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

Data-driven distributionally robust risk parity portfolio optimization

Giorgio Costa · Roy H. Kwon

Received: date / Accepted: date

Abstract We propose a distributionally robust formulation of the traditional risk parity portfolio optimization problem. The probability distribution of the asset returns is considered to be discrete to align with data-driven parameter estimation methods. Typically, the nominal distribution assumes that all scenarios in the dataset are equally likely. However, we break away from this assumption and consider an ambiguity set that allows us to find the most adversarial probability distribution within some predetermined distance of the nominal distribution based on the investor's confidence. Thus, this distributionally robust optimization problem is formulated as a minimax problem. We exploit the risk parity formulation to attain a convex–concave minimax problem. Our approach is financially meaningful, as we attempt to be fully risk diverse with respect to the worst-case instance of the portfolio risk measure. We propose a novel algorithmic approach to solve this minimax problem, which blends projected gradient ascent with sequential convex programming. By design, this algorithm is highly flexible and allows the user to choose among several alternative statistical distance measures to define the ambiguity set. Moreover, the algorithm is highly tractable and scalable. Our numerical experiments suggest that a distributionally robust risk parity portfolio can yield a higher risk-adjusted rate of return when compared against the nominal portfolio.

Keywords Portfolio selection · Distributionally robust optimization · Risk parity · Statistical ambiguity · Saddle-point problem · Gradient descent

G. Costa · R. H. Kwon (✉)
Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, M5S 3G8, Canada
E-mail: {gcosta, rkwon}@mie.utoronto.ca

1 Introduction

Portfolio selection can be aptly presented as an optimal decision-making problem. Such problems have become prevalent in computational finance since the introduction of modern portfolio theory (MPT) by Markowitz [38]. MPT posits that a portfolio's financial reward is quantified by its rate of return, while financial risk is quantified by the portfolio's variance. However, these two parameters are typically unknown to an investor and must be estimated from observable data, leading to estimation errors. In turn, these errors may have a profound impact on the portfolio's ex post financial performance. In the context of computational finance, the sensitivity of portfolio optimization to errors in estimated parameters has been widely explored in the literature [7, 40, 11], leading to what is sometimes referred to as 'error maximization' given the poor out-of-sample performance of these (ex ante) optimal portfolios. In particular, Chopra and Ziemba [15] found that the impact of estimation errors in the expected returns can be an order of magnitude larger than estimation errors in the covariance matrix.

Accounting for uncertainty during optimization has become paramount in any decision-making problem where parameters are non-deterministic. If we assume we have complete knowledge of the underlying uncertainty (or a high degree of confidence in our estimate of the distribution), we can choose an appropriate probability distribution to represent the uncertainty of our parameters, leading to a collection of problems known as stochastic programs [8, 53]. Solving stochastic programs attempts to quantify the expectation of the uncertain parameters during optimization by solving multiple competing instances of the same problem under different scenarios. We then analyze the resulting collection of solutions to make a well-informed decision, typically addressing both optimality and stability [46, 47].

On the other hand, when we have no distributional knowledge (or if we do not have confidence in our estimates) of the uncertain parameters, we can ignore any distributional estimates and instead solve the worst-case instance of the problem to ensure we retain feasibility in our solution. This is the basis of robust optimization [4, 6, 3]. Robust optimization allows us to frame the problem deterministically by taking the most extreme estimates of our uncertain parameters within some distributional bounds when solving our optimization problem. Some examples of robust optimization applications in the context of portfolio optimization are presented in [56, 36, 29, 31, 24].

This manuscript is based on a class of problems that sits somewhere in-between stochastic programming and robust optimization. Such problems attempt to use distributional information during optimization, but accept that the underlying probability distribution is unknown. Instead, the distribution is said to lie within an ambiguity set of probability distributions. Similar to robust optimization, we then take a worst-case approach, but with the distinction that we do this at the distributional level. Such a robust formulation for stochastic programs was proposed by Scarf [51]. Since then, this class of problems has often been referred to as *minimax* problems or, more recently, as

distributionally robust optimization (DRO) problems [20]. A detailed survey paper on DRO is presented in [48].

The minimax problem has its roots in game theory [42]. In the context of this manuscript, we seek to minimize our cost function with respect to our decision variable, while the secondary player, i.e., 'nature', is adversarial and seeks to maximize our cost with respect to our uncertain parameters. Thus, our true goal is to minimize our cost within the decision space against the most adversarial instance of the underlying distribution of the uncertain parameters. Minimax problems have been widely studied in literature in both theory and applications [57, 22, 10, 54, 52]. We note that minimax problems are sometimes referred to as saddle-point problems [33, 50] due to the 'saddle' shape of the cost function when viewed in the higher-dimensional space created by the decision variable and the uncertain parameters. In particular, our manuscript focuses on the well-behaved subset of convex–concave saddle-point problems.

1.1 Motivation

MPT prescribes that the asset allocation problem can be solved as an optimal trade-off between risk and return, or, more specifically, between the mean and variance of the expected portfolio returns. This so-called mean–variance optimization (MVO) problem spurred the creation of many alternative quantitative asset allocation strategies. One such strategy that has gained traction in the last decade both in academia and industry is risk parity. Risk parity is sometimes referred to as 'equal risk contribution' since the objective is to construct a portfolio such that the risk contribution of each constituent towards the overall portfolio risk is equalized. In other words, each asset in the portfolio contributes the same level of risk.

The American investment management firm Bridgewater Associates is often cited as the first to pioneer such a portfolio in 1996, calling it an 'all-weather' fund because it was designed to be fully risk diverse in order to avoid the negative impacts associated with risk concentration. Qian [45] was the first to refer to this asset allocation strategy as 'risk parity', and this terminology has since then been widely adopted. The nominal risk parity optimization problem is based solely on the asset risk measure. Thus, by design, risk parity does not require the estimation of expected returns, which in turn reduces the most significant source of estimation error in portfolio optimization. Maillard et al. [37] carefully explain how to partition a portfolio's variance through Euler's homogeneous function theorem to calculate the risk contribution per asset. Roncalli [49] introduces estimated expected returns into the problem in order to construct portfolios where both the asset risk and return contributions are equalized. Moreover, Costa and Kwon [16] propose a generalized framework that embeds some of the more desirable properties of MVO into risk parity. Finally, Costa and Kwon [18] propose a robust risk parity framework that takes the worst-case estimate of the risk measure but do not consider any distributional information. Our manuscript focuses on designing a distri-

butionally robust counterpart of the nominal risk parity problem, where the nominal problem is the convex optimization problem proposed by Bai et al. [1].

Distributional robustness can be introduced into portfolio optimization problems by targeting the first two moments of the probability distribution governing the random asset returns, namely the expected returns and covariance matrix. In particular, the problem of moment uncertainty in portfolio optimization is addressed in [28, 33, 20].

The problem of optimization under moment uncertainty typically requires that the asset random returns are modelled using continuous distributions. However, our paper departs from this condition and instead focuses on discrete probability distributions. The use of discrete distributions stems from a purely practical necessity rather than a desired assumption. In other words, we do not mean to suggest that asset returns are, in fact, discrete. Instead, we acknowledge that the true latent distribution is continuous, but a parametric optimization model, like MVO or risk parity, necessitates the estimation of these parameters. In turn, a data-driven estimation process is based on a finite sample size that assumes each scenario is equally likely. Thus, our DRO model targets the scenario-based estimation process by breaking the assumption that all scenarios are equally likely and instead attempts to find the most adversarial discrete distribution within the bounds of some measure of statistical distance. This modelling framework is similar to the one proposed in [13] and, more recently, in [44]. In particular, many of our findings and contributions build upon the seminal work by Calafiore and El Ghaoui [13] in the area of portfolio optimization. Nevertheless, as we will discuss in Section 1.2, our contribution focuses on risk parity and carefully elaborates on the application of multiple alternative statistical distance measures.

The deviation of our adversarial discrete probability distribution from some nominal estimated distribution is bounded by a user-defined statistical distance measure. Therefore, the choice of distance measure is what defines our distributional ambiguity set, which in turn defines our distributional robustness. Given the conditions of our problem, we are limited to statistical measures for discrete probability. Calafiore and El Ghaoui [13] proposed a DRO problem using the Kullback–Leibler (KL) divergence, while Postek et al. [44] explored a comprehensive list of different measures, including any generic f-divergence. Our manuscript focuses on a subset of statistical measures that satisfy two conditions: the measure must be a formal distance metric[1] with finite bounds, and we must be able to formulate it as a computationally-tractable convex function.

[1] A metric or 'distance function' must be non-negative and satisfy the axioms of symmetry, identity of indiscernibles, and the triangle inequality.

1.2 Contribution

The objective of this paper is to introduce a distributionally robust risk parity (DRRP) portfolio optimization problem with a discrete probability ambiguity set on the portfolio risk measure. We emphasize the use of discrete probability distributions because they align naturally with data-driven parameter estimation procedures. Typically, each scenario from a dataset is assumed to be equally likely. We seek to break this assumption and instead design a maximization problem to find the most adversarial probability distribution implied by data such that we attain the worst-case instance of the portfolio risk measure.

In the context of risk parity, we have a minimax problem that seeks to equalize the asset risk contributions against the worst-case instance of the portfolio risk measure. The distributional ambiguity is defined by bounded statistical metrics, which allow us to embed the investor's confidence level as a design parameter (i.e., we use the investor's confidence level to limit the distance between the 'equally likely' nominal distribution and the adversarial distribution). Specifically, we discuss the following three measures of statistical distance: the Jensen–Shannon (JS) divergence, Hellinger distance, and total variation (TV) distance. However, we note that our framework extends naturally to any finite statistical distance measure that can be represented as a convex function.

The risk parity framework discussed in this manuscript uses the portfolio variance as the risk measure, modelled as a convex minimization problem. In turn, the corresponding DRRP problem is a convex–concave minimax problem, where we seek to maximize our objective by finding the most adversarial instance of a discrete probability distribution. Our asset allocation variable is pragmatically constrained by the set of admissible portfolios, while the adversarial distribution is fundamentally constrained both by the axioms of probability and by the measure of distance from the nominal distribution. The result is a constrained minimax problem, which can be solved via projected gradient method [5]. Specifically, we propose a projected gradient descent–ascent (PGDA) algorithm that alternates between the descent and ascent steps to reach the saddle point. The projection onto the probability ambiguity set grants the user a high degree of flexibility to choose their preferred statistical distance measure. We note that alternative methods to solve a constrained optimization problem could be used instead, such as the Frank–Wolfe method [25] or a Lagrangian approach. However, for the sake of brevity, our algorithmic development focuses primarily on projection methods.

We proceed to introduce a novel algorithm to solve the DRRP minimax problem that is grounded in projected gradient descent and sequential convex programming. The previous PGDA algorithm requires that we take alternating descent and ascent steps as we move towards the saddle point of our problem. In turn, this doubles the number of design parameters that must be set by the user a priori (e.g., initial point, step sizes). Moreover, this iterative process in two directions increases the possibility of numerical divergence. Instead, we

exploit the existence of a unique optimal risk parity portfolio for any given discrete probability distribution. Thus, our proposed algorithm operates by iteratively ascending in the direction of the gradient within the probability space while solving a risk parity minimization problem in the asset weight space. Conversely, we can interpret this algorithm as a sequential convex programming (SCP) method where we improve our estimate of the portfolio variance after every iteration through projected gradient ascent (PGA). Thus, we refer to this algorithm as SCP–PGA. Compared to the previous PGDA algorithm, each iteration is computationally more expensive. However, the exactness of each step translates to significantly fewer iterations until convergence. Additionally, we will see that the structure of the problem, combined with modern optimization software packages, allows for a computationally tractable and scalable implementation. Finally, from the user's perspective, this algorithm is conceptually easier to understand and implement: we iteratively ascend in the probability space while maintaining the risk parity condition in the asset weight space.

Numerical experiments show that our SCP–PGA algorithm is computationally efficient and scales well for problems with a large number of assets and scenarios. Moreover, the in-sample experiments show that the DRRP problem behaves as expected, while the out-of-sample experiments demonstrate good ex post performance. Financially, the DRRP portfolio is able to attain a higher risk-adjusted rate of return when compared to the nominal risk parity portfolio.

1.3 Outline

The outline of this paper is the following. Section 2 introduces the preliminaries that serve as a foundation for the development of this paper. These preliminaries include an introduction the risk parity asset allocation problem and DRO. Our main contribution is presented in Section 3, where we propose a DRRP problem and present two alternative gradient-based algorithms to find the optimal risk parity portfolio. The corresponding numerical experiments are shown in Section 4, which evaluate the proposed model's computational tractability, as well as its in-sample and out-of-sample financial performance. Finally, Section 5 summarizes the findings and contribution of this paper.

1.4 Notation

A general description of the notation used in this paper follows. We denote a real space of dimension n by $\mathbb{R}^n$ and the corresponding non-negative orthant by $\mathbb{R}^n_+$. Moreover, symmetric matrices of dimension n with real-valued elements are denoted by $\mathbb{S}^n$, while the subset of positive semi-definite (PSD) matrices are denoted by $\mathbb{S}^n_+$. Bold lowercase letters refer to vectors, while bold uppercase letters refer to matrices. A bold number refers to a vector of appropriate

dimension (e.g., $\mathbf{1}$ refers to a vector where each element is equal to one). This 'appropriate dimension' is based on the context of some reference vector. For example, if we have $\mathbf{1}^\top \boldsymbol{z} = b$, then the vector $\mathbf{1}$ has the same dimension as the vector $\boldsymbol{z}$. Moreover, as shown in the previous example, the superscript '$\top$' indicates the transpose of a vector or matrix. If we need to reference some specific element i within a vector, we do this by using the subscript i. For example, if we wish to reference the i^{th} element of vector $\boldsymbol{z}$, we denote this as z_i. In certain circumstances, if we define a vector as the product between a matrix and a vector, and we wish to reference its i^{th} element, then we use square brackets and the subscript i to do so. For example, if we have some matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and some vector $\boldsymbol{z} \in \mathbb{R}^n$, and the resulting product between them is the vector $\boldsymbol{Az} \in \mathbb{R}^m$, then its i^{th} element is denoted as $[\boldsymbol{Az}]_i$. The ℓ_p-norm of an arbitrary vector $\boldsymbol{v} \in \mathbb{R}^n$ is denoted by $\|\cdot\|_p$, where $\|\boldsymbol{v}\|_p \triangleq \left(\sum_{i=1}^n |v_i|^p\right)^{1/p}$. As exemplified in the previous sentence, we also use the notation '$\triangleq$' to mean 'equal by definition'. Some definitions will be expressed as functions of the form $\boldsymbol{a}(\boldsymbol{b}, \boldsymbol{c}) \in \mathbb{R}^n$, where, in this example, $\boldsymbol{b} \in \mathbb{R}^l$ and $\boldsymbol{c} \in \mathbb{R}^m$ are arguments and $\boldsymbol{a} : \mathbb{R}^l \times \mathbb{R}^m \to \mathbb{R}^n$. Note that neither the arguments nor the output need to be vectors, and this will be indicated by their bold font type when appropriate. Moreover, functions may serve to denote a specific parameter, such as the previous example that defined some parameter $\boldsymbol{a}(\cdot) \in \mathbb{R}^n$. Sets are denoted by uppercase letters using a calligraphic font (e.g., $\mathcal{X}$). Additionally, when relational operators are used in the context of vectors, they are applied in an element-wise fashion (e.g., $\boldsymbol{z} \geq \mathbf{0}$ implies that every element of the vector $\boldsymbol{z}$ is greater than or equal to zero). The expectation operator is denoted as $\mathbb{E}[\cdot]$. In the case where multiple variables are specified within the expectation operator, we indicate which variable the operator is acting upon with a corresponding subscript (e.g., $\mathbb{E}_{\boldsymbol{\zeta}}[f(\boldsymbol{z}, \boldsymbol{\zeta})]$ for some variables $\boldsymbol{z}$ and $\boldsymbol{\zeta}$ and some function $f(\boldsymbol{z}, \boldsymbol{\zeta})$). Finally, we denote the optimal solution to an optimization problem with an asterisk superscript (e.g., $\boldsymbol{z}^*$ is the optimal solution corresponding to some decision variable $\boldsymbol{z}$).

2 Preliminaries

2.1 Estimation of parameters

We begin by discussing the measures of financial reward and risk that will be used in this manuscript. As defined in MPT [38], the reward is measured by the portfolio rate of return (or simply the 'return'). The portfolio return is a weighted linear combination of the returns of the n assets that constitute the portfolio. The asset returns are random variables which we define as the vector $\boldsymbol{\xi} \in \mathbb{R}^n$. These random returns are governed by some probability density function $f_{\boldsymbol{\xi}}$ over some support $\mathcal{S}$ and with first and second moments defined as the expected returns $\boldsymbol{\mu} \in \mathbb{R}^n$ and the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{S}_+^n$, respectively. It follows that, at the asset-level, the financial reward is measured by $\boldsymbol{\mu}$ and

the financial risk by $\boldsymbol{\Sigma}$. In particular, the first two moments of $f_{\boldsymbol{\xi}}$ are assumed to be latent and are typically estimated from data, meaning they are prone to suffer from estimation error [15, 7, 40].

