PRIVATE RISK AND VALUATION: A DISTRIBUTIONALLY ROBUST OPTIMIZATION VIEW

J. BLANCHET AND U. LALL

ABSTRACT. This chapter summarizes a body of work which has as objective the quantification of risk exposures that are particularly important in the context of industries such as the mining industry and which are inherently difficult to calibrate against a probabilistic model due to lack of information. In order to address this problem, we propose an approach based on game theoretic representations which are known in the operations research literature as distributionally robust optimization (DRO) problems. Our goal in this chapter is to provide a high level and conceptual discussion of this approach and explain how we extended it and applied it to the mining setting which is the topic of this project.

1. INTRODUCTION

Concerns with climate change are now raising the question of how best to address the exposure of companies to physical climate risk that may be experienced at their physical assets and through their supply chain, see for example ([17]). The processes which dictate how such risks can be assessed, disclosed and used for the valuation of the companies are still at an evolutionary stage. A critical aspect is how to identify and price the exposure to, for example, extreme climate events, under current or future climate.

Even where companies design infrastructure to protect against extreme climate events such as droughts and floods, the design criteria may not be disclosed, and in a nonstationary climate the actual period of record used to estimate the likelihood of extreme events may alter the potential exposure risk. This level of detail and the re-appraisal of such risks is unlikely to happen. Calibrating this exposure is also difficult since by definition extreme events are rare, and their impacts may be highly location dependent. Hence, very little data may be available for relevant exposure and model calibration.

The body of research that we summarize in this chapter proposes to address these types of challenges and we choose the mining industry for an illustration of the ideas that we develop. This industry has a high concentration of its valuation in a relatively small number of assets or mines, with high potential exposure to climate risk, making it a particularly relevant sector for the initial application and insights of our theory. However,

Date: December 24, 2017.

Key words and phrases. Distributionally Robust Optimization, Real Options, Optimal Transport, Extreme Value Analysis, Robustness.

we believe that our methodology can be applied to a wide range of settings; any environment in which private risk, which is difficult to quantify, can have a significant impact in valuations and exposures.

At the core of our methodology lies a game theoretic formulation. There are two players, one is the manager who wishes to maximize the value of the company. The second one is an adversary which is introduced to recognize that the probabilistic model which may be assumed by the manager is not fully known.

We postulate that managers will maximize the value of their company by making rational operational decisions, including investments and extraction policies, among others. Managers face a stochastic environment influenced by financial and physical variables. One such physical variable, for example, involves the occurrence of important environmental shocks. While there may be enough information in the financial markets to calibrate a probabilistic model for the price fluctuation of an underlying mineral, such as cooper or gold, there may not be enough information (due to the circumstances described earlier) to calibrate a probabilistic model for extreme climatic events. The manager may be forced to *assume* what may seem to be a reasonable probabilistic description, which, nevertheless, is subject to model error and thus might yield a valuation error.

In order to cope with this lack of information, we introduce an adversary, which observes the manager's decisions and perturbs the assumed probabilistic description of the model in order to adversarily affect the overall value of the manager's company (or mine).

The estimate of the value under our framework is therefore obtained by solving a max-min game. The max-player corresponds to the manager and the min-player corresponds to the adversary which is artificially introduced to recognize the lack of information discussed earlier and which perturbs a reasonable baseline model. The max-player is able to make operational decisions to maximize the value of the mine. The min-player (i.e. the adversary) is endowed with "features" (i.e. the shape of the perturbations allowed) and "power" (the amount of perturbation allowed). In the end, we think of the min-player only as a tool which systematically allows us to explore the impact in value which is derived by the lack of knowledge of the agent who is evaluating the mine in question.

The body of research that we produce studies questions such as:

a) How to construct a reasonable adversary?

b) How to calibrate the amount of power that should be given to the adversary to avoid overly conservative value estimates?

c) How to connect questions such as b) to well accepted statistical methodology?

d) How to efficiently solve the postulated games?

e) How to translate all of these insights into a technological tool that can aid managers in assessing the financial impact of climatic events?

Throughout the rest of this chapter, we will elaborate on the research that we have produced to address items a) to e), but before doing so, we will provide more details on the merits of our approach from a decision theory standpoint. For example, we will first address questions such as:

I) Aren't we being too conservative by introducing an adversary who chooses a worst-case distribution?

II) Why not just use a Bayesian perspective, impose a parametric family of models, and recognize the uncertainty in the model by imposing a prior on the parameters? What is wrong with that approach?

III) Is this min-max approach supported in any set of basic axioms rooted in rational decision making?

The rest of this chapter is organized as follows. In Section 2 we discuss our framework from a decision theoretic standpoint and address questions such as I) to III).

In Section 3 we discuss the use of entropy as a way to describe the shape of the perturbations which the artificial adversary applies, this section is intended to provide insights into question a).

In Section 4 we discuss a method to calibrate the power given to the adversary when using entropy to describe the shape of the perturbations and in the context of extreme value statistics, this relates to questions b) and c).

Section 5 relates to d) and e), we explain the use of the methodology discussed in Sections 3 and 4 applied to the particular setting of mining by means of the development of a real options valuation tool.

In Section 6 we discuss the problems that arise when using entropy to describe the shape of the perturbations utilized by the adversary and we also introduce a different approach, based on optimal mass transportation, which in turn has natural economic interpretations; this section revisits question a) and d).

In Section 7 we discuss a methodology to calibrate the amount of perturbations allowed for the adversary when optimal mass transportation is used to describe distributional uncertainty, this section revisits questions b) and c). In particular, we establish connections to machine learning methodology and deep learning.

Finally, in Section 8, we present a diagram which summarizes conceptually the body of work produced in this project and its motivations. We conclude with a discussion of potential future research directions.

