Data-driven distributionally robust risk parity portfolio optimization

Giorgio Costa[a] and Roy H. Kwon[a]

[a]Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, Ontario M5S 3G8, Canada

ARTICLE HISTORY

Compiled September 12, 2020

ABSTRACT

We propose a distributionally robust formulation of the traditional risk parity portfolio optimization problem. Distributional robustness is introduced by targeting the discrete probabilities attached to each observation used during parameter estimation. Instead of assuming that all observations are equally likely, we consider an ambiguity set that provides us with the flexibility to find the most adversarial probability distribution based on the investor's confidence level. This allows us to derive robust estimates to parametrize the distribution of asset returns without having to impose any particular structure on the data. The resulting distributionally robust optimization problem is a constrained convex–concave minimax problem. Our approach is financially meaningful and attempts to attain full risk diversification 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 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; Risk parity; Distributionally robust optimization; Statistical ambiguity; Saddle-point problem; Gradient descent

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 [32]. 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, 11, 34], leading to what is sometimes referred to as 'error maximization' given the

Giorgio Costa. Email: gcosta@mie.utoronto.ca
Roy H. Kwon. Email: rkwon@mie.utoronto.ca

poor out-of-sample performance of these (ex ante) optimal portfolios.

Accounting for uncertainty during optimization has become paramount in any decision-making problem where parameters are non-deterministic. If we have knowledge of the underlying probability distribution that governs these parameters, then we can formulate this optimization problem as a stochastic program [8, 41]. 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, which frames the problem deterministically by taking the most extreme estimates of our uncertain parameters within some confidence level[3, 4, 6]. Some examples of robust optimization in the context of portfolio selection are presented in [22, 25, 27, 30, 45].

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, a worst-case approach is taken, but with the distinction that we do this at the distributional level. Such a robust formulation for stochastic programs was proposed by Scarf [39]. Since then, this class of problems has often been referred to as *minimax* problems or, more recently, as distributionally robust optimization (DRO) problems [18]. A detailed survey paper on DRO is presented in [37].

The minimax problem has its roots in game theory [36]. 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 [10, 20, 40, 42, 47]. We note that minimax problems are sometimes referred to as saddle-point problems [28, 38] 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 minimax problems.

The main objective of this manuscript is to introduce a distributionally robust portfolio selection problem. Specifically, we address a portfolio selection strategy known as *risk parity*, which has gained popularity over the last decade among academics and practitioners. Risk parity seeks to construct a portfolio where the risk contribution of its constituent assets is equalized. In other words, each asset in the portfolio contributes the same level of risk towards the portfolio. Thus, by design, the risk parity problem is solely concerned with the portfolio risk measure, and does not necessitate the estimation of a reward measure. Maillard et al. [31] carefully explain how to partition a portfolio's variance to find the risk contribution per asset. Directly optimizing the problem with respect to the asset risk contributions leads to a non-convex optimization problem [1, 14], but some convex reformulations exist. For example, Mausser and Romanko [33] cast the risk parity problem as a second order cone program (SOCP), while Bai et al. [1] casts it as an unconstrained convex optimization problem. To address uncertainty in the estimated risk measure, Costa and Kwon [16] propose a robust risk parity framework built on the SOCP formulation, which takes the worst-case estimate of the risk measure but ignores any distributional information. For the purpose of this manuscript, we will use the convex risk parity problem from [1].

As shown by Calafiore [12], distributional robustness can be introduced into a portfolio selection problem by targeting the scenarios from which we derive our estimated parameters. When parameters are estimated from data, it is typically assumed that each scenario in the dataset is equally likely (i.e., we implicitly assume a uniform discrete probability distribution to describe the probability of each scenario). Instead, Calafiore [12] breaks this assumption and allows the scenarios to have different probabilities. In turn, this discrete probability distribution can be modelled as a set of decision variables, allowing us to design a maximization problem to find the most adversarial discrete probability distribution such that we attain the worst-case instance of the estimated parameters. Given that only the portfolio risk measure is pertinent for risk parity, this manuscript focuses solely on the derivation of the asset covariance matrix from data.