We define a portfolio as a vector of asset weights $\boldsymbol{x} \in \mathbb{R}^n$, where x_i represents the proportion of wealth invested in asset i. From an asset management perspective, $\boldsymbol{x}$ is our vector of decision variables that represents our asset allocation strategy. Thus, at the portfolio-level, the portfolio random return is defined as $\pi(\boldsymbol{x}) = \boldsymbol{\xi}^\top \boldsymbol{x}$. The corresponding measures of financial reward and risk are

$$\mu_\pi(\boldsymbol{x}) \triangleq \boldsymbol{\mu}^\top \boldsymbol{x}, \tag{1}$$

$$\sigma_\pi^2(\boldsymbol{x}) \triangleq \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}, \tag{2}$$

where the portfolio expected return is $\mu_\pi \in \mathbb{R}$ while the portfolio variance is $\sigma_\pi^2 \in \mathbb{R}_+$. The portfolio $\boldsymbol{x}$ is generally constrained by the set of admissible portfolios $\mathcal{X}$, which, in our case, disallows short sales and imposes a unit budget constraint. Short sales are prohibited due to a fundamental limitation of risk parity, which we discuss in greater detail in Section 2.2. It follows that the set of admissible portfolios is the following simplex

$$\mathcal{X} \triangleq \left\{ \boldsymbol{x} \in \mathbb{R}_+^n : \mathbf{1}^\top \boldsymbol{x} = 1 \right\}. \tag{3}$$

Restricting $\boldsymbol{x}$ to the non-negative orthant of the real n-dimensional space serves to disallow short sales. The equality constraint in $\mathcal{X}$ is necessary to ensure that the entirety of our available budget is invested in the assets.

The first two moments of our asset returns, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, are typically estimated from data (e.g., historical scenarios of asset returns). Assume our asset return data consist of T discrete scenarios for n assets (i.e., we have $\hat{\boldsymbol{\xi}} \in \mathbb{R}^{n \times T}$ scenarios and these scenarios suffice to appropriately represent the possible outcomes of the random variable $\boldsymbol{\xi}$). Next, assume there exists some probability p_t associated with each scenario t. In vector notation, this is the probability mass function (PMF) $\boldsymbol{p} \in \mathcal{P}$, where

$$\mathcal{P} \triangleq \left\{ \boldsymbol{p} \in \mathbb{R}_+^T : \mathbf{1}^\top \boldsymbol{p} = 1 \right\} \tag{4}$$

is the simplex defined by the axioms of probability. We note that the true distribution $f_{\boldsymbol{\xi}}$ is latent and typically assumed to be continuous. However, if we use a scenario-based estimation method we implicitly approximate such a distribution through a discrete counterpart, $\boldsymbol{p}_t$ for our T scenarios. Thus, we shift our focus towards discrete probability distributions not because we believe the distribution of the random vector $\boldsymbol{\xi}$ is discrete, but rather because this aligns well with our data-driven framework.

If we have knowledge of $\boldsymbol{p}$, then we can statistically derive the estimated parameters corresponding to the first two moments of the joint probability distribution of asset returns. Let $\hat{\boldsymbol{\xi}}_t \in \mathbb{R}^n$ be the t^{th} scenario of the dataset $\hat{\boldsymbol{\xi}}$.

The first two moments of the asset returns distribution are

$$\hat{\boldsymbol{\mu}}(\boldsymbol{p}) \triangleq \mathbb{E}[\boldsymbol{\xi}] = \sum_{t=1}^{T} p_t \cdot \hat{\boldsymbol{\xi}}_t, \tag{5}$$

$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{p}) \triangleq \mathbb{E}\big[\big(\boldsymbol{\xi} - \hat{\boldsymbol{\mu}}(\boldsymbol{p})\big)^2\big] = \sum_{t=1}^{T} p_t \cdot \big(\hat{\boldsymbol{\xi}}_t - \hat{\boldsymbol{\mu}}(\boldsymbol{p})\big)\big(\hat{\boldsymbol{\xi}}_t - \hat{\boldsymbol{\mu}}(\boldsymbol{p})\big)^\top, \tag{6}$$

where $\hat{\boldsymbol{\mu}} \in \mathbb{R}^n$ and $\hat{\boldsymbol{\Sigma}} \in \mathbb{S}_+^n$ are the data-driven estimates of the latent parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. Our estimates are shown as functions of some discrete probability distribution $\boldsymbol{p}$. If we assume each scenario is equally likely, then (5) and (6) are simply the standard sample arithmetic mean and sample covariance matrix[2] typically derived from data. Finally, we note that $\hat{\boldsymbol{\Sigma}}(\boldsymbol{p})$ in (6) is the result of the weighted sum of T rank-1 symmetric matrices, which means $\hat{\boldsymbol{\Sigma}}(\boldsymbol{p})$ is guaranteed to be a PSD matrix.

The estimated portfolio expected return and variance follow the same logic as (1) and (2), except we replace the latent parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with the estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ from (5) and (6). We break down the derivation as follows. Assume we have a portfolio $\boldsymbol{x}$. For a given dataset of asset returns $\hat{\boldsymbol{\xi}}$, the corresponding vector of portfolio return scenarios is $\hat{\boldsymbol{\pi}}(\boldsymbol{x}) = \hat{\boldsymbol{\xi}}^\top \boldsymbol{x}$. Thus, the estimated portfolio expected return $\hat{\mu}_\pi \in \mathbb{R}$ and variance $\hat{\sigma}_\pi^2 \in \mathbb{R}_+$ are

$$\hat{\mu}_\pi(\boldsymbol{x}, \boldsymbol{p}) \triangleq \boldsymbol{x}^\top \hat{\boldsymbol{\mu}}(\boldsymbol{p}) \tag{7a}$$

$$= \boldsymbol{p}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x}), \tag{7b}$$

$$\hat{\sigma}_\pi^2(\boldsymbol{x}, \boldsymbol{p}) \triangleq \boldsymbol{x}^\top \hat{\boldsymbol{\Sigma}}(\boldsymbol{p}) \boldsymbol{x} \tag{8a}$$

$$= \mathbb{E}\Big[\big(\pi(\boldsymbol{x}) - \mathbb{E}[\pi(\boldsymbol{x})]\big)^2\Big] = \mathbb{E}[\pi^2(\boldsymbol{x})] - \big(\mathbb{E}[\pi(\boldsymbol{x})]\big)^2$$

$$= \boldsymbol{p}^\top \hat{\boldsymbol{\pi}}^2(\boldsymbol{x}) - \boldsymbol{p}^\top \hat{\boldsymbol{\Theta}}(\boldsymbol{x}) \boldsymbol{p}, \tag{8b}$$

where $\hat{\boldsymbol{\pi}}^2(\boldsymbol{x}) \in \mathbb{R}_+^T$ denotes the element-wise square of the vector of portfolio return scenarios, and $\hat{\boldsymbol{\Theta}}(\boldsymbol{x}) \triangleq \hat{\boldsymbol{\pi}}(\boldsymbol{x}) \hat{\boldsymbol{\pi}}(\boldsymbol{x})^\top$. By definition, we have that $\hat{\boldsymbol{\Theta}}(\boldsymbol{y}) \in \mathbb{S}_+^T$ for any vector $\boldsymbol{y} \in \mathbb{R}^n$, meaning the portfolio variance in (8b) is concave over $\boldsymbol{p} \in \mathcal{P}$. Moreover, since $\hat{\boldsymbol{\Sigma}}(\boldsymbol{p}) \in \mathbb{S}_+^n$ for any probability distribution $\boldsymbol{p} \in \mathcal{P}$, the portfolio variance in (8a) is convex over $\boldsymbol{x} \in \mathcal{X}$. As we will see in Section 3, the convexity over $\boldsymbol{x} \in \mathcal{X}$ and concavity over $\boldsymbol{p} \in \mathcal{P}$ of the portfolio variance $\hat{\sigma}_\pi^2(\boldsymbol{x}, \boldsymbol{p})$ will allow us to formulate a convex–concave minimax problem.

[2]We note that $\hat{\boldsymbol{\Sigma}}(\boldsymbol{q})$, where $q_t = 1/T$ for $t = 1, \ldots, T$ is the 'equally likely' nominal distribution, yields the standard scenario-based estimate of the covariance matrix. However, if we wish to recover the standard *unbiased* estimate of the covariance matrix, we should multiply $\hat{\boldsymbol{\Sigma}}(\boldsymbol{q})$ by $T/(T-1)$. For the purpose of this paper, we will see that this distinction has no impact in the formulation of our distributionally robust optimization problem.

2.2 Risk parity

Risk parity is a modern asset allocation strategy that aims to construct a portfolio where every asset has the same level of risk contribution towards the overall portfolio risk. Conceptually, it is similar to the popular '$1/n$' portfolio, except that instead of fully diversifying our wealth distribution, we aim to fully diversify our risk distribution. By design, we are solely concerned with the portfolio risk, which in our case is measured as a function of the asset covariance matrix $\boldsymbol{\Sigma}$. Thus, the risk parity portfolio optimization problem does not require a reward measure, such as the estimated portfolio expected return.

As shown in [37], we can measure the individual risk contribution of each asset by applying Euler's homogeneous function theorem to partition the portfolio risk measure. Assume we have perfect knowledge of the distribution of the asset random returns (i.e., we have knowledge of the true covariance matrix $\boldsymbol{\Sigma}$). The portfolio standard deviation can be found by taking the square root of Equation (2). Applying Euler's theorem, the portfolio standard deviation can be partitioned as follows

$$\sigma_\pi = \sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}} = \sum_{i=1}^{n} x_i \frac{\partial \sigma_p}{\partial x_i} = \sum_{i=1}^{n} x_i \frac{[\boldsymbol{\Sigma}\boldsymbol{x}]_i}{\sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}}. \tag{9}$$

The latter part of (9) shows the partitions of the portfolio standard deviation for each asset $i = 1, \ldots, n$. Note that the denominator in this expression is consistent for all partitions, and it is equal to the portfolio standard deviation. As shown in [16], we can rearrange this expression such that we partition the portfolio variance instead. Thus, we can express the portfolio variance as the sum of n parts,

$$\sigma_\pi^2 = \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} = \sum_{i=1}^{n} x_i [\boldsymbol{\Sigma}\boldsymbol{x}]_i = \sum_{i=1}^{n} R_i, \tag{10}$$

where $R_i \triangleq x_i[\boldsymbol{\Sigma}\boldsymbol{x}]_i$ is the individual risk contribution of asset i. Now that we are able to measure the individual risk contributions, we can formulate an optimization problem to construct a risk parity portfolio such that $R_i = R_j \ \forall \ i, j$.

As prescribed by Bai et al. [1], we can design an unconstrained convex optimization problem that, at optimality, attains the desired risk parity condition. The problem is the following

$$\min_{\boldsymbol{y} \in \mathbb{R}_+^n} \quad \frac{1}{2} \boldsymbol{y}^\top \boldsymbol{\Sigma} \boldsymbol{y} - \kappa \sum_{i=1}^{n} \ln(y_i) \tag{11}$$

where $\kappa > 0$ is some arbitrary constant[3] and the auxiliary variable $\boldsymbol{y} \in \mathbb{R}_+^n$ serves as a placeholder for our asset weights. The auxiliary variable $\boldsymbol{y}$ will most

[3]We note that, in theory, we can assign any positive value to κ without loss of generality. For any value $\kappa > 0$, we will be able to find some optimal $\boldsymbol{y}^*$ such that we can recover our unique risk parity solution $\boldsymbol{x}^* \in \mathcal{X}$. In practice, it is preferable to avoid assigning extremely large or small values to κ to avoid numerical errors during optimization.

likely violate the set of admissible portfolios $\mathcal{X}$ given that we do not impose a budget equality constraint. Note that the first term in the objective function of (11) attempts to minimize the portfolio variance. However, the logarithmic barrier ensures that $y_i > 0 \; \forall \; i$ and leads to the following optimal solution

$$[\boldsymbol{\Sigma} \boldsymbol{y}^*]_i = \frac{\kappa}{y_i^*} \; \forall \; i \quad \iff \quad y_i^* [\boldsymbol{\Sigma} \boldsymbol{y}^*]_i = \kappa \; \forall \; i.$$

Thus, we can see that if the risk contribution from each component i is equal to κ, then they must all be equal to each other. Since Problem (11) does not impose the budget constraint, we cannot claim $\boldsymbol{y}^*$ is an admissible portfolio. However, we can recover the optimal risk parity portfolio $\boldsymbol{x}^*$ as follows

$$x_i^* = \frac{y_i^*}{\sum_{i=1}^n y_i^*}. \tag{12}$$

Traditionally, the risk parity asset allocation strategy restricts itself to 'long-only' portfolios where short sales are disallowed. This aligns well with restrictions typically observed in the asset management industry. However, this restriction stems from a fundamental limitation of risk parity portfolio optimization. As shown in (11), risk parity can be formulated as a strictly convex optimization problem with a unique global solution. Other equivalent convex formulations can be found in [37] and [39]. However, once short sales are allowed the problem becomes non-convex and the uniqueness of our solution is no longer guaranteed [1, 16]. For the sake of computational tractability, this paper restricts itself to the long-only condition imposed by traditional risk parity asset allocation strategies.

Thus far, we have assumed we have knowledge of the true (but latent) covariance matrix $\boldsymbol{\Sigma}$. In practice, we can use the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}(\boldsymbol{p})$ from (6), which corresponds to some discrete probability distribution $\boldsymbol{p}$. We can find the risk parity portfolio corresponding to an instance of $\boldsymbol{p} \in \mathcal{P}$ through the following system of equations,

$$f_{\text{RP}}(\boldsymbol{y}, \boldsymbol{p}) \triangleq \frac{1}{2} \boldsymbol{y}^\top \hat{\boldsymbol{\Sigma}}(\boldsymbol{p}) \boldsymbol{y} - \kappa \sum_{i=1}^n \ln(y_i), \tag{13a}$$

$$\boldsymbol{y}^*(\boldsymbol{p}) \triangleq \underset{\boldsymbol{y} \in \mathbb{R}_+^n}{\operatorname{argmin}} \; f_{\text{RP}}(\boldsymbol{y}, \boldsymbol{p}), \tag{13b}$$

$$\boldsymbol{x}^*(\boldsymbol{p}) \triangleq \varPi_{\mathcal{X}}\big(\boldsymbol{y}^*(\boldsymbol{p})\big), \tag{13c}$$

where $f_{\text{RP}} : \mathbb{R}_+^n \times \mathcal{P} \to \mathbb{R}$ is our risk parity objective function, while $\varPi_{\mathcal{X}}\big(\boldsymbol{y}^*(\boldsymbol{p})\big)$ is the projection of the vector $\boldsymbol{y}^*(\boldsymbol{p})$ onto the set of admissible portfolios $\mathcal{X}$. The logarithmic barrier in (13a) ensures that, at optimality, we converge to a non-negative optimal solution $\boldsymbol{y}^*(\boldsymbol{p})$. Therefore, the projection onto the set $\mathcal{X}$ only needs to enforce the budget equality constraint. For some arbitrary vector $\boldsymbol{z} \in \mathbb{R}_+^n$, the projection is

$$\varPi_{\mathcal{X}}(\boldsymbol{z}) \triangleq \frac{1}{\sum_{i=1}^n z_i} \boldsymbol{z}.$$

In other words, the projection step in (13c) is the vectorized equivalent of (12). We conclude this subsection by highlighting that we can use the optimization problem and projection in (13) to find the optimal risk parity portfolio for any estimate of the covariance matrix $\hat{\boldsymbol{\Sigma}}(\boldsymbol{p})$ with respect to an arbitrary instance of $\boldsymbol{p} \in \mathcal{P}$.

2.3 Distributionally robust optimization

Here we present a brief overview of DRO and its applications to the asset allocation problem. Our decision variable is the vector of asset weights, $\boldsymbol{x} \in \mathcal{X}$. Additionally, we have the vector of asset random returns $\boldsymbol{\xi} \in \mathbb{R}^n$. Let us momentarily depart from the notion of risk parity, and assume our objective is defined by some generic cost function $f(\boldsymbol{x}, \boldsymbol{\xi})$. We can define following the optimization problem to minimize the expected value of our cost function

$$\min_{\boldsymbol{x} \in \mathcal{X}} \quad \mathbb{E}_{\boldsymbol{\xi}}\big[f(\boldsymbol{x}, \boldsymbol{\xi})\big].$$

Given that this optimization problem involves a decision variable $\boldsymbol{x}$ and a random variable $\boldsymbol{\xi}$, this generic problem is a stochastic program. Note that the function $f(\boldsymbol{x}, \boldsymbol{\xi})$ may embed both measures of financial risk and reward. However, formulating our problem in this fashion assumes we have perfect knowledge of the probability density function $f_{\boldsymbol{\xi}}$, which is generally not true. Instead, we can introduce robustness to protect ourselves against plausible adversarial forms of $f_{\boldsymbol{\xi}}$.

We can do this by formulating a minimax stochastic program [51, 54]. By design, the minimax problem seeks to minimize our objective with respect to our decision variable $\boldsymbol{x}$ while simultaneously defining the most adversarial form of the objective with respect to $f_{\boldsymbol{\xi}}$,

$$\min_{\boldsymbol{x} \in \mathcal{X}} \max_{f_{\boldsymbol{\xi}} \in \mathcal{U}_f} \quad \mathbb{E}_{\boldsymbol{\xi}}\big[f(\boldsymbol{x}, \boldsymbol{\xi})\big] \tag{14}$$

where, in this example, $\mathcal{U}_f$ is a set of plausible probability measures. This type of minimax problem is sometimes defined as DRO [14, 43, 20].

