\}$, we obtain,

$$\begin{aligned}
\underline{c} &:= \int_{\{x:c(x,A)<1/\lambda^*\}} c(x,A)d\pi_n(x,y) \leq \int_{D_n^{(1)}} c(x,A)d\pi_n(x,y) + \frac{1}{\lambda^*}\pi_n\big((x,y) \notin D_n^{(1)} : c(x,A) < 1/\lambda^*\big) \\
&\leq \int_{D_n^{(1)}} c(x,A)d\pi_n(x,y) + \frac{1}{\lambda^*}\left(\pi_n\big((x,y) \in C_n : (1-1/n)/\lambda^* \leq c(x,A) < 1/\lambda^*\big) + \pi_n\big((S\times S)\setminus C_n\big)\right) \\
&\leq \int_{D_n^{(1)}} c(x,A)d\pi_n(x,y) + \frac{1}{\lambda^*}\left(\mu\big(x \in S : (1-1/n)/\lambda^* \leq c(x,A) < 1/\lambda^*\big) + \frac{1}{n}\right),
\end{aligned}$$

thus verifying the second inequality in (19). Now that we have verified both the inequalities in (19), the conclusion that $\underline{c} \leq \delta \leq \bar{c}$ is automatic if we send $n \to \infty$ in (17), (18) and use (19).

Case 2: When $\lambda^* = 0$. It follows from the definitions of sets $C_n^{(i)}$ and $D_n^{(i)}$ that $C_n^{(1)} = D_n^{(1)} = S \times A$, and $C_n^{(2)} = D_n^{(2)} = D_n^{(3)} = \varnothing$ when $\lambda^* = 0$. Recall that $C_n := C_n^{(1)} \cup C_n^{(2)}$ and $\pi_n \in \Phi_{\mu,\delta}$ are such that $\pi_n(C_n) = \pi_n(S \times A) \geq 1 - 1/n$. Let $h^-(u) := \int_{\{x:c(x,A)<u\}} c(x,A)d\mu(x)$ and $u_n := \inf\{u \geq 0 : \mu(x : c(x,A) \leq u) \geq 1-1/n\}$. As $\pi_n(S\times A) \geq 1-1/n$ and $c(x,y) \geq c(x,A)$ for any $x \in A$, we have $\int cd\pi_n \geq \int_{S\times A} c(x,A)d\pi_n(x,y) \geq h^-(u_n)$. Next, as $\mu(x : c(x,A) \leq \sup_n u_n) = 1$, we have

$$\underline{\lim}_{n\to\infty} \int cd\pi_n \geq \underline{\lim}_{n\to\infty} h^-(u_n) = \lim_{n\to\infty} \int_{\{x:c(x,A)<u_n\}} c(x,A)(x)d\mu(x) = \int c(x,A)d\mu(x), \tag{20}$$

as a consequence of monotone convergence theorem. Further, as $\int cd\pi_n \leq \delta$ for every n, it follows from (20) that $\bar{c} = \underline{c} = \int c(x,A)d\mu(x) \leq \delta$. This concludes the proof. $\square$

Proof. Proof of Theorem 3. Consider a collection $\{\pi_n : n > 1\} \subseteq \Phi_{\mu,\delta}$ such that $\pi_n(C_n) \geq 1-1/n$, $\int_{(S\times S)\setminus C_n} cd\pi_n = 0$, and $I(\pi_n) := \pi_n(S \times A) \geq I - 2/n$. Such a collection exists because of Lemma 4. Similar to the proof of Lemma 2, we recall the definitions of subsets $D_n^{(i)}, i = 1,2,3$ introduced before stating Lemma 4, to observe that $D_n^{(i)} \subseteq C_n^{(i)}$, $i = 1,2$, and consequently,

$$\mathbf{1}_A(y) = 1 \text{ if } (x,y) \in D_n^{(1)}, \qquad \text{and } \mathbf{1}_A(y) = 0 \text{ if } (x,y) \in D_n^{(2)}.$$

As a result, $\pi_n\left(D_n^{(1)}\right) \leq \pi_n(S \times A) \leq \pi_n\left((S\times S)\setminus D_n^{(2)}\right)$. Combining this with the observation that $I - 2/n \leq \pi_n(S \times A) \leq I$, we obtain,

$$\pi_n\left(D_n^{(1)}\right) \leq I \quad \text{and} \quad \pi_n\left((S\times S)\setminus D_n^{(2)}\right) \geq I - \frac{2}{n}, \tag{21}$$

for every $n > 1$. Similar to the proof of Lemma 2, the rest of the proof is dedicated towards proving

$$\overline{\lim_n}\, \pi_n\left((S\times S)\setminus D_n^{(2)}\right) \;\leq\; \mu\big(x \in S : c(x,A) \leq 1/\lambda^*\big) \;\leq\; \underline{\lim}_n\, \pi_n\left(D_n^{(1)}\right), \tag{22}$$

when $\underline{c} = \bar{c}$. The first inequality is immediate when we observe that every $(x,y) \notin D_n^{(2)}$ is such that $c(x,A) \leq (1+1/n)/\lambda^*$, and therefore,

$$\overline{\lim_n}\, \pi_n\left((S\times S)\setminus D_n^{(2)}\right) \leq \overline{\lim_n}\, \pi_n\big((x,y) : c(x,A) \leq (1+1/n)/\lambda^*\big) = \lim_n \mu\big(x : c(x,A) \leq (1+1/n)/\lambda^*\big),$$

which equals $\mu(x : c(x,A) \leq 1/\lambda^*)$, due to bounded convergence theorem. This verifies the first inequality in (21). Next, as $D_n^{(1)} := \{(x,y) \in C_n : c(x,A) \leq (1-1/n)/\lambda^*\}$, the probability $\pi_n(\{(x,y) : c(x,A) \leq 1/\lambda^*\})$ equals,

$$\pi_n\left(D_n^{(1)}\right) + \pi_n\left(\{(x,y) \in C_n : (1-1/n)/\lambda^* < c(x,A) \leq 1/\lambda^*\}\right) + \pi_n\left(\{(x,y) \notin C_n : c(x,A) \leq 1/\lambda^*\}\right).$$

Since $\pi_n(C_n) \geq 1 - 1/n$ and $\mu(x : c(x,A) \leq 1/\lambda^*) = \pi_n(\{(x,y) : c(x,A) \leq 1/\lambda^*\})$, we obtain that

$$\pi_n\left(D_n^{(1)}\right) \geq \mu\big(x : c(x,A) \leq 1/\lambda^*\big) - \mu\left(\{x : (1-1/n)/\lambda^* < c(x,A) \leq 1/\lambda^*\}\right) - 1/n.$$

Further, as $\underline{c} = \bar{c}$, we have $\mu(x : c(x,y) = 1/\lambda^*) = 0$, and therefore, $\underline{\lim}_n \pi_n\left(D_n^{(1)}\right) \geq \mu\big(x : c(x,A) \leq 1/\lambda^*\big)$, thus verifying the inequality in the right hand side of (22). Now that both the inequalities in (22) are verified, we combine (21) with (22) to obtain that $I = \mu(x \in S : c(x,A) \leq 1/\lambda^*)$. □

3. Computing ruin probabilities: A first example.

In this section, we consider an example with the objective of computing a worst-case estimate of ruin probabilities in a ruin model whose underlying probability measure is not completely specified. As mentioned in the Introduction, such a scenario can arise for various reasons, including the lack of data necessary to pin down a satisfactory model. A useful example to keep in mind would be the problem of pricing an exotic insurance contract (for example, a contract covering extreme climate events) in a way the probability of ruin is kept below a tolerance level.

Example 1. In this example, we consider the celebrated Cramer-Lundberg model, where a compound Poisson process is used to calculate ruin probabilities of an insurance risk/reserve process. The model is specified by 4 primitives: initial reserve u, safety loading $\eta > 0$, the rate ν at which claims arrive, and the distribution of claim sizes F with first and second moments m_1 and m_2 respectively. Then the stochastic process

$$R(t) = u + (1+\eta)\nu m_1 t - \sum_{i=1}^{N_t} X_i,$$

specifies a model for the reserve available at time t. Here, X_n denotes the size of the n-th claim, and the collection $\{X_1, X_2, \ldots\}$ is assumed to be independent samples from the distribution F. Further, $\tilde{p} := (1+\eta)\nu m_1$ is the rate at which a premium is received, and N_t is taken to be a Poisson process with rate λ. Then one of the important problems in risk theory is to calculate the probability

$$\psi(u,T) := \Pr\left\{\inf_{t\in[0,T]} R(t) \leq 0\right\}$$

that the insurance firm runs into bankruptcy before a specified duration T.

Despite the simplicity of the model, existing results in the literature do not admit simple methods for the computation of $\psi(u,T)$ (see Asmussen and Albrecher (2010), Embrechts et al. (1997), Rolski et al. (1999) and references therein for a comprehensive collection of results). In addition, if the historical data is not adequately available to choose an appropriate distribution for claim sizes, as is typically the case in an exotic insurance situation, it is not uncommon to use a diffusion approximation

$$R_B(t) := u + (1+\eta)\nu m_1 t - (\nu m_1 t + \sqrt{\nu m_2} B(t))$$

$$= u + \eta \nu m_1 t - \sqrt{\nu m_2} B(t)$$

that depends only on first and second moments of the claim size distribution F. Here, the stochastic process $(B(t) : 0 \leq t \leq T)$ denotes the standard Brownian motion, and

$$\psi_B(u,T) := \Pr\left\{\sup_{t\in[0,T]} \left(\sqrt{\nu m_2} B(t) - \eta \nu m_1 t\right) \geq u\right\},$$

is to serve as a diffusion approximation based substitute for ruin probabilities $\psi(u,T)$. See the seminal works of Iglehart and Harrison (Iglehart (1969); Harrison (1977)) for a justification and some early applications of diffusion approximations in computing insurance ruin. Such diffusion approximations have enabled approximate computations of various path-dependent quantities, which may be otherwise intractable. However, as it is difficult to verify the accuracy of the Brownian approximation of ruin probabilities, in this example, we use the framework developed in Section 2 to compute worst-case estimates of ruin probabilities over all probability measures in a neighborhood around the baseline Brownian motion $B(t)$ driving the ruin model $R_B(t)$.

For this purpose, we identify the Polish space where the stochastic processes of our interest live as the Skorokhod space $S = D([0,T],\mathbb{R})$ equipped with the J_1 topology. In other words, $S = D([0,T],\mathbb{R})$ is simply the space of real-valued right-continuous functions with left limits (rcll) defined on the interval $[0,T]$, equipped with the J_1-metric d_{J_1}. Please refer Lemma 11 in Appendix B for an expression of d_{J_1} and Chapter 3 of Whitt (2002) for an excellent exposition on the space $D([0,T],\mathbb{R})$. Next, we take the transportation cost (corresponding to p-th order Wasserstein distance) as

$$c(x,y) = \Big(d_{J_1}(x,y)\Big)^p \qquad x,y \in S,$$

for some $p \geq 1$, and the baseline measure μ as the probability measure induced in the path space by the Brownian motion $B(t)$. In addition, if we let

$$A_u := \left\{ x \in S : \sup_{t\in[0,T]} \left(\sqrt{\nu m_2} x(t) - \nu m_1 t\right) \geq u \right\},$$

then the following observations are in order: First, the Brownian approximation for ruin probability $\psi_B(u,T)$ equals $\mu(A_u)$. Next, the set A_u is closed (verified in Lemma 10 in Appendix B), and hence the function $\mathbf{1}_{A_u}(\cdot)$ is upper semicontinuous. Further, it is verified in Lemma 11 in Appendix B that

$$c(x,A_u) := \inf\{c(x,y) : y \in A_u\} = \left(u - \sup_{t\in[0,T]} \left(\sqrt{\nu m_2} x(t) - \eta\nu m_1 t\right)\right)^p, \qquad \text{for } x \notin A_u. \tag{23}$$

As $h(s) = E_\mu[c(X,A); c(X,A_u) \leq s]$ is continuous, for any given $\delta > 0$, it follows from Theorem 3 that

$$\psi_{rob}(u,T) := \sup\{P(A_u) : d_c(P,\mu) \leq \delta\} = \mu\left\{ x \in S : c(x,A_u) \leq \frac{1}{\lambda^*} \right\}, \tag{24}$$

where, the level $1/\lambda^*$ is identified as $h^{-1}(\delta) = \inf\{s \geq 0 : h(s) \geq \delta\}$. As $c(x,A_u)$ admits a simple form as in (23), following (24), the problem of computing $\psi_{rob}(u,T)$ becomes as elementary as

$$\psi_{rob}(u,T) = \Pr\left\{ \sup_{t\in[0,T]} \left(\sqrt{\nu m_2} B_t - \eta\nu m_1 t\right) > u - \left(\frac{1}{\lambda^*}\right)^{1/p} \right\} = \psi_B(\tilde{u},T), \tag{25}$$

with $\tilde{u} := u - (\lambda^*)^{-1/p}$. Thus, the computation of worst-case ruin probability remains the problem of evaluating the probability that a Brownian motion with negative drift crosses a positive level, $\tilde{u}$, smaller than the original level u. In other words, the presence of model ambiguity has manifested itself only in reducing the initial capital to a new level $\tilde{u}$. Apart from the tractability, such interpretations of the worst-case ruin probabilities in terms of level crossing a modified ruin set is an attractive feature of using Wasserstein distances to model distributional ambiguities.

Numerical illustration: To make the discussion concrete, we consider a numerical example where $T = 100, \nu = 1, p = 2$, and the safety loading $\eta = 0.1$. The claim sizes are taken from a distribution F that is not known to the entire estimation procedure. Our objective is to compute the Brownian approximation to the ruin probability $\psi_B(u,T) = \mu(A_u)$, and its worst-case upper bound $\psi_{rob}(u,T) := \sup\left\{P(A_u) : d_c(\mu,P) \leq \delta\right\}$. The estimation methodology is data-driven in the following sense: we derive the estimates of moments m_1, m_2 and δ from the observed realizations $\{X_1,\ldots,X_N\}$ of claim sizes. While obtaining estimates for moments m_1 and m_2 are straightforward, to compute δ, we use the claim size samples $\{X_1,\ldots,X_N\}$ to embed realizations of compound Poisson risk process in a Brownian motion, as in Khoshnevisan (1993). Specific details of the algorithm that estimates δ can be found in Appendix A. The estimate of δ obtained is such that the compensated compound Poisson process of interest that models risk in our setup lies within the δ-neighborhood $\{P : d_c(\mu, P) \leq \delta\}$ with high probability. Next, given the knowledge of Lagrange multiplier λ^*, the evaluation of $\psi_B(u,T)$ and $\psi_{rob}(u,T) = \psi_B(\tilde{u},T)$ (with $\tilde{u} = u - 1/\lambda^*$) are straightforward, because of the closed-form expressions for level crossing probabilities of Brownian motion. Computation of λ^* is also elementary, and can be accomplished in multiple ways. We resort to an elementary sample average approximation scheme that solves for λ^*. The estimates of Brownian approximation to ruin probability $\psi_B(u,T)$ and the worst-case ruin probability $\psi_{rob}(u,T)$, for various values of u, are plotted in Table 1. To facilitate comparison, we have used a large deviation approximation of the ruin probabilities $\psi_{_{LD}}(u,T)$ that satisfy $\psi_{_{LD}}(u,T) \sim \psi(u,T)$, as $u \to \infty$, as a common denominator to compare magnitudes of the Brownian estimate $\psi_B(u,T)$ and the worst-case bound $\psi_{rob}(u,T)$. In particular, we have drawn samples of claim sizes from the Pareto distribution $1 - F(x) = 1 \wedge x^{-2.2}$. While the Brownian approximation $\psi_B(u,T)$ remarkably underestimates the ruin probability $\psi(u,T)$, the worst-case upper bound $\psi_{rob}(u,T)$ appears to yield estimates that are conservative, yet of the correct order of magnitude.

TABLE 1. Comparison of ruin probability estimate $\psi_B(u,T)$ and the worst-case upper bound $\psi_{rob}(u,T)$: Example 1. The denominator $\psi_{_{LD}}(u,T)$ is such that $\psi_{_{LD}}(u,t)/\psi(u,T) \to 1$, as $u \to \infty$.

u	$\psi_B(u,T)$	$\tilde{u}$	$\psi_{rob}(u,T)$	$\frac{\psi_B(u,T)}{\psi_{_{LD}}(u,T)}$	$\frac{\psi_{rob}(u,T)}{\psi_{_{LD}}(u,T)}$
50	5.18×10^{-2}	28.23	0.2294	3.32	14.71
100	4.26×10^{-4}	50.78	4.88×10^{-2}	1.07×10^{-1}	12.28
150	4.36×10^{-7}	62.76	1.84×10^{-2}	2.52×10^{-4}	10.65
200	5.05×10^{-11}	69.40	1.02×10^{-2}	5.35×10^{-8}	10.80
250	6.75×10^{-16}	74.29	6.50×10^{-3}	1.15×10^{-12}	10.98

For further discussion on this toy example, let us say that an insurer paying for the claims distributed according to X_i has a modest objective of keeping the probability of ruin before time T at a level below 0.01. The various combinations of initial capital u and safety loading factors η that achieve this objective are shown in Figure 1. While the (η, u) combinations that work for the Brownian approximation model is drawn in red, the corresponding (η, u) combinations that keep $\psi_{rob}(u,T) \leq 0.01$ is shown in blue.

A discussion on regulatory capital requirement. For the safety loading $\eta = 0.1$ we have considered, the Brownian approximation model $R_B(t)$ requires that the initial capital u be at least 60 to achieve $\psi_B(u,T) \leq 0.01$. On the other hand, to keep the robust ruin probability estimates $\psi_{rob}(u,T)$ below 0.01, it is required that $u \approx 200$, roughly 3 times more than the capital requirement of the Brownian model. From the insurer's point of view, due to the model uncertainty the contract is faced with, it is not uncommon to increase the premium income (or) initial capital by a factor of 3 or 4 (referred as "hysteria factor", see Embrechts et al. (1998), Lewis (2007)). This increase in premium can be thought of as a guess for the price for statistical uncertainty. Choosing higher premium and capital (larger η and u), along the blue curve in Figure 1, as dictated by the robust estimates of ruin probabilities $\psi_{rob}(u,T)$, instead provides a mathematically sound way of doing the same. Unlike the Brownian approximation model, the blue curve demonstrates that one cannot decrease the probability of ruin by arbitrarily increasing the premium alone, the initial capital also has to be sufficiently large. This prescription is consistent with the behaviour of markets with heavy-tailed claims where, in the absence of initial capital, a few initial large claims that occur before enough premium income gets accrued are enough to cause ruin. The minimum

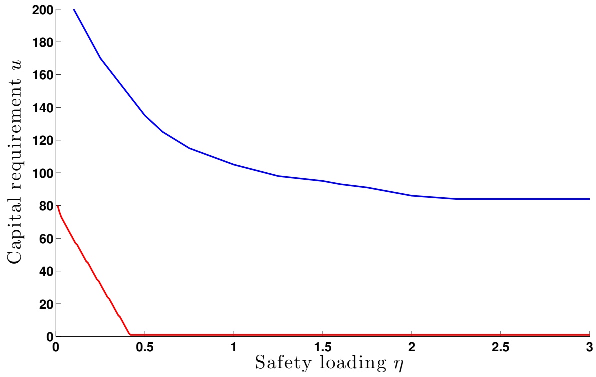


FIGURE 1. Safety loading vs capital requirement for the Brownian model $R_B(t)$ (in red) and its robust counterpart (in blue) in Example 1. The objective is to keep the probability of ruin below 0.01.

capital requirement u prescribed according to the worst-case estimates $\psi_{rob}(u, T)$ can be thought of as the regulatory minimum capital requirement.