Addressing distributional robustness through a discrete probability distribution aligns naturally with a data-driven parameter estimation process. This assumes that market efficiency holds and that raw market data suffices to accurately represent the set of possible future returns. More importantly, this avoids making any assumptions about the underlying probability distribution of the asset returns, as well as avoiding assigning a structured process (such as a factor model) to model the returns. Thus, we are not required to impose a structure on the raw market data, which fully exempts us from the risk of model misspecification. This follows a similar rationale to another popular scenario-based portfolio risk measure known as historical value-at-risk, which assumes that raw market data suffices to represent the set of possible future outcomes. For the purpose of robustness in this manuscript, the application of a discrete probability distribution avoids the biases that could arise from assuming the returns have a specific structure, and provides us with the flexibility to derive a robust estimate of the asset covariance matrix implied by the raw market data themselves.

1.1. Contribution

This manuscript presents a distributionally robust risk parity (DRRP) portfolio optimization problem with a discrete probability ambiguity set on the portfolio risk measure. The introduction of distributional robustness through a discrete probability distribution in the same fashion as [12] allows us to design a minimax risk parity problem. Specifically, this minimax problem seeks to equalize the asset risk contributions against the worst-case instance of the portfolio variance.

The distributional ambiguity can be modelled as a convex set. This convex set is defined by constraints corresponding to the axioms of probability and, in particular, by a constraint that bounds the statistical distance between a nominal (i.e., assumed) probability distribution and its adversarial counterpart. The nominal distribution can be defined as any reasonable discrete probability distribution, but this amounts to a uniform distribution when we assume that all scenarios are equally likely. Thus, the adversarial distribution is allowed to deviate from the 'equally likely' nominal distribution by a maximum permissible limit defined by our choice of statistical distance measure and our confidence level.

Given the conditions of our problem, we are limited to statistical distance measures for discrete probability distributions. The statistical distance measure used in [12] was the Kullback–Leibler (KL) divergence. However, the KL divergence is not a proper distance metric,[1] making it difficult for an investor to appropriately define this distance based on a given confidence level. Thus, our manuscript focuses on statistical distance measures that satisfy the following two conditions: the measure must be a proper distance metric with finite bounds, and we must be able to formulate it as a computationally-tractable convex function. Therefore, the distributional ambiguity set is predominantly defined by our choice of distance measure and confidence level, which in turn defines our distributional robustness. Specifically, this manuscript discusses the following three statistical distance measures: 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.

This manuscript models the nominal risk parity problem 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 statistical distance from the nominal distribution. Our modelling framework gives the user the flexibility to choose their preferred measure of statistical distance, provided this can be modelled as a convex function. The result is a constrained minimax problem, which we can solve with a projected gradient method [5]. Specifically, we present a projected gradient descent–ascent (PGDA) algorithm that alternates between the descent and ascent steps to reach the saddle point. The standard approaches to solve constrained minimax problems are projection-type methods [35, 46].

We proceed by introducing a novel algorithm to solve the DRRP minimax problem that is grounded in projected gradient descent and sequential convex programming. The standard PGDA algorithm requires that we take alternating descent and ascent steps as we move towards the saddle point of our problem. Such an approach typically requires double the number of design parameters when compared to algorithms that move in a single direction. Moreover, these design parameters must be defined by the user a priori (e.g., initial point, step sizes). Finally, iterating 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 iteratively through gradient ascent in the probability space while solving a risk parity minimization problem in the asset weight space after every iteration. We can interpret our proposed algorithm as an implementation of sequential convex programming (SCP), where we ascend in the probability space towards the most adversarial instance of the portfolio risk measure after every iteration using a projected gradient ascent (PGA) method. Thus, we refer to our proposed algorithm as SCP–PGA. Compared to the PGDA algorithm, each iteration of the SCP–PGA algorithm 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. Exploiting the convex minimization step at each iteration simplifies the algorithmic development since we are only required to iteratively ascend in the probability space while maintaining the risk parity condition

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

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. Specifically, the DRRP portfolio is able to attain a higher risk-adjusted rate of return when compared to the nominal risk parity portfolio.