Unlike the generic problem in (14), our problem focuses specifically on discrete probability distributions to align with traditional data-driven estimation procedures. Thus, for the purpose of this manuscript, we assume that we have some PMF $\boldsymbol{p} \in \mathcal{P}$ associated with scenarios that describe the finite set of possible outcomes of our asset returns $\boldsymbol{\xi}$. Therefore, we present the minimax problem as

$$\min_{\boldsymbol{x} \in \mathcal{X}} \max_{\boldsymbol{p} \in \mathcal{U}_{\boldsymbol{p}}} \quad \mathbb{E}_{\boldsymbol{\xi}}\big[f(\boldsymbol{x}, \boldsymbol{\xi})\big] \tag{15}$$

where $\mathcal{U}_{\boldsymbol{p}}$ is the ambiguity set of plausible discrete probability distributions. In addition to complying with the axioms of probability, the set $\mathcal{U}_{\boldsymbol{p}}$ includes the distributional ambiguity of $\boldsymbol{p}$. In a similar fashion to Calafiore [12], we define this ambiguity relative to some nominal distribution $\boldsymbol{q} \in \mathcal{P}$. For example, standard estimation procedures typically assume that every scenario in a dataset

is equally likely. In other words, we have that $q_t = 1/T$ for $t = 1, \ldots, T$. In this case, if we use $\boldsymbol{q}$ in (5) and (6), then the asset expected returns $\hat{\boldsymbol{\mu}}(\boldsymbol{q})$ is simply the arithmetic mean and the covariance matrix $\hat{\boldsymbol{\Sigma}}(\boldsymbol{q})$ is the standard scenario-based estimate. Thus, as we will see in Section 2.4 below, we can use a measure of statistical distance to limit the relative ambiguity between our nominal distribution $\boldsymbol{q}$ and its adversarial counterpart $\boldsymbol{p}$.

2.4 Statistical distance measures

Statistical distances serve to measure the level of similarity or discrepancy between two probability distributions. We limit our choice of statistical distance measures to a subset of convex functions that operate on discrete distributions. As we move forward into Section 3, this will allow us to retain computational tractability. In particular, the remainder of this section presents a non-exhaustive list of five different statistical distance measures that are applicable to our framework.

The first distance measure we present is the KL divergence [34], sometimes referred to as 'relative entropy'. The KL divergence is not a formal metric since it is not symmetric and does not respect the triangle inequality. Nevertheless, the KL divergence has become an increasingly popular tool to measure the difference between two probability distributions. For two discrete probability distributions $\boldsymbol{p},\ \boldsymbol{q} \in \mathcal{P}$, the KL divergence is defined as

$$D_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \sum_{t=1}^{T} p_t \ln\left(\frac{p_t}{q_t}\right). \tag{16}$$

The asymmetry of the KL divergence becomes apparent if we reverse the order of the arguments $\boldsymbol{p}$ and $\boldsymbol{q}$ (i.e., $D_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{q}) \neq D_{\mathrm{KL}}(\boldsymbol{q}, \boldsymbol{p})$). By definition, the KL divergence is always non-negative. However, the upper bound of the KL divergence is not properly defined, making it difficult to define an appropriate permissible distance between $\boldsymbol{p}$ and $\boldsymbol{q}$. In spite of these issues, the KL divergence is a useful tool to constrain the divergence between our nominal distribution $\boldsymbol{q}$ and our adversarial distribution $\boldsymbol{p}$. An example of its application to the asset allocation problem is shown in [12].

The KL divergence can be made symmetric by averaging its forward and reverse forms. We can define the symmetric KL (SKL) divergence as

$$D_{\mathrm{SKL}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \frac{1}{2} D_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{q}) + \frac{1}{2} D_{\mathrm{KL}}(\boldsymbol{q}, \boldsymbol{p}). \tag{17}$$

Although the SKL divergence is symmetric, it still does not respect the triangle inequality and cannot be defined as a true metric. Moreover, it lacks a properly defined upper bound.

A measure closely related to the KL divergence is the JS divergence, which was introduced by Lin [35]. Unlike the KL divergence, the JS divergence is

symmetric and has finite bounds. The JS divergence is defined as

$$D_{\mathrm{JS}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \frac{1}{2} D_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{m}) + \frac{1}{2} D_{\mathrm{KL}}(\boldsymbol{q}, \boldsymbol{m}). \tag{18}$$

where $\boldsymbol{m} = \frac{1}{2}(\boldsymbol{p} + \boldsymbol{q}) \in \mathcal{P}$. Given that our definition of the KL divergence in (16) uses the natural logarithm, our definition of the JS divergence has the useful property of being bounded between zero and ln(2), i.e.,

$$0 \leq D_{\mathrm{JS}}(\boldsymbol{p}, \boldsymbol{q}) \leq \ln(2).$$

Additionally, we can take the square root of the JS divergence to derive a formal distance metric known as the JS distance [23, 27] (i.e., the JS distance is $\sqrt{D_{\mathrm{JS}}(\boldsymbol{p}, \boldsymbol{q})}$). This distance measure is bounded between zero and $\sqrt{\ln(2)}$. As we will see in Section 3, the finite bounds on the JS distance will allow us to properly define the permissible distance d between the nominal distribution $\boldsymbol{q}$ and its adversarial counterpart $\boldsymbol{p}$.

Next, we present the Hellinger distance

$$D_{\mathrm{H}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \frac{1}{\sqrt{2}} \sqrt{\sum_{t=1}^{T} \left(\sqrt{p_t} - \sqrt{q_t}\right)^2}. \tag{19}$$

The Hellinger distance is a formal distance metric and, by its definition, is bounded between zero and one. As with the JS distance, this will allow us to define a permissible distance d between $\boldsymbol{q}$ and $\boldsymbol{p}$.

The last distance measure we discuss is the TV distance,

$$D_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \frac{1}{2} \|\boldsymbol{p} - \boldsymbol{q}\|_1, \tag{20}$$

which is also a formal distance metric. The TV distance is bounded between zero and one. As with the JS and Hellinger distances, its formal definition as a bounded metric will allow us to define a permissible distance d between $\boldsymbol{q}$ and $\boldsymbol{p}$.

3 Distributionally robust risk parity

This section presents our two main contributions: a data-driven DRRP portfolio optimization problem and the SCP–PGA algorithm to solve it. The 'data-driven' aspect arises from using a discrete probability distribution to assign different weights to our scenarios, emphasizing the scenarios that are more likely to adversely impact our estimate of the covariance matrix. As discussed in Section 2, we are not suggesting that the true distribution of asset returns is discrete. Instead, we choose to model the asset returns with a discrete probability distribution because it aligns well with traditional parameter estimation methods that would otherwise assume all scenarios are equally likely. Thus, we aim to assign an adversarial set of weights to each scenario such that we can derive the most robust set of parameters implied by the data themselves.

Consider the minimax problem from (15), and recall the set of plausible discrete probability distributions $\mathcal{U}_{\boldsymbol{p}}$. Our immediate goal is twofold: to design an appropriate ambiguity set $\mathcal{U}_{\boldsymbol{p}}$ for our adversarial probability distribution $\boldsymbol{p}$, and to recast the generic minimax problem from (15) into our DRRP problem. We address these two objectives in the following two subsections, before proceeding into the algorithmic development. Finally, we will conclude this section by discussing a variant of the risk parity problem where the investor can incorporate estimated expected returns into the optimization problem.

3.1 Probability distribution ambiguity set

Our adversarial probability distribution $\boldsymbol{p}$ belongs to the ambiguity set $\mathcal{U}_{\boldsymbol{p}}$, and we now proceed to formally define this set. A probability distribution must adhere to the simplex $\mathcal{P}$ defined by the axioms of probability. Moreover, our goal is to define an ambiguity set where the adversarial distribution $\boldsymbol{p}$ must lie within a maximum permissible distance d from the nominal distribution $\boldsymbol{q}$. Thus, the ambiguity set is

$$\mathcal{U}_{\boldsymbol{p}}(\boldsymbol{q}, d) \triangleq \big\{\boldsymbol{p} \in \mathcal{P} : \psi(\boldsymbol{p}, \boldsymbol{q}) \leq d\big\} \tag{21}$$

where $\psi(\boldsymbol{p}, \boldsymbol{q})$ is a convex function that describes any of the statistical distance measures defined in (16–20), while $d \in \mathbb{R}_+$ is a user-defined bound on the maximum distance between $\boldsymbol{p}$ and $\boldsymbol{q}$. We note that, by definition, $\mathcal{U}_{\boldsymbol{p}} \subseteq \mathcal{P}$.

For the purpose of computational tractability, our five statistical distance measures are restated or, where appropriate, redefined as follows,

$$\psi_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \sum_{t=1}^{T} p_t \ln\left(\frac{p_t}{q_t}\right), \tag{22a}$$

$$\psi_{\mathrm{SKL}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \frac{1}{2}\sum_{t=1}^{T} (p_t - q_t) \ln\left(\frac{p_t}{q_t}\right), \tag{22b}$$

$$\psi_{\mathrm{JS}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \frac{1}{2}\sum_{t=1}^{T} p_t \ln(p_t) + q_t \ln(q_t) - (p_t + q_t) \ln\left(\frac{p_t + q_t}{2}\right), \tag{22c}$$

$$\psi_{\mathrm{H}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \frac{1}{2}\sum_{t=1}^{T} p_t - 2\sqrt{p_t}\sqrt{q_t} + q_t, \tag{22d}$$

$$\psi_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \frac{1}{2}\|\boldsymbol{p} - \boldsymbol{q}\|_1. \tag{22e}$$

We make the distinction in notation between $\psi(\boldsymbol{p}, \boldsymbol{q})$ in (22) and $D(\boldsymbol{p}, \boldsymbol{q})$ in (16–20) because, in some cases, the original distance measure was not presented in a computationally-tractable form. In particular, we have simplified the expressions of the SKL and JS divergences, and we have squared the Hellinger distance. The distance measures in (22) are convex functions over $\boldsymbol{p} \in \mathcal{P}$ for any $\boldsymbol{q} \in \mathcal{P}$. In turn, this means the ambiguity set $\mathcal{U}_{\boldsymbol{p}}(\boldsymbol{q}, d)$ is convex.

Moreover, with the exception of (22d) and (22e), these definitions of distance can be easily implemented and solved by any modern non-linear optimization software package. The functions in (22d) and (22e) can be computationally implemented by introducing auxiliary variables during optimization, but this does not fundamentally alter the problem. An example of how to computationally implement them shown in Appendix A.

In Section 2.3, we defined the nominal probability distribution as $q_t = 1/T$ for $t = 1, \ldots, T$, which is merely a probabilistic representation of the assumption that each scenario in our dataset is equally likely. Our modelling framework provides sufficient flexibility for the user to prescribe their own choice of $\boldsymbol{q} \in \mathcal{P}$. However, given the data-driven nature of our manuscript, we formally reiterate our definition of the nominal probability distribution as a discrete uniform distribution, i.e.,

$$\boldsymbol{q} \triangleq \begin{bmatrix} 1/T \\ \vdots \\ 1/T \end{bmatrix} \in \mathbb{R}^T. \tag{23}$$

This falls in line with our goal to define the most adversarial distribution $\boldsymbol{p}$ relative to the distribution implied by the data.

To finalize the definition of $\mathcal{U}_{\boldsymbol{p}}$, we must determine the value of the maximum permissible distance d based on the investor's confidence level. Given that the KL divergence in (22a) and SKL divergence in (22b) are not formal statistical distance metrics, we will not discuss how to appropriately determine d for these two measures. However, our modelling framework will retain sufficient flexibility in case the user is able to specify an appropriate maximum permissible distance d for the KL or SKL divergences.

The remainder of this subsection describes how to define an appropriate value for d for the JS, Hellinger and TV measures. As discussed in Section 2.4, these three distance measures have theoretical lower and upper bounds. In particular, the upper bounds are only attainable if the nominal distribution $\boldsymbol{q}$ differs the most from our adversarial distribution $\boldsymbol{p}$. For a discrete probability distribution, this happens when both the nominal and adversarial distributions assign a probability $q_i = p_j = 1$ for scenarios $i \neq j$, with all other scenarios having a probability of zero. In practice, the upper bounds presented in Section 2.4 are unattainable under our the assumption that $\boldsymbol{q}$ is a discrete uniform distribution. Consider the following example of an extreme probability distribution

$$\boldsymbol{s} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^T,$$

which assigns all of its weight to a single scenario. The distribution $\boldsymbol{s}$ is the most we can differ from the uniform distribution $\boldsymbol{q}$. Thus, in practice, the true upper bound is defined as $B(T) \triangleq \psi(\boldsymbol{s}, \boldsymbol{q}) \in \mathbb{R}_+$, where the argument T

corresponds to the dimension of the fixed distributions $\boldsymbol{s}$ and $\boldsymbol{q}$. We define the practical upper bounds of our three distance measures as

$$B_{\mathrm{JS}}(T) = \psi_{\mathrm{JS}}(\boldsymbol{s}, \boldsymbol{q}), \tag{24a}$$
$$B_{\mathrm{H}}(T) = \psi_{\mathrm{H}}(\boldsymbol{s}, \boldsymbol{q}), \tag{24b}$$
$$B_{\mathrm{TV}}(T) = \psi_{\mathrm{TV}}(\boldsymbol{s}, \boldsymbol{q}). \tag{24c}$$

For example, if our data consist of ten scenarios, $T = 10$, then the upper bounds of the JS divergence in (22c), the squared Hellinger distance in (22d), and the TV distance in (22e) are

$$B_{\mathrm{JS}}(10) = \frac{1}{2}\Big((0.1)\ln(0.1) - (1.1)\ln(0.55) + (9)(0.1)\ln(2)\Big) \approx 0.5256,$$
$$B_{\mathrm{H}}(10) = \frac{1}{2}\Big(1 - 2\sqrt{0.1} + (10)(0.1)\Big) \approx 0.6838,$$
$$B_{\mathrm{TV}}(10) = \frac{1}{2}\Big(0.9 + (9)(0.1)\Big) = 0.9.$$

Furthermore, as T increases, the upper bounds approach their theoretical values (i.e., as $T \to \infty$, we have $B_{\mathrm{JS}} \to \ln(2)$, $B_{\mathrm{H}} \to 1$ and $B_{\mathrm{TV}} \to 1$). The purpose of this exercise is to avoid defining d in terms of a theoretical upper bound. Instead, we seek to define it relative to the number of scenarios in our dataset, which is more relevant in practice.

In Section 2.4 we saw that the square root of the JS divergence is a true distance metric. We can exploit this property to define an appropriate maximum permissible distance d_{JS} between our nominal and adversarial distributions based on the upper bound B_{JS} and the user-defined confidence level $0 \leq \omega \leq 1$. In turn, we can use this to constrain the statistical distance between $\boldsymbol{p}$ and $\boldsymbol{q}$ (i.e., $\psi_{\mathrm{JS}}(\boldsymbol{p}, \boldsymbol{q}) \leq d_{\mathrm{JS}}$). Recall that we must square the investor's confidence level since the JS divergence is the square of the JS distance. Thus, for a given confidence level ω and number of scenarios T, the maximum permissible distance is

$$d_{\mathrm{JS}}(\omega, T) \triangleq \omega^2 B_{\mathrm{JS}}(T). \tag{25}$$

Similarly, since $\psi_{\mathrm{H}}(\boldsymbol{p}, \boldsymbol{q})$ in (22d) is defined as the square of the Hellinger distance, the appropriate maximum permissible distance is

$$d_{\mathrm{H}}(\omega, T) \triangleq \omega^2 B_{\mathrm{H}}(T). \tag{26}$$

Finally, given that the TV distance in (22e) is already a true metric of statistical distance, we can define the maximum permissible distance as

$$d_{\mathrm{TV}}(\omega, T) \triangleq \omega B_{\mathrm{TV}}(T). \tag{27}$$

To properly define the ambiguity set $\mathcal{U}_{\boldsymbol{p}}$ in (21), the user must choose their preferred distance measure and define $\psi(\boldsymbol{p}, \boldsymbol{q})$ as one of the options in (22), while defining d accordingly. In particular, if the selected distance measure is either the JS divergence, Hellinger distance or TV distance, then d can be defined from the options in (25–27).

3.2 Minimax problem

For a given dataset $\hat{\boldsymbol{\xi}}$, our problem is defined by the investor's choice of statistical distance measure (e.g., JS or TV) and confidence level ω. Given this information, we aim to construct an optimal DRRP portfolio $\boldsymbol{x}^*$. The nominal risk parity problem in (13) is strictly convex for any given estimate of the covariance matrix $\hat{\boldsymbol{\Sigma}}(\boldsymbol{p})$. Therefore, as discussed in Section 2.2, there exists a unique risk parity portfolio $\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p})$ for every instance of $\boldsymbol{p} \in \mathcal{P}$.