Remark 7. Defining the candidate set of ambiguous probability measures via KL-divergence (or) other likelihood based divergence has been the most common approach while studying distributional robustness (see, for example, Hansen and Sargent (2001); Ben-Tal et al. (2013); Breuer and Csiszár (2013)). This would not have been useful in this setup, as the compensated Poisson process that models risk is not absolutely continuous with respect to Brownian motion. In general, it is normal to run into absolute continuity issues when we deal with continuous time stochastic processes. In such instances, as demonstrated in Example 1, modeling via optimal transport costs (or) Wasserstein distances not only offers a tractable alternative, but also yields insightful equivalent reformulations.

4. Proof of the duality theorem.

4.1. **Duality in compact spaces.** We prove Theorem 1 by first proving the following progressively strong duality results in Polish spaces S that are compact. Proofs of all the technicalities that are not central to the argument are relegated to Appendix B.

Proposition 5. *Suppose that S is a compact Polish space, and $f : S \to \mathbb{R}$ satisfies Assumption (A2). In addition to satisfying Assumption (A1), suppose that $c : S \times S \to \mathbb{R}_+$ is continuous. Then $I = J$, and a primal optimizer $\pi^* \in \Phi_{\mu,\delta}$ satisfying $I(\pi^*) = I$ exists.*

Proposition 6. *Suppose that S is a compact Polish space, $f : S \to \mathbb{R}$ satisfies Assumption (A2), and $c : S \times S \to \mathbb{R}_+$ satisfies Assumption (A1). Then $I = J$, and a primal optimizer $\pi^* \in \Phi_{\mu,\delta}$ satisfying $I(\pi^*) = I$ exists.*

As in the proof of the standard Kantorovich duality in compact spaces in Villani (2003), we use Fenchel duality theorem (see Theorem 1 in Chapter 7, Luenberger (1997)) to prove Proposition 5. However, the similarity stops there, owing to the reason that the set of primal feasible measures $\Phi_{\mu,\delta}$ is not tight when S is non-compact (contrast this with Kantorovich duality, where the collection of feasible measures $\Pi(\mu, \nu)$, the set of all joint distributions between any two probability measures μ and ν, is tight).

Proof of Proposition 5. As a preparation for applying Fenchel duality theorem, we first let $X = C_b(S \times S)$, and identify $X^* = M(S \times S)$ as its topological dual. Here, the spaces X and X^*, representing the vector space of bounded continuous functions and finite Borel measures on $S \times S$, respectively, are equipped with the supremum and total variation norms. The fact that X^* is the dual space of X is a consequence of Riesz representation theorem (see, for example, Rudin (1986)). Next, let C be the collection of all functions g in X that are of the form

$$g(x,y) = \phi(x) + \lambda c(x,y), \qquad \text{for all } x, y,$$

for some $\phi \in C_b(S)$ and $\lambda \geq 0$. In addition, let D denote the collection of all functions g in X that satisfy $g(x,y) \geq f(y)$, for every x, y. Every g in the convex subset C is defined by the pair (λ, ϕ), which, in turn, can be uniquely identified by,

$$\phi(x) = g(x,x) \quad \text{and} \quad \lambda = \frac{g(x,y) - \phi(x)}{c(x,y)},$$

for some x, y in S such that $c(x,y) \neq 0$. Keeping this invertible relationship in mind, define the functionals $\Phi : C \to \mathbb{R}$ and $\Gamma : D \to \mathbb{R}$ as below:

$$\Phi(g) := \lambda\delta + \int \phi d\mu \quad \text{and} \quad \Gamma(g) := 0.$$

Evidently, the functional Φ is convex, Γ is concave, and we are interested in

$$\inf_{g \in C \cap D} \big(\Phi(g) - \Gamma(g)\big) = \inf \big\{ J(\lambda, \phi) : \lambda \geq 0,\ \phi \in C_b(S),\ \phi(x) + \lambda c(x,y) \geq f(y) \text{ for all } x, y \big\}. \tag{26}$$

The next task is to identify the conjugate functionals Φ^*, Γ^* and their respective domains C^*, D^* as below:

$$C^* = \left\{ \pi \in X^* : \sup_{g \in C} \left\{ \int g d\pi - \Phi(g) \right\} < \infty \right\} \text{ and } D^* = \left\{ \pi \in X^* : \inf_{g \in D} \int g d\pi > -\infty \right\}.$$

The conjugate functionals $\Phi^* : C^* \to \mathbb{R}$ and $\Gamma^* : D^* \to \mathbb{R}$ are defined accordingly as,

$$\Phi^*(\pi) := \sup_{g \in C} \left\{ \int g d\pi - \Phi(g) \right\} \quad \text{and} \quad \Gamma^*(\pi) := \inf_{g \in D} \int g d\pi.$$

First, to determine C^*, we see that for every $\pi \in M(S \times S)$,

$$\begin{aligned}
\sup_{g \in C} \left\{ \int g d\pi - \Phi(g) \right\} &= \sup_{(\lambda,\phi)\ \in\ \mathbb{R}_+ \times C_b(S)} \left\{ \int \big(\phi(x) + \lambda c(x,y)\big) d\pi(x,y) - \left(\lambda\delta + \int \phi(x) d\mu(x)\right) \right\} \\
&= \sup_{(\lambda,\phi)\ \in\ \mathbb{R}_+ \times C_b(S)} \left\{ \lambda \left(\int c(x,y) d\pi(x,y) - \delta \right) + \int \phi(x) \big(d\pi(x,y) - d\mu(x)\big) \right\}, \\
&= \begin{cases} 0 & \text{if } \int c d\pi \leq \delta \text{ and } \pi(A \times S) = \mu(A) \text{ for all } A \in \mathcal{B}(S), \\ \infty & \text{otherwise.} \end{cases}
\end{aligned}$$

Therefore,

$$C^* = \left\{ \pi \in M(S \times S) : \int c d\pi \leq \delta,\ \pi(A \times S) = \mu(A) \text{ for all } A \in \mathcal{B}(S) \right\} \quad \text{and} \quad \Phi^*(\pi) = 0$$

Next, to identify D^*, we show in Lemma 15 in Appendix B that $\inf_{g \in D} \int g d\pi = -\infty$ if a measure $\pi \in M(S \times S)$ is not non-negative. On the other hand, if $\pi \in M(S \times S)$ is non-negative, then

$$\inf \left\{ \int g(x,y) d\pi(x,y) : g(x,y) \geq f(y) \text{ for all } x, y \right\} = \int f(y) d\pi(x,y).$$

This is because, f being an upper semicontinuous function that is bounded from above, it can be approximated pointwise by a monotonically decreasing sequence of continuous functions[6]. Then

[6] Let $f : X \to \mathbb{R}$ be an upper semicontinuous function, that is bounded from above, defined on a Polish space X. If $d(\cdot,\cdot)$ is a function that metrizes the Polish space X, then, for instance, the choice $f_n(x) = \sup_{y \in X}\{f(y) - nd(x,y)\}$ is continuous, and satisfies $f_n \downarrow f$ pointwise

the above equality follows as a consequence of monotone convergence theorem. As a result,

$$D^* = \left\{ \pi \in M_+(S \times S) : \int f(y) d\pi(x,y) > -\infty \right\} \quad \text{and} \quad \Gamma^*(\pi) = \int f(y) d\pi(x,y).$$

Then, $\Gamma^*(\pi) - \Phi^*(\pi) = \int f(y)d\pi(x,y)$ on $C^* \cap D^* = \{\pi \in \cup_{\nu \in P(S)} \Pi(\mu,\nu) : \int c d\pi \le \delta, \int f(y)d\pi(x,y) > -\infty\}$. As I is defined to equal $\sup\{\int f d\nu : d_c(\mu,\nu) \le \delta, \int f d\nu > -\infty\}$, it follows that

$$\sup \left\{ \Gamma^*(\pi) - \Phi^*(\pi) : \pi \in C^* \cap D^* \right\} = I. \tag{27}$$

As the set $C \cap D$ contains points in the relative interiors of C and D (consider, for example, the candidate $h(x,y) = c(x,y) + \sup_{x \in S} f(x)$ where $\sup_{x\in S} f(x) < \infty$ because f is upper semicontinuous and S is compact) and the epigraph of the function Γ has non-empty interior, it follows as a consequence of Fenchel's duality theorem (see, for example, Theorem 1 in Chapter 7, Luenberger (1997)) that

$$\inf_{g \in C \cap D} \big(\Phi(g) - \Gamma(g) \big) = \sup \left\{ \Gamma^*(\pi) - \Phi^*(\pi) : \pi \in C^* \cap D^* \right\},$$

where the supremum in the right is achieved by some $\pi^* \in \Phi_{\mu,\delta}$. In other words (see (26) and (27)),

$$\inf \left\{ J(\lambda,\phi) : \lambda \ge 0,\ \phi \in C_b(S),\ \phi(x) + \lambda c(x,y) \ge f(y) \text{ for all } x,y \right\} = \max \left\{ I(\pi) : \pi \in \Phi_{\mu,\delta} \right\} =: I.$$

As $C_b(S) \subseteq \mathfrak{m}_{\mathcal{U}}(S;\overline{\mathbb{R}})$, it follows from the definition of J that

$$J \le \inf \left\{ J(\lambda,\phi) : \lambda \ge 0,\ \phi \in C_b(S),\ \phi(x) + \lambda c(x,y) \ge f(y) \text{ for all } x,y \right\} = I.$$

However, due to weak duality (5), we have $J \ge I$. Therefore, $J = I = \max\{I(\pi) : \pi \in \Phi_{\mu,\delta}\}$. □

Proof of Proposition 6. First, observe that c is a nonnegative lower semicontinuous function defined on the Polish space $S \times S$. Therefore, we can write $c = \sup_n c_n$, where $(c_n : n \ge 1)$ is a nondecreasing sequence of continuous cost functions $c_n : S \times S \to \mathbb{R}_+$.[7] With c_n as cost function, define the corresponding primal and dual problems,

$$I_n = \sup_{\pi \in \Phi^n_{\mu,\delta}} I(\pi) \qquad \text{and} \qquad J_n = \inf_{(\lambda,\phi) \in \Lambda_{c_n,f}} J(\lambda,\phi).$$

Here, the feasible sets $\Phi^n_{\mu,\delta}$ and $\Lambda^n_{c,f}$ are obtained by suitably modifying the cost function in sets $\Phi_{\mu,\delta}$ and $\Lambda_{c,f}$ (defined in (6a) and (6b)) as below:

$$\Phi^n_{\mu,\delta} := \left\{ \pi \in \bigcup_{\nu \in P(S)} \Pi(\mu,\nu) : \int c_n d\pi \le \delta \right\} \text{ and}$$

$$\Lambda_{c_n,f} := \left\{ (\lambda,\phi) : \lambda \ge 0,\ \phi \in \mathfrak{m}_{\mathcal{U}}(S;\overline{\mathbb{R}}),\ \phi(x) + \lambda c_n(x,y) \ge f(y) \text{ for all } x,y \in S \right\}.$$

As the cost functions c_n are continuous, due to Proposition 5, there exists a sequence of measures $(\pi_n : n \ge 1)$ such that

$$I(\pi_n) = I_n = J_n. \tag{28}$$

As S is compact, the set $(\pi_n : n \ge 1)$ is automatically tight, and due to Prokhorov's theorem, there exists a subsequence $(\pi_{n_k} : k \ge 1)$ weakly converging to some $\pi^* \in P(S \times S)$. First, we check that π^* is feasible:

$$\int c d\pi^* = \int \lim_n c_n d\pi^* = \lim_n \int c_n d\pi^* = \lim_n \lim_k \int c_n d\pi_{n_k},$$

where the second and third equalities, respectively, are consequences of monotone convergence and weak convergence. In addition, since $(c_n : n \ge 1)$ is a nondecreasing sequence of functions,

$$\lim_k \int c_n d\pi_{n_k} \le \underline{\lim}_k \int c_{n_k} d\pi_{n_k} \le \delta.$$

[7] For instance, one can choose $c_n(x,y) = \inf_{\tilde{x},\tilde{y} \in S} \{c(\tilde{x},\tilde{y}) + nd((x,y),(\tilde{x},\tilde{y}))\}$, where the function d metrizes the Polish space $S \times S$. Then $c_n(\cdot,\cdot)$ is continuous, and satisfies $c_n(x,y) = 0$ if and only if $x = y$.

Therefore, $\int c d\pi^* \leq \delta$. Again as a simple consequence of weak convergence, we have $\int g(x)d\pi^*(x,y) = \lim_n \int g(x)d\pi_{n_k}(x,y) = \int g(x)d\mu(x)$, for every $g \in C_b(S)$. Therefore, $\pi^* \in \Phi_{\mu,\delta}$, and is indeed feasible. Next, as f is upper semicontinuous and S is compact, f is bounded from above. Then, due to the weak convergence $\pi_{n_k} \Rightarrow \pi^*$, the objective function $I(\pi^*)$ satisfies

$$I(\pi^*) = \int f d\pi^* \geq \overline{\lim_k} \int f d\pi_{n_k} = \overline{\lim_k} I(\pi_{n_k}) = \overline{\lim_k} J_{n_k},$$

because $I(\pi_{n_k}) = J_{n_k}$ as in (28). Further, as $c_n \leq c$, it follows that $\Lambda_{c_n,f}$ is a subset of $\Lambda_{c,f}$ and hence $J_n \geq J$, for all n. As a result, we obtain,

$$I(\pi^*) \geq \overline{\lim_k} J_{n_k} \geq J.$$

Since I never exceeds J (see (5)), it follows that $I(\pi^*) = I = J$. □

4.2. **A note on additional notations and measurability.** Given that we have established duality in compact spaces, the next step is to establish the same when S is not compact. For this purpose, we need the following additional notation. For any probability measure $\pi \in P(S \times S)$, let

$$S_\pi := \mathrm{Spt}(\pi_{_X}) \cup \mathrm{Spt}(\pi_{_Y}),$$

where $\mathrm{Spt}(\pi_{_X})$ and $\mathrm{Spt}(\pi_{_Y})$ denote the respective supports of marginals $\pi_{_X}(\cdot) := \pi(\,\cdot \times S)$ and $\pi_{_Y}(\cdot) := \pi(S \times \cdot\,)$. As every probability measure defined on a Polish space has σ-compact support, the set $S_\pi \times S_\pi$, which is a subset of $S \times S$, is σ-compact. As we shall see in the proof of Proposition 7, $S_\pi \times S_\pi$ can be written as the union of an increasing sequence of compact subsets $(S_n \times S_n : n \geq 1)$. It will then be easy to make progress towards strong duality in noncompact sets such as $S_\pi \times S_\pi$ by utilizing the strong duality results derived in Section 4.1 (which are applicable for compact sets $S_n \times S_n$) via a sequential argument; this is accomplished in Section 4.3 after introducing additional notation as follows. For any closed $K \subseteq S$, let

$$\Lambda(K \times K) := \big\{(\lambda, \phi) : \lambda \geq 0,\ \phi \in \mathfrak{m}_{\mathcal{U}}(K; \overline{\mathbb{R}}),\ \phi(x) + \lambda c(x,y) \geq f(y) \text{ for all } x, y \in K\big\}, \tag{29}$$

where $\mathfrak{m}_{\mathcal{U}}(K;\overline{\mathbb{R}})$ is used to denote the collection of measurable functions $\phi : (K, \mathcal{U}(K)) \to (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$, with $\mathcal{U}(K)$ denoting the universal $\sigma-$algebra of the Polish space K. With this notation, the dual feasible set $\Lambda_{c,f}$ is nothing but $\Lambda(S \times S)$. The next issue we address is the measurability of functions of the form $\phi_\lambda(x) = \sup_{y \in S} \big\{f(y) - \lambda c(x,y)\big\}$. For any $u \in \mathbb{R}$,

$$\big\{\phi_\lambda(x) > u\big\} = \mathrm{Proj}_1\left(\{(x,y) \in S \times S : f(y) - \lambda c(x,y) > u\}\right),$$

which is analytic, because projections $\mathrm{Proj}_i(A) := \{x_i \in S : (x_1, x_2) \in A\}, i = 1, 2$, of any Borel measurable set A are analytic (see, for example, Proposition 7.39 in Bertsekas and Shreve (1978)). As analytic subsets of S lie in $\mathcal{U}(S)$ (see Corollary 7.42.1 in Bertsekas and Shreve (1978)), the function $\phi_\lambda(x) : (S, \mathcal{U}(S)) \to (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$ is measurable (that is, universally measurable; refer Chapter 7 of Bertsekas and Shreve (1978) for an introduction to universal measurability and analytic sets).

4.3. **Extension of duality to non-compact spaces.** Proposition 7, presented below in this section, is an important step towards extending the strong duality results proved in Section 4.1 to non-compact sets.