2. Economic and Statistical Foundations of Distributionally Robust Optimization

Our methodology is based on the celebrated work on economic robustness due to Hansen and Sargent[1]. In their book on robustness, [11], Section 1.9, Hansen and Sargent address the merits of a max-min formulation such as the one that we propose, and also questions I) to III). Here we simply summarize the arguments provided by Hansen and Sargent.

First, in connection to the use of II), we believe that such a Bayesian approach is reasonable, but one may be in a position in which there is just not enough data to calibrate a

[1]Hansen and Sargent each won the Nobel Price in Economics in 2013 and 2011, respectively.

reasonable prior distribution. So, the results may be too sensitive to the particular choice of a prior distribution or model, which is precisely what the distributionally robust approach is trying to mitigate.

In settings involving scarce data, it is well known that different priors will results in substantially different inference. In turn, there may be multiple Bayesian priors that could be used as a reasonable representation of parametric uncertainty, but resulting in significantly different posterior distributions.

In addition, the lack of information may manifest itself in ways that are not necessarily captured by a Bayesian specification which typically involves finitely many hyperparameters on which one imposes a prior distribution. In other words, one would need to involve a prior distribution which is supported in an infinite dimensional space and this creates computational complications.

There is a substantial amount of theory devoted to inference in the context of multiple priors and this is precisely the setting that leads to the max-min formulation that we consider here.

Now, in connection to statistical decision theory, the work of Savage, [13], studied a set of axioms based on utility theory which lead to Bayesian formulations as optimal statistical decision rules, based on expected utfility maximization problems. But some objections have been raised with regard to Savage's formulation and one of them appears particularly relevant to our current setting. Following Gilboa and Schmeidler, [10], we discuss an experiment from [9] in which the issue of lack of information is exposed. Consider two urns, A and B, each containing 100 balls. Each ball is either black or white. Urn A contains 50 black balls and 50 white balls. No additional information is available for urn B. One ball is drawn at random from each urn and we consider the following propositions: P1) the ball drawn from urn A is black, P2) the ball drawn from urn A is white, P3) the ball drawn from urn B is black, and P4) the ball drawn from urn B is white. Perhaps not surprisingly, in empirical experiments, the bets are typically ranked as P1 = P2 $>$ P3 = P4. There is no way in which a Bayesian approach, leading to a utility maximization formulation, would support these preferences. An approach in which one considers all possible priors and minimizes the expected utility over such priors, on the other hand, would support precisely such ranking of preferences.

In contrast to Savage's approach, which is the foundation of a Bayesian framework, the approach that we take here is more related to the work of Wald, [16], game-theoretic paradigm in which the decision maker faces an adversary leading to a max-min utility formulation as the one that we consider in our approach.

In [10] it is shown that by altering only one of the axioms imposed in [13] one arrives to an axiomatic foundation of the type of criterion that we use here (i.e. the one prescribed

by Wald in [16]) and which is motivated, following [16], by a setting in which "... an apriori distribution does not exist or it is unknown to the observer".[2]

So, in summary, the approach that we follow here is also rooted in a rational decision making theory, just as the Bayesian approach. In fact, it is fundamentally similar, but it appears more appropriate in a setting in which many priors may be reasonable and they may lead to significantly different inference outcomes.

Regarding question I), it is conceivable that the choice of available perturbations given to the adversary and the size of such perturbations might result in conservative estimates. In order to mitigate this problem we need to provide a solid foundation which justifies the use of a "reasonable" adversary and the calibration of a "reasonable choice" of perturbation size.

What constitutes reasonable has to do with our perception of the phenomena which is most difficult to model probabilistically. Most of our methodological research is directed precisely to this problem. We shall illustrate the guiding principles that we obtained through a range of different examples.

3. Distributionally Robust Optimization and Entropy

At this point, we must introduce mathematical notation to explain the concept of entropy. It is useful to keep in mind the high-level explanation given in the Introduction and contrast such a description with the mathematical formulations provided next. We assume that P_0 represents the probabilistic model assumed by the manager.

A decision made by the manager is represented by a parameter θ, which is assumed to belong to a set of admissible decisions, Θ, that the manager is allowed to consider. Each decision may encode a complex set of actions (e.g. the extraction policy, the decision of when to close and re-open a mine, etc.) We are not concerned at the moment with the problem of how to computationally obtain an optimal decision θ, we simply are interested in conceptually explaining the approach that we study.

The manager recognizes that P_0 is a plausible, but imperfect, representation of reality and wishes to choose a decision θ which performs well over a range of possible models P which are in some neighborhood of P_0 (representing also plausible descriptions of the reality).

So, we define such a neighborhood, called the distributional uncertainty set, via

$$\mathcal{U}_\delta(P_0) = \{P : D(P, P_0) \leq \delta\},$$

where $D(P, P_0)$ is a notion of discrepancy between P and P_0 and $\delta > 0$ is the size of the uncertainty set. The description of $D(\cdot)$ dictates the types of perturbations (or shape) that are allowed by the artificial player that we will introduce, and δ represents the size

[2]Gilboa and Schmeidler consider a weakening of the so-called independence action, relaxing it to the so-called C-independence action. The C-independence assumes a certain consistency in preferences involving any two acts and their convex combination with constant acts only (as opposed to general acts). A discussion on the benefits of this relaxation in practice is given in Gilboa and Schmeidler (1989), p. 145.

of such perturbations. So, both the choice of $D(\cdot)$ and δ are important elements in our modeling framework.