The distinction between $\boldsymbol{x}^*$ and $\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p})$ is the following. The latter is the optimal risk parity portfolio for some arbitrary instance of $\boldsymbol{p} \in \mathcal{P}$, as shown in (13). On the other hand, $\boldsymbol{x}^*$ arises from the most adversarial instance of $\boldsymbol{p} \in \mathcal{U}_{\boldsymbol{p}}$ such that it maximizes our risk parity objective function. Thus, our optimal DRRP portfolio $\boldsymbol{x}^*$ can be formulated as a minimax problem where we seek an optimal portfolio against an optimally adversarial discrete probability distribution.

The risk parity problem in (13) requires that we first optimize an unconstrained problem and then project it onto the set of admissible portfolios. However, for simplicity, let us ignore the projection step and treat the unconstrained auxiliary variable $\boldsymbol{y}$ as a proxy[4] for our asset weights $\boldsymbol{x}$. Thus, for now, let our minimax variables be $\boldsymbol{y}$ and $\boldsymbol{p}$.

Recall our original definition of the portfolio variance, which was expressed in two equivalent forms in (8a) and (8b). Moreover, recall our original definition of the risk parity objective function $f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p})$ in (13a). Using both expressions of the portfolio variance, restate our risk parity objective function in two equivalent forms

$$f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p}) \triangleq \frac{1}{2}\boldsymbol{y}^\top \hat{\boldsymbol{\Sigma}}(\boldsymbol{p})\boldsymbol{y} - \kappa \sum_{i=1}^{n} \ln(y_i) \tag{28a}$$

$$\triangleq \frac{1}{2}\left(\boldsymbol{p}^\top \hat{\boldsymbol{\pi}}^2(\boldsymbol{y}) - \boldsymbol{p}^\top \hat{\boldsymbol{\Theta}}(\boldsymbol{y})\boldsymbol{p}\right) - \kappa \sum_{i=1}^{n} \ln(y_i), \tag{28b}$$

where (28a) is exactly the same as (13a) and is restated for clarity, while (28b) presents $f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p})$ explicitly in terms of $\boldsymbol{p}$. By formulating the objective function in these two equivalent forms allows us to observe how the function acts upon both the decision variable $\boldsymbol{y}$ and the adversarial probability $\boldsymbol{p}$.

The corresponding DRRP problem, stated as a minimax problem, is the following

$$\min_{\boldsymbol{y} \in \mathbb{R}^n_+} \max_{\boldsymbol{p} \in \mathcal{U}_{\boldsymbol{p}}} \quad f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p}). \tag{29}$$

As we saw in Section 2.1, both $\hat{\boldsymbol{\Sigma}}(\boldsymbol{p})$ and $\hat{\boldsymbol{\Theta}}(\boldsymbol{y})$ are PSD for any $\boldsymbol{p} \in \mathcal{P}$ and $\boldsymbol{y} \in \mathbb{R}^n_+$, respectively. Therefore, the function $f_{\mathrm{RP}}(\cdot, \boldsymbol{p}) : \mathbb{R}^n_+ \to \mathbb{R}$ is strictly convex for every $\boldsymbol{p} \in \mathcal{P}$, while $f_{\mathrm{RP}}(\boldsymbol{y}, \cdot) : \mathcal{P} \to \mathbb{R}$ is concave for every $\boldsymbol{y} \in \mathbb{R}^n_+$.

[4] Projecting the auxiliary variable $\boldsymbol{y}$ onto the set of of admissible portfolios such that we find a portfolio $\boldsymbol{x}$ is a trivial step, as shown in Equation (11).

Moreover, the sets $\mathcal{X}$ and $\mathcal{U}_{\boldsymbol{p}}$ are convex. This means we have a convex–concave DRRP problem, which means that any local optimum is a global optimum. In turn, this leads to the following observation

$$f_{\mathrm{RP}}(\boldsymbol{y}^*, \boldsymbol{p}) \leq f_{\mathrm{RP}}(\boldsymbol{y}^*, \boldsymbol{p}^*) \leq f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p}^*) \;\; \forall\, \boldsymbol{y} \in \mathbb{R}^n_+,\; \boldsymbol{p} \in \mathcal{U}_{\boldsymbol{p}}. \tag{30}$$

The maximization step in (29) is also meaningful in a financial context. Consider the second definition of $f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p})$ in (28b) where, without loss of generality, we have defined the portfolio variance using the unnormalized proxy variable $\boldsymbol{y}$. The maximization step in (29) pertains solely to the portfolio variance given that the logarithmic barrier term only acts on the variable $\boldsymbol{y}$. Thus, intuitively, the maximization step aims to find the most adversarial probability distribution $\boldsymbol{p}$ such that we attain the worst-case instance of the portfolio variance. This leads to the following conclusion: the minimax problem in (29) seeks the optimal risk parity portfolio with respect to the worst-case portfolio variance.

3.3 Projected gradient descent–ascent

Here we present and discuss the optimization algorithm that will be used to solve our DRRP problem. We can exploit the convex–concave structure of our minimax problem to attain global optimality through a gradient-based algorithm. In particular, we discuss a PGDA algorithm that sequentially alternates between descending in $\boldsymbol{y}$ and ascending in $\boldsymbol{p}$ until convergence.

To retain feasibility after each iteration, we project each step in $\boldsymbol{y}$ and $\boldsymbol{p}$ onto the sets $\mathbb{R}^n_+$ and $\mathcal{U}_{\boldsymbol{p}}$, respectively. In particular, the non-linearity of the statistical distance measure means that the projection onto the ambiguity set $\mathcal{U}_{\boldsymbol{p}}$ is non-trivial and cannot be solved in closed form. Instead, the projection must be solved as a constrained optimization problem. A Euclidean projection ensures the problem is strictly convex, guaranteeing the uniqueness of our solution. We define the projection of some arbitrary vector $\boldsymbol{u} \in \mathbb{R}^T$ onto the set $\mathcal{U}_{\boldsymbol{p}}$ as follows,

$$\begin{aligned}
\varPi_{\mathcal{U}_{\boldsymbol{p}}}(\boldsymbol{u}) \triangleq \;\; & \underset{\boldsymbol{p}}{\operatorname{argmin}} \quad \|\boldsymbol{u} - \boldsymbol{p}\|_2^2 && \text{(31a)}\\
& \text{s.t.} \quad\;\; \mathbf{1}^T\boldsymbol{p} = 1, && \text{(31b)}\\
& \qquad\;\; \psi(\boldsymbol{p}, \boldsymbol{q}) \leq d, && \text{(31c)}\\
& \qquad\;\; \boldsymbol{p} \geq 0, && \text{(31d)}
\end{aligned}$$

where the constraints (31b–31d) arise from the ambiguity set $\mathcal{U}_{\boldsymbol{p}}(\boldsymbol{q}, d)$. In particular, constraint (31c) is shown with respect to a generic distance measure $\psi(\boldsymbol{p}, \boldsymbol{q})$, which can be defined by the user as any of the measures in (22a–22e) with an appropriate maximum permissible distance d. For example, d can be set using (25–27). Thus, for some point $\boldsymbol{u}$, the projection $\varPi_{\mathcal{U}_{\boldsymbol{p}}}(\boldsymbol{u})$ finds the closest solution within the ambiguity set $\mathcal{U}_{\boldsymbol{p}}(\boldsymbol{q}, d)$.

Likewise, we retain feasibility in the descent step by projecting each iteration in the descent direction onto onto the set $\mathbb{R}^n_+$. This projection is trivial and, for some arbitrary point $\boldsymbol{z} \in \mathbb{R}^n$, can be computed as follows

$$\Pi_{\mathbb{R}^n_+}(\boldsymbol{u}) \triangleq \begin{cases} u_i & \text{if } u_i > 0 \\ 0^+ & \text{otherwise} \end{cases} \quad \text{for } i = 1, \dots, n, \tag{32}$$

where we inspect every element of $\boldsymbol{z}$ and set any non-positive element to an arbitrarily small positive value.[5]

We proceed to describe the PGDA algorithm. Like an unconstrained gradient descent–ascent algorithm, we take steps to descend in $\boldsymbol{y}$ and ascend in $\boldsymbol{p}$ in the direction of the respective gradients of $f_{\text{RP}}(\boldsymbol{y}, \boldsymbol{p})$. The gradients of $f_{\text{RP}}(\boldsymbol{y}, \boldsymbol{p})$ are

$$\nabla_{\boldsymbol{y}} f_{\text{RP}}(\boldsymbol{y}, \boldsymbol{p}) = \hat{\boldsymbol{\Sigma}}(\boldsymbol{p})\boldsymbol{y} - \kappa \boldsymbol{y}^{-1}, \tag{33}$$

$$\nabla_{\boldsymbol{p}} f_{\text{RP}}(\boldsymbol{y}, \boldsymbol{p}) = \frac{1}{2}\hat{\boldsymbol{\pi}}^2(\boldsymbol{y}) - \hat{\boldsymbol{\Theta}}(\boldsymbol{y})\boldsymbol{p} \tag{34}$$

where $\boldsymbol{y}^{-1} = [1/y_1 \cdots 1/y_n]^\top$.

Given that the feasible sets $\mathbb{R}^n_+$ and $\mathcal{U}_{\boldsymbol{p}}$ are convex, we can design the search directions in both $\boldsymbol{y}$ and $\boldsymbol{p}$ such that we retain feasibility after each iteration. Assume we have some feasible solutions $\boldsymbol{y}^k \in \mathbb{R}^n_+$ and $\boldsymbol{p}^k \in \mathcal{U}_{\boldsymbol{p}}$. The search directions at iteration k are

$$\begin{aligned} \boldsymbol{g}^k &\triangleq \Pi_{\mathbb{R}^n_+}\Big(\boldsymbol{y}^k - \alpha_k \nabla_{\boldsymbol{y}} f_{\text{RP}}(\boldsymbol{y}^k, \boldsymbol{p}^k)\Big) - \boldsymbol{y}^k, \\ \boldsymbol{h}^k &\triangleq \Pi_{\mathcal{U}_{\boldsymbol{p}}}\Big(\boldsymbol{p}^k + \gamma_k \nabla_{\boldsymbol{p}} f_{\text{RP}}(\boldsymbol{y}^{k+1}, \boldsymbol{p}^k)\Big) - \boldsymbol{p}^k, \end{aligned}$$

where α_k and γ_k are the step sizes in each direction. To ensure that our next iteration remains within the feasible set, we define the search parameters $\eta_{\boldsymbol{y}}, \eta_{\boldsymbol{p}} \in [0, 1]$. Thus, our next iterations in each direction are

$$\begin{aligned} \boldsymbol{y}^{k+1} &= \boldsymbol{y}^k + \eta_{\boldsymbol{y}} \boldsymbol{g}^k, \\ \boldsymbol{p}^{k+1} &= \boldsymbol{p}^k + \eta_{\boldsymbol{p}} \boldsymbol{h}^k. \end{aligned}$$

Thus, the resulting points $\boldsymbol{y}^{k+1}$ and $\boldsymbol{p}^{k+1}$ are linear combinations between two feasible points in each set, respectively. Since the sets are convex, these points are also convex.

We defer to the Barzilai–Borwein method [2] to define our step sizes. Specifically, we use the following definition of the Barzilai–Borwein method to define

[5] We must replace any non-positive value with a strictly positive value, 0^+. This stems from the derivative of the logarithm barrier in our objective function, where $(d/dz_i)\ln(z_i) = 1/z_i$. Thus, if z_i is not strictly positive, this will lead to a numerical error when we implement the algorithm. When implementing this numerically, 0^+ can be set to some small positive value.

our descent and ascent step sizes. For any iteration $k \geq 1$ where $k = 0, 1, \ldots$, we have

$$\alpha_k = \frac{\left|(\boldsymbol{y}^k - \boldsymbol{y}^{k-1})^\top \left(\nabla_{\boldsymbol{y}} f_{\mathrm{RP}}(\boldsymbol{y}^k, \boldsymbol{p}^k) - \nabla_{\boldsymbol{y}} f_{\mathrm{RP}}(\boldsymbol{y}^{k-1}, \boldsymbol{p}^{k-1})\right)\right|}{\left\|\nabla_{\boldsymbol{y}} f_{\mathrm{RP}}(\boldsymbol{y}^k, \boldsymbol{p}^k) - \nabla_{\boldsymbol{y}} f_{\mathrm{RP}}(\boldsymbol{y}^{k-1}, \boldsymbol{p}^{k-1})\right\|_2^2}, \tag{35}$$

$$\gamma_k = \frac{\left|(\boldsymbol{p}^k - \boldsymbol{p}^{k-1})^\top \left(\nabla_{\boldsymbol{p}} f_{\mathrm{RP}}(\boldsymbol{y}^k, \boldsymbol{p}^k) - \nabla_{\boldsymbol{p}} f_{\mathrm{RP}}(\boldsymbol{y}^{k-1}, \boldsymbol{p}^{k-1})\right)\right|}{\left\|\nabla_{\boldsymbol{p}} f_{\mathrm{RP}}(\boldsymbol{y}^k, \boldsymbol{p}^k) - \nabla_{\boldsymbol{p}} f_{\mathrm{RP}}(\boldsymbol{y}^{k-1}, \boldsymbol{p}^{k-1})\right\|_2^2}. \tag{36}$$

The Barzilai–Borwein step sizes are sometimes referred to as the 'spectral step size'. In conjunction with projected gradient methods, this class of algorithms is sometimes called 'spectral projected gradient descent' [9].

Next, we discuss how to determine the parameters $\eta_{\boldsymbol{y}}, \eta_{\boldsymbol{p}} \in (0, 1]$. Specifically, we favour the non-monotone Grippo–Lampariello–Lucidi (GLL) line search proposed in [30]. The GLL line search method has been shown to work well with spectral projected gradient descent and ensures global convergence on closed convex sets [9, 19].

A brief overview of this line search method follows. Consider the descent step in $\boldsymbol{y}$. For a given integer $m \geq 1$, we are searching for $\eta_{\boldsymbol{y}} \in (0, 1]$ such that

$$f_{\mathrm{RP}}(\boldsymbol{y}^k + \eta_{\boldsymbol{y}} \boldsymbol{g}^k, \boldsymbol{p}^k) \leq \max_{j \in \mathcal{J}} f_{\mathrm{RP}}(\boldsymbol{y}^{k-j}, \boldsymbol{p}^{k-j}) + \beta \eta_{\boldsymbol{y}} (\boldsymbol{g}^k)^\top \nabla_{\boldsymbol{y}} f_{\mathrm{RP}}(\boldsymbol{y}^k, \boldsymbol{p}^k) \tag{37}$$

where $\mathcal{J} \triangleq \left\{j \in \mathbb{Z} : 0 \leq j \leq \min\{k, m-1\}\right\}$ and $\beta \in (0, 1)$ is a given constant. Intuitively, a more aggressive descent step corresponds to a larger value of $\eta_{\boldsymbol{y}}$. This typically means we initially set $\eta_{\boldsymbol{y}} = 1$ and shrink it appropriately by some fixed factor $\tau \in (0, 1)$, resulting in an inexact but fast method to determine an appropriate value for $\eta_{\boldsymbol{y}}$.

The GLL method stems from an Armijo-type line search, but it allows us to be greedier with our step sizes. For example, if we set $m = 1$, then we revert back to a traditional Armijo-type line search method and the condition in (37) causes our objective function to decrease monotonically. Thus, by considering multiple previous iterations of the objective value we allow for a non-monotonic decrease.

For the purpose of PGDA, the ascent direction follows the same logic. However, we do not discuss it in detail for the sake of brevity. Instead, the complete PGDA algorithm is presented in Algorithm 1, which shows how to calculate the steps in the descent and ascent directions.

Although the global convergence of spectral projected gradient descent with a GLL line search has been established [9, 19], we purposely avoid making any claims this is true for PGDA. However, we note that for appropriate step sizes, the convergence of convex–concave constrained minimax problems has been previously established [41]. Nevertheless, the algorithmic development of

the PGDA algorithm will help to guide our development of the SCP–PGA algorithm in Section 3.4 below.