In summary, our contributions are the following. First, we introduce the DRRP problem, which seeks risk parity with respect to the most adversarial estimate of the portfolio risk measure through a purely data-driven process (i.e, the probabilistic ambiguity is implied by the data themselves). Second, we explicitly define how to construct this ambiguity set using different statistical distance metrics and we show how to use an investor's confidence level to size ambiguity set. Finally, we propose the SCP–PGA algorithm to solve the resulting DRRP minimax problem. We note that the flexible structure of the SCP–PGA algorithm means that it may be applied to solve other portfolio selection problems, as well as other types of constrained convex–concave minimax problems from other disciplines.

1.2. 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. 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 problem'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.

1.3. Notation

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 $\mathbb{S}^n_+$. If we need to reference some specific element i within a vector $\boldsymbol{z}$, we we denote this as z_i. If we define a vector as the product between a matrix and a vector, $\boldsymbol{Az} \in \mathbb{R}^m$, then we reference its i^{th} element as $[\boldsymbol{Az}]_i$. Finally, the ℓ_p-norm of an arbitrary vector $\boldsymbol{z} \in \mathbb{R}^n$ is denoted by $\| \cdot \|_p$, where $\|\boldsymbol{z}\|_p \triangleq (\sum_{i=1}^n |z_i|^p)^{1/p}$.

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 [32], 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 distribution with first and second moments defined as the expected returns $\boldsymbol{\mu} \in \mathbb{R}^n$ and 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 true moments are assumed to be latent and are typically estimated from data, meaning they are prone to suffer from estimation error [7, 13, 34].

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 the joint probability distribution of 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 $\boldsymbol{\xi}$). In a similar fashion to [12], we assume there exists some probability p_t associated with each scenario t. In vector notation, this is the probability mass function $\boldsymbol{p} \in \mathcal{P}$, 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.

If we have knowledge of $\boldsymbol{p}$, then we can statistically estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Let $\hat{\boldsymbol{\xi}}_t \in \mathbb{R}^n$ be the t^{th} scenario of the dataset $\hat{\boldsymbol{\xi}}$. The first two moments are

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

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

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.

Estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in this fashion means we are not required to impose any particular structure to model the latent asset returns distribution, avoiding the any biases and errors arising from model misspecification. Our only assumption is that market efficiency holds, meaning we can derive adequate estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ directly from a given raw market dataset $\hat{\boldsymbol{\xi}}$.

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 $\hat{\boldsymbol{\mu}}(\boldsymbol{p})$ and $\hat{\boldsymbol{\Sigma}}(\boldsymbol{p})$ from (5) and (6). We break down the derivation as follows. Assume we have a portfolio $\boldsymbol{x}$. For a given dataset $\hat{\boldsymbol{\xi}}$, 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

$$\begin{aligned} \hat{\mu}_\pi(\boldsymbol{x}, \boldsymbol{p}) &\triangleq \boldsymbol{x}^\top \hat{\boldsymbol{\mu}}(\boldsymbol{p}) && \text{(7a)} \\ &= \boldsymbol{p}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x}), && \text{(7b)} \end{aligned}$$

$$\begin{aligned} \hat{\sigma}_\pi^2(\boldsymbol{x}, \boldsymbol{p}) &\triangleq \boldsymbol{x}^\top \hat{\boldsymbol{\Sigma}}(\boldsymbol{p}) \boldsymbol{x} && \text{(8a)} \\ &= \mathbb{E}\Big[\big(\pi(\boldsymbol{x}) - \mathbb{E}[\pi(\boldsymbol{x})]\big)^2\Big] = \mathbb{E}[\pi^2(\boldsymbol{x})] - \big(\mathbb{E}[\pi(\boldsymbol{x})]\big)^2 \\ &= \boldsymbol{p}^\top \hat{\boldsymbol{\pi}}^2(\boldsymbol{x}) - \boldsymbol{p}^\top \hat{\boldsymbol{\Theta}}(\boldsymbol{x}) \boldsymbol{p}, && \text{(8b)} \end{aligned}$$

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.