Suppose that X represents all of the random risk factors that affect the value of a mine in a given time horizon representing the life of the mine, which might be random itself. Let $u(X,\theta)$ be the discounted net present value at time zero of the cash flow generated by the mine, given that X is observed in a given time horizon, and given that the manager has implemented a decision encoded by θ. A traditional valuation approach (using real option pricing methodology) consists in calibrating $P_0(\cdot)$ and evaluating

$$v_0 = \max_{\theta\in\Theta} E_0\left(u(X,\theta)\right).$$

Instead, we postulate solving a so-called distributionally robust optimization (DRO) problem, obtaining

$$v_\delta^- = \max_{\theta\in\Theta} \min_{P\in\mathcal{U}_\delta(P_0)} E_0\left(u(X,\theta)\right) < v_0.$$

This approach produces valuation estimates which are lower than traditional valuation approaches because we incorporate a risk premium derived from the amount of distributional uncertainty. If $\delta = 0$, and mild continuity assumptions are imposed on the discrepancy $D(\cdot)$, then $v_\delta^- = v_\delta$. In Section 5 we discuss how to produce an interval for the value; such an interval can be used to assess if the market value of an asset underestimates or overestimates the exposure to the type of private risk discussed earlier in the Introduction.

Hansen and Sargent [11] advocate the use of the relative entropy (or Kullback-Leibler divergence) as a notion of discrepancy, that is,

$$D(P,P_0) := D(P||P_0) = E_P\left[\log\left(\frac{dP}{dP_0}\right)\right] > 0. \tag{1}$$

We do not believe that there is enough evidence which supports the use of the relative entropy from a structural standpoint, but we discuss the main reasons for using relative entropy.

First, we wish to use a non-parametric notion because we want to minimize as much as possible introducing bias in valuation (that is why we are not following a Bayesian approach). There are not many non-parametric notions of discrepancy that lead to a tractable minimization problem for the adversary. So, one of the main reasons provided in the literature for the use of entropy is tractability. For example, when the underlying random factor, X, is Gaussian under P_0 and $u(X,\theta)$ is a quadratic form in X, then the worst case probability model resulting from (1) is also Gaussian and therefore the maximization problem for the manager is very similar to the one that he would solve without introducing the min-player.

Another reason for choosing relative entropy is that it possesses a compelling invariance property. In particular, the value of the minimization under the relative entropy

unchanged under different parameterizations of the problem (see [7]). Actually this invariance also holds in other discrepancies (which we also consider), the key fact is the dependence on the likelihood ratio dP/dP_0.

Finally, the relative entropy concept has been studied in statistics, robust control, information theory and in economics and this is also a pragmatic reason for employing it.

In the case of capturing discrepancy in financial valuations, the use of a discrepancy based on the likelihood ratio (i.e. dP/dP_0) is compelling because the concept of likelihood ratio lies at the core of pricing theory. The fundamental theorem of asset pricing (see [6]) asserts that there is no arbitrage (i.e. free lunch) if and only if prices can be computed as the expected net present value according to some probability distribution which is "equivalent" to the probability model which dictates the physical dynamics of the underlying risk factors in the economy. The notion of equivalence is a mathematical concept defined in the theory of probability (see [8]), two probability models P and P_0 are equivalent if and only if the corresponding likelihood ratios dP/dP_0 and dP_0/dP are well defined.

In order to calibrate P_0 we use an approach suggested in the real options treatment from [12]. Suppose that $X = (Y, Z)$ and Y corresponds to market risks which can be hedged and Z corresponds to private risks whose distribution is difficult to assess. In the case of climate risk it seems reasonable to assume that Y and Z are stochastically independent under P_0. A reasonable approach is to calibrate the distribution of Y under P_0 using risk neutral valuation techniques and construct the specification of Z under P_0 use a statistical procedure. Then utilize the distributionally robust approach by allowing perturbations on the risk factor Z.

4. Calibrating Entropy in Distributionally Robust (DR) Extreme Value Analysis

The advantages discussed for the relative entropy can be extended to more general entropy notions, in particular, by the so-called Renyi divergence of degree $\alpha > 1$, defined via

$$D_\alpha\left(P||P_0\right) = \frac{1}{\alpha - 1}\log E_{P_0}\left(\left(\frac{dP}{dP_0}\right)^\alpha\right).$$

It turns out that as $\alpha \downarrow 1$, $D_\alpha\left(P||P_0\right) \to D\left(P||P_0\right)$ so the Renyi divergence contains the relative entropy as a special case.

In the mining setting, most of the private risk that affect the valuation of a mine, which are denoted as Y in the notation introduced in Section 3, can be idealized in terms of an environmental event (e.g. tailing dams failure due to high precipitation) which has a very small probability (in the order of 1/1000 or smaller) of impacting a particular mine. (There are about two significant tailing dams failures events per year around the world and thousands of mines which are exposed.)

In order to use the min-max approach, we revisit a classical statistical problem involving the extrapolation of extreme quantiles. The problem is the following, from a small sample, say of about 150 observations, which can be used to estimate quantiles corresponding to probabilities not larger than, say, $1/12$, one is interested in estimating quantiles corresponding to probabilities of size $1/1000$ or even smaller. Non parametric estimation of such quantiles even in the context of independent and identically distributed (i.i.d.) observations is impossible. So, one resorts to a semi-parametric theory which is the corner-stone of statistical extreme value analysis (EVA). Such theory postulates (i.e. assumes) that if n data points, $Z_1, ..., Z_n$, are roughly independent and identically distributed (i.i.d), then

$$\max(Z_1, ..., Z_n) \approx_d a(n) + b(n) M, \tag{2}$$

where $\approx_d$ means "approximately equal in distribution". The key insight is that $a(n)$ and $b(n)$ are deterministic quantities, so the n sources of randomness which are present on the left hand side of (2), can be summarized in a single source of randomness, represented by the random variable M.

Under assumption (2) and also assuming that the Z_i's are i.i.d. then it can be shown that M follows a generalized extreme value distribution with some parameter $\gamma \in (-\infty, \infty)$. The parameter γ has a qualitative interpretation which is both intuitive and important from a modeling standpoint.