Algorithm 1: PGDA for DRRP portfolio optimization

Input: Data $\hat{\boldsymbol{\xi}} \in \mathbb{R}^{n \times T}$; Confidence level $\omega \in (0,1)$; Distance measure {JS, Hellinger, TV}; Nominal distribution $\boldsymbol{q} \in \mathcal{P}$; Risk parity constant $\kappa > 0$; Initial step sizes $\alpha_0, \gamma_0 > 0$; Initial proxy portfolio $\boldsymbol{y}^0$; Convergence tolerance ε_0; Search control parameters $\beta, \tau \in (0,1)$; GLL parameter $m \geq 1$

Output: Optimal DRRP portfolio $\boldsymbol{x}^*$

Find the distance limit $d(\omega, T)$ as shown in either of (25–27)
Initialize the adversarial distribution: $\boldsymbol{p}^0 = \boldsymbol{q}$
Initialize the convergence measure: $\varepsilon = 1$
Initialize the counter: $k = 0$
while $\varepsilon > \varepsilon_0$ **do**
  if $k \geq 1$ **then**
    Update α_k as shown in (35)
    Update γ_k as shown in (36)
  $\boldsymbol{g}^k = \varPi_{\mathbb{R}^n_+}\left(\boldsymbol{y}^k - \alpha_k \nabla_{\boldsymbol{y}} f_{\mathrm{RP}}(\boldsymbol{x}^k, \boldsymbol{p}^k)\right) - \boldsymbol{y}^k$
  $\eta_{\boldsymbol{y}} = 1$
  $\bar{\boldsymbol{y}} = \boldsymbol{y}^k + \eta_{\boldsymbol{y}} \boldsymbol{g}^k$
  while $f_{RP}(\bar{\boldsymbol{y}}, \boldsymbol{p}^k) > \max\limits_{j \in \mathcal{J}} f_{RP}(\boldsymbol{y}^{k-j}, \boldsymbol{p}^{k-j}) + \beta \eta_{\boldsymbol{y}} (\boldsymbol{g}^k)^\top \nabla_{\boldsymbol{y}} f_{RP}(\boldsymbol{x}^k, \boldsymbol{p}^k)$ **do**
    $\eta_{\boldsymbol{y}} = \eta_{\boldsymbol{y}} \tau$
    $\bar{\boldsymbol{y}} = \boldsymbol{y}^k + \eta_{\boldsymbol{y}} \boldsymbol{g}^k$
  $\boldsymbol{y}^{k+1} = \bar{\boldsymbol{y}}$
  $\boldsymbol{h}^k = \varPi_{\mathcal{U}_{\boldsymbol{p}}}\left(\boldsymbol{p}^k + \gamma_k \nabla_{\boldsymbol{p}} f_{\mathrm{RP}}(\boldsymbol{x}^k, \boldsymbol{p}^k)\right) - \boldsymbol{p}^k$
  $\eta_{\boldsymbol{p}} = 1$
  $\bar{\boldsymbol{p}} = \boldsymbol{p}^k + \eta_{\boldsymbol{p}} \boldsymbol{h}^k$
  while $f_{RP}(\boldsymbol{y}^k, \bar{\boldsymbol{p}}) < \min\limits_{j \in \mathcal{J}} f_{RP}(\boldsymbol{y}^{k-j}, \boldsymbol{p}^{k-j}) + \beta \eta_{\boldsymbol{p}} (\boldsymbol{h}^k)^\top \nabla_{\boldsymbol{p}} f_{RP}(\boldsymbol{x}^k, \boldsymbol{p}^k)$ **do**
    $\eta_{\boldsymbol{p}} = \eta_{\boldsymbol{p}} \tau$
    $\bar{\boldsymbol{p}} = \boldsymbol{p}^k + \eta_{\boldsymbol{p}} \boldsymbol{h}^k$
  $\boldsymbol{p}^{k+1} = \bar{\boldsymbol{p}}$
  if $k \geq 1$ **then**
    $\varepsilon = \|\boldsymbol{p}^{k+1} - \boldsymbol{p}^k\|_2$
  $k = k + 1$
Find the optimal portfolio: $\boldsymbol{x}^* = \boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^k)$

Result: Optimal DRRP portfolio $\boldsymbol{x}^*$

3.4 Sequential convex programming with projected gradient ascent

Our discussion of the PGDA algorithm served two purposes. First, It provided a straightforward approach to solve a convex–concave minimax problem. More importantly, it showed the steps required to navigate such a problem and highlighted some structural weaknesses. In particular, the PGDA algorithm requires that we determine two appropriate step sizes, α_k and γ_k, during each

iteration. Moreover, the set $\mathbb{R}^n_+$ in which $\boldsymbol{y}$ exists is not compact. Therefore, the PGDA algorithm necessitates careful initialization and, in general, may be prone to diverge.

The PGDA has three structural weaknesses. First, the design of the risk parity problem in (13) means that the descent step in the PGDA algorithm must ignore the budget equality constraint. In other words, the algorithm operates in the unbounded set $\mathbb{R}^n_+$ instead of the compact set $\mathcal{X}$. Only after convergence of the PGDA algorithm do we project our solution onto $\mathcal{X}$.

Second, the PGDA algorithm is twice as susceptible the problem of vanishing gradients. As we approach a saddle point, the gradient information in both directions starts to vanish, slowing the convergence of the algorithm to an optimal saddle point.

The third and final weakness is the burden placed on the user to define an initial step size in both $\boldsymbol{y}$ and $\boldsymbol{p}$ directions, as well as the initial guess $\boldsymbol{y}^0$. Since the proxy variable $\boldsymbol{y}$ does not have an upper bound, an improperly sized $\boldsymbol{y}^0$ may slow down convergence.

These three weaknesses can be remediated by redesigning the algorithm to operate directly on the set of admissible portfolios $\mathcal{X}$. The strict convexity of the nominal risk parity problem in (13) means that there exists a unique risk parity portfolio $\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p})$ for every $\boldsymbol{p} \in \mathcal{U}_{\boldsymbol{p}}$. Assume we have an ascent algorithm and let $\boldsymbol{p}^k$ be the the k^{th} iteration of our adversarial probability. Then, for every $k = 0, 1, \ldots$, there exists a corresponding risk parity portfolio $\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^k)$. Thus, we can formulate an algorithm that ascends in $\boldsymbol{p} \in \mathcal{U}_{\boldsymbol{p}}$ while enforcing the risk parity condition in $\boldsymbol{x} \in \mathcal{X}$ after every iteration.

Conversely, we can interpret this algorithm as solving a sequence of convex problems. Specifically, we solve the risk parity problem $\boldsymbol{x}^k = \boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^k)$, where we update the covariance matrix $\hat{\boldsymbol{\Sigma}}(\boldsymbol{p}^k)$ after every iteration k. Thus, the resulting algorithm needs only to ascend in $\boldsymbol{p} \in \mathcal{U}_{\boldsymbol{p}}$, meaning it can be solved using PGA. In turn, this means that the user no longer needs to define any of the initial conditions and updates associated with the proxy variable $\boldsymbol{y}$. Given that the proposed PGA algorithm involves iteratively solving a sequence of convex problems, we refer to it as the SCP–PGA algorithm.

By the definition of the set of admissible portfolios in (3), we have that $\mathcal{X} \subset \mathbb{R}^n_+$. Therefore, from (30), it follows that

$$f_{\mathrm{RP}}(\boldsymbol{x}^*, \boldsymbol{p}) \leq f_{\mathrm{RP}}(\boldsymbol{x}^*, \boldsymbol{p}^*) \leq f_{\mathrm{RP}}(\boldsymbol{x}, \boldsymbol{p}^*) \;\; \forall\, \boldsymbol{x} \in \mathcal{X},\; \boldsymbol{p} \in \mathcal{U}_{\boldsymbol{p}}.$$

As prescribed by the SCP–PGA algorithm, we have $\boldsymbol{x}^k = \boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^k)$ for every iteration k. At convergence, we must have that $\boldsymbol{x}^* = \boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^*)$. Consequently, the inequality above can be restated as follows

$$f_{\mathrm{RP}}\big(\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}), \boldsymbol{p}\big) \leq f_{\mathrm{RP}}\big(\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^*), \boldsymbol{p}^*\big) \;\; \forall\, \boldsymbol{x} \in \mathcal{X},\; \boldsymbol{p} \in \mathcal{U}_{\boldsymbol{p}},$$

indicating that we can use PGA to maximize $f_{\mathrm{RP}}\big(\boldsymbol{x}^{\mathrm{RP}}(\cdot), \cdot\big) : \mathcal{U}_{\boldsymbol{p}} \to \mathbb{R}$. Since a closed-form solution to $\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p})$ does not exist, we must proceed iteratively by solving $\boldsymbol{x}^k = \boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^k)$ at every iteration k. Although this increases the computational cost per iteration when compared against our previous PGDA

algorithm, we note that convex optimization problems can be efficiently solved by modern optimization algorithms and software packages.

The SCP–PGA algorithm follows the same logic as the PGDA algorithm, except we are only concerned with the ascent step. Our maximization problem is quadratic concave over a compact convex set, and, the gradient of the objective function is Lipschitz continuous since its Hessian is PSD for all $\boldsymbol{p}$. Therefore, using an appropriate line search can guarantee convergence. Specifically, the GLL line search in our algorithm means that, by design, each iteration achieves a sufficient increase in $f_{\mathrm{RP}}\big(\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}), \boldsymbol{p}\big)$ such that we converge to the global maximum. We finish this subsection by presenting the complete SCP–PGA algorithm in Algorithm 2.

Algorithm 2: SCP–PGA for DRRP portfolio optimization

Input: Data $\hat{\boldsymbol{\xi}} \in \mathbb{R}^{n\times T}$; Confidence level $\omega \in (0,1)$; Distance measure {JS, Hellinger, TV}; Nominal distribution $\boldsymbol{q} \in \mathcal{P}$; Risk parity constant $\kappa > 0$; Initial step size $\gamma_0 > 0$; Convergence tolerance ε_0; Search control parameters $\beta, \tau \in (0,1)$; GLL parameter $m \geq 1$

Output: Optimal DRRP portfolio $\boldsymbol{x}^*$

Find the distance limit $d(\omega, T)$ as shown in either of (25–27)
Initialize the adversarial distribution $\boldsymbol{p}^0 = \boldsymbol{q}$
Initialize the convergence measure: $\varepsilon >> 1$
Initialize the counter: $k = 0$
while $\varepsilon > \varepsilon_0$ **do**
  if $k \geq 1$ **then**
    Update γ_k as shown in (36)
  $\boldsymbol{x}^k = \boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^k)$
  $\boldsymbol{h}^k = \Pi_{\mathcal{U}_{\boldsymbol{p}}}\Big(\boldsymbol{p}^k + \gamma_k \nabla_{\boldsymbol{p}} f_{\mathrm{RP}}(\boldsymbol{x}^k, \boldsymbol{p}^k)\Big) - \boldsymbol{p}^k$
  $\eta_{\boldsymbol{p}} = 1$
  $\bar{\boldsymbol{p}} = \boldsymbol{p}^k + \eta_{\boldsymbol{p}} \boldsymbol{h}^k$
  while $f_{RP}(\boldsymbol{x}^k, \bar{\boldsymbol{p}}) < \min\limits_{j\in\mathcal{J}} f_{RP}(\boldsymbol{x}^{k-j}, \boldsymbol{p}^{k-j}) + \beta\eta_{\boldsymbol{p}}(\boldsymbol{h}^k)^\top \nabla_{\boldsymbol{p}} f_{RP}(\boldsymbol{x}^k, \boldsymbol{p}^k)$ **do**
    $\eta_{\boldsymbol{p}} = \eta_{\boldsymbol{p}} \tau$
    $\bar{\boldsymbol{p}} = \boldsymbol{p}^k + \eta_{\boldsymbol{p}} \boldsymbol{h}^k$
  $\boldsymbol{p}^{k+1} = \bar{\boldsymbol{p}}$
  if $k \geq 1$ **then**
    $\varepsilon = \|\boldsymbol{p}^{k+1} - \boldsymbol{p}^k\|_2$
  $k = k + 1$

Result: Optimal DRRP portfolio $\boldsymbol{x}^*$

3.5 Model variants

We complete this section by introducing a variant to our DRRP problem and stating how this variant can be similarly solved using the SCP–PGA algorithm. By design, risk parity considers only the portfolio risk measure.

However, there exist risk-based diversification portfolio optimization problems that also consider the portfolio expected return [32, 16].

We specifically focus on the problem proposed by Roncalli [49], where we wish to construct a risk–return parity (RRP) portfolio. The RRP problem seeks to equalize the asset risk and return contributions. The specific risk–return profile is subjectively defined by the investor.

The problem in [49] is defined using the portfolio standard deviation. However, we prefer to use of the portfolio variance for the sake of computational tractability. In the context of MPT, the investor's risk–return profile depicts a trade-off between maximizing the portfolio expected return while simultaneously minimizing its variance. In the RRP problem, the investor seeks to equalize the individual asset contributions towards this risk–return profile. This objective can be partitioned as follows,

$$\begin{aligned}\hat{\sigma}_\pi^2(\boldsymbol{x},\boldsymbol{p}) - \lambda\hat{\mu}_\pi(\boldsymbol{x},\boldsymbol{p}) &= \boldsymbol{x}^\top\hat{\boldsymbol{\Sigma}}(\boldsymbol{p})\boldsymbol{x} - \lambda\boldsymbol{x}^\top\hat{\boldsymbol{\mu}}(\boldsymbol{p}) \\ &= \sum_{i=1}^{n} x_i[\hat{\boldsymbol{\Sigma}}(\boldsymbol{p})\boldsymbol{x}]_i - \lambda x_i\hat{\mu}_i(\boldsymbol{p}) \\ &= \sum_{i=1}^{n} R_i - \lambda r_i\end{aligned}$$

where $\lambda \in \mathbb{R}_+$ is a risk–return trade-off coefficient subjectively determined by the investor's risk appetite, and $r_i \triangleq x_i\hat{\mu}_i(\boldsymbol{p})$ is the return contribution of asset i towards the portfolio return. Thus, we seek to construct a portfolio such that $R_i - \lambda r_i = R_j - \lambda r_j \;\forall\; i, j$. We can do this by introducing the RRP objective function $f_{\text{RRP}}(\boldsymbol{y},\boldsymbol{p})$ and its corresponding gradients

$$f_{\text{RRP}}(\boldsymbol{y},\boldsymbol{p}) \triangleq \frac{1}{2}\boldsymbol{y}^\top\hat{\boldsymbol{\Sigma}}(\boldsymbol{p})\boldsymbol{y} - \lambda\boldsymbol{y}^\top\hat{\boldsymbol{\mu}}(\boldsymbol{p}) - \kappa\sum_{i=1}^{n}\ln(y_i) \tag{38a}$$

$$\triangleq \frac{1}{2}\Big(\boldsymbol{p}^\top\hat{\boldsymbol{\pi}}^2(\boldsymbol{y}) - \boldsymbol{p}^\top\hat{\boldsymbol{\Theta}}(\boldsymbol{y})\boldsymbol{p}\Big) - \lambda\boldsymbol{p}^\top\hat{\boldsymbol{\pi}}(\boldsymbol{y}) - \kappa\sum_{i=1}^{n}\ln(y_i), \tag{38b}$$

$$\nabla_{\boldsymbol{y}} f_{\text{RRP}}(\boldsymbol{y},\boldsymbol{p}) = \hat{\boldsymbol{\Sigma}}(\boldsymbol{p})\boldsymbol{y} - \lambda\hat{\boldsymbol{\mu}}(\boldsymbol{p}) - \kappa\boldsymbol{y}^{-1}, \tag{39}$$

$$\nabla_{\boldsymbol{p}} f_{\text{RRP}}(\boldsymbol{y},\boldsymbol{p}) = \frac{1}{2}\hat{\boldsymbol{\pi}}^2(\boldsymbol{y}) - \hat{\boldsymbol{\Theta}}(\boldsymbol{y})\boldsymbol{p} - \lambda\hat{\boldsymbol{\pi}}(\boldsymbol{y}). \tag{40}$$

We can solve the RRP problem by using the RRP objective function in (38) and its gradients in (39) and (40) to replace the risk parity function in (13) and the SCP–PGA algorithm.

Moreover, the RRP minimax problem can be easily converted into a distributionally robust MVO problem if we discard the logarithmic barrier in (38) and remove the one-half factor multiplying the portfolio variance. Doing this would address the original problem posed by Calafiore [12], except we can solve the MVO problem using the gradient-based Algorithms 1 or 2 using different measures of statistical distance. Both the RRP and MVO problems

serve to exemplify how the DRRP framework can be naturally extended to other portfolio optimization problems.

4 Numerical experiments

This section consists of three separate experiments. The first experiment serves to evaluate the numerical performance of the SCP–PGA algorithm (Algorithm 2) and is conducted using synthetic data to generate increasingly larger datasets. As a benchmark, this experiment also includes results from the PGDA algorithm (Algorithm 1). The second experiment assesses the in-sample performance of the DRRP portfolio in a financial context and uses historical data. The third and last experiment assesses the out-of-sample financial performance of the portfolio. To limit our scope, the experiments are conducted on DRRP portfolios based on the JS, Hellinger and TV distance measures.

The second and third experiments share the same data. The data consist of historical observations ranging from the start of 1998 until the end of 2016 for 30 industry portfolios. These industry portfolios serve as our financial assets and are akin to many popular exchange traded funds. The data were obtained from Kenneth R. French's data library [26]. Table 1 lists the 30 industry portfolios.

Table 1: List of assets

Food Products	Tobacco	Beer and Liquor	Recreation
Household Products	Apparel	Healthcare	Chemicals
Fabricated Products	Construction	Steel Works	Electrical Equip.
Aircraft, Ships, Rail Equip.	Mining	Coal	Oil and Gas
Communication	Services	Business Equip.	Paper
Restaurants and Hotels	Wholesale	Retail	Financials
Printing	Textiles	Automobiles	Utilities
Transportation	Other		

All experiments were conducted on an Apple MacBook Pro computer (2.8 GHz Intel Core i7, 16 GB 2133 MHz DDR3 RAM) running macOS 'Catalina'. The computer script was written in the Julia programming language (version 1.4.0) using the modelling language 'JuMP' [21] with IPOPT (version 3.12.6) as the optimization solver.

4.1 Numerical performance and tractability

The second part of the numerical performance experiment evaluates the numerical performance of the SCP–PGA algorithm under different statistical

distance measures to construct DRRP portfolios. In particular, we test the JS, Hellinger and TV distance measures. Additionally, we compare the SCP–PGA to the PGDA algorithm.