2.2. Risk parity

Risk parity is a modern asset allocation strategy that aims to construct a portfolio where every asset contributes the same amount of risk. Thus, risk parity is fully diversified from a risk perspective. In turn, the risk parity problem is only concerned with financial risk, and does not require a measure of financial reward during optimization.

As shown in [31], 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,

[2] We note that $\hat{\boldsymbol{\Sigma}}(\boldsymbol{q})$, where $q_t = 1/T$ for $t = 1, \ldots, T$, yields the standard scenario-based estimate of the covariance matrix. To recover the *unbiased* estimate of the covariance matrix, we should multiply $\hat{\boldsymbol{\Sigma}}(\boldsymbol{q})$ by $T/(T-1)$. However, this distinction has no effect for the purpose of this paper.

the portfolio standard deviation can be partitioned as follows

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

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 [14], 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,

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

where $R_i \triangleq x_i[\boldsymbol{\Sigma x}]_i$ 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 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

$$[\boldsymbol{\Sigma y}^{\mathrm{RP}}]_i = \frac{\kappa}{y_i^{\mathrm{RP}}} \ \forall \ i \qquad \Longleftrightarrow \qquad y_i^{\mathrm{RP}} [\boldsymbol{\Sigma y}^{\mathrm{RP}}]_i = \kappa \ \forall \ i.$$

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 $\boldsymbol{y}^{\mathrm{RP}}$ is an admissible portfolio. However, we can recover the optimal risk parity portfolio $\boldsymbol{x}^{\mathrm{RP}}$ as follows

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

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

[3] In theory, we can assign any positive value to κ without loss of generality. In practice, we should avoid assigning extremely large or small values to κ to avoid numerical instability.

parity can be formulated as a strictly convex optimization problem with a unique global solution. Other equivalent convex formulations can be found in [31] and [33]. However, once short sales are allowed the problem becomes non-convex and the uniqueness of our solution is no longer guaranteed [1, 14]. 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 $\mathbf{\Sigma}$. In practice, we can use the estimated covariance matrix $\hat{\mathbf{\Sigma}}(\boldsymbol{p})$ 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,

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

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

$$\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}) \triangleq \Pi_{\mathcal{X}}\big(\boldsymbol{y}^{\mathrm{RP}}(\boldsymbol{p})\big), \tag{13c}$$

where $f_{\mathrm{RP}} : \mathbb{R}_{+}^{n} \times \mathcal{P} \to \mathbb{R}$ is our risk parity objective function, while $\Pi_{\mathcal{X}}\big(\boldsymbol{y}^{\mathrm{RP}}(\boldsymbol{p})\big)$ is the projection of the vector $\boldsymbol{y}^{\mathrm{RP}}(\boldsymbol{p})$ 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 $\boldsymbol{y}^{\mathrm{RP}}(\boldsymbol{p})$. Therefore, the projection onto the set $\mathcal{X}$ only needs to enforce the budget equality constraint. Finally, the risk parity portfolio for an arbitrary instance of $\boldsymbol{p} \in \mathcal{P}$ is the vector $\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p})$.

For some arbitrary vector $\boldsymbol{z} \in \mathbb{R}_{+}^{n}$, the projection is

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

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 an optimal risk parity portfolio for any estimate of the covariance matrix $\hat{\mathbf{\Sigma}}(\boldsymbol{p})$ with respect to any instance of $\boldsymbol{p} \in \mathcal{P}$.