Proposition 7. *Suppose that Assumptions (A1) and (A2) are in force. Let $\pi \in P(S \times S)$ be a probability measure satisfying*
(a) $\int c(x,y)d\pi(x,y) < \infty$,
(b) $\int f(y)d\pi(x,y) \in (-\infty, \infty)$, *and*
(c) $\pi(A \times S) = \mu(A)$ *for every* $A \in \mathcal{B}(S)$. *Then*

$$\inf_{(\lambda,\phi) \in \Lambda(S_\pi \times S_\pi)} J(\lambda, \phi) \leq I$$

Proof of Proposition 7. As any Borel probability measure on a Polish space is concentrated on a σ-compact set (see Theorem 1.3 of Billingsley (1999)), the sets $\mathrm{Spt}(\pi_X)$ and $\mathrm{Spt}(\pi_Y)$ are σ-compact, and therefore $S_\pi \times S_\pi$ is σ-compact as well. Then, by the definition of σ-compactness, the set $S_\pi \times S_\pi$ can be written as the union of an increasing sequence of compact subsets $(C_n : n \geq 1)$ of $S_\pi \times S_\pi$. If we let S_n to be the union of the projections of C_n over its two coordinates, then it follows that $(S_n \times S_n : n \geq 1)$ is an increasing sequence of subsets of $S_\pi \times S_\pi$ satisfying $\mathrm{Spt}(\pi) \subseteq S_\pi \times S_\pi = \cup_{n\geq 1}(S_n \times S_n)$. As $\int |f(y)| d\pi(x,y)$ and $\int c(x,y)d\pi(x,y)$ are finite, one can pick the increasing sequence $(S_n \times S_n : n \geq 1)$ with $S_\pi \times S_\pi = \cup_{n\geq 1}(S_n \times S_n)$ to be satisfying,

$$p_n := \pi(S_n \times S_n) \ \geq 1 - \frac{1}{n}, \tag{30a}$$

$$\int c(x,y)\mathbf{1}_{(S_n\times S_n)^c} d\pi(x,y) \ \leq \ \frac{\delta}{n} \quad \text{and} \tag{30b}$$

$$\int |f|(y)\mathbf{1}_{(S_n\times S_n)^c} d\pi(x,y) \leq \frac{1}{n}, \tag{30c}$$

where $(S_n \times S_n)^c := (S \times S) \setminus (S_n \times S_n)$, and $\delta \in (0,\infty)$ is introduced while defining the primal feasibility set $\Phi_{\mu,\delta}$ (recall that $I := \sup\{I(\pi) : \pi \in \Phi_{\mu,\delta}\}$) and the dual objective $J(\lambda,\phi) = \lambda\delta + \int \phi d\mu$ in Section 2.2. Having chosen the compact subsets $(S_n : n \geq 1)$, define a joint probability measure $\pi_n \in P(S_n \times S_n)$ and its corresponding marginal $\mu_n \in P(S_n)$ as below:

$$\pi_n(\cdot) = \frac{\pi\big(\ \cdot\ \cap (S_n \times S_n)\big)}{p_n} \quad \text{and} \quad \mu_n(\cdot) = \pi_n(\ \cdot \times S_n),$$

for every $n \geq 1$. Additionally, let $\delta_n := \delta(1 - 1/n)$. For every $n \geq 1$, these new definitions can be used to define "restricted" primal and dual optimization problems I_n and J_n supported on the compact set $S_n \times S_n$:

$$I_n := \sup\left\{ \int f(y)\gamma(dx,dy) : \gamma \in \bigcup_{\nu\in P(S_n)} \Pi(\mu_n,\nu), \int c(x,y)d\gamma(x,y) \leq \delta_n \right\} \text{ and} \tag{31}$$

$$J_n := \inf\left\{ \lambda\delta_n + \int \phi d\mu_n : \ (\lambda,\phi) \in \Lambda(S_n \times S_n) \right\}. \tag{32}$$

To comprehend (32), refer the definition of $\Lambda(\cdot)$ in (29). As S_n is compact, due to the duality result in Proposition 6, we know that there exists a $\gamma_n^* \in P(S_n \times S_n)$ that is feasible for optimization of I_n and satisfies

$$\int f(y)\gamma_n^*(dx,dy) = I_n = J_n. \tag{33}$$

Using this optimal measure γ_n^*, one can, in turn, construct a measure $\tilde{\pi} \in P(S \times S)$ by stitching it together with the residual portion of π as below:

$$\tilde{\pi}(\cdot) = p_n\gamma_n^*\big(\ \cdot\ \cap (S_n \times S_n)\big) + \pi\big(\ \cdot\ \cap (S_n \times S_n)^c\big).$$

Having tailored a candidate measure $\tilde{\pi} \in P(S \times S)$, next we check its feasibility that $\tilde{\pi} \in \Phi_{\mu,\delta}$ as follows: Since $\gamma_n^* \in \Pi(\mu_n,\nu)$ for some $\nu \in P(S_n)$, it is easy to check, as below, that $\tilde{\pi}$ has μ as its marginal for the first component: first, it follows from the definition of $\tilde{\pi}$ that

$$\begin{aligned}\tilde{\pi}(\ \cdot \times S) &= p_n\gamma_n^*((\ \cdot\ \cap S_n) \times S_n) + \pi((\ \cdot \times S)\ \cap\ (S_n \times S_n)^c)\\ &= p_n\mu_n(\ \cdot\ \cap S_n) + \pi((\ \cdot \times S)\ \cap\ (S_n \times S_n)^c).\end{aligned}$$

As $\mu_n(\cdot) = \pi(\ \cdot \times S_n)/p_n$, it follows that $p_n\mu_n(\ \cdot \cap S_n) = \pi((\ \cdot \times S) \cap (S_n \times S_n))$, and consequently,

$$\tilde{\pi}(\ \cdot \times S) = \pi((\ \cdot \times S) \cap (S_n \times S_n)) + \pi((\ \cdot \times S)\ \cap\ (S_n \times S_n)^c) = \pi(\ \cdot \times S) = \mu(\cdot).$$

Therefore, $\tilde{\pi} \in \Pi(\mu,\nu)$ for some $\nu \in P(S)$. Further, as γ_n^* is a feasible solution for the optimization problem in (31),

$$\int c d\tilde{\pi} = p_n \int_{S_n\times S_n} c d\gamma_n^* + \int_{(S_n\times S_n)^c} c d\pi \ \leq\ \delta_n + \frac{\delta}{n},$$

which does not exceed δ. To see this, refer (30b) and recall that $\delta_n := \delta(1-1/n)$. Therefore, the stitched measure $\tilde{\pi}$ is primal feasible, that is $\tilde{\pi} \in \Phi_{\mu,\delta}$. Consequently,

$$I \geq I(\tilde{\pi}) = \int f(y) d\tilde{\pi}(x,y) = p_n \int f(y) d\gamma_n^*(x,y) + \int f(y) \mathbf{1}_{(S_n \times S_n)^c} d\pi(x,y).$$

As S_n is chosen to satisfy (30c), we have $I \geq p_n I_n - n^{-1}$. Therefore, it is immediate from (33) that

$$J_n \leq \left(1 - \frac{1}{n}\right)^{-1} \left(I + \frac{1}{n}\right), \tag{34}$$

which, if $I < \infty$, is a useful upper bound for all of $J_n, n \geq 1$. See that the statement of Proposition 7 already holds when I equals ∞. Therefore, let us take I to be finite. Next, given $n \geq 1$ and $\varepsilon > 0$, take an ε-optimal solution (λ_n, ϕ_n) for J_n : that is, $(\lambda_n, \phi_n) \in \Lambda(S_n \times S_n)$ and

$$\lambda_n \delta_n + \int \phi_n d\mu_n \leq J_n + \varepsilon.$$

For any pair (λ_n, ϕ) that belongs to $\Lambda(S_n \times S_n)$, we have from the definition of $\Lambda(S_n \times S_n)$ that $\phi(x)$ is larger than $\sup_{z \in S_n}\{f(z) - \lambda_n c(x,z)\}$, for every $x \in S_n$.

$$\lambda_n \delta_n + \int \sup_{z \in S_n} \big\{f(z) - \lambda_n c(x,z)\big\} d\mu_n(x) \leq J_n + \varepsilon,$$

for every n. As $\mu_n(\cdot) = \pi(\cdot \times S_n)/p_n$, combining the above expression with (34) results in

$$\overline{\lim_{n\to\infty}} \left(\lambda_n \delta_n + \int \sup_{z \in S_n} \big\{f(z) - \lambda_n c(x,z)\big\} \frac{\mathbf{1}_{S_n}(x)\mathbf{1}_{S_n}(y)}{p_n} d\pi(x,y) \right) \leq I + \varepsilon. \tag{35}$$

Since $c(x,x) = 0$, the integrand admits $f(x)$ as a lower bound on $S_n \times S_n$; further, as $f \in L^1(d\mu)$, $-f^-(x)$ serves as a common integrable lower bound for all $n \geq 1$. This has two consequences: First, because of the finite upper bound in (35), $\underline{\lim}_n \lambda_n$ and $\overline{\lim}_n \lambda_n$ are finite (recall that $\lambda_n \geq 0$ as well), and there exists a subsequence $(\lambda_{n_k} : k \geq 1)$ such that $\lambda_{n_k} \to \lambda^*$, as $k \to \infty$, for some $\lambda^* \in [0, \infty)$. The second consequence of the existence of a common integrable lower bound is that we can apply Fatou's lemma in (35) along the subsequence $(n_k : k \geq 1)$ to obtain

$$\begin{aligned} I + \varepsilon \ \geq\ & \underline{\lim_{k\to\infty}} \left(\lambda_{n_k} \delta_{n_k} + \int \sup_{z \in S_{n_k}} \big\{f(z) - \lambda_n c(x,z)\big\} \frac{\mathbf{1}_{S_{n_k}}(x)\mathbf{1}_{S_{n_k}}(y)}{p_{n_k}} d\pi(x,y) \right) \\ & \geq \lambda^* \delta + \int_{S_\pi \times S_\pi} \sup_{z \in \cup_k S_{n_k}} \big\{f(z) - \lambda^* c(x,z)\big\} d\pi(x,y) \\ & = \lambda^* \delta + \int_S \sup_{z \in S_\pi} \big\{f(z) - \lambda^* c(x,z)\big\} d\mu(x). \end{aligned}$$

Here, we have used that $\delta_n \to \delta$, $p_n \to 1$ as $n \to \infty$, $\cup_{n\geq 1} S_n = \cup_{k \geq 1} S_{n_k} = S_\pi$, and the fact that π is supported on a subset of $S_\pi \times S_\pi$. The fact that $\underline{\lim}_k \sup_{z \in S_{n_k}}\{f(z) - \lambda_{n_k} c(x,z)\}$ is at least $\sup_{z \in \cup_k S_{n_k}}\{f(z) - \lambda^* c(x,z)\}$, for every $x \in S$, is carefully checked in Lemma 16, Appendix B. If we let $\phi^*(x) := \sup_{y \in S_\pi} \big\{f(y) - \lambda^* c(x,y)\big\}$, then $(\lambda^*, \phi^*) \in \Lambda(S_\pi \times S_\pi)$, and as $\varepsilon > 0$ is arbitrary, it follows from the above inequality that $J(\lambda^*, \phi^*) \leq I$. This completes the proof. □

Proof of Theorem 1(a). If $I = \infty$, then as I never exceeds J, J equals ∞ as well, and there is nothing to prove. Next, consider the case where I is finite: that is, $I \in [\int f d\mu, \infty)$. Let E denote the convex set of probability measures $\pi \in P(S \times S)$ that satisfy conditions (a)-(c) of Proposition 7. Then, due to Proposition 7, for every $\pi \in E$,

$$I \geq \inf_{(\lambda,\phi) \in \Lambda(S_\pi \times S_\pi)} \left\{ \lambda\delta + \int \phi(x) d\mu(x) \right\} \geq \inf_{\lambda \geq 0} \left\{ \lambda\delta + \int \sup_{y \in S_\pi} \big\{f(y) - \lambda c(x,y)\big\} d\mu(x) \right\}.$$

For any $\pi \in E$ and $\lambda \geq 0$, define

$$T(\lambda, \pi) := \lambda\delta + \int \sup_{y \in S_\pi} \big\{f(y) - \lambda c(x,y)\big\} d\mu(x).$$

As $c(x,x) = 0$, it is easy to see that $T(\lambda,\pi) \geq \lambda\delta + \int f d\mu$. Since $T(\lambda,\pi) > I$ for every $\lambda > \lambda_{\max} := (I - \int f d\mu)/\delta$, one can rather restrict attention to the compact subset $[0, \lambda_{\max}]$, as follows:

$$I \geq \inf_{\lambda \in [0,\lambda_{\max}]} T(\lambda, \pi). \tag{36}$$

Being a pointwise supremum of a family of affine functions, $\sup_{y \in S_\pi}\{f(y) - \lambda c(x,y)\}$ is a lower semicontinuous with respect to the variable λ. Further, recall that $T(\lambda,\pi) \geq \lambda\delta + \int f d\mu$. Thus, for any $\lambda_n \to \lambda$, due to Fatou's lemma,

$$\begin{aligned}\varliminf_{n\to\infty} T(\lambda_n, \pi) &\geq \lambda\delta + \int \varliminf_n \sup_{y\in S_\pi} \{f(y) - \lambda_n c(x,y)\} \, d\mu(x) \\ &\geq \lambda\delta + \int \sup_{y\in S_\pi} \{f(y) - \lambda c(x,y)\} \, d\mu(x) = T(\lambda,\pi),\end{aligned}$$

which means that $T(\lambda,\pi)$ is lower semicontinuous in λ. Along with this lower semicontinuity, for every fixed π, $T(\lambda,\pi)$ is also convex in λ. In addition, for any $\alpha \in (0,1)$ and $\pi_1, \pi_2 \in E$,

$$\begin{aligned}T(\lambda, \alpha\pi_1 + (1-\alpha)\pi_2) &= \lambda\delta + \int \sup_{y \in S_{\alpha\pi_1 + (1-\alpha)\pi_2}} \big\{f(y) - \lambda c(x,y)\big\} d\mu(x) \\ &\geq \max_{i=1,2} \left\{ \lambda\delta + \int \sup_{y \in S_{\pi_i}} \big\{f(y) - \lambda c(x,y)\big\} d\mu(x) \right\} \geq \ \alpha T(\lambda,\pi_1) + (1-\alpha) T(\lambda,\pi_2).\end{aligned}$$

Therefore, $T(\lambda,\pi)$ is a concave function in π for every fixed λ. One can apply a standard minimax theorem (see, for example, Sion (1958)) to conclude

$$\sup_{\pi\in E} \inf_{\lambda\in[0,\lambda_{\max}]} T(\lambda,\pi) = \inf_{\lambda\in[0,\lambda_{\max}]} \sup_{\pi \in E} T(\lambda,\pi).$$

This observation, in conjunction with (36), yields

$$I \geq \sup_{\pi\in E} \inf_{\lambda\in[0,\lambda_{\max}]} T(\lambda,\pi) = \inf_{\lambda\in[0,\lambda_{\max}]} \left\{ \lambda\delta + \sup_{\pi\in E} \int \sup_{y\in S_\pi} \big\{f(y) - \lambda c(x,y)\big\} d\mu(x) \right\}$$

Lemma 8 below is the last piece of technicality that completes the proof of Part (a) of Theorem 1.

Lemma 8. *Suppose that Assumptions (A1) and (A2) are in force. Then, for any* $\lambda \geq 0$,

$$\sup_{\pi\in E} \int \sup_{y \in S_\pi} \big\{f(y) - \lambda c(x,y)\big\} d\mu(x) = \int \sup_{y\in S} \big\{f(y) - \lambda c(x,y)\big\} d\mu(x).$$

If Lemma 8 holds, then it is automatic that

$$I \geq \inf_{\lambda\in[0,\lambda_{\max}]} \left\{ \lambda\delta + \int \sup_{y\in S} \big\{f(y) - \lambda c(x,y)\big\} d\mu(x) \right\} \geq J,$$

which will complete the proof of Part (a) of Theorem 1.

Proof. **Proof of Lemma 8.** For brevity, let $g(x,y) := f(y) - \lambda c(x,y)$ and $\phi_\lambda(x) := \sup_{y\in S} g(x,y)$. For any $n \geq 1$ and $k \leq n^2$, define the sets,

$$A_{k,n} := \{(x,y) \in S\times S : (k-1)/n \leq g(x,y) \leq k/n\} \quad \text{and} \quad B_{k,n} := \mathrm{Proj}_1(A_{k,n}) \setminus \cup_{j>k} \mathrm{Proj}_1(A_{j,n}).$$

Here, recall that for any $A \subset S \times S$, $\mathrm{Proj}_1(A) = \{x_1 : (x_1,x_2) \in A\}$. The sequence $(B_{k,n} : k \leq n^2)$ also admits the following equivalent backward recursive definition,

$$B_{n^2,n} = \mathrm{Proj}_1\left(A_{n^2,n}\right), \text{ and } B_{k,n} = \mathrm{Proj}_1\left(A_{k,n}\right) \setminus \cup_{j>k} B_{j,n} \text{ for } k < n^2,$$

that renders the collection $(B_{k,n} : k \leq n^2)$ disjoint. As the function g is upper semicontinuous, it is immediate that the sets $A_{k,n}$ are Borel measurable and their corresponding projections $B_{k,n}$ are analytic subsets of S. Then, due to Jankov-von Neumann selection theorem[8], there exists, for each $k \leq n^2$ such that $A_{k,n} \neq \varnothing$, an universally measurable function $\gamma_k(x) : \mathrm{Proj}_1(A_{k,n}) \to S$ such that $(x, \gamma_k(x)) \in A_{k,n}$, or, in other words,

$$\frac{k-1}{n} \leq g(x, \gamma_k(x)) \leq \frac{k}{n}. \tag{37}$$

Next, let us define the function $\Gamma_n : S \to S$ as below:

$$\Gamma_n(x) = \begin{cases} \gamma_k(x), & \text{if } x \in B_{k,n} \text{ for some } k \leq n^2 \\ x, & \text{otherwise} . \end{cases}$$

This definition is possible because $B_{k,n} \subseteq \mathrm{Proj}_1(A_{k,n})$, and the collection $(B_{k,n} : k \leq n^2)$ is disjoint. As each γ_k is universally measurable, the composite function $\Gamma_n(x)$ is universally measurable as well, and satisfies that

$$\sup\left\{g(x,y) : g(x,y) \leq n, y \in S\right\} - \frac{1}{n} \leq g(x, \Gamma_n(x)) \leq f(x) \vee n,$$

for each x. Both sides of the inequality above can be inferred from (37) after recalling that $g(x,x) = f(x) - \lambda c(x,x) = f(x)$. Letting $n \to \infty$, we see that

$$\phi_\lambda(x) := \sup\{g(x,y) : y \in S\} \;\leq\; \varliminf_{n} g(x, \Gamma_n(x)) \;\leq\; \infty. \tag{38}$$