Because one can directly derive approximation (2) for many examples known in practice (e.g. Beta distributions, Exponential, Gaussian, Pareto, etc.) we are able to gain intuition about the meaning of γ. In particular, $\gamma < 0$ is obtained for super-light tailed observations, that is, observations with bounded support; $\gamma = 0$ is obtained when considering light-tailed observations (i.e. observations with infinite support, but with exponential-type decay, like Exponentials and Gaussians), and $\gamma > 0$ is associated to heavy-tailed observations (i.e. observations with a polynomially decaying density, such as the Pareto distribution). If $\gamma < 0$, then Z_i is said to belong the the "domain of attraction" of a Weibull distribution

It is important to recognize in mind that (2) is assumed. The mathematical result characterizes M under assumptions which are impossible to verify (e.g. the data may not even be stationary) or may not hold even in simple examples. For instance, in [2], we show provide several examples in which the use of EVA would underestimate risk exposure to extremes.

It is also important to keep in mind that, in the end, we wish to extrapolate the behavior from "typical" observations far out into the tail information (which is unavailable). In the absence of a physical mechanism which informs such an extrapolation, a statistical approach based on a linear-type regression as in (2) is reasonable because it is parsimonious and therefore easy to calibrate.

So, our point of view is that a reasonable model, P_0, might be constructed from the assumptions that support the standard application of statistical EVA. But then, in [2], we

use a distributionally robust (DR) approach to correct from the fact that the assumptions behind (2) might not hold.

The point in DRO is choosing a reasonable distributional uncertainty set $\mathcal{U}_\delta(P_0)$ and the size of the uncertainty δ. At the very least, the choice should preserve some coarse knowledge that the modeler typically possess about the uncertainty (e.g. are we facing a heavy tails vs light tails phenomena). Since the modeler (i.e. the statistician or the manager who is interested in the estimation problem) chooses P_0 using standard EVA, then the intuition involving the type of tail phenomena and its connection to the parameter $\gamma \in (-\infty, \infty)$ should typically be preserved, even after introducing a distributionally robust correction. That is, unless the modeler has little confidence on the type of tail behavior that he is facing, in such case it is necessary to choose a more powerful adversary that quantifies the potential of qualitatively higher-than-expected extremes (relative to standard EVA).

The paper [2] shows that choosing $D_\alpha(P||P_0)$ to define $\mathcal{U}_\delta(P_0)$ preserves the domain of attraction of the resulting worst case distribution. That is, defining

$$\bar{F}_{\alpha,\delta}(x) = \max_{P:D_\alpha(P||P_0)\leq\delta} P_0(M \geq x)$$

for $\alpha > 1$ results in a distribution that preserves the same domain of attraction as that of M under P_0. Moreover, if $\alpha = 1$, and the support of the underlying data is unbounded, the domain of attraction becomes qualitatively different, in particular much heavier tails are induced. This results provides a structural reason for choosing the Renyi divergence with $\alpha > 1$ and only select the relative entropy if one wishes to be "extra careful" because we are not even sure of the type of tail phenomena that we are facing, even qualitatively.

Next, in [2] we propose a data-driven approach which can be used to "automatically" select both α and δ for extrapolation in a setting which might be somewhat close to the standard EVA case (but we still are able to provide valid upper bounds in the context of inhomogeneous data). The idea is the following. If P_n represents empirically the data, $D_\alpha(P_n||P_0)$ can be estimated, together with a and upper confidence level (for instance using Bootstrap). This (i.e. the upper confidence level obtain with say 95% confidence) results in a value $\delta_{n,\alpha}$. Note that this estimation is relatively insensitive to lack of information on the tails of the distribution. The value of α is then chosen to obtain the smallest upper bound for the quantites that can be well estimated using a non-parametric procedure (such as Bootstrap).

This approach is then shown in [2] using simple examples that the DR correction alleviates problems with underestimation of extreme quantiles in traditional statistical EVA.

5. Application of DRO and Real Options in Mining Operations

We now go back to the problem of mining valuation with probabilistic mispecification in private risks. The methodology summarized in this section is documented in [5].

The starting point is to use a real options valuation methodology which assumes that some probabilistic model, P_0, has been calibrated and the value of the mine is obtained

via

$$v_\delta = \max_{\theta \in \Theta} E_0 \left(u \left(Y, Z, \theta \right) \right),$$

where the maximization is solved using a numerical technique for optimal stochastic control detailed in [5], in this case θ includes decisions such as opening and closing the mine at each point in time, and the rate of extraction of the underlying mineral. We do not discuss in this summary the advantages of using real options instead of discounted net present value. This debate has actually little to do with our contributions. Any valuation methodology that takes advantage of probabilistic modeling is subject to the type of model error that arises due to lack of data and which is precisely what we are trying to quantify.

The risk factor $X = (Y, Z)$ is split into two components, market risk (Y) and private risk (Z), which are assumed to be independent. The contribution of Z, which corresponds to financial risks, is calibrated under P_0 using market information (such as options, for instance, in the case of the underlying mineral - e.g. gold) and we do not apply distributional uncertainty over Z.

The contribution of the private risk is modeled according to a Poisson process with intensity λ under P_0, where the intensity is expressed in annual basis and it is interpreted as the probability of a substantial disaster with devastating consequences for the mine. We assume that $\lambda = .01$ (i.e. the chance of a catastrophic event is about $1/100$).

Next, we provide a mechanism to robustify the value of the mine using our distributionally robust extreme value theory as we explain next.