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. Our immediate goal is twofold: to design an appropriate ambiguity set $\mathcal{U}_{\boldsymbol{p}}$ for our adversarial probability distribution $\boldsymbol{p}$, and to formulate the DRRP minimax problem. We address these two issues 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 an 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 $\mathcal{U}_{\boldsymbol{p}}$, which we proceed to formally define. 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

$$\mathcal{U}_{\boldsymbol{p}}(\boldsymbol{q}, d) \triangleq \left\{\boldsymbol{p} \in \mathcal{P} : D(\boldsymbol{p}, \boldsymbol{q}) \leq d\right\} \tag{14}$$

where $D(\boldsymbol{p}, \boldsymbol{q})$ is a convex function that models a given statistical distance measure, while $d \in \mathbb{R}_+$ is a user-defined bound on the maximum permissible distance between $\boldsymbol{p}$ and $\boldsymbol{q}$. We note that, by definition, $\mathcal{U}_{\boldsymbol{p}} \subseteq \mathcal{P}$.

A statistical distance measure can be used to quantify the similarity between two probability distributions. We limit our choice of statistical distance measures to a subset of convex functions that operate on discrete distributions.

The distributionally robust portfolio selection problem in [12] used the KL divergence to define the ambiguity set. However, the KL divergence is not a proper metric since it is not symmetric and does not respect the triangle inequality. For two discrete probability distributions $\boldsymbol{p},\ \boldsymbol{q} \in \mathcal{P}$, the KL divergence is defined as

$$D_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \sum_{t=1}^{T} p_t \ln\left(\frac{p_t}{q_t}\right) \tag{15}$$

The asymmetry of the KL divergence becomes apparent if we reverse the order of the arguments $\boldsymbol{p}$ and $\boldsymbol{q}$ (i.e., $D_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{q}) \neq D_{\mathrm{KL}}(\boldsymbol{q}, \boldsymbol{p})$). Moreover, the upper bound of the KL divergence is not properly defined, making it difficult to define an appropriate maximum permissible distance between $\boldsymbol{p}$ and $\boldsymbol{q}$.

A measure closely related to the KL divergence is the JS divergence, which was introduced by Lin [29]. Unlike the KL divergence, the JS divergence is symmetric and has finite bounds. The JS divergence is defined as

$$\begin{aligned} D_{\mathrm{JS}}(\boldsymbol{p}, \boldsymbol{q}) &\triangleq \frac{1}{2} D_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{m}) + \frac{1}{2} D_{\mathrm{KL}}(\boldsymbol{q}, \boldsymbol{m}) \\ &= \frac{1}{2} \sum_{t=1}^{T} p_t \ln(p_t) + q_t \ln(q_t) - (p_t + q_t) \ln\left(\frac{p_t + q_t}{2}\right), \end{aligned} \tag{16}$$

where $\boldsymbol{m} = \frac{1}{2}(\boldsymbol{p} + \boldsymbol{q}) \in \mathcal{P}$. Given that our definition of the KL divergence in (15) uses the natural logarithm, our definition of the JS divergence has the useful property of being bounded between zero and ln(2), i.e.,

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

We can derive a proper metric from the JS divergence by taking its square root, which is known as the JS distance [21, 24] (i.e., the JS distance is $\sqrt{D_{\mathrm{JS}}(\boldsymbol{p}, \boldsymbol{q})}$). This distance measure is bounded between zero and $\sqrt{\ln(2)}$.

Next, we present the square of the Hellinger distance

$$D_{\mathrm{H}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \frac{1}{2} \sum_{t=1}^{T} \left(\sqrt{p_t} - \sqrt{q_t} \right)^2. \tag{17}$$

The Hellinger distance is a proper distance metric and, by its definition, is bounded between zero and one. We define $D_{\mathrm{H}}(\boldsymbol{p}, \boldsymbol{q})$ as the squared Hellinger distance to improve computational tractability in practice.

The last distance measure we discuss is the TV distance,

$$D_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \frac{1}{2} \|\boldsymbol{p} - \boldsymbol{q}\|_1, \tag{18}$$

which is a proper distance metric. The TV distance is bounded between zero and one.