Next, define the family of probability measures $(\pi_n : n \geq 1)$ as below:

$$d\pi_n(x,y) := d\mu(x) \times d\delta_{\Gamma_n(x)}(y),$$

with $\delta_a(\cdot)$ representing the dirac measure centred at $a \in S$. As $g(x,y) \leq f(x) \vee n$, π_n almost surely, we have that the functions $c(x,y)$ and $f(y)$ are integrable with respect to π_n. This means that $\pi_n \in E$ (here, recall that E is the set of probability measures satisfying conditions (a)-(c) in the statement of Proposition 7). Further, as $S_{\pi_n} \supseteq \mathrm{Spt}(\mu)$,

$$\sup_{y \in S_{\pi_n}} g(x,y) = \sup_{y \in S_{\pi_n}} \big\{f(y) - \lambda c(x,y)\big\} \geq f(x) - \lambda c(x,x) = f(x), \qquad \text{for all } x,\ \mu \text{ a.s.},$$

which is integrable with respect to the measure μ. Therefore, due to Fatou's lemma,

$$\begin{aligned} \varliminf_{n\to\infty} \int \sup_{y \in S_{\pi_n}} g(x,y) d\mu(x) &\geq \int \varliminf_{n\to\infty} \sup_{y \in S_{\pi_n}} g(x,y) d\mu(x) \\ &\geq \int \varliminf_{n\to\infty} g(x, \Gamma_n(x)) d\mu(x) \geq \int \phi_\lambda(x) d\mu(x), \end{aligned}$$

where the last inequality follows from (38). Since π_n is a member of set E, for every n, it is then immediate that

$$\sup_{\pi \in E} \int \sup_{y \in S_\pi} g(x,y) d\mu(x) \geq \varliminf_{n\to\infty} \int \sup_{y \in S_{\pi_n}} g(x,y) d\mu(x) \geq \int \phi_\lambda(x) d\mu(x).$$

$$\text{Thus, } \sup_{\pi \in E} \int \sup_{y \in S_\pi} \big\{f(y) - \lambda c(x,y)\big\} d\mu(x) \geq \int \sup_{y \in S} \big\{f(y) - \lambda c(x,y)\big\} d\mu(x).$$

As $S_\pi \subseteq S$, for every $\pi \in E$, the inequality in the reverse direction is trivially true. This completes the proof of Lemma 8. □

[8]See, for example, Chapter 7 of Bertsekas and Shreve (1978) for an introduction to analytic subsets and Jankov-von Neumann measurable selection theorem

Proof of Theorem 1(b). To summarize, in the proof of Part (a) of Theorem 1, we established that

$$I = \inf_{\lambda \geq 0} \left\{ \lambda\delta + \int \phi_\lambda(x) d\mu(x) \right\},$$

where $\phi_\lambda(x) := \sup_{y \in S}\{f(y) - \lambda c(x,y)\}$ is lower-semicontinuous in λ, and is bounded by $f(x)$ from below. Recall that $\int f d\mu$ is finite. If we let

$$g(\lambda) := \lambda\delta + \int \phi_\lambda(x) d\mu(x),$$

then due to a routine application of Fatou's lemma, we have that $\underline{\lim}_{n\to\infty} g(\lambda_n) \geq g(\lambda)$, whenever $\lambda_n \to \lambda$. In other words, $g(\cdot)$ is lower semicontinuous. In addition, as $g(\lambda) \geq \lambda\delta + \int f(x)d\mu(x)$, we have that $g(\lambda) \to \infty$ if $\lambda \to \infty$. This means that the level sets $\{\lambda \geq 0 : g(\lambda) \leq u\}$ are compact for every u, and therefore, $g(\cdot)$ attains its infimum. In other words, there exists a $\lambda^* \in [0, \infty)$ such that

$$I = \lambda^*\delta + \int \phi_{\lambda^*}(x) d\mu(x).$$

Thus, we conclude that a dual optimizer of the form $(\lambda^*, \phi_{\lambda^*})$ always exists. Next, if the primal optimizer π^* satisfying $I = I(\pi^*)$ also exist, then

$$\int f(y) d\pi^*(x,y) = \lambda^*\delta + \int \phi_{\lambda^*}(x) d\mu(x),$$

which gives us complementary slackness conditions (8a) and (8b), as a consequence of equality getting enforced in the following series of inequalities:

$$\begin{aligned}
\int f(y) d\pi^*(x,y) &= \int \big(f(y) - \lambda^* c(x,y)\big) d\pi^*(x,y) + \lambda^* \int c(x,y) d\pi^*(x,y) \\
&\leq \int \phi_{\lambda^*}(x) d\pi^*(x,y) + \lambda^* \int c(x,y) d\pi^*(x,y) \\
&\leq \int \phi_{\lambda^*}(x) d\mu(x) + \lambda^*\delta.
\end{aligned}$$

Alternatively, if the complementary slackness conditions (8a) and (8b) are satisfied by any $\pi^* \in \Phi_{\mu,\delta}$ and $(\lambda^*, \phi^*_\lambda) \in \Lambda_{c,f}$, then automatically,

$$\begin{aligned}
\int f(y) d\pi^*(x,y) &= \int \big(f(y) - \lambda^* c(x,y)\big) d\pi^*(x,y) + \lambda^* \int c(x,y) d\pi^*(x,y) \\
&= \int \sup_{z \in S} \big\{f(z) - \lambda^* c(x,z)\big\} d\pi^*(x,y) + \lambda^*\delta \\
&= \int \phi_{\lambda^*}(x) d\mu(x) + \lambda^*\delta = J(\lambda^*, \phi_{\lambda^*}),
\end{aligned}$$

thus proving that π^* and $(\lambda^*, \phi_{\lambda^*})$ are, respectively, the primal and dual optimizers. This completes the proof of Theorem 1. □

5. On the existence of a primal optimal transport plan.

Unlike the well-known Kantorovich duality where an optimizer that attains the infimum in (2) exists if the transport cost function $c(\cdot,\cdot)$ is nonnegative and lower semicontinuous (see, for example, Theorem 4.1 in Villani (2008)), a primal optimizer π^* satisfying $I(\pi^*) = I = J$ need not exist for the primal problem, $I = \sup\{I(\pi) : \pi \in \Phi_{\mu,\delta}\}$, that we have considered in this paper. The feasible set for the problem (2) in Kantorovich duality, which is the set of all joint distributions $\Pi(\mu,\nu)$ with given μ and ν as marginal distributions, is compact in the weak topology. On the other hand, the primal feasibility set $\Phi_{\mu,\delta}$, for a given μ and $\delta > 0$, that we consider in this paper is not necessarily compact, and the existence of a primal optimizer $\pi^* \in \Phi_{\mu,\delta}$, satisfying $I(\pi^*) = I$, is not guaranteed. Example 2 below demonstrates a setting where there is no primal optimizer $\pi^* \in \Phi_{\mu,\delta}$ satisfying $I(\pi^*) = I$ even if the transport cost $c(\cdot,\cdot)$ is chosen as a metric defined on the Polish space S.

Example 2. *Consider $S = \mathbb{R}$. Let $\mu = \delta_{\{0\}}$, the dirac measure at 0, be the reference measure defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Let $f(x) = (1-\exp(-x))_+$ and $c(x,y) = |x-y|/(1+|x-y|)$ for all $x, y \in \mathbb{R}$. The primal problem of interest is $I = \sup\{I(\pi) : \pi \in \Phi_{\mu,\delta}\}$, where $I(\pi) := \int f(y)d\pi(x,y)$, for the choice $\delta = 2$. We first argue that $I = 1$ here: As $f(x) \leq 1$ for every $x \in \mathbb{R}$, we have $I \leq 1$. For $\lambda \geq 0$, recall the definition $\phi_\lambda(x) := \sup_{y\in\mathbb{R}}\{f(y) - \lambda c(x,y)\}$. As Assumptions (A1) and (A2) hold, an application of Theorem 1(a) results in,*

$$
\begin{aligned}
I &= \inf_{\lambda\geq 0}\left\{\lambda\delta + \int \phi_\lambda(x)d\mu(x)\right\} = \inf_{\lambda\geq 0}\left\{2\lambda + \phi_\lambda(0)\right\} \\
&= \inf_{\lambda\geq 0}\left\{2\lambda + \sup_{y\geq 0}\left\{1 - e^{-y} - \lambda\frac{y}{1+y}\right\}\right\} \geq \sup_{y\geq 0}\inf_{\lambda\geq 0}\left\{1 - e^{-y} + 2\lambda - \lambda\frac{y}{1+y}\right\} = 1.
\end{aligned}
$$

Since we had already verified that $I \leq 1$, the above lower bound results in $I = 1$, with the infimum being attained at $\lambda^* = 0$. However, as $f(x) < 1$ for every $x \in \mathbb{R}$, it is immediate that $I(\pi) = \int f(y)d\pi(x,y)$ is necessarily smaller than 1 for all $\pi \in \Phi_{\mu,\delta}$. Therefore, there does not exist a primal optimizer $\pi^* \in \Phi_{\mu,\delta}$ such that $I(\pi^*) = I$. A careful examination of Part (b) of Theorem 1 also results in the same conclusion: A primal optimizer, if it exists, is concentrated on $\{(x,y) : x \in S, y \in \arg\max\{f(y) - \lambda^* c(x,y)\}\}$, which is empty in this example, because $\arg\max\{f(y) - \lambda^* c(x,y)\} = \arg\max_{y\geq 0}\{1 - \exp(-y) - 0 \times c(x,y)\} = \varnothing$, for every $x \in S$. □

Recall from Proposition 6 that a primal optimizer always exists whenever the underlying Polish space S is compact. In the absence of compactness of S, Proposition 9 below presents additional topological properties (P-Compactness) and (P-USC) under which a primal optimizer exists. Corollary 1, that follows Proposition 9, discusses a simple set of sufficient conditions for which these properties listed in Proposition 9 are easily verified.

Additional notation. Define $\Phi'_{\mu,\delta} := \left\{\pi \in \Phi_{\mu,\delta} : \pi\big((x,y) \in S \times S : f(x) \leq f(y)\big) = 1\right\}$. For every $\pi \in \Phi_{\mu,\delta}$, it follows from Lemma 17 in Appendix B that there exists $\pi' \in \Phi'_{\mu,\delta}$ such that $I(\pi') \geq I(\pi)$, whenever $I(\pi)$ is well-defined. As a result,

$$I := \sup\{I(\pi) : \pi \in \Phi_{\mu,\delta}\} = \sup\{I(\pi) : \pi \in \Phi'_{\mu,\delta}\}.$$

Throughout this section, we assume that $(\lambda^*, \phi_{\lambda^*}) \in \Lambda_{c,f}$ is a dual-optimal pair satisfying $I = J = J(\lambda^*, \phi_{\lambda^*}) < \infty$. As $J(\lambda^*, \phi_{\lambda^*}) = \lambda^*\delta + \int \phi_{\lambda^*} d\mu < \infty$, it follows that $\phi_{\lambda^*}(x) < \infty$, $\mu-$almost surely.

Proposition 9. *Suppose that the Assumptions (A1) and (A2) are in force. In addition, suppose that the functions $c(\cdot,\cdot)$ and $f(\cdot)$ are such that the following properties are satisfied:*

Property 1 (P-Compactness)**.** For any $\varepsilon > 0$, there exists a compact $K_\varepsilon \subseteq S$ such that $\mu(K_\varepsilon) > 1-\varepsilon$, and the closure $\mathrm{cl}(\Gamma_\varepsilon)$ of the set, $\Gamma_\varepsilon := \left\{(x,y) \in K_\varepsilon \times S : f(y) - \lambda^* c(x,y) \geq \phi_{\lambda^*}(x) - \gamma\right\}$, is compact, for some $\gamma \in (0,\infty)$.

Property 2 (P-USC)**.** For every collection $\{\pi_n : n \geq 1\} \subseteq \Phi'_{\mu,\delta}$ such that $\pi_n \Rightarrow \pi^*$ for some $\pi^* \in \Phi_{\mu,\delta}$, we have $\overline{\lim}_n I(\pi_n) \leq I(\pi^*)$.

Then there exists a primal optimizer $\pi^* \in \Phi_{\mu,\delta}$ satisfying $I(\pi^*) = I = J = J(\lambda^*, \phi_{\lambda^*})$.

It is well-known that upper semicontinuous functions defined on compact sets attain their supremum. As Property 1 (P-Compactness) and Property 2 (P-USC), respectively, enforce a specific type of compactness and upper semicontinuity requirements that are relevant to our setup, we use the suggestive labels (P-Compactness) and (P-USC), respectively, to identify Properties 1 and 2 throughout the rest of this section.

Proof. Proof of Proposition 9. As $I = \sup\{I(\pi) : \pi \in \Phi'_{\mu,\delta}\}$, consider a collection $\{\pi_n : n \geq 1\} \subseteq \Phi'_{\mu,\delta}$ such that $I(\pi_n) \geq I - 1/n$. We first show that Property (P-Compactness) guarantees the existence of a weakly convergent subsequence $(\pi_{n_k} : k \geq 1)$ of $(\pi_n : n \geq 1)$ satisfying $\pi_{n_k} \Rightarrow \pi^*$, for some $\pi^* \in \Phi'_{\mu,\delta}$.

Step 1 (Verifying tightness of $(\pi_n : n \geq 1)$): As $I(\pi_n) \geq I - n^{-1}$, it follows from (11) that

$$\int \big(\phi_{\lambda^*}(x) - f(y) + \lambda^* c(x,y)\big)\, d\pi_n(x,y) \leq \frac{1}{n} \quad \text{and} \quad \lambda^* \left(\delta - \int c\, d\pi_n\right) \leq \frac{1}{n}. \tag{39}$$

for every $n \geq 1$. For any $a > 0$, Markov's inequality results in,

$$\pi_n\left(\big\{(x,y) \in S \times S : f(y) - \lambda^* c(x,y) < \phi_{\lambda^*}(x) - a\big\}\right) \leq \frac{1}{a}\int \big(\phi_{\lambda^*}(x) - f(y) + \lambda^* c(x,y)\big)\, d\pi_n(x,y),$$

which is at most $1/(na)$. Then, given any $\varepsilon > 0$ and for the choice $a = \gamma$ specified in (P-Compactness), it follows from union bound that $\pi_n((S \times S) \setminus \Gamma_{\varepsilon/2})$ is at least,

$$\pi_n\big(\big\{(x,y) : x \notin K_{\varepsilon/2}, y \in S\big\}\big) + \pi_n\left(\big\{(x,y) \in K_{\varepsilon/2} \times S :\ f(y) - \lambda^* c(x,y) < \phi_{\lambda^*}(x) - \gamma\big\}\right) \leq \frac{\varepsilon}{2} + \frac{1}{n\gamma},$$

because $\pi_n(\{(x,y) : x \notin K_{\varepsilon/2}, y \in S\}) = \mu(S \setminus K_{\varepsilon/2}) \leq \varepsilon/2$. Consequently, $\pi_n(\Gamma_{\varepsilon/2}) \geq 1 - \varepsilon/2 - 1/(n\gamma)$. As the closure of $\Gamma_{\varepsilon/2}$ is compact for the choice of γ in (P-Compactness), we have $\mathrm{cl}(\Gamma_{\varepsilon/2})$ is a compact subset of $S \times S$ for given $\varepsilon > 0$. Therefore, for all $n \geq 2\varepsilon^{-1}\gamma^{-1}$, we obtain that $\pi_n(\mathrm{cl}(\Gamma_\varepsilon)) \geq 1 - \varepsilon$ for any given $\varepsilon > 0$, thus rendering that the collection $\{\pi_n : n \geq 1\}$ is tight.

Then, as a consequence of Prokhorov's theorem, we have a subsequence $(\pi_{n_k} : k \geq 1)$ such that $\pi_{n_k} \Rightarrow \pi^*$ for some $\pi^* \in P(S \times S)$.

Step 2 (Verifying $\pi^* \in \Phi_{\mu,\delta}$): As $\pi_{n_k}(\,\cdot \times S) = \mu(\cdot)$, we obtain $\pi^*(\cdot \times S) = \mu(\cdot)$ as a consequence of the weak convergence $\pi_{n_k} \Rightarrow \pi^*$ (see that $\int g(x) d\pi^*(x,y) = \lim_k \int g(x) d\pi_n(x,y) = \int g(x) d\mu(x)$ for all $g \in C_b(S)$). Further, as c is a nonnegative lower semicontinuous function, it follows from a version of Fatou's lemma for weakly converging probabilities (see Theorem 1.1 in Feinberg et al. (2014) and references therein) that $\delta \geq \underline{\lim}_k \int c\, d\pi_{n_k} \geq \int c\, d\pi^*$. Therefore, $\pi^* \in \Phi_{\mu,\delta}$.

Step 3 (Verifying optimality of π^*): As Property (P-USC) guarantees that $\overline{\lim}_k I(\pi_{n_k}) \leq I(\pi^*)$, we obtain $I = \overline{\lim}_k I(\pi_{n_k}) \leq I(\pi^*)$, thus yielding $I(\pi^*) = I$. The fact that $I = J = J(\lambda^*, \phi_{\lambda^*})$ follows from Theorem 1. ☐

The following assumptions imposing certain growth conditions on the functions $c(\cdot,\cdot)$ and $f(\cdot)$ provide a simple set of sufficient conditions required to verify properties (P-Compactness) and (P-USC) stated in Proposition 9.

Assumption 3 (A3)**.** *There exist a nondecreasing function $g : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $g(t) \uparrow \infty$ as $t \to \infty$, and a positive constant C such that $c(x,y) \geq g(\|x - y\|)$ whenever $\|x - y\| > C$.*

Assumption 4 (A4)**.** *There exist an increasing function $h : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $h(t) \uparrow \infty$ as $t \to \infty$, and a positive constant K such that $\sup_{x,y \in S} \frac{f(y) - f(x)}{1 + h(\|x-y\|)} \leq K$. In addition, given $\varepsilon > 0$, there exists a positive constant C_ε such that $f(y) - f(x) \leq \varepsilon(1 + c(x,y))$, for every $x, y \in S$ such that $\|x - y\| > C_\varepsilon$.*

The growth condition imposed in Assumption (A4) is similar to the requirement that the growth rate parameter κ, defined in Gao and Kleywegt (2016), be equal to 0.

Corollary 1. *Let S be locally compact when equipped with the topology induced by a norm $\|\cdot\|$ defined on S. Suppose that the functions c and f satisfy Assumptions (A1) - (A4). Then, whenever the dual optimal pair $(\lambda^*, \phi_{\lambda^*})$ satisfying $J(\lambda^*, \phi_{\lambda^*}) = J < \infty$ is such that $\lambda^* > 0$, there exists a primal optimizer $\pi^* \in \Phi_{\mu,\delta}$ satisfying $I(\pi^*) = I = J = J(\lambda^*, \phi_{\lambda^*})$.*

Proof. We verify the properties (P-COMPACTNESS) and (P-USC) in the statement of Proposition 9 in order to establish the existence of a primal optimizer.