We assume that the mine in question builds a tailing dam which is tall and strong enough to withstand extreme precipitation events for over 100 yrs. We then consider the historical precipitation recorded in the geographical region where the mine is located. We use standard EVA to compute the quantile x_p such that the maximum precipitation over 100 yrs is guaranteed to be less than x_p with 95% confidence. Then, we apply our DR extreme value analysis described in Section 4, thereby obtaining a DR tail distribution $\bar{F}_{\alpha,\delta}(\cdot)$. Then we let $\bar{\lambda} = \bar{F}_{\alpha,\delta}(x_p)$ and we evaluate

$$v_\delta^- = \max_{\theta \in \Theta} \bar{E}_0 \left(u \left(Y, Z, \theta \right) \right),$$

where the probability model $\bar{P}_0$ retains the same distribution of Y calibrated from market specifications to obtain P_0, but Z is now a Poisson process with intensity $\bar{\lambda}$, and both Y and Z remain independent.

This valuation procedure is equivalent to a pricing procedure of the form

$$v_\delta^- = \max_{\theta \in \Theta} \min_{\mathcal{U}_\delta(P_0)} \bar{E}_0 \left(u \left(Y, Z, \theta \right) \right),$$

where the robustification is only applied to Z.

In [5] there is another approach which is used to calibrate δ, using the relative entropy as the underlying discrepancy criterion. Such an approach depends on the existence of a basket of mines which are assumed to be well valued by a group of analysts. This basket

is used to calibrate the value of δ, given the specification of a baseline model, P_0. In simple words, we obtain the value of every mine in the basket applying the real options pricing methodology given model P_0, then we robustify changing δ in order to match the valuation provided by the analyst. This is repeated for every company in the basket, thus generating a set of δ's. We can then sort these values of δ and select the 95% quantile, for example. Such value would provide a choice of δ which is informed by expert opinion.

The paper [5] applies this methodology using a group of mines, showing reasonable empirical performance.

6. The Problem with Entropy: Optimal Transport as a Notion of Model Uncertainty

The problem with using entropy-type notions to describe the uncertainty set $\mathcal{U}_\delta(P_0)$ arises in the context of risk quantification, as opposed to valuation. In the context of valuation, the use of relative entropy can be motivated by invoking the fundamental theorem of asset pricing, which in particular implies that, to avoid arbitrage, the likelihood ratio between the physical model and the risk neutral model must be well defined. This necessary condition is very weak, but it at least motivates using relative entropy as a notion of discrepancy. In particular, if P_0 is chosen to satisfy such a weak necessary condition, then the correct risk neutral specification will eventually make it into the uncertainty set by choosing δ sufficiently large.

The use of entropy-type notions for the evaluation of model error impact in risk quantification cannot be motivated from the standpoint of the existence of a likelihood ratio. Risk quantification is performed using the physical probability model, so this type of calculation is more statistical in nature. From the standpoint of risk analysis one might specify a probabilistic description based, for example, on purely empirical data (i.e. fully non-parametric).[3]

In fact, this type of approach is often used in the estimation of Value-at-Risk using historical simulation. Clearly, an empirical distribution (which is supported in finitely many points) is not equivalent (in the mathematical sense described earlier) to a true underlying distribution which is continuous.

So, given that risk quantification is more of a statistical estimation problem (as opposed to a calibration problem, which is the case in valuation), familiar notions such as overfitting and out-of-sample performance are important to keep in mind. If P_0 corresponds to the empirical distribution induced by, say, n observations, then $D(P||P_0) \leq \delta < \infty$ implies that P must be supported on those n samples. Therefore, if we insist in using entropy notions to quantify discrepancy, the max-min approach that we advocate would

[3]Actually, even in the context of valuation, the fundamental theorem of asset pricing, becomes more technical in the case of an economy with uncountably many possible random outcomes. In that case, there are many reasonable ways in which one might interprete an arbitrage free market and these interpretations might lead to risk neutral models which are not necessarily equivalent to the underlying physical model.

not induce strong out-of-sample performance of the risk estimates (because every perturbation allowed must remain the sample intact, just the weights associated to each data point are allowed to be changed by the adversary in our game formulation).

In order to cope with this problem, we consider a different notion of discrepancy called "optimal transport" discrepancy. The optimal transport discrepancy between two probability models, P and P_0, depends on a cost function $c(\cdot)$ (to be discussed momentarily) and it is denoted by $D_c(P, P_0)$. The cost function is evaluated at a pair (x, y), $c(x, y)$, represents the cost of transporting a unit of mass from position x to position y. For example, one might consider, $c(x, y) = \|x - y\|$, the Euclidean distance between x and y. In general, it is only assumed that $c(x, x) = 0$, that $c(x, y) \geq 0$ and that $c(\cdot)$ is continuous (although this condition can be relaxed).

In simple words, $D_c(P, P_0)$ is the minimum cost of transporting the mass described by the histogram represented by P into the histogram represented by P_0. So, for example, if $\sum_{i=1}^{m} P(x_i) = 1$ and $\sum_{j=1}^{n} P_0(y_i) = 1$ for sets of points $\{x_1, ..., x_m\} \subset R^d$ and $\{y_1, ..., y_n\} \subset R^d$ then $D_c(P, P_0)$ is computed by solving the linear program,

$$
\begin{aligned}
D_c(P, P_0) &= \min \sum_{i,j} \pi_{i,j} c(x_i, y_j) \\
s.t. \ \sum_j \pi_{i,j} &= P(x_i) \text{ for all } i = 1, ..., m \\
\sum_i \pi_{i,j} &= P_0(y_j) \text{ for all } j = 1, ..., n \\
\pi_{i,j} &\geq 0 \text{ for all } i, j.
\end{aligned}
$$

The optimal solution, say $\left\{\pi^*_{i,j} : 1 \leq i \leq m, 1 \leq j \leq n\right\}$ represents a joint distribution which in particular has as its marginales P and P_0, respectively. The optimal solution to the previous linear program always exists because the feasible region is compact and non-empty (note that $\pi_{i,j} = P(x_i) P_0(y_j)$ is feasible).