The JS, Hellinger and TV distances are proper metrics and have finite bounds, which will allow us to define a maximum permissible distance d between $\boldsymbol{q}$ and $\boldsymbol{p}$. Moreover, (16–18) 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. We note that the functions (17) and (18) can be implemented computationally by introducing auxiliary variables during optimization, but this does not fundamentally alter the problem. An example of how to computationally implement them is shown in Appendix A.

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 define the nominal probability distribution as a discrete uniform distribution, i.e., $\boldsymbol{q} \triangleq [1/T \ \cdots \ 1/T]^\top \in \mathbb{R}^T$. 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. The distance measures in (16–18) 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 theoretical upper bounds are unattainable under the assumption that $\boldsymbol{q}$ is a discrete uniform distribution. Consider the following example of an extreme probability distribution, $\boldsymbol{s} \triangleq [1 \ 0 \ \cdots \ 0]^\top \in \mathbb{R}^T$, 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 $B(T) \triangleq D(\boldsymbol{s}, \boldsymbol{q}) \in \mathbb{R}_+$, 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

$$B_{\mathrm{JS}}(T) \triangleq D_{\mathrm{JS}}(\boldsymbol{s}, \boldsymbol{q}), \tag{19a}$$

$$B_{\mathrm{H}}(T) \triangleq D_{\mathrm{H}}(\boldsymbol{s}, \boldsymbol{q}), \tag{19b}$$

$$B_{\mathrm{TV}}(T) \triangleq D_{\mathrm{TV}}(\boldsymbol{s}, \boldsymbol{q}). \tag{19c}$$

For example, if our data consist of ten scenarios ($T = 10$), then the upper bounds of the measures in (16–18) are

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

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.

We define an appropriate maximum permissible distance d between our nominal and adversarial distributions based on the upper bound B 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., $D(\boldsymbol{p}, \boldsymbol{q}) \leq d_{\mathrm{JS}}$).

Recall that, in the case of the JS distance, 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

$$d_{\mathrm{JS}}(\omega, T) \triangleq \omega^2 B_{\mathrm{JS}}(T). \tag{20}$$

Similarly, since $D_{\mathrm{H}}(\boldsymbol{p}, \boldsymbol{q})$ in (17) is defined as the square of the Hellinger distance, the maximum permissible distance is

$$d_{\mathrm{H}}(\omega, T) \triangleq \omega^2 B_{\mathrm{H}}(T). \tag{21}$$

Finally, given that the TV distance in (18) is already a proper metric, we define the maximum permissible distance as

$$d_{\mathrm{TV}}(\omega, T) \triangleq \omega B_{\mathrm{TV}}(T). \tag{22}$$

To properly define the ambiguity set $\mathcal{U}_{\boldsymbol{p}}$ in (14), the we must choose a single distance measure and define $D(\boldsymbol{p}, \boldsymbol{q})$ as one of the options in (16–18), with d defined accordingly.

3.2. Minimax problem

For a given dataset $\hat{\boldsymbol{\xi}}$, our problem is defined by the investor's choice of statistical distance measure and confidence level. 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, 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, we use $\boldsymbol{x}^*$ to denote the portfolio resulting 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 the variables of our minimax problem 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, we can restate our risk parity objective function in two equivalent forms

$$f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p}) \triangleq \frac{1}{2}\boldsymbol{y}^{\top}\hat{\boldsymbol{\Sigma}}(\boldsymbol{p})\boldsymbol{y} - \kappa \sum_{i=1}^{n} \ln(y_i) \tag{23a}$$

$$\triangleq \frac{1}{2}\left(\boldsymbol{p}^{\top}\hat{\boldsymbol{\pi}}^{2}(\boldsymbol{y}) - \boldsymbol{p}^{\top}\hat{\boldsymbol{\Theta}}(\boldsymbol{y})\boldsymbol{p}\right) - \kappa \sum_{i=1}^{n} \ln(y_i), \tag{23b}$$

where (23a) is exactly the same as (13a) and is restated for clarity, while (23b) presents $f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p})$ explicitly in terms of $\boldsymbol{p}$. 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

$$\min_{\boldsymbol{y} \in \mathbb{R}^{n}_{+}} \max_{\boldsymbol{p} \in \mathcal{U}_{\boldsymbol{p}}} \quad f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p}). \tag{24}$$