Step 1 (*Verification of* (P-COMPACTNESS)): As any Borel probability measure on a Polish space is tight, there exists a compact $K_\varepsilon \subseteq S$ such that $\mu(K_\varepsilon) \geq 1-\varepsilon$, for any $\varepsilon > 0$. Let $a_\varepsilon := \sup_{x\in K_\varepsilon} \|x\|$. Given $\varepsilon' > 0$, it follows from Assumption (A4) that there exists $C_{\varepsilon'}$ large enough satisfying $f(y) - f(x) \leq \varepsilon' c(x,y)$ for all x, y such that $\|y-x\| > C_{\varepsilon'}$. Then, for any $x \in K_\varepsilon$ and $\gamma > 0$, $f(y) - \lambda^* c(x,y) < f(x) - \gamma$ when $\|y - x\| > C \vee C_{\varepsilon'}$ is large enough such that $g(\|x-y\|) \geq \gamma/(\lambda^*-\varepsilon')$; this is because, $f(y)-f(x)-\lambda^*c(x,y) < (\varepsilon'-\lambda^*)c(x,y) \leq (\varepsilon'-\lambda^*)g(\|x-y\|)$ when $\varepsilon' < \lambda^*$ and $\|x-y\| > C \vee C_{\varepsilon'}$. In particular, if $\|y\| > c_\varepsilon := C \vee C_{\varepsilon'} + a_\varepsilon + g^{-1}(\gamma/(\lambda^* - \varepsilon'))$, we have $f(y) - \lambda^* c(x,y) < f(x) - \gamma$ for every $x \in K_\varepsilon$. As $\phi_{\lambda^*}(x) \geq f(x)$, it follows that,

$$\Gamma_\varepsilon := \left\{(x,y) \in K_\varepsilon \times S : f(y) - \lambda^* c(x,y) \geq \phi_{\lambda^*}(x) - \gamma\right\} \subseteq \{(x,y) : x \in K_\varepsilon, \|y\| \leq c_\varepsilon\},$$

is a compact subset of $S \times S$. Therefore, $\mathrm{cl}(\Gamma_\varepsilon)$ is compact as well, thus verifying Assumption (A3) in the statement of Proposition 9.

Step 2 (*Verification of* P-USC). Let $\{\pi_n : n \geq 1\} \subseteq \Phi'_{\mu,\delta}$ be such that $\pi_n \Rightarrow \pi^*$ for some $\pi^* \in \Phi_{\mu,\delta}$. Our objective is to show that

$$\overline{\lim_n}\, I(\pi_n) = \overline{\lim_n} \int f(y) d\pi_n(x,y) \leq \int f(y) d\pi^*(x,y) = I(\pi^*) \tag{40}$$

While (40) follows directly from the properties of weak convergence when f is bounded upper semicontinuous, an additional asymptotic uniform integrability condition that,

$$\overline{\lim_{M\to\infty}} \sup_n \int_{|f(y)|>M} |f(y)| d\pi_n(x,y) = 0,$$

is sufficient to guarantee (40) in the absence of boundedness (see, for example, Corollary 3 in Zapała (2008)). In order to demonstrate this asymptotic uniform integrability property, we proceed as follows: Given $\varepsilon > 0$, Assumption (A4) guarantees that $f(y) - f(x) \leq \varepsilon c(x,y)$ for every x, y satisfying $\|x - y\| > C_\varepsilon$. Let A_1, A_2 and $B_{(M)}$ be subsets of $S \times S$ defined as follows:

$$A_1 := \{(x,y) : \|x-y\| > C_\varepsilon\},\ A_2 := (S \times S) \setminus A_1, \text{ and } B_{(M)} := \{(x,y) : |f(y)| > M\} \text{ for } M > 0.$$

As $\pi_n((x,y) : f(x) \leq f(y)) = 1$ for every n (recall that $\pi_n \in \tilde{\Phi}_{\mu,\delta}$), it follows from the growth conditions in Assumption (A4) that

$$\int |f(y) - f(x)| d\pi_n(x,y) \leq \varepsilon \int_{A_1} (1 + c(x,y)) d\pi_n(x,y) + K \int_{A_2} (1 + h(\|x-y\|) d\pi_n(x,y).$$

Further, as any $(x,y) \in A_2$ satisfies $\|x-y\| \leq C_\varepsilon$, it follows from the nondecreasing nature of $h(\cdot)$ that $h(\|x-y\|) \leq h(C_\varepsilon) < \infty$, for every for every $(x,y) \in A_2$. Combining this observation with the fact that $\int c d\pi_n \leq \delta$, we obtain,

$$\int_{B_{(M)}} |f(y) - f(x)| d\pi_n(x,y) \leq \varepsilon(1+\delta) + K(1 + h(C_\varepsilon))\pi_n(A_3 \cap B_{(M)}).$$

As $|f| = f + 2f^-$, it follows from Markov's inequality that, $\pi_n(B_{(M)}) = \pi_n(y : |f(y)| > M)$ is at most,

$$\frac{1}{M}\int |f(y)| d\pi_n(x,y) = \frac{1}{M}\left(\int f(y) d\pi_n(x,y) + 2\int f^-(y) d\pi_n(x,y)\right).$$

Since $\pi_n((x,y) : f(x) \leq f(y)) = 1$, we have $f^-(x) \geq f^-(y)$, π_n-almost surely, for every n. Consequently, $\int f^-(y) d\pi_n(x,y) \leq \int f^-(x) d\pi_n(x,y) = \int f^-(x) d\mu(x)$, and therefore,

$$\sup_n \pi_n\left(B_{(M)}\right) \leq \frac{1}{M}\left(I + 2\int f^- d\mu\right),$$

where $I = J < \infty$, and $\int f^- d\mu < \infty$. As a result, we obtain

$$\sup_n \int_{B_{(M)}} |f(y) - f(x)| d\pi_n(x,y) \leq \varepsilon(1+\delta) + \frac{K}{M}(1 + h(C_\varepsilon))\left(I + 2\int f^- d\mu\right). \tag{41}$$

Further, as $\int |f(x)| d\pi_n(x,y) = \int |f(x)| d\mu(x)$ is finite and $\sup_n \pi_n(B_{(M)}) \leq M^{-1}(I + 2\int f^- d\mu) \to 0$ when $M \to \infty$, we have $\int_{B_{(M)}} |f(x)| d\mu(x) \to 0$ as $M \to \infty$. Consequently, letting $M \to \infty$ and $\varepsilon \to 0$ in (41), we obtain

$$\varlimsup_{M\to\infty} \sup_n \int_{\{|f(y)|>M\}} |f(y)| d\pi_n(x,y) \leq \varlimsup_{M\to\infty} \sup_n \int_{B_{(M)}} |f(y) - f(x)| d\pi_n(x,y) + \varlimsup_{M\to\infty} \int_{B_{(M)}} |f(x)| d\mu(x) = 0,$$

thus verifying the desired uniform integrability property. This observation, in conjunction with the upper semicontinuity of f and weak convergence $\pi_n \Rightarrow \pi^*$, results in (40). As all the requirements stated in Proposition 9 are verified, a primal optimizer satisfying $I(\pi^*) = I$ exists. □

Remark 8. In addition to the assumptions made in Corollary 1, suppose that $c(x,y)$ is a convex function in y for every $x \in S$, and f is a concave function. Then, due to Corollary 1, a primal optimizer $\pi^* \in \Phi_{\mu,\delta}$ exists. Further, as the supremum in $\sup_{y\in S}\{f(y) - \lambda^* c(x,y)\}$, is attained at a unique maximizer for every $x \in S$, it follows from Remark 3 that the primal optimizer π^* is unique.

6. A few more examples.

6.1. **Applications to computing general first passage probabilities.** Computing probabilities of first passage of a stochastic process into a target set of interest has been one of the central problems in applied probability. The objective of this section is to demonstrate that, similar to the one dimensional level crossing example in Section 3, one can compute general worst-case probabilities of first passage into a target set B by simply computing the probability of first passage of the baseline stochastic process into a suitably inflated neighborhood of set B. The goal of such a demonstration is to show that the worst-case first passage probabilities can be computed with no significant extra effort.

Example 3. Let $R(t) = (R_1(t), R_2(t)) \in \mathbb{R}^2$ model the financial reserve, at time t, of an insurance firm with two lines of business. Ruin occurs if the reserve process R_t hits a certain ruin set B within time $T > 0$. In the univariate case, the ruin set is usually an interval of the form $(-\infty, 0)$ or its translations (as in Example 1). However, in multivariate cases, the ruin set B can take a variety of shapes based on rules of capital transfers between the different lines of businesses. For example, if capital can be transferred between the two lines without any restrictions, a natural choice is to declare ruin when the total reserve $R_1(t) + R_2(t)$ becomes negative. On the other hand, if no capital transfer is allowed between the two lines, ruin is declared immediately when either $R_1(t)$ or $R_2(t)$ becomes negative. In this example, let us consider the case where the regulatory requirements allow only $\beta \in [0,1]$ fraction of reserve, if positive, to be transferred from one line of business to the other. Such a restriction will result in a ruin set of the form

$$B := \left\{(x_1, x_2) \in \mathbb{R}^2 : \beta x_1 + x_2 \leq 0\right\} \cup \left\{(x_1, x_2) \in \mathbb{R}^2 : \beta x_2 + x_1 \leq 0\right\}.$$

It is immediately clear that capital transfer is completely unrestricted when $\beta = 1$ and altogether prohibited when $\beta = 0$. The intermediate values of $\beta \in (0,1)$ softly interpolates between these two extreme cases. Of course, one can take the fraction of capital transfer allowed from line 1 to line 2 to be different from that allowed from line 2 to line 1, and various other modifications. However, to keep the discussion simple we focus on the model described above, and refer the readers to Hult and Lindskog (2006) for a general specification of ruin models allowing different rules of capital transfers between d lines of businesses.

As in Example 1, we take the space in which the reserve process $(R(t) : 0 \leq t \leq T)$ takes values to be $S = D([0,T], \mathbb{R}^2)$, the set of all $\mathbb{R}^2$-valued right continuous functions with left limits defined on the interval $[0,T]$. The space S, again, as in Example 1, is equipped with the standard J_1-metric (see Chapter 3 of Whitt (2002)), and consequently, the cost function

$$c(x,y) = \inf_{\lambda\in\Lambda} \left(\sup_{t\in[0,T]} ||x(t) - y(\lambda(t))||_2^2 + \sup_{t\in[0,T]} |\lambda(t) - t| \right), \quad x, y \in S \tag{42}$$

is continuous. Here, $||x||_2 = (|x_1|^2 + |x_2|^2)^{1/2}$ is the standard Euclidean norm for $x = (x_1, x_2)$ in $\mathbb{R}^2$, and Λ is the set of all strictly increasing functions $\lambda : [0,T] \to [0,T]$ such that both λ and λ^{-1} are continuous. Let $\mu \in \mathcal{P}(S)$ denote a baseline probability measure that models the reserve process $R(t)$ in the path space. Given $\delta > 0$, our objective is to characterize the worst-case first passage probability, $\sup\{P(x(t) \in B \text{ for some } t \in [0,T]) : d_c(\mu, P) \leq \delta\}$, of the reserve process hitting the ruin set B. As the set $\{x \in S : x(t) \in B \text{ for some } t \in [0,T]\}$ is not closed, we consider its topological closure

$$A := \left\{ x \in S : \inf_{t\in[0,T]} (\beta x_1(t) + x_2(t)) \leq 0 \right\} \cup \left\{ x \in S : \inf_{t\in[0,T]} (x_1(t) + \beta x_2(t)) \leq 0 \right\},$$

which, apart from the paths that hit ruin set B, also contains the paths that come arbitrarily close to B without hitting B due to the presence of jumps; the fact that A is the desired closure is verified in Lemma 13 and Corollary 2 in Appendix B. Further, it is verified in Lemma 14 in Appendix B that

$$c(x, A) = \frac{1}{1+\beta^2} \left[\inf_{t\in[0,T]} \left(\beta x_1(t) + x_2(t)\right)^2 \wedge \inf_{t\in[0,T]} \left(\beta x_2(t) + x_1(t)\right)^2 \right],$$

If the reference distribution $\mu(\cdot)$ is such that $h(u) := E_\mu[c(X,A); c(X,A) \leq u]$ is continuous, then

$$\begin{aligned} \sup\{P(x(t) \in B \text{ for some } t \in [0,T]) : d_c(\mu, P) \leq \delta\} &\leq \sup\left\{P(A) : d_c(\mu, P) \leq \delta\right\} \\ &= \mu\left\{ x \in S : c(x,A) \leq \frac{1}{\lambda^*} \right\}, \end{aligned} \tag{43}$$

as an application of Theorem 3; here, as in Section 2.4, $1/\lambda^* = h^{-1}(\delta) := \inf\{u \geq 0 : h(u) \geq \delta\}$. Following the expression for $c(x,A)$ in Lemma 14, we also have that $\{x \in S : c(x,A) \leq 1/\lambda^*\}$ equals

$$\left\{ x \in S : \inf_{t\in[0,T]} (\beta x_1(t) + x_2(t)) \leq \sqrt{\frac{1+\beta^2}{\lambda^*}} \right\} \cup \left\{ x \in S : \inf_{t\in[0,T]} (x_1(t) + \beta x_2(t)) \leq \sqrt{\frac{1+\beta^2}{\lambda^*}} \right\}. \tag{44}$$

Next, if we let $c^* = \sqrt{(1+\beta^2)/\lambda^*}$ and

$$A_{(c)} = \left\{ x \in S : \inf_{t\in[0,T]} (\beta x_1(t) + x_2(t)) \leq c \right\} \cup \left\{ x \in S : \inf_{t\in[0,T]} (x_1(t) + \beta x_2(t)) \leq c \right\} \tag{45}$$

as a family of sets parameterized by $c \in \mathbb{R}$, then it follows from (44) that

$$\sup\{P\left(A_{(0)}\right) : d_c(\mu, P) \leq \delta\} = \mu\left(A_{(c^*)}\right); \tag{46}$$

in words, the worst-case probability that the reserve process hits (or) comes arbitrarily close to the ruin set B is simply equal to the probability under reference measure μ that the reserve process hits (or) comes arbitrarily close to an inflated ruin set $B_{(c^*)}$, where for any $c \in \mathbb{R}$ we define

$$B_{(c)} := \left\{(x_1, x_2) \in \mathbb{R}^2 : \beta x_1 + x_2 \leq c\right\} \cup \left\{(x_1, x_2) \in \mathbb{R}^2 : \beta x_2 + x_1 \leq c\right\}.$$

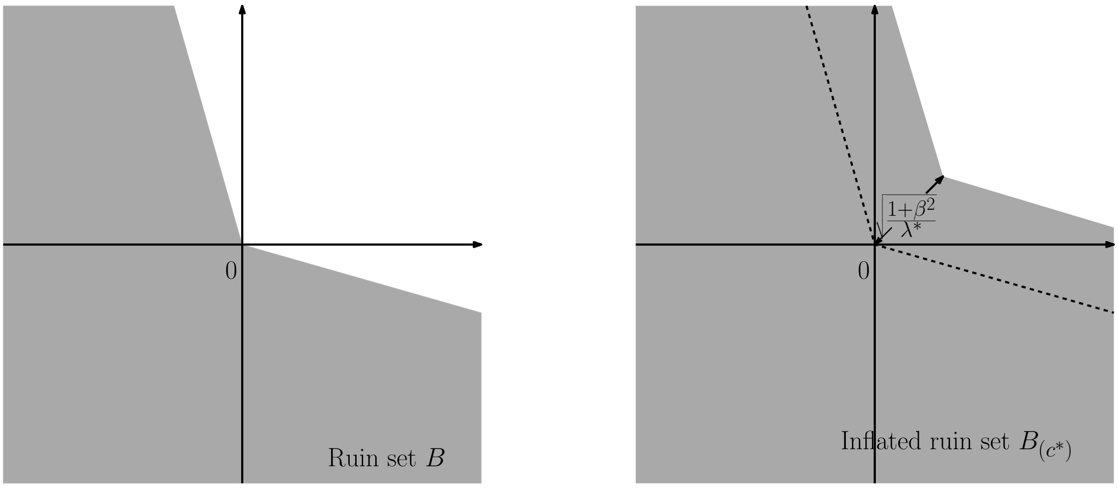
This conclusion is very similar to (25) derived for 1-dimensional level crossing in Section 3. The original ruin set $B = B_{(0)}$ and the suitably inflated ruin set $B_{(c^*)}$ are respectively shown in the Figures 2(a) and (b).

Next, as an example, let us take the reserve process satisfying the following dynamics as our baseline model:

$$dR(t) = ub + mdt + \Sigma dB(t),$$

where $m \in \mathbb{R}^2_-$ is the drift vector with negative components, Σ is a 2×2 positive definite covariance matrix, and $(B(t) : 0 \leq t \leq T)$ a 2-dimensional standard Brownian motion. The real number u denotes the total initial capital and $b = (b_1, b_2) \in \mathbb{R}^2_+$ such that $b_1 + b_2 = 1$ denotes the proportion of initial capital set aside for the two different lines of businesses. The probability measure induced in the path space by the process $R(t)$ is taken as the baseline measure μ. Further, for purposes

FIGURE 2. Comparison of computation of ruin under baseline measure (in Fig(a)) and worst-case ruin (in Fig(b))

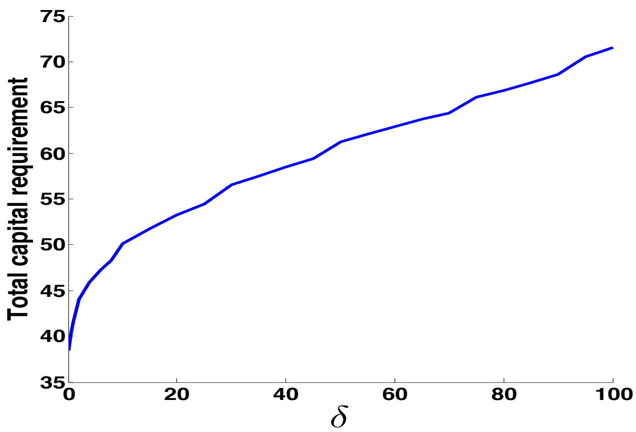


(a) Computation of ruin under baseline measure

(b) Computation of worst-case ruin using the baseline measure

of numerical illustration, we take $m = [-0.1, -0.1], b = [0.5, 0.5]$ and $\Sigma = I_2$, the 2×2 identity matrix. Our aim is to find total initial capital u such that $\sup\{P(A_{(0)}) : d_c(\mu, P) \leq \delta\} \leq 0.01$ (recall the definition of the ruin event $A_{(0)}$ in 45). This is indeed possible due to the equivalent characterization in (46), and the resulting capital requirement for various values of δ are displayed in Figure 3.