The definition is completely analogous for arbitrary distribution functions P and P_0 (even distributions describing random elements infinite dimensions, such as random functions like Brownian motion). In the general case, the summations become integrals, but the resulting linear program has infinitely many variables and, therefore, the associated duality theory is much more complicated than in the standard finite dimensional linear programming. But this theory is well understood in the literature on optimal mass transportation, as explained in [3].

In order to gain intuition for the use of optimal transport costs, Figure 1 idealizes the optimal transport discrepancy between two densities depicted by the curves μ and v. Density μ represents a pile of sand, normalized to contain a total mass equal to one (say 1 ton), and v represents a sinkhole, which can be completely covered by a 1 ton of sand.

If the solution to the optimization problem defining $D_c(\mu, v)$ is a joint distribution π^* such that $\pi^*(y|X = x)$ (with X following distribution μ and Y following distribution

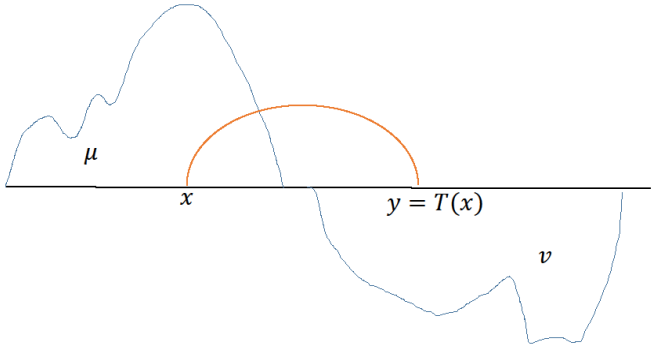

FIGURE 1. Idealization of optimal transport and an optimal mass transportation strategy.

v) is concentrated at a single point, we may write $y = T(x)$ (the mapping $T(\cdot)$ is implicitly given by the conditional distribution $\pi^*(y|X = x)$). In such case, the cheapest way to transport the mass described by the mountain μ to cover the sinkhole profiled by v is to move all the mass in x to position $y = T(x)$, and the total transportation cost $\int \mu(x)\, c(x, T(x))\, dx$ is therefore minimized.

As mentioned in Section 3, one of the main reasons for choosing relative entropy as tool for describing distributionally robust uncertainty sets is tractability. So, to enable the use of optimal transport in defining distributional uncertainty sets, we need to provide a tractable way for computing quantities such as

$$\sup_{P:D(P,P_0)\leq\delta} E_P[f(X)]. \tag{3}$$

The paper [3] provides a complete duality theory for the evaluation of quantities such as (3). In particular, it is shown in (3) that under mild conditions on $f(\cdot)$ (in particular if $f(\cdot)$ is continuous and $E_{P_0}|f(X)| < \infty$, then

$$\sup_{P:D(P,P_0)\leq\delta} E_P[f(X)] = \inf_{\lambda\geq 0}(\lambda\delta + E_{P_0}[\sup_y\{f(y) - \lambda c(X, y)\}]).$$

For example, if $f(X) = I(X \in A)$, so that $P(X \in A)$, and $c_A(x) = \inf\{C(x, y) : y \in A\}$ then

$$\sup_{P:D(P,P_0)\leq\delta} E_P[f(X)] = P_0(c_A(X) \leq 1/\lambda_\delta),$$

where $\lambda_\delta > 0$ is a Lagrange multiplying and if $c_A(X)$ has a continuous distribution and $E_{P_0}(c_A(X)) > \delta$, then λ_δ is the unique solution to

$$E_{P_0}[c_A(X)\, I(c_A(X) \leq 1/\lambda_\delta)] = \delta.$$

This result is applied to various stylized risk evaluation settings in [3].

In summary, paper [3] *provides a fundamental theory for the systematic use of optimal transport discrepancies as a tool for distributionally robust optimization beyond the use of entropy discrepancies.*

7. Calibrating Optimal Transport in Statistical Problems

The goal of the paper [1] is to develop the statistical theory required to use optimal transport and DRO for statistical estimation (and, therefore, for risk estimation). For example, how to select δ in the definition of the uncertainty set $\mathcal{U}_\delta(P_0)$ from a statistical standpoint?

The paper also explains why DRO is actually a desirable approach by connecting it to well established and successful statistical procedures. In particular, we revisit standard statistical problems, such as linear regression and generalized linear models. For example, consider the estimation of a linear regression parameter, β_*, in a linear regression model of the form

$$Y_i = \beta_*^T X_i + \epsilon_i,$$

where the Y_is are responses, X_is are predictors and ϵ_is are errors. The standard estimation approach consists in minimizing the mean squared errors,

$$\min_\beta E_{P_n}^{1/2}\left[\left(Y-\beta^T X\right)^2\right] = \min_\beta \left(n^{-1}\sum_{i=1}^n \left(Y_i-\beta^T X\right)^2\right)^{1/2},$$

the notation E_{P_n} represents averages with respect to the empirical data (as shown in the right hand side of the previous display). In [1] we showed that if

$$c\left((x,y),(x',y')\right) = \begin{cases} \|x-x'\|_q & \text{if} \quad y=y' \\ \infty & \text{if} \quad y\neq y' \end{cases}, \tag{4}$$

with $\|x\|_q = \left(\sum_{i=1}^d x_i^q\right)^{1/q}$ for a vector $x=(x_1,...,x_d)$ and $q\geq 1$, then

$$\min_\beta \max_{P:D_c(P,P_n)\leq\delta} E_{P_n}^{1/2}\left[\left(Y-\beta^T X\right)^2\right] = \min_\beta \left[E_{P_n}^{1/2}\left[\left(Y-\beta^T X\right)^2\right] + \sqrt{\delta}\,\|\beta\|_p\right]. \tag{5}$$

This is result is quite significant because of three main reasons.