For the comparison between the two algorithms, the experimental setup is the following. We randomly generate synthetic datasets with $n = 50$, 200 assets and $T = 1{,}000$, 5,000 scenarios for a total of four different datasets. The largest dataset, with $n = 200$ and $T = 5,000$, simulates the conditions to create a portfolio with 200 constituents using approximately 20 years worth of daily scenarios. We construct DRRP portfolios for three different confidence levels $\omega = 0.15$, 0.3, 0.45.

The remainder of the user-defined parameters are the following. The risk parity constant is set to $\kappa = 1$, while the convergence tolerance is set to $\varepsilon_0 = 10^{-6}$. As recommended in [9], we set $m = 10$. Moreover, we set the search parameters to $\beta = 10^{-6}$ and $\tau = 0.9$. The initial ascent step size is $\gamma_0 = 0.1$. In addition, we set the following values for the PGDA algorithm: $\alpha_0 = 30$, $y_i^0 = 5$ for $i = 1, \ldots, n$.

The SCP–PGA and PGDA algorithms are evaluated based on their runtime in seconds, the number of iterations until convergence, and the resulting portfolio variance. The two algorithms aim to construct risk parity portfolios with the worst-case estimate of the portfolio variance with respect to the probability ambiguity set $\mathcal{U}_{\boldsymbol{p}}$. Since, by design, both algorithms yield portfolios that satisfy the risk parity condition,[6] we evaluate the convergence quality of the two algorithms by comparing the corresponding portfolio variances. The numerical results are presented in Table 2.

The results in Table 2 show that the SCP–PGA algorithm is able to attain an equal or higher portfolio variance than the PGDA algorithm in every single instance, indicating that the SCP–PGA algorithm converges to a higher quality solution. This highlights the sensitivity of the PGDA algorithm to its initial conditions, where the solution appears to converge if the step sizes become ill-conditioned after a few iterations. Moreover, for every instance where the variance of both algorithms is the same, the runtime of the SCP–PGA algorithm is significantly faster than the PGDA algorithm (e.g., see the results for $\omega = 0.15$, $n = 50$ and $T = 5,000$ with the JS divergence as the distance measure).

The second part of the numerical performance experiment focuses solely on the SCP–PGA algorithm and extends our previous experiment to include synthetic datasets with $T = 100$ and $n = 500$. Thus, including the original values of T and n, we have a total of nine different datasets.

As before, the DRRP portfolios from the SCP–PGA algorithm are evaluated based on their runtime, the number of iterations until convergence, and the portfolio variance. We also include the runtime per iteration. Finally, the results also show the variance of the nominal risk parity portfolio for the same dataset. The nominal portfolio variance serves as a benchmark. When looking

[6]The SCP–PGA algorithm finds a risk parity portfolio $\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^k)$ during each iteration k, while the last line of the PGDA algorithm also enforces $\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^k)$ after convergence in $\boldsymbol{p}$.

Table 2: Comparison of numerical performance between the PGDA algorithm (A.1) and the SCP–PGA algorithm 2 (A.2). The maximum number of iterations is limited to 1,000, after which the algorithms terminate.

	$n = 50$						$n = 200$					
	JS		Hellinger		TV		JS		Hellinger		TV	
	A.1	A.2	A.1	A.2	A.1	A.2	A.1	A.2	A.1	A.2	A.1	A.2
$\omega = 0.15$												
$\boldsymbol{T = 1,000}$												
Time (s)	26.3	1.49	3.49	2.40	57.9	8.32	5.96	5.94	18.2	6.11	7.84	10.5
Iterations	221	12	13	11	185	27	13	13	25	11	11	17
Var. ($\times 10^4$)	2.40	6.61	3.84	7.13	3.90	9.38	2.07	6.17	3.49	6.70	5.81	9.13
$\boldsymbol{T = 5,000}$												
Time (s)	335	9.15	72.3	17.9	191	33.9	79.6	35.3	4,172	33.2	322	104
Iterations	442	12	29	12	90	19	17	14	1,000	11	44	29
Var. ($\times 10^4$)	5.64	5.64	1.68	6.03	3.49	9.89	4.43	5.23	5.51	5.68	2.76	8.17
$\omega = 0.3$												
$\boldsymbol{T = 1,000}$												
Time (s)	7.04	2.98	9.42	7.44	7.69	10.9	436	7.99	143	14.2	19.9	14.9
Iterations	76	34	39	28	23	40	1000	16	272	25	31	22
Var. ($\times 10^4$)	10.2	10.8	11.7	11.9	5.62	14.4	1.16	10.5	0.97	11.7	3.28	14.2
$\boldsymbol{T = 5,000}$												
Time (s)	31.9	13.1	60.5	36.5	275	50.1	157	79.9	569	99.6	235	128
Iterations	22	24	37	23	127	32	59	33	132	28	48	37
Var. ($\times 10^4$)	5.80	10.8	11.6	11.6	1.28	16.2	3.97	9.06	3.67	9.98	0.71	12.9
$\omega = 0.45$												
$\boldsymbol{T = 1,000}$												
Time (s)	8.43	4.16	5.70	8.58	24.5	21.5	5.27	13.0	31.7	15.8	61.4	23.6
Iterations	87	45	22	38	86	72	11	27	51	27	108	35
Var. ($\times 10^4$)	16.1	16.1	0.51	17.8	4.87	19.3	10.2	16.1	10.3	17.9	9.50	19.1
$\boldsymbol{T = 5,000}$												
Time (s)	34.5	20.6	198	64.6	98.3	66.2	310	99.4	320	122	127	224
Iterations	51	37	78	38	47	40	141	41	42	38	14	62
Var. ($\times 10^4$)	17.9	17.9	1.81	19.3	22.0	22.5	14.3	14.3	1.70	15.7	4.87	17.6

at the DRRP portfolio variance, we must keep in mind that this corresponds to the worst-case estimate of the variance as defined by the ambiguity set $\mathcal{U}_{\boldsymbol{p}}$. Therefore, we expect the DRRP portfolio variance to be larger than the nominal. The results are presented in Table 3.

Table 3: Numerical performance of the SCP–PGA algorithm

	$n = 50$				$n = 200$				$n = 500$			
	Nom.	JS	H	TV	Nom.	JS	H	TV	Nom.	JS	H	TV
$\omega = 0.15$												
$\boldsymbol{T = 100}$												
Time (s)	-	0.24	0.34	0.98	-	1.26	0.98	2.24	-	5.57	5.19	8.57
Iterations	-	10	9	27	-	10	9	22	-	10	9	14
Time/Iter. (s)	-	0.02	0.04	0.04	-	0.13	0.11	0.1	-	0.56	0.57	0.61
Var. ($\times 10^4$)	3.17	4.90	5.19	5.64	2.98	4.96	5.31	5.96	3.29	5.91	6.37	7.44
$\boldsymbol{T = 1,000}$												
Time (s)	-	1.49	2.40	8.32	-	5.94	6.11	10.45	-	27.7	23.9	117
Iterations	-	12	11	27	-	13	11	17	-	12	11	42
Time/Iter. (s)	-	0.12	0.22	0.31	-	0.46	0.56	0.61	-	2.31	2.18	2.79
Var. ($\times 10^4$)	4.03	6.61	7.13	9.38	3.55	6.17	6.70	9.13	3.07	5.02	5.42	6.24
$\boldsymbol{T = 5,000}$												
Time (s)	-	9.15	17.9	33.9	-	35.3	33.2	104	-	147	132	277
Iterations	-	12	12	19	-	14	11	29	-	12	12	21
Time/Iter. (s)	-	0.76	1.50	1.78	-	2.52	3.02	3.6	-	12.2	11.0	13.2
Var. ($\times 10^4$)	3.11	5.64	6.03	9.89	3.11	5.23	5.68	8.17	3.39	5.91	6.37	9.76
$\omega = 0.3$												
$\boldsymbol{T = 100}$												
Time (s)	-	0.36	0.63	0.94	-	1.65	1.34	1.97	-	9.55	8.91	13.0
Iterations	-	17	917	28	-	17	13	18	-	16	15	22
Time/Iter. (s)	-	0.02	0.04	0.03	-	0.10	0.10	0.11	-	0.6	0.59	0.59
Var. ($\times 10^4$)	3.17	6.92	7.51	7.69	2.98	7.38	8.06	8.46	3.29	9.07	9.89	9.75
$\boldsymbol{T = 1,000}$												
Time (s)	-	2.98	7.44	10.9	-	7.99	14.2	14.9	-	64.9	62.3	65.1
Iterations	-	34	28	40	-	16	25	22	-	31	25	29
Time/Iter. (s)	-	0.09	0.27	0.27	-	0.5	0.57	0.67	-	2.09	2.49	2.24
Var. ($\times 10^4$)	4.03	10.8	11.9	14.4	3.55	10.5	11.7	14.2	3.07	7.57	8.52	9.25
$\boldsymbol{T = 5,000}$												
Time (s)	-	13.1	36.5	50.1	-	79.9	99.6	128	-	251	250	273
Iterations	-	24	23	32	-	33	28	37	-	29	23	30
Time/Iter. (s)	-	0.54	1.29	1.56	-	2.42	3.56	3.45	-	8.65	10.89	9.10
Var. ($\times 10^4$)	3.11	10.8	11.6	16.2	3.11	9.06	9.98	12.9	3.39	10.7	11.7	15.4
$\omega = 0.45$												
$\boldsymbol{T = 100}$												
Time (s)	-	0.66	0.93	1.32	-	2.16	1.87	2.25	-	12.7	11.1	16.3
Iterations	-	22	21	37	-	22	18	21	-	20	17	25
Time/Iter. (s)	-	0.03	0.04	0.04	-	0.1	0.1	0.11	-	0.64	0.65	0.65
Var. ($\times 10^4$)	3.17	9.04	9.84	9.61	2.98	9.69	10.5	10.5	3.29	11.1	12.0	10.7
$\boldsymbol{T = 1,000}$												
Time (s)	-	4.16	8.58	21.5	-	13.0	15.8	23.6	-	87.5	87.6	98.4
Iterations	-	45	38	72	-	27	27	35	-	44	36	44
Time/Iter. (s)	-	0.09	0.23	0.3	-	0.48	0.58	0.67	-	1.99	2.43	2.24
Var. ($\times 10^4$)	4.03	16.1	17.8	19.3	3.55	16.1	17.89	19.1	3.07	10.7	12.1	12.2
$\boldsymbol{T = 5,000}$												
Time (s)	-	20.6	64.6	66.2	-	99.4	122	224	-	288	501	510
Iterations	-	37	38	40	-	41	38	62	-	34	41	48
Time/Iter. (s)	-	0.56	1.70	1.66	-	2.43	3.22	3.61	-	8.47	12.2	10.6
Var. ($\times 10^4$)	3.11	17.87	19.3	22.5	3.11	14.3	15.7	17.6	3.39	16.9	18.5	20.7

The results in Table 3 indicate that, overall, the SCP–PGA algorithm converges within reasonable time, even for the largest dataset tested. We note that the largest dataset, with $n = 500$ and $T = 5,000$, exaggerates the number of scenarios T that we would normally consider for parameter estimation in a conventional environment.[7] Most financial data service providers tend to use anywhere from 10 days to five years when calculating risk metrics such as the portfolio variance or the CAPM 'beta' [55], and rely on daily, weekly or monthly scenarios for these calculations.

As shown by the different runtimes in Table 3, the SCP–PGA algorithm converged the fastest when we used the JS divergence to construct the ambiguity set (with a few exceptions where the Hellinger distance was faster). The runtime per iteration is relatively similar for all three distance measures. Thus, this suggests that having a faster convergence rate is mostly dependent on the number of iterations required until convergence.

If we inspect the portfolio variance, we can see that the JS divergence is more restrictive than either the Hellinger distance or the TV distance. In other words, for the same confidence level, the JS divergence provides a smaller feasible region in our variance maximization step, leading to portfolios with lower variances. Conversely, the TV distance is the most permissive, consistently having the highest variance for all trials. This suggests that, although the three distance measures have been scaled proportionally, their intrinsic differences suggest some are fundamentally more permissive than others once we convert the probability distribution into the portfolio variance.

4.2 In-sample experiment

To better understand how the distributionally robust framework operates on our dataset, we present a set of in-sample trials over multiple confidence levels. The DRRP portfolio is, by design, a risk parity portfolio under the worst-case estimate of the risk measure (i.e., the portfolio variance). This means that the portfolio risk is perfectly diversified among the constituent assets with respect to a given estimate of the covariance matrix. However, it is paramount to understand that only one risk parity portfolio exists for a specific instance of the covariance matrix. For example, if we have two estimates of the covariance matrix, $\boldsymbol{\Sigma}^a$ and $\boldsymbol{\Sigma}^b$, and we find the corresponding risk parity portfolios for each matrix, $\boldsymbol{x}^a$ and $\boldsymbol{x}^b$, then we have that $\boldsymbol{x}^a = \boldsymbol{x}^b$ if and only if $\boldsymbol{\Sigma}^a = c \cdot \boldsymbol{\Sigma}^b$ for any $c > 0$. It follows that if our covariance estimates differ, $\boldsymbol{\Sigma}^a \neq c \cdot \boldsymbol{\Sigma}^b \; \forall \; c > 0$, then $\boldsymbol{x}^a$ will not be a risk parity portfolio with respect to $\boldsymbol{\Sigma}^b$, and $\boldsymbol{x}^b$ will not be a risk parity portfolio with respect to $\boldsymbol{\Sigma}^a$.

With that said, our goal is to evaluate how the asset allocations and risk contributions differ between the robust portfolios and the nominal portfolio. We use the 30 industry portfolios listed in Table 1 as our assets ($n = 30$), and we use two years of weekly returns to estimate the covariance matrix, meaning

[7] This may exclude high frequency trading environments.

we have 104 historical scenarios ($T = 104$). Specifically, the data corresponds to the time period 01–Jan–2008 to 31–Dec–2009.

We begin by visually inspecting the asset weights and risk contributions for robust portfolios built with a confidence level $\omega = 0.3$. The asset weights are shown in Figure 1, and show that the DRRP portfolios exhibit a similar behaviour under all three statistical distance measures. Not only do the DRRP portfolios differ from the nominal portfolio, but the asset weights of all three DRRP portfolios also follow a similar pattern (i.e., the peaks and troughs in Figure 1 are similar for all four portfolios). We also note that the wealth allocation of the DRRP portfolios is less pronounced than the nominal portfolio, with the DRRP portfolios exhibiting a more even distribution of wealth.

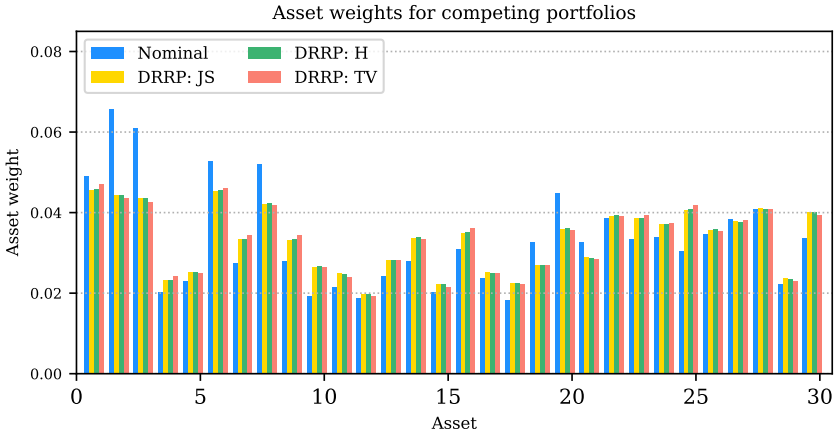

Fig. 1: Asset weights of the nominal and DRRP portfolios with $\omega = 0.3$.

Figure 2 presents a similar analysis, except now we compare the risk contribution per asset with respect to some estimate of the covariance matrix. For convenience, we restate that the risk contribution per asset is defined as $R_i \triangleq x_i[\hat{\boldsymbol{\Sigma}}\boldsymbol{x}]_i$ for some estimate $\hat{\boldsymbol{\Sigma}}$. For a fair comparison, the top plot in Figure 2 shows the risk contribution per asset for all portfolios relative to the nominal estimate of the covariance matrix, $\boldsymbol{\Sigma}^{\text{nom}} \triangleq \boldsymbol{\Sigma}(\boldsymbol{q})$. The remaining three plots compares the risk contributions of the nominal portfolio with respect to $\boldsymbol{\Sigma}^{\text{JS}}$, $\boldsymbol{\Sigma}^{\text{H}}$ and $\boldsymbol{\Sigma}^{\text{TV}}$, respectively. These three matrices correspond to the estimated covariance matrix at convergence of Algorithm 2 for the respective distance measure.

The top plot in Figure 2 confirms the similarity between all three distributionally robust portfolios. For all risk contributions per asset, the three DRRP portfolios together are either lower or higher than the nominal portfolio (i.e., there is no asset where its risk contribution under a DRRP portfolio is higher than the nominal portfolio while simultaneously lower under another DRRP portfolio). The DRRP portfolios choose the same assets to over- or under-

contribute risk relative to the nominal portfolio, highlighting the structural similarity between the DRRP portfolios.