FIGURE 3. Capital requirement for various values of δ. The capital requirement is calculated to keep the worst-case probability of ruin under 0.01



It may also be useful to note that one may as well choose $S = C([0,T], \mathbb{R}^2)$, the space of continuous functions taking values in $\mathbb{R}^2$, as the underlying space to work with if the modeler decides to restrict the distributional ambiguities to continuous stochastic process; in that case, the first passage set $\{x \in S : x(t) \in B \text{ for some } t \in [0,T]\}$ itself is closed, and the inequality in (43) holds with equality.

As a final remark, if the modeler believes that model ambiguity is more prevalent in one line of business over other, she can perhaps quantify that effect by instead choosing $||x||_2 = (|x_1|^2 + \alpha|x_2|^2)^{1/2}$, for $x = (x_1, x_2) \in \mathbb{R}^2$ in the cost function $c(x, y)$ defined in (42). This would penalise moving probability mass in one direction more than the other and result in the ruin set being inflated asymmetrically along different directions. We identify studying the effects of various choices of cost functions, the corresponding δ and the appropriateness of the resulting inflated ruin set for real world ruin problems as an important direction towards applying the proposed framework in quantitative risk management. For example, suppose that the regulator has interacted with the insurance company, and both the company and the regulator have negotiated a certain level of capital requirement multiple times in the past. Then the insurance company can calibrate δ from these previous interactions with the regulator. That is, find the value of δ which implies that the negotiated capital requirement in a given past interaction is necessary for the bound on ruin probability to be lesser than a fixed (regulatory driven) acceptance level (for instance, we use 0.01 as the acceptance level for ruin probability in our earlier numerical illustration in this example). An appropriate quantile of these 'implied' δ values may be used to choose δ based on risk preference.

6.2. **Applications to ambiguity-averse decision making.** In this section, we consider a stochastic optimization problem in the presence of model uncertainty. It has been of immense interest recently to search for distributionally robust optimal decisions, that is, to find a decision variable b that solves

$$\mathrm{OPT} = \inf_{b \in B} \sup_{P \in \mathcal{P}} E\left[f(X, b)\right].$$

Here, f is a performance/risk measure that depends on a random element X and a decision variable b that can be chosen from an action space B. The solution to the above problem minimizes worst-case risk over a family of ambiguous probability measures $\mathcal{P}$. Such an ambiguity-averse optimal choice is also referred as a distributionally robust choice, because the performance of the chosen decision variable is guaranteed to be better than OPT irrespective of the model picked from the family $\mathcal{P}$.

There has been a broad range of ambiguity sets $\mathcal{P}$ that has been considered: For examples, refer Delage and Ye (2010); Goh and Sim (2010); Wiesemann et al. (2014) for moment-based uncertainty sets, Hansen and Sargent (2001); Iyengar (2005); Nilim and El Ghaoui (2005); Lim and Shanthikumar (2007); Jain et al. (2010); Ben-Tal et al. (2013); Wang et al. (2015); Jiang and Guan (2015); Hu and Hong (2012); Bayraksan and Love (2015) for KL-divergence and other likelihood based uncertainty sets, Pflug and Wozabal (2007); Wozabal (2012); Esfahani and Kuhn (2015); Zhao and Guan (2015); Gao and Kleywegt (2016) for Wasserstein distance based neighborhoods, Erdoğan and Iyengar (2006) for neighborhoods based on Prokhorov metric, Bandi et al. (2015); Bandi and Bertsimas (2014) for uncertainty sets based on statistical tests and Ben-Tal et al. (2009); Bertsimas and Sim (2004) for a general overview. As most of the works mentioned above assume the random element X to be $\mathbb{R}^d$-valued, it is of our interest in the following example to demonstrate the usefulness of our framework in formulating and solving distributionally robust optimization problems that involve stochastic processes taking values in general Polish spaces as well.

Example 4. We continue with the insurance toy example considered in Section 3. In practice, as the risk left to the first-line insurer is too large, reinsurance is usually adopted. In proportional reinsurance, one of the popular forms of reinsurance, the insurer pays only for a proportion b of the incoming claims, and a reinsurer pays for the remaining $1 - b$ fraction of all the claims received. In turn, the reinsurer receives a premium at rate $p_r = (1 + \theta)(1 - b)\nu c m_1$ from the insurer. Here, $\theta > \eta$, otherwise, the insurer could make riskless profit by reinsuring the whole portfolio. The problem we consider here is to find the reinsurance proportion b that minimises the expected maximum loss that happens within duration T. In the extensive line of research that studies optimal reinsurance proportion, diffusion models have been particularly recommended for tractability reasons (see, for example Hojgaard and Taksar (1998); Schmidli (2001)). As in Example 1, if we take $\sqrt{\nu m_2} B(t) + \nu m_1 t$ to be the diffusion process that approximates the arrival of claims, then

$$L_b(t) := p_r t + b\big(\sqrt{\nu m_2} B(t) + \nu m_1 t\big) - pt$$
$$= b\sqrt{\nu m_2} B(t) - \big(b\theta - (\theta - \eta)\big)\nu m_1 t$$

is a suitable model for losses made by the firm. Here, $p_r = (1+\theta)(1-b)\nu m_1$ is the rate of payout for the reinsurance contract, and $p = (1+\eta)\nu m_1$ is the rate at which a premium income is received by the insurance firm. The quantity of interest is to determine the reinsurance proportion b that minimises the maximum expected losses,

$$L := \inf_{b\in[0,1]} E\left[\max_{t\in[0,T]} L_b(t)\right]. \tag{47}$$

However, as we saw in Example 1, conclusions based on diffusion approximations can be misleading. Following the practice advocated by the rich literature of robust optimization, we instead find a reinsurance proportion b that performs well against the family of models specified by $\mathcal{P} := \{P : d_c(\mu, P) \leq \delta\}$. In other words, we attempt to solve for

$$L' := \inf_{b\in[0,1]} \sup_{P\in\mathcal{P}} E_P\left[\sup_{t\in[0,T]} \big(a_2(b)X(t) - a_1(b)t\big)\right],$$

where X is a random element in space $D[0,T]$ following measure P, $a_2(b) := b\sqrt{\nu m_2}$ and $a_1(b) := (b\theta - (\theta-\eta))\nu m_1$. Here, as in Section 2.4, we have taken $S = D[0,T]$ and

$$c(x,y) = \Big(d_{J_1}(x,y)\Big)^2.$$

For $x \in D[0,T]$, if we take

$$f(x,b) := \sup_{t\in[0,T]} \big(a_2(b)x(t) - a_1(b)t\big) \quad \text{and} \quad \phi_{\lambda,b}(x) := \sup_{y\in S}\{f(y,b) - \lambda c(x,y)\},$$

then due to the application of Theorem 1, we obtain

$$L' = \inf_{b\in[0,1]} \sup\{E_P[f(X,b)] : d_c(\mu,P) \leq \delta\} = \inf_{b\in[0,1]} \inf_{\lambda\geq 0} \left\{\lambda\delta + \int \phi_{\lambda,b}(x)d\mu(x)\right\},$$

To keep this discussion terse, it is verified in Lemma 12 in Appendix B that the inner infimum evaluates simply to

$$E\left[\sup_{t\in[0,T]} \big(a_2(b)B(t) - a_1(b)t\big)\right] + a_2(b)\sqrt{\delta}.$$

As a result,

$$L' = \inf_{b\in[0,1]} \left\{E\left[\sup_{t\in[0,T]} \big(a_2(b)B(t) - a_1(b)t\big)\right] + a_2(b)\sqrt{\delta}\right\}, \tag{48}$$

which is an optimization problem that involves the same effective computational effort as the non-robust counterpart in (47). For the specific numerical values employed in Example 1, if we additionally take the new parameter $\theta = 0.3$, the Brownian approximation model evaluates to the optimal choice $b = 0.66$ and the corresponding loss $L = 17.63$, whereas the robust counterpart in (48) evaluates to worst-case $L' = 28.86$ for the ambiguity-averse optimal choice $b = 0.42$. For a collection of various other examples of using Wasserstein based ambiguity sets in the context of distributionally robust optimization, refer Esfahani and Kuhn (2015); Zhao and Guan (2015) and Gao and Kleywegt (2016).

Acknowledgments.

The authors would like to thank the anonymous referees whose valuable suggestions have been immensely useful in supplementing the proof of the strong duality theorem with crisp arguments at various instances. The authors would also like to thank Garud Iyengar and David Goldberg for helpful discussions, and Rui Gao for providing a comment that led us to add Remark 6 in the paper. The authors gratefully acknowledge support from Norges Bank Investment Management and NSF grant CMMI 1436700.

References

C. Aliprantis and K. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Studies in Economic Theory. Springer, 1999. ISBN 9783540658542. URL https://books.google.com/books?id=6jjY2Vi3aDEC.

C. D. Aliprantis and O. Burkinshaw. *Principles of Real Analysis*. Gulf Professional Publishing, 1998.

L. Ambrosio and L. A. Caffarelli. *Optimal transportation and applications: Lectures given at the CIME summer school 2001*. Springer Science & Business Media, 2003.

S. Asmussen and H. Albrecher. *Ruin probabilities*. World Scientific Publishing Co. Pvt. Ltd., 2010. ISBN 978-981-4282-52-9; 981-4282-52-9. doi: 10.1142/9789814282536. URL http://dx.doi.org/10.1142/9789814282536.

R. Atar, K. Chowdhary, and P. Dupuis. Robust bounds on risk-sensitive functionals via Rényi divergence. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):18–33, 2015. ISSN 2166-2525. doi: 10.1137/130939730. URL http://dx.doi.org/10.1137/130939730.

C. Bandi and D. Bertsimas. Robust option pricing. *European J. Oper. Res.*, 239(3):842–853, 2014. ISSN 0377-2217. doi: 10.1016/j.ejor.2014.06.002. URL http://dx.doi.org/10.1016/j.ejor.2014.06.002.

C. Bandi, D. Bertsimas, and N. Youssef. Robust queueing theory. *Operations Research*, 63(3):676–700, 2015. doi: 10.1287/opre.2015.1367. URL http://dx.doi.org/10.1287/opre.2015.1367.

A. D. Barbour and A. Xia. On Stein's factors for poisson approximation in wasserstein distance. *Bernoulli*, 12(6):943–954, 2006. ISSN 13507265. URL http://www.jstor.org/stable/25464848.

G. Bayraksan and D. K. Love. Data-driven stochastic programming using phi-divergences. *The Operations Research Revolution*, pages 1–19, 2015.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton Series in Applied Mathematics. Princeton University Press, 2009. ISBN 978-0-691-14368-2. doi: 10.1515/9781400831050. URL http://dx.doi.org/10.1515/9781400831050.

A. Ben-Tal, D. den Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013. doi: 10.1287/mnsc.1120.1641. URL http://dx.doi.org/10.1287/mnsc.1120.1641.

D. P. Bertsekas and S. E. Shreve. *Stochastic optimal control: The discrete time case*. Academic Press New York, 1978.

D. Bertsimas and M. Sim. The price of robustness. *Oper. Res.*, 52(1):35–53, 2004. ISSN 0030-364X. doi: 10.1287/opre.1030.0065. URL http://dx.doi.org/10.1287/opre.1030.0065.

P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 1999. ISBN 0-471-19745-9. doi: 10.1002/9780470316962. URL http://dx.doi.org/10.1002/9780470316962.

J. Blanchet, Y. Kang, and K. Murthy. Robust wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627*, 2016.

T. Breuer and I. Csiszár. Measuring distribution model risk. *Mathematical Finance*, 2013. ISSN 1467-9965. doi: 10.1111/mafi.12050. URL http://dx.doi.org/10.1111/mafi.12050.

G. Canas and L. Rosasco. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, pages 2492–2500, 2012.

E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.*, 58(3):595–612, 2010. ISSN 0030-364X. doi: 10.1287/opre.1090.0741. URL http://dx.doi.org/10.1287/opre.1090.0741.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events*. Springer-Verlag, New York, 1997. ISBN 3-540-60931-8. doi: 10.1007/978-3-642-33483-2. URL http://dx.doi.org/10.1007/978-3-642-33483-2.

P. Embrechts, S. Resnick, and G. Samorodnitsky. Living on the edge. *Risk*, 11(1):96–100, 1998.

P. Embrechts, G. Puccetti, L. Rschendorf, R. Wang, and A. Beleraj. An academic response to basel 3.5. *Risks*, 2(1):25–48, 2014. ISSN 2227-9091. doi: 10.3390/risks2010025. URL http://www.mdpi.com/2227-9091/2/1/25.

E. Erdoğan and G. Iyengar. Ambiguous chance constrained problems and robust optimization. *Math. Program.*, 107(1-2, Ser. B):37–61, 2006. ISSN 0025-5610. doi: 10.1007/s10107-005-0678-0. URL http://dx.doi.org/10.1007/s10107-005-0678-0.

P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*, 2015.

E. A. Feinberg, P. O. Kasyanov, and N. V. Zadoianchuk. Fatou's lemma for weakly converging probabilities. *Theory of Probability & Its Applications*, 58(4):683–689, 2014.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2014. ISSN 1432-2064. doi: 10.1007/s00440-014-0583-7. URL http://dx.doi.org/10.1007/s00440-014-0583-7.

C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems (NIPS) 28*, 2015.

R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199v1*, 2016.

P. Glasserman and X. Xu. Robust risk measurement and model risk. *Quantitative Finance*, 14(1):29–58, 2014. doi: 10.1080/14697688.2013.822989. URL http://dx.doi.org/10.1080/14697688.2013.822989.

J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Oper. Res.*, 58(4, part 1):902–917, 2010. ISSN 0030-364X. doi: 10.1287/opre.1090.0795. URL http://dx.doi.org/10.1287/opre.1090.0795.

N. Gozlan and C. Leonard. Transport inequalities. A survey. *Markov Processes and Related Fields*, 16(4):707–738, 2016. ISSN 635-736.

L. P. Hansen and T. J. Sargent. Robust control and model uncertainty. *The American Economic Review*, 91(2):pp. 60–66, 2001. ISSN 00028282. URL http://www.jstor.org/stable/2677734.

J. Harrison. Ruin problems with compounding assets. *Stochastic Processes and their Applications*, 5(1):67 – 79, 1977. ISSN 0304-4149. doi: http://dx.doi.org/10.1016/0304-4149(77)90051-5. URL http://www.sciencedirect.com/science/article/pii/0304414977900515.

B. Hojgaard and M. Taksar. Optimal proportional reinsurance policies for diffusion models. *Scandinavian Actuarial Journal*, 1998(2):166–180, 1998. doi: 10.1080/03461238.1998.10414000. URL http://dx.doi.org/10.1080/03461238.1998.10414000.

Z. Hu and L. J. Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available on optimization online*, 2012.

H. Hult and F. Lindskog. Heavy-tailed insurance portfolios: buffer capital and ruin probabilities. Technical report, Cornell University, 2006.

D. L. Iglehart. Diffusion approximations in collective risk theory. *Journal of Applied Probability*, 6(2):285–292, 1969. ISSN 00219002. URL http://www.jstor.org/stable/3211999.

G. N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

A. Jain, A. E. B. Lim, and J. G. Shanthikumar. On the optimality of threshold control in queues with model uncertainty. *Queueing Systems*, 65(2):157–174, 2010. ISSN 1572-9443. doi: 10.1007/s11134-010-9172-3. URL http://dx.doi.org/10.1007/s11134-010-9172-3.

R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, pages 1–37, 2015. ISSN 1436-4646. doi: 10.1007/s10107-015-0929-7. URL http://dx.doi.org/10.1007/s10107-015-0929-7.

D. Khoshnevisan. An embedding of compensated compound poisson processes with applications to local times. *Ann. Probab.*, 21(1):340–361, 01 1993. doi: 10.1214/aop/1176989408. URL http://dx.doi.org/10.1214/aop/1176989408.

H. Lam. Robust sensitivity analysis for stochastic systems. *arXiv preprint arXiv:1303.0326*, 2013.

M. Lewis. In nature's casino. *The New York Times Magazine*, page 626, Aug 2007.

A. E. Lim and J. G. Shanthikumar. Relative entropy, exponential utility, and robust dynamic pricing. *Operations Research*, 55(2):198–214, 2007.

D. G. Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.*, 41(1):370–400, 02 2013. doi: 10.1214/12-AOS1065. URL http://dx.doi.org/10.1214/12-AOS1065.

A. Nilim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Oper. Res.*, 53(5):780–798, 2005.

G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quant. Finance*, 7(4):435–442, 2007. ISSN 1469-7688. doi: 10.1080/14697680701455410. URL http://dx.doi.org/10.1080/14697680701455410.

S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems. Volume I: Theory*. Springer Science & Business Media, 1998a.

S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems. Volume II: Applications*. Springer Science & Business Media, 1998b.

T. Rolski, H. Schmidli, V. Schmidt, and J. Teugels. *Stochastic processes for insurance and finance*. John Wiley & Sons, Ltd., 1999. ISBN 0-471-95925-1. doi: 10.1002/9780470317044. URL http://dx.doi.org/10.1002/9780470317044.

W. Rudin. *Real and complex analysis (3rd)*. McGraw-Hill Inc, 1986.

H. Schmidli. Optimal proportional reinsurance policies in a dynamic setting. *Scandinavian Actuarial Journal*, 2001(1):55–68, 2001. doi: 10.1080/034612301750077338. URL http://dx.doi.org/10.1080/034612301750077338.

M. Sion. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958. URL http://projecteuclid.org/euclid.pjm/1103040253.

J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *Proceedings of The 31st International Conference on Machine Learning*, pages 306–314, 2014.

C. Villani. *Topics in optimal transportation*. American Mathematical Soc., 2003.

C. Villani. *Optimal transport: old and new*. Springer Science & Business Media, 2008.

F. Wang and L. J. Guibas. Supervised earth movers distance learning and its computer vision applications. In *Computer Vision–ECCV 2012*, pages 442–455. Springer, 2012.

Z. Wang, P. W. Glynn, and Y. Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, pages 1–21, 2015. doi: 10.1007/s10287-015-0240-3. URL http://dx.doi.org/10.1007/s10287-015-0240-3.