First, the right hand side is a celebrated statistical estimation procedure called sqrt-Lasso. In addition, analogous representations for many other machine learning estimators (such as regularized logistic regression and support vector machines) are also obtained in [1]. So, we are able to show that many successful estimators are just a particular case of the DRO framework using optimal transport as we propose. In addition, our representation provides insight on the interpretation of traditional regularization. For example, (5) shows that the use of direct regularization in the parameter β, as it is traditionally applied, amounts to assuming that there is no distributional uncertainty in the response variable. This is precisely the interpretation of an infinite cost associated to perturbing the responses as prescribed in the definition of $c(\cdot)$ in 4. It is important to emphasize

that representation (5) is the first exact representation of machine learning estimators using DRO. Our paper is among the very first that studies these types of representations, a discussion of the literature is given in [1].

We note that the definition of $c(\cdot)$ in 4 is given just to recover the form of sqrt-Lasso. We do not necessarily advocate such a choice, although it may seem reasonable in cases in which distributional uncertainty in the response are assumed to be the result exclusively of uncertainty in predictors. If this is the case it makes perfect sense to use 4, but this is rarely discussed in standard applications of regularized estimators.

Second, the penalty term of the form $\sqrt{\delta}\left\|\beta\right\|_p$, which appears in the representations that we derive explains, from a distributionally robust perspective, the role of regularization in machine learning. Regularization is a technique that is often used to recognize that a complex model will tend to overfit the data, so penalizing "the complexity" of a model using the norm of the parameter is a sensible approach to reduce overfitting. In contrast, the game-theoretic representation that we obtain shows that regularization arises as the result of recognizing that the statistician is selecting an estimator which is intended to perform well for models which are a perturbation of the empirical data. This novel view leads to a natural and data-driven way for choosing the regularization parameter, namely $\sqrt{\delta}$. This leads to the third reason which advocated for the significance of our approach, namely, optimally choosing δ.

The regularization parameter in machine learning estimators (i.e. the coefficient which multiplies the penalty term $\left\|\beta\right\|_p$) is typically chosen using cross validation. Basically, cross validation consists in splitting the data in training and testing sets, and the parameter δ is chosen to maximize the performance on the testing set. Cross validation is a time-consuming estimation strategy and, strictly speaking, if one wishes to avoid overfitting biases, one should have a substantial amount of data so that choosing the parameter δ is performed truly independently from the ultimate testing phase. So, cross validation, although traditionally used in practice, is suboptimal both in computational time and usage of the information. In fact, if not done property, cross validation might not even be consistent (see [15]).

We propose a different (optimal) approach for choosing the parameter δ, which we explain next. Owing to the min-max representation (5) we are able to define a sensible criterion for the selection of δ. The probabilistic model P_n is plausible, yet imperfect, representation of the true underlying probability model, which is unknown, and therefore there are many models which are also plausible representations of the data, those are precisely the ones conceptualized by the set

$$\mathcal{U}_\delta(P_n) = \{P : D_c(P, P_n) \leq \delta\}.$$

So, consider a thought experiment in which we use each $P \in \mathcal{U}_\delta(P_n)$ to compute the optimal choice of regression parameter estimate β_P such that

$$\beta_P = \text{argmin}\, E_P\left[\left(Y - \beta^T X\right)^2\right].$$

By convexity of the quadratic loss function, β_P is characterized by the equation

$$E_P\left[\left(Y-\beta_P^T X\right)X\right]=0.$$

Each such β_P is therefore a plausible estimate of β_* and, hence, the set

$$\Lambda_\delta(n)=\{\beta_P : P\in\mathcal{U}_\delta(P_n)\}$$

forms a confidence region for the parameter β_*. Note that $\Lambda_\delta(n)$ is a random set because it depends on the sample encoded via P_n. Such confidence region is increasing in δ because the set $\mathcal{U}_\delta(P_n)$ is increasing in δ and therefore, given a confidence level $1-\alpha$, say $1-\alpha=.95$, it is natural to choose δ solving the following optimization problem

$$\delta_n^*=\min\{\delta : P_*(\beta_*\in\Lambda_\delta(n))\geq 1-\alpha\},$$

P_* is the true probability distribution underlying the generation of the sample data encoded in P_n. Under the assumption the data is independent and identically distributed (an assumption that can be weakened to require only stationarity and weak dependence), in [1] we establish that $\delta_n^*\to 0$ and that $n\delta_n^*\to\chi_{1-\alpha}$, where $\chi_{1-\alpha}$ is the $(1-\alpha)$-quantile of an explicit distribution which can be calibrated directly from the testing data set.

In summary, the paper [1] *established the fundamental statistical theory for the use of DRO via optimal transport costs as a comprehensive statistical tool.*

8. Conclusions and Future Work

The diagram in Figure 2 summarizes the research activities that were pursued on this project. At the top of the diagram we explain the motivation for using the game-theoretic approach for valuation and risk quantification. As explained in the Introduction that this is justified by the severe lack of data available in mining applications. We then note that risk and valuation are different types of estimations problems, one of them, valuation is mostly a calibration problem and it motivates using a entropy-based specification of distributional uncertainty. There is a wealth of theory in economics, statistics, and robust optimal control which makes the use of entropy-based discrepancies relatively easy to adapt to applications, in this case, the mining setting. We adapt the study of entropy and divergence notions to develop a distributionally robust extreme value analysis approach for studying extreme quantiles in [2], and apply this approach to the context of distributionally robust valuation of mines using an extended real option valuation methodology in [5]. The papers [2] and [5] correspond to the bottom two items on the left of the diagram.