The three remaining plots in Figure 2 serve to show that all robust portfolios are true risk parity portfolios with respect to their corresponding estimate of the covariance matrix. As shown in the plots, the bars for the robust portfolios are of equal height.

The last component of the in-sample experiment replicates the same procedure but for varying levels of confidence ω. For brevity, these results are summarized the results in Table 4. The table shows the total variance of the four competing portfolios with respect to the nominal estimate of the covariance matrix, as well as a pairwise comparison of the total variance of the robust portfolios against the nominal using the worst-case estimates of the covariance matrix. In addition, we report on the risk concentration through the coefficient of variation (CV) of the risk contributions. The CV is calculated by taking the standard deviation of the risk contributions and dividing them by their average, i.e.,

$$\mathrm{CV} = \frac{\mathrm{SD}\big(\boldsymbol{x} \odot [\hat{\boldsymbol{\Sigma}}\boldsymbol{x}]\big)}{\frac{1}{n}\boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x}}, \tag{41}$$

where '$\odot$' is the element-wise multiplication operator and $\mathrm{SD}(\cdot)$ computes the standard deviation of the corresponding vector. In theory, an optimal risk parity portfolio should have a CV of zero.

The results in Table 4 show that all portfolios have perfect risk diversification with respect to their corresponding estimates of the covariance matrix (i.e., the CV of the portfolios is approximately zero with respect to their corresponding instance of $\hat{\boldsymbol{\Sigma}}$). An interesting observation from Table 4 is that the nominal portfolio has the lowest total variance when compared against the robust portfolios for all instances of the covariance matrix. We note that this observation does not fundamentally conflict with our objective, as our robust portfolios aim to diversify risk, and not minimize it. Nevertheless, the results suggest that these robust portfolios incur more ex ante risk when compared to the nominal portfolio.

4.3 Out-of-sample experiment

For the out-of-sample experiment, we test the risk parity problem and the RRP variant. The latter problem allows us to evaluate the ex post impact of including the estimated expected returns during optimization.

An overview of the experimental setup follows. Our portfolio constituents are the 30 assets listed in Table 1 (i.e., $n = 30$ for all portfolios). The dataset consists of weekly historical returns from 01–Jan–1998 to 31–Dec–2016, with the data obtained from [26]. This is a rolling window experiment, where we use two years of weekly scenarios to calibrate our portfolios ($T = 104$) and we hold these portfolios for six month before rebalancing takes place. The portfolios are recalibrated every time we rebalance our portfolios. To exemplify

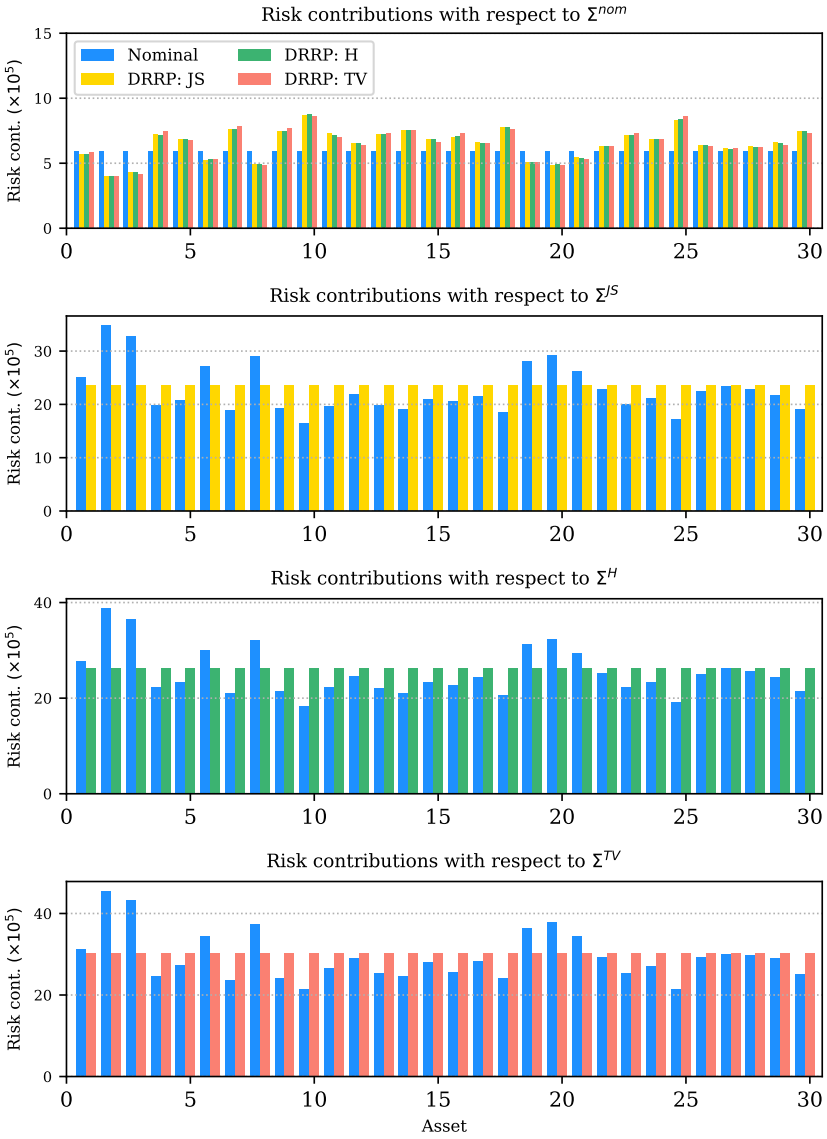

Fig. 2: Risk contributions per asset of the nominal and DRRP portfolios with $\omega = 0.3$. The top plot shows the risk contributions with respect to the nominal estimate of the covariance matrix. The remaining plots show the risk contributions of the nominal portfolio and a single DRRP portfolio for the corresponding robust estimate of the covariance matrix.

Table 4: Portfolio variance and CV based on the nominal and worst-case estimates of the asset covariance matrix

	$\boldsymbol{\Sigma}^{\text{nom}}$				$\boldsymbol{\Sigma}^{\text{JS}}$		$\boldsymbol{\Sigma}^{\text{H}}$		$\boldsymbol{\Sigma}^{\text{TV}}$	
	$\boldsymbol{x}^{\text{nom}}$	$\boldsymbol{x}^{\text{JS}}$	$\boldsymbol{x}^{\text{H}}$	$\boldsymbol{x}^{\text{TV}}$	$\boldsymbol{x}^{\text{nom}}$	$\boldsymbol{x}^{\text{JS}}$	$\boldsymbol{x}^{\text{nom}}$	$\boldsymbol{x}^{\text{H}}$	$\boldsymbol{x}^{\text{nom}}$	$\boldsymbol{x}^{\text{TV}}$
$\boldsymbol{\omega = 0.1}$										
Var. ($\times 10^3$)	1.77	1.87	1.87	1.99	2.91	3.03	3.13	3.25	4.35	4.54
CV	7e-16	0.10	0.10	0.19	0.10	6e-16	0.11	2e-16	0.22	2e-16
$\boldsymbol{\omega = 0.2}$										
Var. ($\times 10^3$)	1.77	1.93	1.94	1.97	4.63	4.83	5.12	5.34	6.59	6.83
CV	7e-16	0.15	0.15	0.19	0.17	3e-16	0.17	3e-16	0.21	4e-16
$\boldsymbol{\omega = 0.3}$										
Var. ($\times 10^3$)	1.77	1.96	1.96	1.95	6.80	7.06	7.59	7.87	8.81	9.09
CV	7e-16	0.18	0.18	0.188	0.20	6e-16	0.20	4e-16	0.20	3e-16
$\boldsymbol{\omega = 0.4}$										
Var. ($\times 10^3$)	1.77	1.96	1.95	1.95	9.27	9.57	10.4	10.7	11.0	11.32
CV	7e-16	0.18	0.18	0.18	0.20	3e-16	0.20	3e-16	0.20	2e-16
$\boldsymbol{\omega = 0.5}$										
Var. ($\times 10^3$)	1.77	1.95	1.95	1.94	11.9	12.3	13.3	13.6	13.2	13.5
CV	7e-16	0.19	0.19	0.18	0.20	3e-16	0.20	3e-16	0.20	2e-16
$\boldsymbol{\omega = 0.6}$										
Var. ($\times 10^3$)	1.77	1.94	1.94	1.94	14.6	15.0	16.1	16.5	15.3	15.67
CV	7e-16	0.19	0.19	0.19	0.20	3e-16	0.20	6e-16	0.20	5e-16

our approach, consider the first investment period. We use the data from 01–Jan-1998 to 31–Dec-1999 to calibrate our initial portfolios, and then we hold and observe the out-of-sample performance from 01–Jan–2000 to 30–Jun–2000. Afterwards, we roll the calibration window forward and recalibrate and rebalance our portfolios using the preceding two-year period (01–Jul–1998 to 30–Jun–2000). We then observe the out-of-sample performance from 01–Jul–2000 to 31–Dec–2000. We repeat these steps until the end of the investment horizon. Our out-of-sample experiment runs from 01–Jan–2000 until 31–Dec–2016, meaning we have a total of 34 six-month out-of-sample investment periods. We record the wealth evolution of the portfolios over the entire horizon. Finally, we note that this experiment is non-exhaustive since the portfolio performance is highly dependent on our choice of assets and historical time period. However, we hope that having a diverse basket of assets representative of major U.S. industries and a 17-year out-of-sample investment period will suffice for our analysis.

The first set of results, shown in Table 5 and Figure 3, correspond to risk parity portfolios with confidence level $\omega = 0.15, 0.3, 0.45$. The top plot in Figure 3 presents the total wealth evolution of the nominal portfolio plus robust portfolios with $\omega = 0.3$. We note that, from a total wealth perspective, these four portfolios behave quite similarly. This plot serves mostly to visualize the general trend over the 17-year investment period. The remaining three plots in Figure 3 present the relative wealth of the robust portfolios. The 'relative wealth' is defined as a percentage, $(W_i^t/W_{\text{nom}}^t - 1) \times 100$, where W_i^t is the wealth of portfolio i at each weekly time step t, while W_{nom}^t is the nominal portfolio's wealth.

The robust portfolios exhibit a drop in their relative wealth over the bear market periods of 2000–2003 and 2008–2009. However, we note that the risk parity portfolios are not designed to minimize a portfolio's risk, but rather to be fully risk diverse. In turn, this suggests that the robust portfolios are in a better position to take advantage of the subsequent bull market periods, where we can see sustained growth relative to the nominal. Moreover, this portfolio performance aligns with our findings from the in-sample experiment, where we saw that the robust portfolios had a somewhat higher risk appetite given that they had a higher total variance when compared to the nominal portfolio.

We summarize the portfolio ex post financial performance in Table 5, where we show the annualized average excess return, annualized volatility,[8] Sharpe ratio and average turnover rate over the entire investment horizon (2000–2016). Moreover, we also provide the results over a bear market period and its subsequent recovery (2007–2011). The results consistently show that the robust portfolios are able to attain a higher average excess return while maintaining a similar level of volatility, leading to higher Sharpe ratios. However, as the Sharpe ratio increases so does the average turnover rate, which serves as a proxy of transaction costs. With that said, transaction costs for large asset managers are very low in modern financial markets. Moreover, we note that the turnover rates of risk parity portfolios are typically very low when compared against other asset allocation strategies such as MVO (e.g., see [17]), and our results in 5 are no exception. Thus, the increased transaction costs of the robust portfolios are somewhat negligible. We note that our observations are consistent over the 2007–2011, with the robust portfolios having a higher Sharpe ratio over this time period.

Next, we present the out-of-sample results for the RRP variant of the problem. The results correspond to RRP portfolios with confidence level $\omega = 0.15, 0.3, 0.45$ and with the risk–return trade-off coefficient $\lambda = 0.25$. We note that we do not intend to compare the risk parity asset allocation strategy against the RRP strategy, nor do we intend to argue that one strategy is more favourable than the other one. The results are self-contained and are meant to be illustrative, and it is up to the investor to determine which strategy aligns best with their financial objective.

[8] The portfolio volatility is the square root of the ex post portfolio variance.

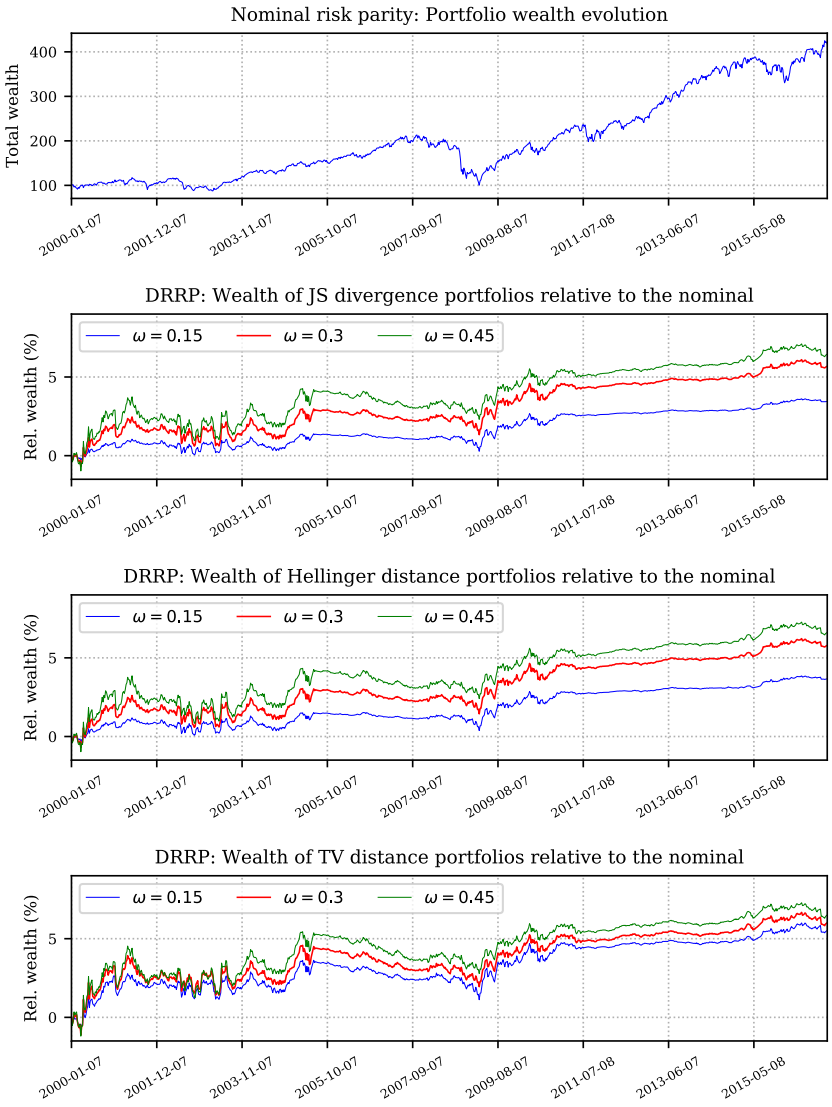

Fig. 3: Wealth evolution for DRRP portfolios. The top plot shows the total wealth evolution of the nominal portfolio. The remaining three plots present the relative wealth evolution of the DRRP portfolios with respect to the nominal for varying confidence levels.

Table 5: Summary of financial performance of the risk parity portfolios over the periods 2000–2016 and 2007–2011

	Nom.	JS			Hellinger			TV		
$\omega =$		0.15	0.3	0.45	0.15	0.3	0.45	0.15	0.3	0.45
2000 − 2016										
Return (%)	6.64	6.84	6.96	7.01	6.85	6.97	7.02	6.95	6.98	7.01
Vol. (%)	17.0	17.2	17.2	17.2	17.2	17.2	17.2	17.2	17.2	17.2
Sharpe Ratio	0.390	0.398	0.405	0.407	0.399	0.405	0.408	0.404	0.406	0.407
Turnover	0.100	0.126	0.151	0.167	0.128	0.152	0.168	0.160	0.172	0.180
2007 − 2011										
Return (%)	2.39	2.67	2.77	2.73	2.69	2.77	2.73	2.71	2.66	2.62
Vol. (%)	23.3	23.7	23.8	23.8	23.7	23.8	23.8	23.9	23.8	23.8
Sharpe Ratio	0.102	0.113	0.116	0.115	0.113	0.116	0.115	0.114	0.112	0.110
Turnover	0.098	0.123	0.150	0.166	0.125	0.151	0.166	0.162	0.172	0.177

Figure 4 presents the wealth evolution of the portfolios. As before, the top plot shows the total wealth evolution over the entire investment period, while the remaining three plots show the wealth of the robust portfolios relative to the nominal portfolio wealth. The relative wealth plots show a more pronounced behaviour than before, with a more pronounced drop during the 2008 financial crisis. However, these losses are reversed over the remainder of the investment period and the robust RRP portfolios are actually able to attain greater wealth relative to the nominal when compared against the risk parity results from before. We note that, although the robust RRP portfolios in Figure 4 may appear to be more volatile, the optics may be skewed since these plots measure the relative wealth. For example, if the nominal portfolio is very volatile while the robust portfolios are consistent, the corresponding relative wealth plot may appear to be quite erratic. Thus, we refer to the summary of financial performance in Table 6.