W. Whitt. *Stochastic process limits*. Springer-Verlag, New York, 2002. ISBN 0-387-95358-2.

W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Oper. Res.*, 62(6):1358–1376, 2014. ISSN 0030-364X. doi: 10.1287/opre.2014.1314. URL http://dx.doi.org/10.1287/opre.2014.1314.

D. Wozabal. A framework for optimization under ambiguity. *Ann. Oper. Res.*, 193:21–47, 2012. ISSN 0254-5330. doi: 10.1007/s10479-010-0812-0. URL http://dx.doi.org/10.1007/s10479-010-0812-0.

A. M. Zapała. Unbounded mappings and weak convergence of measures. *Statistics & Probability Letters*, 78(6):698 – 706, 2008. ISSN 0167-7152. doi: http://doi.org/10.1016/j.spl.2007.09.033. URL http://www.sciencedirect.com/science/article/pii/S0167715207003203.

C. Zhao and Y. Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Available on Optimization Online*, 2015.

Appendix A. Brownian embeddings for the estimation of δ in Example 1

Recall that the definition of optimal transport cost between two probability measures μ and ν, denoted by $d_c(\mu,\nu)$, involves computing the minimum expected cost over all possible couplings between μ and ν (see Section 2.1 for a definition). Though it may not always be possible to identify the optimal coupling (with the lowest expected cost) between μ and ν, one can perhaps employ a 'good' coupling to derive an upper bound on $d_c(\mu,\nu)$. In this section, we describe one such coupling, popularly referred as Skorokhod embedding, between the risk process $R(t)$ and its Brownian motion based diffusion approximation $R_B(t)$ in Example 1. More specifically, we 'embed' the compensated compound Poisson process

$$Z(t) = \frac{1}{\sqrt{m_2}}\left(\sum_{i=1}^{N_t} X_i - m_1 t\right)$$

in a Brownian motion $B(t)$ to obtain a coupling between the risk processes $R(t)$ and $R_B(t)$ in order to choose a δ in Example 1. Here, the symbols m_1 and m_2 denote the first and second moments of claim sizes X_i, and N_t is a unit rate Poisson process. Please refer Example 1 in Section 3 for a thorough review of notations. The procedure is data-driven in the sense that we do not assume the knowledge of the distribution of claim sizes X_i. Instead, we simply assume access to an oracle that provides independent realizations of claim sizes $\{X_1, X_2, \ldots\}$. Given this access to claim size information, Algorithm 1 below specifies the coupling that embeds the process $Z(t)$ in Brownian motion $B(t)$.

Algorithm 1 To embed the process $(Z(t) : t \geq 0)$ in Brownian motion $(B(t) : t \geq 0)$
Given: Brownian motion $B(t)$, moment m_1 and independent realizations of claim sizes $X_1, X_2, \ldots$

Initialize $\tau_0 := 0$ and $\Psi_0 := 0$. For $j \geq 1$, recursively define,

$$\tau_{j+1} := \inf\left\{ s \geq \tau_j : \sup_{\tau_j \leq r \leq s} B_r - B_s = X_{j+1} \right\}, \text{ and } \Psi_j := \Psi_{j-1} + X_j.$$

Define the auxiliary processes

$$\tilde{S}(t) := \sum_{j>0} \sup_{\tau_j \leq s \leq t} B(s) \mathbf{1}\left(\tau_j \leq t < \tau_{j+1}\right) \text{ and } \tilde{N}(t) := \sum_{j \geq 0} \Psi_j \mathbf{1}(\tau_j \leq t < \tau_{j+1}).$$

Let $A(t) := \tilde{N}(t) + \tilde{S}(t)$, and identify the time change $\sigma(t) := \inf\{s : A(s) = m_1 t\}$. Next, take the time changed version $Z(t) := \tilde{S}(\sigma(t))$.

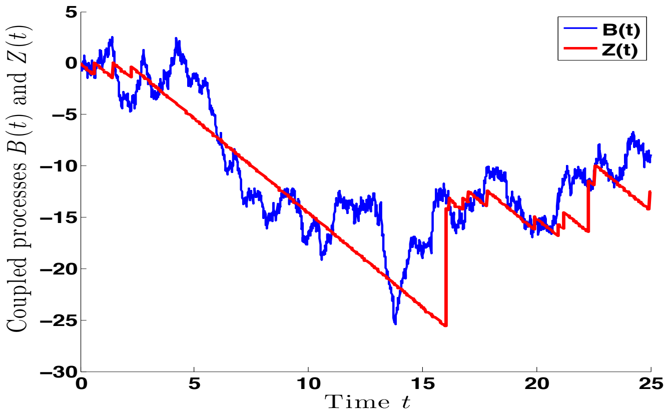
Replace $Z(t)$ by $-Z(t)$ and $B(t)$ by $-B(t)$.

Algorithm 1 is a brief description of the coupling developed in Khoshnevisan (1993), where it is also proved that the process $Z(t)$ output by the construction is indeed the desired compound Poisson process 'closely' coupled with the Brownian motion $B(t)$. Figure 4 below shows a typical coupled path output by Algorithm 1. Several independent replications of such coupled paths are used to simulate the coupled risk processes $R(t)$ and $R_B(t)$ in Example 1, and δ is chosen as prescribed by the confidence intervals (obtained due to a straightforward application of central limit theorem) for the empirical average cost of the simulated coupling.

Appendix B. Some technical proofs

In this section, we first state and prove all the technical results that are utilized in Examples 1, 3 and 4. Then we prove Lemma 15 and 16, which are technical results used in Section 4 to

FIGURE 4. A coupled path output by Algorithm 1



complete the proof of Theorem 1. We conclude the section with Lemma 17 and Corollary 3 that add more clarity to the notation,

$$\sup\left\{\int f d\nu : d_c(\mu,\nu) \leq \delta\right\} := \sup\left\{\int f d\nu : d_c(\mu,\nu) \leq \delta, \int f^- d\nu < \infty\right\} \tag{49}$$

adopted while defining the primal problem I in Section 2.2.

B.1. **Technical results used in Examples 1 and 4.** For the results (Lemma 10, 11 and 12) stated and proved in this section, let $S = D([0,T],\mathbb{R})$ be equipped with the J_1-topology, and define

$$A_u := \left\{ x \in D([0,T],\mathbb{R}) : \sup_{t\in[0,T]} x(t) \geq u \right\}, \qquad u \in \mathbb{R}.$$

Lemma 10. *For every $u \in \mathbb{R}$, the set A_u is closed.*

Proof. Pick any $x \notin A_u$, and let $\varepsilon := (u - \sup_{t\in[0,T]} x(t))/2$. As $\sup_{t\in[0,T]} x(t) < u$, ε is strictly positive. Then, for every $y \in S$ such that $d_{J_1}(x,y) < \varepsilon$, we have that

$$\sup_{t\in[0,T]} y(t) \leq \sup_{t\in[0,T]} x(t) + d_{J_1}(x,y) \leq \sup_{t\in[0,T]} x(t) + \varepsilon,$$

which, in turn, is smaller than u because of the way ε is chosen. Thus, $\{y \in S : d_{J_1}(x,y) < \varepsilon\}$ is a subset of $S \setminus A_u$, and hence $S\setminus A_u$ is open. This automatically means that the set A_u is closed. $\square$

Lemma 11. *Let $c(x,y) = d_{J_1}(x,y)$, for x,y in $D([0,T],\mathbb{R})$. For any $u \in \mathbb{R}$, $c(x,A_u) := \inf\{c(x,y) : y \in A_u\}$ is given by*

$$c(x,A_u) = u - \sup_{t\in[0,T]} x(t),$$

for every $x \notin A_u$.

Proof. Given $\varepsilon > 0$, let $z(t) := x(t) + \inf_{t\in[0,T]}(u - x(t)) + \varepsilon$, for all $t \in [0,T]$. As $z \in A_u$, it is immediate from the definition of $c(x,A_u)$ that

$$c(x,A_u) \leq d_{J_1}(x,z) \leq \sup_{t\in[0,T]} |z(t) - x(t)| = \inf_{t\in[0,T]}(u - x(t)) + \varepsilon = u - \sup_{t\in[0,T]} x(t) + \varepsilon. \tag{50}$$

Next, let Λ be the set of strictly increasing functions λ mapping the interval $[0,T]$ onto itself, such that both λ and λ^{-1} are continuous. Also, let e be the identity map, that is, $e(t) = t$, for all t in $[0,T]$. Then,

$$d_{J_1}(x,y) := \inf_{\lambda\in\Lambda} \{||x - y \circ \lambda||_\infty + ||\lambda - e||_\infty\},$$

where $||z||_\infty = \sup_{t\in[0,T]} |z(t)|$, for any z in S. Since $c(x,y) = d_{J_1}(x,y)$, we have

$$c(x, A_u) = \inf_{y\in A_u} \inf_{\lambda\in\Lambda} \big\{ ||x - y\circ\lambda||_\infty + ||\lambda - e||_\infty \big\}.$$

Next, as $y \circ \lambda \in A_u$ for any $y \in A_u$, it is immediate that one can restrict to $\lambda = e$ without loss of generality, and subsequently, the inner infimum is not necessary. In other words, for any $\varepsilon > 0$,

$$\begin{aligned} c(x, A_u) &= \inf_{y\in A_u} ||x-y||_\infty = \inf_{y\in A_u} \sup_{t\in[0,T]} |x(t) - y(t)| \\ &\geq \inf_{y\in A_u} \left(\inf_{t\in[0,T]:y_t > u-\varepsilon} y(t) - \sup_{t\in[0,T]} x(t) \right) \geq u - \sup_{t\in[0,T]} x(t) - \varepsilon. \end{aligned}$$

As ε is arbitrary, this observation, when combined with (50), concludes the proof of Lemma 11. $\square$

Lemma 12. *Let $c(x,y) = d^2_{J_1}(x,y)$ for x, y in $D([0,T],\mathbb{R})$. Given nonnegative constants a_1, a_2 and λ, define the functions,*

$$f(x) = \sup_{t\in[0,T]} \big(a_2 x(t) - a_1 t\big) \text{ and } \phi_\lambda(x) = \sup_{y\in S} \{ f(y) - \lambda c(x,y) \},$$

for every $x \in S$. Then, for any $\delta > 0$,

$$\inf_{\lambda\geq 0} \left\{ \lambda\delta + \int \phi_\lambda(x) d\mu(x) \right\} = \int f(x) d\mu(x) + a_2\sqrt{\delta}.$$

Proof. Fix any $x \in S$ and $\lambda > 0$. For any positive constant b, as $y(t) = x(t) + b$ is also a member of S, it follows from the definition of $\phi_\lambda(x)$ that

$$\phi_\lambda(x) \geq \sup_{\substack{y:y=x+b\\ b\,\geq\, 0}} \left\{ \sup_{t\in[0,T]} \big(a_2 y(t) - a_1 t\big) - \lambda d^2_{J_1}(x,y) \right\} = \sup_{b\,\geq\,0} \left\{ \sup_{t\in[0,T]} \big(a_2(x(t)+b) - a_1 t\big) - \lambda b^2 \right\}.$$

As the function $a_2 b - \lambda b^2$ is maximized at $b = a_2/2\lambda$, the above lower bound simplifies to

$$\phi_\lambda(x) \;\geq\; f(x) + \frac{a_2^2}{4\lambda}. \tag{51}$$

To obtain an upper bound, we first observe from the definition of the metric d_{J_1} that

$$\phi_\lambda(x) = \sup_{y\in S} \inf_{\nu\in\Lambda} \left\{ f(y) - \lambda\big(||x - y\circ\nu||_\infty + ||\nu - e||_\infty\big)^2 \right\},$$

where the quantities Λ and e are defined as in the proof of Lemma 11. Again, as $y \circ \nu$ lies in S, for every time change $\nu \in \Lambda$, one can take $\nu = e$ without loss of generality. Consequently,

$$\phi_\lambda(x) = \sup_{y\in S} \{ f(y) - \lambda||x-y||^2_\infty \} = \sup_{z\in S} \{ f(x+z) - \lambda||z||^2_\infty \},$$

where we have also changed the variable from $y - x$ to z. In particular,

$$\begin{aligned} \phi_\lambda(x) &= \sup_{z\in S} \left\{ \sup_{t\in[0,T]} \big(a_2(x(t)+z(t)) - a_1 t\big) - \lambda\|z\|^2_\infty \right\} \\ &\leq \sup_{z\in S} \left\{ \sup_{t\in[0,T]} \big(a_2 x(t) - a_1 t\big) + a_2 \sup_{t\in[0,T]} z(t) - \lambda\|z\|^2_\infty \right\} \\ &= f(x) + \sup_{z\in S} \big(a_2\|z\|_\infty - \lambda\|z\|^2_\infty\big), \end{aligned}$$

which, in turn, is maximized for $||z|| = a_2/2\lambda$. Combining this upper bound with the lower bound in (51), we obtain that $\phi_\lambda(x) = f(x) + a_2^2/4\lambda$. Next, it is a straightforward exercise in calculus to verify that the function $\lambda\delta + \int \phi_\lambda d\mu$ is minimized at $\lambda^* = a_2/2\sqrt{\delta}$, and subsequently,

$$\inf_{\lambda\geq 0}\left\{\lambda\delta + \int \phi_\lambda d\mu\right\} = \int f d\mu + a_2\sqrt{\delta}.$$

This completes the proof. □ □

B.2. **Proofs of results used in Section 6.1.** For the results (Lemma 13, 14 and Corollary 2) stated and proved in this section, let $S = D([0,T],\mathbb{R}^2)$ be equipped with the J_1-topology induced by the metric

$$d_{J_1}(x,y) = \inf_{\lambda\in\Lambda}\left\{\sup_{t\in[0,T]}\max_{i=1,2}|x_i(t) - y_i(\lambda(t))| + \sup_{t\in[0,T]}|\lambda(t) - t|\right\},$$

where the set Λ, as in Lemma 3, is the set of strictly increasing functions λ mapping the interval $[0,T]$ onto itself, such that both λ and λ^{-1} are continuous.

Lemma 13. *For a given vector $a = (a_1, a_2)$ with $a_1, a_2 \geq 0$ and $a_1 a_2 > 0$,*

$$\mathrm{cl}\left(\{x \in S : a^T x(t) \leq 0 \text{ for some } t \in [0,T]\}\right) = \left\{x \in S : \inf_{t\in[0,T]} a^T x(t) \leq 0\right\}.$$

Proof. We first show that the set $C := \{x \in S : \inf_{t\in[0,T]} a^T x(t) \leq 0\}$ is closed by showing that its complement is open. Given any $x \notin C$, let $\varepsilon := \inf_{t\in[0,T]} a^T x(t) > 0$. Then, for every $y \in S$ such that $d_{J_1}(x,y) < \varepsilon/(2a_1 + 2a_2)$, we have that

$$\inf_{t\in[0,T]} y_i(t) \geq \inf_{t\in[0,T]} x_i(t) - d_{J_1}(x,y) > \inf_{t\in[0,T]} x_i(t) - \frac{\varepsilon}{2(a_1+a_2)},$$

for $i = 1, 2$. Further, as a_1, a_2 are non-negative, we obtain that

$$\inf_{t\in[0,T]} a^T y(t) \geq \inf_{t\in[0,T]} a^T x(t) - (a_1 + a_2)\frac{\varepsilon}{2(a_1+a_2)} > 0.$$

Thus, for every $x \in S\setminus C$, we can find an $\varepsilon > 0$ such that $\{y \in S : d_{J_1}(x,y) < \varepsilon/(2a_1 + 2a_2)\}$ is a subset of $S \setminus C$, and hence $S\setminus C$ is open. Therefore, C is closed.

Letting $D := \{x \in S : a^T x(t) \leq 0 \text{ for some } t \in [0,T]\}$, the remaining task is to show that for every $\varepsilon > 0$ and $x \in C$, there exists $y \in D$ such that $d_{J_1}(x,y) < \varepsilon$. For any given $\varepsilon > 0$ and $x \in C$, we first see that there exists $t_0 \in [0,T]$ such that $a^T x(t_0) < \varepsilon(a_1 + a_2)/2$; such a t_0 exists for every $x \in C$ because $\inf_{t\in[0,T]} a^T x(t) \leq 0$. Then $(y(t) = (y_1(t), y_2(t)) : t \in [0,T])$ defined as $y_i(t) = x_i(t) - \varepsilon, i = 1,2$ lies in D because $a^T y(t_0) = a^T x(t_0) - (a_1 + a_2)\varepsilon < 0$. Therefore, every x in C is a closure point of D. Therefore, $\mathrm{cl}(D) = C$. □ □

Corollary 2. *Let $B = \{(x_1,x_2) \in \mathbb{R}^2 : \beta x_1 + x_2 \leq 0\} \cup \{(x_1,x_2) \in \mathbb{R}^2 : x_1 + \beta x_2 \leq 0\}$ for some $\beta \geq 0$. Then $\mathrm{cl}(\{x \in S : x(t) \in B \text{ for some } t \in [0,T]\})$ equals $A_1 \cup A_2$, where*

$$A_1 := \left\{x \in S : \inf_{t\in[0,T]}(\beta x_1(t) + x_2(t)) \leq 0\right\} \text{ and } A_2 := \left\{x \in S : \inf_{t\in[0,T]}(x_1(t) + \beta x_2(t)) \leq 0\right\}. \tag{52}$$

Proof. It follows from Lemma 13 that $\mathrm{cl}(\{x \in S : \beta x_1(t) + x_2(t) \leq 0 \text{ for some } t \in [0,T]\}) = A_1$ and $\mathrm{cl}(\{x \in S : x_1(t) + \beta x_2(t) \leq 0 \text{ for some } t \in [0,T]\}) = A_2$. Then the statement to verify follows from the fact that the closure of the union of two sets equals the union of closures of those two sets. □

Lemma 14. *For $x, y \in D([0,T], \mathbb{R}^2)$, let $c(x,y)$ be defined as in (42). Let $A = A_1 \cup A_2$, where A_1 and A_2 are defined as in (52), and $\beta \geq 0$. Then for $x \in S \setminus A$,*

$$c(x, A) = \frac{1}{1+\beta^2}\left[\inf_{t\in[0,T]}\big(\beta x_1(t) + x_2(t)\big)^2 \wedge \inf_{t\in[0,T]}\big(x_1(t) + \beta x_2(t)\big)^2\right],$$

and for $\lambda^* \geq 0$, $\{x \in S : c(x, A) \leq 1/\lambda^*\}$ equals

$$\left\{x \in S : \inf_{t\in[0,T]}(\beta x_1(t) + x_2(t)) \leq \sqrt{\frac{1+\beta^2}{\lambda^*}}\right\} \cup \left\{x \in S : \inf_{t\in[0,T]}(x_1(t) + \beta x_2(t)) \leq \sqrt{\frac{1+\beta^2}{\lambda^*}}\right\}$$