On the other hand, risk quantification is inherently a statistical problem and we argue that entropy is not necessarily a suitable approach for settings in which we are interested in quantifying out-of-sample performance (because entropy perturbations only affect the likelihood of empirical observations not their actual value). So, we consider optimal transport discrepancies, which are based on infinite dimensional linear programming problems. Because optimal transport is not a conventional approach in DRO, we needed

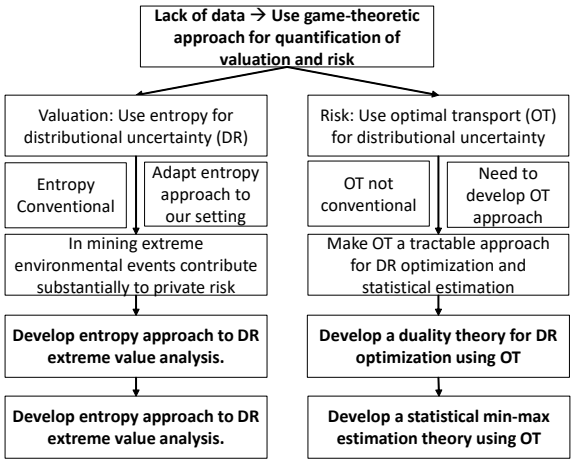

FIGURE 2. Diagram illustrating the motivation and development of our research.

to develop a duality theory which can be used to make optimal transport a tractable method in DRO; this is done in [3]. Then, we also needed to build a theory for the systematic use of optimal transport in statistical inference problems, this was done in [1]. Papers [3] and [1] correspond to the bottom two items on the right of the diagram.

The methodological framework that we have developed, based on game-theoretic considerations, can be applied to more general industries, beyond the mining setting considered.

Current work in progress and future research consists in applying the DRO framework to the mining industry. In this setting, it is important to adapt the theory developed to the setting of distributionally robust extreme value analysis for climate risk. The entropy-based theory is not well suited to study multidimensional extremes and, also, time stationarity is not easy to accommodate without resulting in pessimistic estimates. In contrast, the optimal transport approach is more flexible, but we first need to develop a framework for a sensible choice of the cost function $c(\cdot)$. Initial work in the context of using DRO based on optimal transport for risk valuation in mining applications is discussed in [4].

Recent work (see [14]), takes advantage of our developed theory, based on optimal transport, to train deep neural networks. The authors in (see [14]) that state-of-the-art neural networks are highly susceptible to adversarial attacks. They apply our DRO based on optimal transport to train networks and show that such an approach immunizes the network to false classifications due to a natural class of adversaries. The issue of training against adversarial attacks has been identified a major challenge in artificial intelligence. We believe that the applications of our approach in these types of settings is of great interest and we intend to pursue this line of research in the future.

Acknowledgement: We thank the Norges Bank for the generous support provided for this project.

REFERENCES

[1] Blanchet, J., Kang, Y., and Murthy, K. (2016) Robust Wasserstein profile inference and applications to machine learning. https://arxiv.org/abs/1610.05627.

[2] Blanchet, J. and Murthy, K. (2016) Distributionally robust extreme value theory. https://arxiv.org/abs/1601.06858.

[3] Blanchet, J. and Murthy, K. (2016) Quantifying distributional model risk via optimal transport https://arxiv.org/abs/1604.01446.

[4] Dolan, C. (2017) Distributionally Robust Performance Analysis with Applications to Mine Valuation and Risk. Ph.D. Dissertation in Statistics, Columbia University, NYC, US.

[5] Dolan, C., Blanchet, J., Iyengar, G., and Lall, U. (2017) A model robust real options valuation methodology incorporating climate risk, *submitted*.

[6] Duffie, D. (2001) *Dynamic Asset Pricing Theory*, Princeton University Press, Princeton, NJ, US.

[7] Dupuis P, James MR, Petersen I (2000) Robust properties of risk-sensitive control. *Mathematics of Control, Signals and Systems*, 13, 318-332.

[8] Durrett, R. (2011) *Probability: Theory and Examples*. Cambridge University Press, UK.

[9] Ellsberg, D. (1961) Risk, ambiguity and the Savage axioms, Quarterly *Journal of Economics*, 75, 643-669,

[10] Gilboa, I. and Schmeidler, D. (1989) Maxmin expected utility with non-unique prior, *Journal of Mathematical Economics*, 18, 141-153.

[11] Hansen, L. and Sargent, T. (2008) *Robustness*. Princeton University. Press, Princeton, NJ, US.

[12] Luenberger, D. (1997) Investment Science. Oxford University Press, UK.

[13] Savage, L. (1954) *The Foundations of Statistics*. Wiley, NY, US.

[14] Sinha, A., Namkoong, H., Duchi, J. (2017) Certifiable Distributional Robustness with Principled Adversarial Training. https://arxiv.org/abs/1710.10571.

[15] Yang, Y. Consistency of cross validation for comparing regression procedures, https://arxiv.org/abs/0803.2963.

[16] Wald, A. (1950) *Statistical Decision Functions*. Wiley, NY, US.

[17] https://www.fsb-tcfd.org/publications/final-recommendations-report/

COLUMBIA UNIVERSITY, DEPARTMENT OF INDUSTRIAL ENGINEERING & OPERATIONS RESEARCH AND DEPARTMENT OF STATISTICS. 500 W. 120 STREET, NEW YORK, NY 10027, UNITED STATES.
E-mail address: jose.blanchet@columbia.edu

COLUMBIA UNIVERSITY, DEPARTMENT OF EARTH AND ENVIRONMENTAL ENGINEERING AND DEPARTMENT OF CIVIL ENGINEERING AND ENGINEERING MECHANICS. MUDD BUILDING, 500 W. 120 STREET, NEW YORK, NY 10027, UNITED STATES.
E-mail address: ula2@columbia.edu