Overall, the results of the RRP portfolios align with the results of the risk parity portfolios, leading to similar observations and conclusions. As shown in Table 6, we can see that the robust portfolios consistently attain a higher Sharpe ratio when compared to the nominal portfolio. We observe a similar behaviour for the average turnover ratio, which also increases as the confidence level ω increases. These two trends are consistent over the 2007–2011 period. Again, the results suggest that, overall, the robust portfolios are able to attain a higher Sharpe ratio over a 17-year investment period.

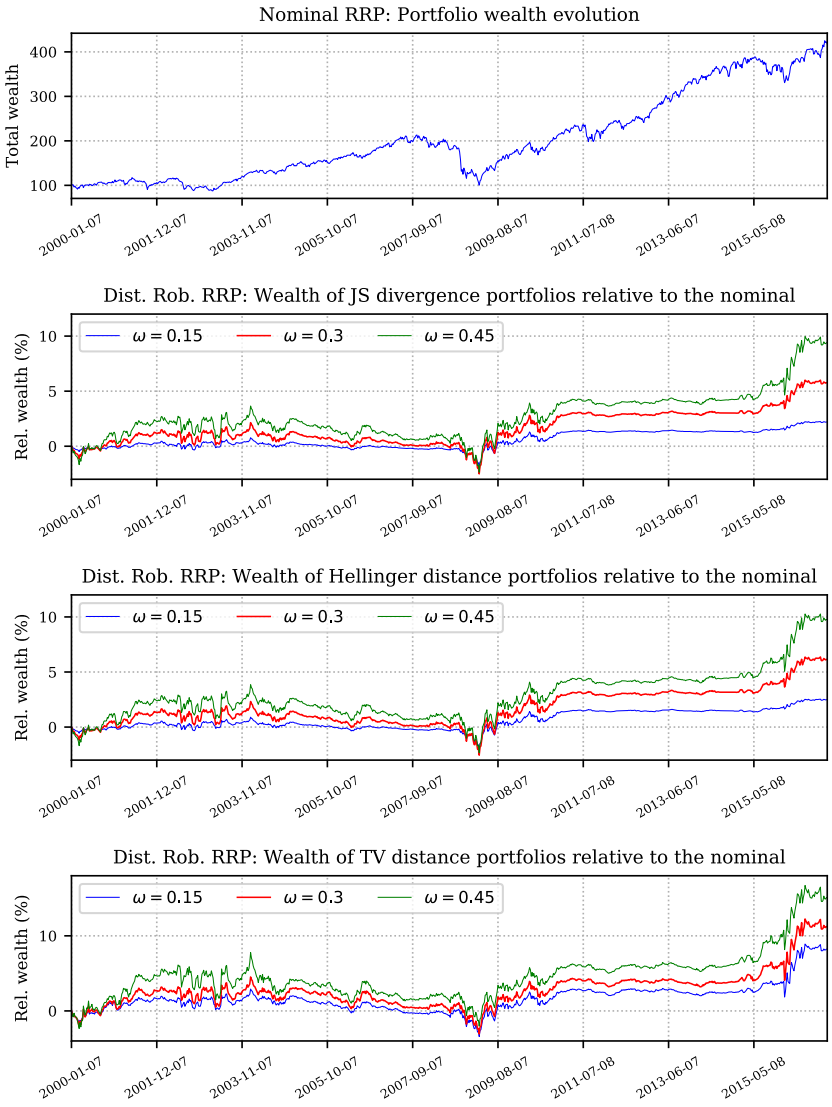

Fig. 4: Wealth evolution for distributionally robust RRP portfolios. The top plot shows the total wealth evolution of the nominal RRP portfolio. The remaining three plots present the relative wealth evolution of the distributionally robust RRP portfolios with respect to the nominal for varying confidence levels.

Table 6: Summary of financial performance of the RRP portfolios over the periods 2000–2016 and 2007–2011

	Nom.	JS			Hellinger			TV		
$\omega =$		0.15	0.3	0.45	0.15	0.3	0.45	0.15	0.3	0.45
2000 − 2016										
Return (%)	6.64	6.77	6.97	7.17	6.78	6.99	7.19	7.11	7.27	7.47
Vol. (%)	17.0	17.2	17.4	17.5	17.2	17.4	17.5	17.5	17.6	17.6
Sharpe Ratio	0.390	0.393	0.401	0.411	0.394	0.402	0.412	0.407	0.414	0.424
Turnover	0.101	0.124	0.158	0.191	0.127	0.162	0.195	0.193	0.233	0.279
2007 − 2011										
Return (%)	2.40	2.66	2.84	2.87	2.68	2.85	2.89	2.76	2.82	2.98
Vol. (%)	23.3	23.8	24.2	24.5	23.9	24.3	24.5	24.4	24.6	24.7
Sharpe Ratio	0.103	0.112	0.117	0.118	0.112	0.117	0.118	0.113	0.115	0.120
Turnover	0.098	0.125	0.156	0.182	0.127	0.160	0.185	0.180	0.205	0.225

5 Conclusion

This paper introduced a distributionally robust optimization problem to construct risk parity portfolios. In particular, distributional robustness is applied on the discrete probability distribution used for scenario-based estimation instead of the true (but latent) underlying continuous probability distribution of asset returns. Specifically, this discrete distribution pertains to the probabilistic weight attached to each scenario when estimating the asset expected returns and covariance matrix. This is a data-driven approach that deviates away from a nominal equally-weighted approach, and instead allows us to derive the most robust set of parameters implied by the data themselves.

In particular, robustness is introduced by finding the most adversarial instance of the discrete distribution such that our risk parity objective function is maximized within a predetermined probability ambiguity set. The ambiguity set is constrained by a statistical distance measure between the nominal and adversarial probabilities, and our modelling framework grants the user with the flexibility to choose any convex function to measure this distance. We exemplify this by implementing three alternative distance measures: the JS divergence, the Hellinger distance, and the TV distance.

Setting the adversarial probability distribution against our risk parity asset allocation strategy leads to a minimax problem. In this context, the investor seeks to construct a portfolio where the asset risk contributions are equalized, while the adversarial distribution seeks to find the worst-case instance of the variance corresponding to that specific portfolio.

Using the portfolio variance as the risk measure has the benefit of yielding a constrained convex–concave minimax problem over convex sets. We emphasize the application of projected gradient methods to optimize our minimax problem in both descent and ascent directions. The projections ensure that we retain feasibility after each iteration. However, iteratively moving in both directions is problematic as it may lead to instability and slow down convergence. Instead, we propose a novel algorithmic framework to solve our minimax optimization problem. This SCP–PGA algorithm exploits the strict convexity of the risk parity problem, which guarantees that we have a unique risk parity portfolio for each instance of the adversarial probability distribution. Thus, we aim to iteratively ascend in the probability space through PGA, while solving the corresponding convex risk parity problem after every iteration. The SCP–PGA algorithm dramatically improves computational runtime and, by design, retains the global convergence properties of general projected gradient methods.

Our numerical results show that the SCP–PGA algorithm is computationally tractable and scalable, even when solving problems with a large number of assets and scenarios. From a financial perspective, the resulting DRRP portfolios behave as expected. The asset weight allocation between the nominal and DRRP portfolios follows a similar pattern, emphasizing that they only differ by the statistical distance limit placed upon the adversarial distribution. Finally, our out-of-sample experiment shows that the DRRP portfolios are able to attain a higher risk-adjusted return when compared to the nominal portfolio.

The DRRP portfolio optimization problem can be adapted to solve other asset allocation problems, such as the RRP problem or MVO discussed in this paper, as well as other investment strategies which may benefit from distributional robustness. Moreover, the general design of the SCP–PGA algorithm should allow it to solve other types of constrained convex–concave minimax problems, including problems in other disciplines. These topics are the subject of ongoing and future research.

Declarations of Interest

The authors report no conflicts of interest.

References

1. Bai, X., Scheinberg, K., and Tütüncü, R. H. (2016). Least-squares approach to risk parity in portfolio selection. *Quantitative Finance*, 16(3):357–376.
2. Barzilai, J. and Borwein, J. M. (1988). Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148.
3. Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust optimization*, volume 28. Princeton University Press.

4. Ben-Tal, A. and Nemirovski, A. (1998). Robust convex optimization. *Mathematics of operations research*, 23(4):769–805.
5. Bertsekas, D. P. (1976). On the goldstein-levitin-polyak gradient projection method. *IEEE Transactions on automatic control*, 21(2):174–184.
6. Bertsimas, D. and Sim, M. (2004). The price of robustness. *Operations research*, 52(1):35–53.
7. Best, M. J. and Grauer, R. R. (1991). On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results. *The review of financial studies*, 4(2):315–342.
8. Birge, J. R. and Louveaux, F. (2011). *Introduction to stochastic programming*. Springer Science & Business Media.
9. Birgin, E. G., Martínez, J. M., and Raydan, M. (2000). Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10(4):1196–1211.
10. Breton, M. and El Hachem, S. (1995). Algorithms for the solution of stochastic dynamic minimax problems. *Computational Optimization and Applications*, 4(4):317–345.
11. Broadie, M. (1993). Computing efficient frontiers using estimated parameters. *Annals of Operations Research*, 45(1):21–58.
12. Calafiore, G. C. (2007). Ambiguous risk measures and optimal robust portfolios. *SIAM Journal on Optimization*, 18(3):853–877.
13. Calafiore, G. C. and El Ghaoui, L. (2006). On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(1):1–22.
14. Chen, X., Sim, M., and Sun, P. (2007). A robust optimization perspective on stochastic programming. *Operations Research*, 55(6):1058–1071.
15. Chopra, V. K. and Ziemba, W. T. (1993). The effect of errors in means, variances, and covariances on optimal portfolio choice. *Journal of Portfolio Management*, pages 6–11.
16. Costa, G. and Kwon, R. H. (2020a). Generalized risk parity portfolio optimization: An ADMM approach. *Journal of Global Optimization*.
17. Costa, G. and Kwon, R. H. (2020b). A regime-switching factor model for mean–variance optimization. *Journal of Risk*, 22(4):31–59.
18. Costa, G. and Kwon, R. H. (forthcoming). A robust framework for risk parity portfolios. *Journal of Asset Management*.
19. Dai, Y.-H. and Fletcher, R. (2005). Projected barzilai-borwein methods for large-scale box-constrained quadratic programming. *Numerische Mathematik*, 100(1):21–47.
20. Delage, E. and Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612.
21. Dunning, I., Huchette, J., and Lubin, M. (2017). Jump: A modeling language for mathematical optimization. *Society for Industrial and Applied Mathematics*, 59(2):295–320.
22. Dupačová, J. (1987). The minimax approach to stochastic programming and an illustrative application. *Stochastics: An International Journal of*

Probability and Stochastic Processes, 20(1):73–88.
23. Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860.
24. Fabozzi, F. J., Kolm, P. N., Pachamanova, D. A., and Focardi, S. M. (2007). Robust portfolio optimization. *Journal of Portfolio Management*, 33(3):40.
25. Frank, M., Wolfe, P., et al. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110.
26. French, K. R. (2020). Data library. `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. [Online; accessed 20-May-2020].
27. Fuglede, B. and Topsoe, F. (2004). Jensen-shannon divergence and hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.
28. Ghaoui, L. E., Oks, M., and Oustry, F. (2003). Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations research*, 51(4):543–556.
29. Goldfarb, D. and Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38.
30. Grippo, L., Lampariello, F., and Lucidi, S. (1986). A nonmonotone line search technique for newtons method. *SIAM Journal on Numerical Analysis*, 23(4):707–716.
31. Guastaroba, G., Mitra, G., and Speranza, M. G. (2011). Investigating the effectiveness of robust portfolio optimization techniques. *Journal of Asset Management*, 12(4):260–280.
32. Haugh, M., Iyengar, G., and Song, I. (2017). A generalized risk budgeting approach to portfolio construction. *Journal of Computational Finance*, 21(2):29–60.
33. Kim, S.-J. and Boyd, S. (2008). A minimax theorem with applications to machine learning, signal processing, and finance. *SIAM Journal on Optimization*, 19(3):1344–1367.
34. Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
35. Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.
36. Lobo, M. S. and Boyd, S. (2000). The worst-case risk of a portfolio. *Technical report. Available from http://web.stanford.edu/∼boyd/papers/pdf/risk_bnd.pdf.*
37. Maillard, S., Roncalli, T., and Teiletche, J. (2010). The properties of equally weighted risk contribution portfolios. *Journal of Portfolio Management*, 36(4):60–70.
38. Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1):77–91.
39. Mausser, H. and Romanko, O. (2014). Computing equal risk contribution portfolios. *IBM Journal of Research and Development*, 58(4):5–1.
40. Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of financial economics*, 8(4):323–361.

41. Nedić, A. and Ozdaglar, A. (2009). Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228.
42. Neumann, J. v. (1928). Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320.
43. Popescu, I. (2007). Robust mean-covariance solutions for stochastic optimization. *Operations Research*, 55(1):98–112.
44. Postek, K., den Hertog, D., and Melenberg, B. (2016). Computationally tractable counterparts of distributionally robust constraints on risk measures. *SIAM Review*, 58(4):603–650.
45. Qian, E. (2005). Risk parity portfolios: Efficient portfolios through true diversification. *Panagora Asset Management*.
46. Rachev, S. T. (1991). *Probability metrics and the stability of stochastic models*, volume 269. John Wiley & Son Ltd.
47. Rachev, S. T. and Römisch, W. (2002). Quantitative stability in stochastic programming: The method of probability metrics. *Mathematics of Operations Research*, 27(4):792–818.
48. Rahimian, H. and Mehrotra, S. (2019). Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.
49. Roncalli, T. (2015). Introducing expected returns into risk parity portfolios: A new framework for asset allocation. *Bankers, Markets & Investors*, (138):18–28.
50. Rustem, B. and Howe, M. (2009). *Algorithms for worst-case design and applications to risk management*. Princeton University Press.
51. Scarf, H. (1958). A min-max solution of an inventory problem. *Studies in the mathematical theory of inventory and production*, pages 201–209.
52. Shapiro, A. and Ahmed, S. (2004). On a class of minimax stochastic programs. *SIAM Journal on Optimization*, 14(4):1237–1249.
53. Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). *Lectures on stochastic programming: modeling and theory*. SIAM.
54. Shapiro, A. and Kleywegt, A. (2002). Minimax analysis of stochastic problems. *Optimization Methods and Software*, 17(3):523–542.
55. Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442.
56. Tütüncü, R. H. and Koenig, M. (2004). Robust asset allocation. *Annals of Operations Research*, 132(1-4):157–187.
57. Žáčková, J. (1966). On minimax solutions of stochastic linear programming problems. *Časopis pro pěstování matematiky*, 91(4):423–430.

Appendix A Numerical implementation of statistical distances

This section describes how to numerically implement the squared Hellinger distance in (22d) and the TV distance in (22e). We use either of these two distance measures to define the ambiguity set $\mathcal{U}_{\boldsymbol{p}}$, and then use the set to construct the corresponding Euclidean projection optimization problem in (31). However, in their current form, most optimization solvers will reject them.

If we wish to use the squared Hellinger distance in (22d), then the projection optimization problem can be implemented as follows.

$$\begin{aligned}
\min_{\boldsymbol{p},\boldsymbol{r}} \quad & \|\boldsymbol{u}-\boldsymbol{p}\|_2^2 \\
\text{s.t.} \quad & \mathbf{1}^T\boldsymbol{p} = 1 \\
& \frac{1}{2}\sum_{t=1}^{T} p_t - 2r_t\sqrt{q_t} + q_t \leq d_{\mathrm{H}}, \\
& p_t \geq r_t^2, \quad \text{for } t = 1,\dots,T \\
& \boldsymbol{p},\ \boldsymbol{r} \geq 0,
\end{aligned}$$

where $\boldsymbol{u} \in \mathbb{R}^T$ is some arbitrary vector that we wish to project onto the set $\mathcal{U}_{\boldsymbol{p}}$, while $\boldsymbol{r} \in \mathbb{R}^T$ is an auxiliary variable that serves as a placeholder for the square root of each element of $\boldsymbol{p}$. As before, $\boldsymbol{q} \in \mathcal{P}$ is the nominal probability distribution, while d_{H} is the maximum permissible distance in (26) and is defined by the number of scenarios T and the investor's subjective confidence level ω.

On the other hand, if we wish to use the TV distance in (22e), then the projection optimization problem can be implemented as follows.

$$\begin{aligned}
\min_{\boldsymbol{p},\boldsymbol{\zeta}} \quad & \|\boldsymbol{u}-\boldsymbol{p}\|_2^2 \\
\text{s.t.} \quad & \mathbf{1}^T\boldsymbol{p} = 1 \\
& \frac{1}{2}\sum_{t=1}^{T} \zeta_t \leq d_{\mathrm{TV}}, \\
& \zeta_t \geq p_t - q_t \quad \text{for } t = 1,\dots,T \\
& \zeta_t \geq q_t - p_t \quad \text{for } t = 1,\dots,T \\
& \boldsymbol{p} \geq 0,
\end{aligned}$$

where $\boldsymbol{\zeta} \in \mathbb{R}^T$ is an auxiliary variable that represents the absolute value of the difference between the elements of $\boldsymbol{p}$ and $\boldsymbol{q}$. The maximum permissible distance d_{TV} is defined by the number of scenarios T and the investor's subjective confidence level ω as shown in (27).