Proof. Let us focus on determining $c(x, A_1) = \inf_{y\in A_1} c(x,y)$ for $x \notin A_1$. Given $x \notin A_1$ and $\varepsilon > 0$, we first define $\tilde{x}_\varepsilon(t) := x(t) - (b+\varepsilon)u$, where $b := \inf_{t\in[0,T]}(\beta x_1(t) + x_2(t))/\sqrt{1+\beta^2} > 0$ and u is the unit vector (in ℓ_2-norm) along the direction $[\beta, 1]$. As

$$\beta \tilde{x}_{\varepsilon,1}(t) + \tilde{x}_{\varepsilon,2}(t) = \beta x_1(t) + x_2(t) - (b+\varepsilon)\sqrt{1+\beta^2} \leq 0 \quad \text{ for some } t \leq T,$$

we have $\tilde{x}_\varepsilon := (\tilde{x}_{\varepsilon,1}, \tilde{x}_{\varepsilon,2}) \in A_1$. As a result,

$$c(x, A_1) \leq c(x, \tilde{x}_\varepsilon) = \|(b+\varepsilon)u\|_2^2 = \left(\inf_{t\in[0,T]} \frac{\beta x_1(t) + x_2(t)}{\sqrt{1+\beta^2}} + \varepsilon\right)^2 \tag{53}$$

for $x \notin A_1$. Next, using the same line of reasoning as in the proof of Lemma 11, one can restrict to time changes $\lambda(t) = t$ in the computation of lower bound. Then,

$$c(x, A_1) = \inf_{y\in A_1} \sup_{t\in[0,T]} ||x(t) - y(t)||_2^2 \geq \inf_{y\in A_1} \sup_{t:\beta y_1(t)+y_2(t)\leq\varepsilon} ||x(t) - y(t)||_2^2,$$

for any $\varepsilon > 0$. For $z \in \mathbb{R}^2$, if we let $g_\varepsilon(z) = \inf_{\{y\in\mathbb{R}^2:\beta y_1 + y_2 \leq \varepsilon\}} ||z - y||_2^2$, then

$$c(x, A_1) \geq \inf_{y\in A_1} \sup_{t:\beta y_1(t)+y_2(t)\leq\varepsilon} g_\varepsilon(x(t)).$$

Next, for $z = (z_1, z_2)$ such that $\beta z_1 + z_2 > \varepsilon$, observe that $g_\varepsilon(z) = (\beta z_1 + z_2 - \varepsilon)^2/(1+\beta^2)$. Therefore,

$$c(x, A_1) \geq \inf_{y\in A_1} \sup_{t:\beta y_1(t)+y_2(t)\leq\varepsilon} \frac{(\beta x_1(t) + x_2(t) - \varepsilon)^2}{1+\beta^2} \geq \inf_{t\in[0,T]} \frac{(\beta x_1(t) + x_2(t) - \varepsilon)^2}{1+\beta^2}$$

for every ε small enough. As ε can be arbitrarily small, combining the upper bound in (53) with the above lower bound results in

$$c(x, A_1) = \inf_{t\in[0,T]} \frac{\big(\beta x_1(t) + x_2(t)\big)^2}{1+\beta^2} \quad \text{ for } x \notin A_1.$$

Consequently, $\{x \in S : c(x, A_1) \leq 1/\lambda^*\} = A_1 \cup \{x \in S : c(x, A_1) \in (0, 1/\lambda^*)\}$, which is simply,

$$\left\{x \in S : \inf_{t\in[0,T]}\big(\beta x_1(t) + x_2(t)\big) \leq \sqrt{\frac{1+\beta^2}{\lambda^*}}\right\} \quad \text{ for } x \notin A_1.$$

Similarly, one can show that $c(x, A_2) = \inf_{t\in[0,T]}(x_1(t) + \beta x_2(t))^2/(1+\beta^2)$ and a similar expression as above for $\{x \in S : c(x, A_2) \leq 1/\lambda^*\}$. As $A = A_1 \cup A_2$, it is immediate that $c(x, A) = c(x, A_1) \wedge c(x, A_2)$ and $\{x \in S : c(x, A) \leq \frac{1}{\lambda^*}\}$ is the union of two sets $\{x \in S : c(x, A_i) \leq 1/\lambda^*\}, i = 1, 2,$ thus verifying both the claims. □ □

B.3. **Technical results used in Section 4 to complete the proof of Theorem 1.**

Lemma 15. *Let $S \times S$ be a Polish space that is compact, and $f : S \to \mathbb{R}$ is upper semicontinuous. Suppose that $\pi \in M(S \times S)$, with the Jordan decomposition $\pi = \pi^+ - \pi^-$ comprising positive measures π^+ and π^-, is such that $\pi^+(A) = 0 < \pi^-(A) < \infty$ for some $A \in \mathcal{B}(S \times S)$. Then,*

$$\inf \left\{ \int g d\pi : g \in C_b(S \times S), g(x,y) \geq f(y) \right\} = -\infty.$$

Proof. We first observe that if π is not As any finite Borel measure on a Polish space is regular, there exists a compact set K_ε and an open set O_ε such that $K_\varepsilon \subseteq A \subseteq O_\varepsilon$ and

$$\pi^-(O_\varepsilon) - \varepsilon \ \leq \ \pi^-(A) \ \leq \ \pi^-(K_\varepsilon) + \varepsilon, \quad \text{ and } \quad \pi^+(O_\varepsilon) \ \leq \ \varepsilon,$$

for any given $\varepsilon > 0$ (see Lemma 18.5 in Aliprantis and Burkinshaw (1998)). Further, as $S \times S$ is compact, Urysohn's lemma (see, for example, Theorem 10.8 in Aliprantis and Burkinshaw (1998)) guarantees us the existence of a continuous function $h : S \times S \to [0,1]$ such that $h(x,y) = 1$ for all $x \in K_\varepsilon$ and $h(x,y) = 0$ for all $x \notin O_\varepsilon$. In that case, choosing $\varepsilon < \pi^-(A)/2$, we have $\inf_{n \geq 1} \int g_n d\pi = -\infty$ for the sequence of continuous functions $g_n(x,y) = nh(x,y) + \sup_{x \in S} f(x)$. This is because, $\sup_{x\in S} f(x) < \infty$ (recall that f is upper semicontinuous and S is compact), and

$$\int h d\pi = \int h d\pi^+ - \int h d\pi^- \ \leq \ \pi^+(O_\varepsilon) - \pi^-(K_\varepsilon) \ \leq \ 2\varepsilon - \pi^-(A) \ < \ 0$$

when $\varepsilon < \pi^-(A)/2$. As $g_n(x,y) \geq f(y)$ for all x, y in S, it follows that $\inf_{g \in D} \int g d\pi \leq \inf_n \int g_n d\pi = -\infty$ as well, whenever there exists A such that $\pi^-(A) > 0$. ☐ ☐

Lemma 16. *Suppose that Assumptions (A1) and (A2) are in force. Let $(S_n : n \geq 1)$ be an increasing sequence of subsets of S, and $(\lambda_n : n \geq 1)$ be a real-valued sequence satisfying $\lambda_n \to \lambda^*$, for some $\lambda^* \geq 0$, as $n \to \infty$. Then for any $x \in S$,*

$$\varliminf_n \sup_{y \in S_n} \left\{ f(y) - \lambda_n c(x,y) \right\} \geq \sup_{y \in \cup_n S_n} \left\{ f(y) - \lambda^* c(x,y) \right\}.$$

Proof. Pick any $x \in S$. For brevity, let $g(y, \lambda) := f(y) - \lambda c(x,y)$ (hiding the dependence on the fixed choice $x \in S$) and $\bar{\lambda}_m := \sup_{k \geq m} \lambda_k$. The observations that
(A) $g(y, \lambda)$ is a non-increasing function in λ, and
(B) $\sup_{y \in S_n} g(y,\lambda) \leq \sup_{y \in S_{n+1}} g(y, \lambda)$ for $n \geq 1$,
will be used repeatedly throughout this proof. While Observation (A) follows from the fact that $S_n \subseteq S_{n+1}$, Observation (B) is true because $c(\cdot,\cdot)$ is non-negative. The quantity of interest,

$$\varliminf_n \sup_{y \in S_n} \left\{ f(y) - \lambda_n c(x,y) \right\} = \sup_{n > 0} \inf_{m \geq n} \sup_{y \in S_m} g(y, \lambda_m) \geq \sup_{n>0} \inf_{m \geq n} \sup_{y \in S_n} g(y,\lambda_m) \geq \sup_{n > 0} \sup_{y \in S_n} \inf_{m \geq n} g(y, \lambda_m),$$

where the first inequality follows from Observation (B), and the second inequality from the exchange of inf and sup operations. As $\inf_{m \geq n} g(y, \lambda_m) = f(y) - (\sup_{m \geq n} \lambda_m)c(x,y) = g(y, \bar{\lambda}_n)$, we obtain

$$\varliminf_n \sup_{y \in S_n} \left\{ f(y) - \lambda_n c(x,y) \right\} \geq \sup_{n > 0} \sup_{y \in S_n} g(y, \bar{\lambda}_n) = \sup_{y \in \cup_n S_n} g(y, \lambda^*), \tag{54}$$

where the rest of this proof is devoted to justify the last equality in (54): As $\sup_{y \in S_n} g(y, \bar{\lambda}_n)$ is non-decreasing in n, it is immediate that $\sup_n \sup_{y \in S_n} g(y, \bar{\lambda}_n) \leq \bar{g} := \sup_{y \in \cup_n S_n} g(y, \lambda^*)$. To show that $\sup_n \sup_{y \in S_n} g(y, \bar{\lambda}_n)$ indeed equals $\bar{g}$, we use $\hat{S}$ to denote $\hat{S} := \cup_n S_n$, and consider distinct cases:
Case 1: $\bar{g} < \infty$. If $\bar{g}$ is finite, then there exists $y_\varepsilon \in \hat{S}$ such that $g(y_\varepsilon, \lambda^*) \geq \bar{g} - \varepsilon/2$ for any $\varepsilon > 0$. Further, as $g(y_\varepsilon, \lambda)$ is continuous in λ, there exists n_0 large enough such that $g(y_\varepsilon, \bar{\lambda}_n) \geq g(y_\varepsilon, \lambda^*) - \varepsilon/2$, and consequently, $g(y_\varepsilon, \bar{\lambda}_n) \geq \bar{g} - \varepsilon$, for all $n \geq n_0$. Therefore, for all $n > n_0$ satisfying $y_\varepsilon \in S_n$, it follows that $\sup_{y \in S_n} g(y, \bar{\lambda}_n) \geq g(y_\varepsilon, \bar{\lambda}_n) \geq \bar{g} - \varepsilon$. Since ε is arbitrary, we obtain $\sup_n \sup_{y\in S_n} g(y, \bar{\lambda}_n) = \bar{g}$ whenever $\bar{g}$ is finite.

Case 2: $\bar g = \infty$. If $\bar g = \sup_{y\in\hat S} g(y,\lambda^*) = \infty$, then there exists a strictly increasing subsequence $(n_k : k \geq 1)$ of natural numbers such that $\sup_{y\in\hat S} g(y,\bar\lambda_{n_k}) > k$ for all k (this is because, being a pointwise supremum of family of continuous functions of λ, $\sup_{y\in\hat S} g(y,\lambda)$ is a lower semicontinuous function of λ). Next, as $\sup_{y\in\hat S} g(y,\bar\lambda_{n_k}) > k$, there exists a sequence $(y_k : k \geq 1)$ comprising elements of $\hat S$ such that $g(y_k,\bar\lambda_{n_k}) \geq k/2$ for all $k \geq 1$. As $(S_n : n \geq 1)$ is a sequence of sets increasing to $\hat S$, one can identify a strictly increasing subsequence $(m_k : k \geq 1)$ of natural numbers such that $y_k \in S_{m_k}$ and consequently, $\sup_{y\in S_{m_k}} g(y,\bar\lambda_{n_k}) \geq k/2$ for all $k \geq 1$. Next, if we let $l_k = n_k \vee m_k$, then $\bar\lambda_{l_k} \leq \bar\lambda_{n_k}, S_{m_k} \subseteq S_{l_k}$, and it follows from Observations (A) and (B) that $\sup_{y\in S_{l_k}} g(y,\bar\lambda_{l_k}) \geq k/2$ as well. This results in a non-decreasing sequence $(l_k : k \geq 1)$ with $l_k \to \infty$ such that $\sup_{y\in S_{l_k}} g(y,\bar\lambda_{l_k}) \geq k/2$, thus yielding $\sup_n \sup_{y\in S_n} g(y,\bar\lambda_n) = \infty = \bar g$.

The proof is complete because $\sup_{y\in\cup_n S_n} g(y,\lambda^*)$ in (54) is the desired right hand side. □ □

B.4. A technical result to add clarity to the notation (49).

Lemma 17. *Suppose that Assumptions (A1) and (A2) are in force. Then, for every $\pi \in \Phi_{\mu,\delta}$, there exists a $\pi' \in \Phi_{\mu,\delta}$ such that π' is concentrated on $\{(x,y) \in S\times S : f(y) \geq f(x)\}$, along with satisfying,*

$$\int f^+(y)d\pi'(x,y) \geq \int f^+(y)d\pi(x,y) \quad \text{and} \quad \int f^-(y)d\pi'(x,y) \leq \int f^-(x)d\mu(x). \tag{55}$$

Further, $\int f(y)d\pi'(x,y) \geq \int f(y)d\pi(x,y)$, whenever the integral $\int f(y)d\pi(x,y)$ is well-defined.

Proof. Let (X,Y) be jointly distributed according to π. Using (X,Y), let us define a new jointly distributed pair (X',Y'), with joint distribution denoted by π', as follows:

$$X' := X \quad \text{and} \quad Y' := YI\left(f(X) \leq f(Y)\right) + XI\left(f(X) > f(Y)\right), \tag{56}$$

where $I(\cdot)$ denotes the indicator function. As $X' = X$, the marginal distribution of X' is μ. Further, it follows from the definition of (X',Y') in (56) that

$$\int c d\pi' = \int_{\{f(x)\leq f(y)\}} c(x,y)d\pi(x,y) + \int_{\{f(x)>f(y)\}} c(x,x)d\pi(x,y) = \int_{\{f(x)\leq f(y)\}} c(x,y)d\pi(x,y),$$

because $c(x,x) = 0$ for every x in S. In addition, as $c(\cdot,\cdot)$ is non-negative, it follows that $\int c d\pi' \leq \int c d\pi \leq \delta$. Therefore, $\pi \in \Phi_{\mu,\delta}$. Next,

$$\begin{aligned}
\int f^+(y)d\pi'(x,y) = &\int_{\{f(x)\leq f(y)\}} f^+(y)d\pi(x,y) + \int_{\{f(x)>f(y)\}} f^+(x)d\pi(x,y)\\
= &\int_{\{f^+(x)\leq f^+(y)\}} f^+(y)d\pi(x,y) + \int_{\{f^+(x)>f^+(y)\}} f^+(x)d\pi(x,y)\\
\geq &\int_{\{f^+(x)\leq f^+(y)\}} f^+(y)d\pi(x,y) + \int_{\{f^+(x)>f^+(y)\}} f^+(y)d\pi(x,y) = \int f^+(y)d\pi(x,y), \qquad \text{and}\\
\int f^-(y)d\pi'(x,y) = &\int_{\{f(x)\leq f(y)\}} f^-(y)d\pi(x,y) + \int_{\{f(x)>f(y)\}} f^-(x)d\pi(x,y)\\
= &\int_{\{f^-(x)\geq f^-(y)\}} f^-(y)d\pi(x,y) + \int_{\{f^-(x)<f^-(y)\}} f^-(x)d\pi(x,y)\\
\leq &\int_{\{f^-(x)\geq f^-(y)\}} f^-(x)d\pi(x,y) + \int_{\{f^-(x)<f^-(y)\}} f^-(x)d\pi(x,y) = \int f^-(x)d\pi(x,y) = \int f^- d\mu,
\end{aligned}$$

As $\int f^- d\mu < \infty$, $\int f(y) d\pi'(x,y)$ is always well-defined. Further, when $\int f(y) d\pi(x,y)$ is also well defined, it is immediate from the definition of π' that,

$$\begin{aligned}\int f(y) d\pi'(x,y) &= \int_{\{f(x)\leq f(y)\}} f(y) d\pi(x,y) + \int_{\{f(x)>f(y)\}} f(x) d\pi(x,y) \\ &\geq \int_{\{f(x)\leq f(y)\}} f(y) d\pi(x,y) + \int_{\{f(x)>f(y)\}} f(y) d\pi(x,y) = \int f(y) d\pi(x,y),\end{aligned}$$

thus verifying all the claims made in the statement. □

Corollary 3. *Suppose that Assumptions (A1) and (A2) are in force. Let $\nu \in P(S)$ be such that $\int f^- d\nu = \infty$ and $d_c(\mu,\nu) \leq \delta$ for a given reference probability measure $\mu \in P(S)$. Then there exists $\nu' \in P(S)$ such that $\int f^+ d\nu' \geq \int f^+ d\nu$, $\int f^- d\nu' < \infty = \int f^- d\nu$, and $d_c(\mu,\nu') \leq \delta$.*

Proof. Let $\pi \in \Pi(\mu,\nu)$ be an optimal coupling that attains the infimum in the definition of $d_c(\mu,\nu)$; in other words, $d_c(\mu,\nu) = \int c d\pi$ (such an optimal coupling π always exists because of the lower semicontinuity of $c(\cdot,\cdot)$, see Theorem 4.1 in Villani (2008)). Since $d_c(\mu,\nu) \leq \delta$ and $\pi(\cdot \times S) = \mu(\cdot)$, we have that $\pi \in \Phi_{\mu,\delta}$. Then, according to Lemma 17, there exists $\pi' \in \Phi_{\mu,\delta}$ satisfying (55). Consequently, the marginal distribution defined by $\nu'(\cdot) := \pi'(S \times \cdot)$ satisfies $d_c(\mu,\nu') \leq \delta$, along with $\int f^+ d\nu' \geq \int f^+ d\nu$ and $\int f^- d\nu' \leq \int f^- d\mu < \infty = \int f^- d\nu$. □

Columbia University, Department of Industrial Engineering & Operations Research, S. W. Mudd Building, 500 W 120th St., New York, NY 10027, United States.

E-mail address: {jose.blanchet, karthyek.murthy}@columbia.edu