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

$$f_{\mathrm{RP}}(\boldsymbol{y}^{*}, \boldsymbol{p}) \leq f_{\mathrm{RP}}(\boldsymbol{y}^{*}, \boldsymbol{p}^{*}) \leq f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p}^{*}) \;\; \forall\; \boldsymbol{y} \in \mathbb{R}^{n}_{+},\; \boldsymbol{p} \in \mathcal{U}_{\boldsymbol{p}}, \tag{25}$$

where $(\boldsymbol{y}^{*}, \boldsymbol{p}^{*})$ is the optimal solution (i.e., the saddle point) of $f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p})$.

The maximization step in (24) is also meaningful in a financial context. Consider the definition of $f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p})$ in (23b) where, without loss of generality, we have defined the portfolio variance using the unnormalized proxy variable $\boldsymbol{y}$. The maximization step in (24) 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 (24) seeks the optimal risk parity portfolio with respect to the worst-case portfolio variance.

[4] Projecting the auxiliary variable $\boldsymbol{y} \in \mathbb{R}^{n}_{+}$ onto $\mathcal{X}$ is a trivial step, as shown in (12).

3.3. Projected gradient descent–ascent

We can exploit the convex–concave structure of our minimax problem to search for 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,

$$\begin{aligned} \Pi_{\mathcal{U}_{\boldsymbol{p}}}(\boldsymbol{u}) \triangleq \quad & \underset{\boldsymbol{p}}{\text{argmin}} \quad \|\boldsymbol{u} - \boldsymbol{p}\|_2^2 && \text{(26a)} \\ & \text{s.t.} \quad \mathbf{1}^T \boldsymbol{p} = 1, && \text{(26b)} \\ & \quad D(\boldsymbol{p}, \boldsymbol{q}) \leq d, && \text{(26c)} \\ & \quad \boldsymbol{p} \geq 0, && \text{(26d)} \end{aligned}$$

where the constraints (26b–26d) arise from the ambiguity set $\mathcal{U}_{\boldsymbol{p}}(\boldsymbol{q}, d)$. In particular, constraint (26c) is shown with respect to a generic distance measure $D(\boldsymbol{p}, \boldsymbol{q})$, which can be defined by the user as any of the measures in (16–18) with an appropriate maximum permissible distance d. Thus, for some point $\boldsymbol{u}$, the projection $\Pi_{\mathcal{U}_{\boldsymbol{p}}}(\boldsymbol{u})$ 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

$$\Pi_{\mathbb{R}^n_+}(\boldsymbol{z}) \triangleq \begin{cases} z_i & \text{if } z_i > 0 \\ 0^+ & \text{otherwise} \end{cases} \quad \text{for } i = 1, \ldots, n, \tag{27}$$

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

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

$$\nabla_{\boldsymbol{y}} f_{\text{RP}}(\boldsymbol{y}, \boldsymbol{p}) = \hat{\boldsymbol{\Sigma}}(\boldsymbol{p})\boldsymbol{y} - \kappa \boldsymbol{y}^{-1}, \tag{28}$$

$$\nabla_{\boldsymbol{p}} f_{\text{RP}}(\boldsymbol{y}, \boldsymbol{p}) = \frac{1}{2}\hat{\boldsymbol{\pi}}^2(\boldsymbol{y}) - \hat{\boldsymbol{\Theta}}(\boldsymbol{y})\boldsymbol{p} \tag{29}$$

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

[5] Any non-positive value must be replaced with a strictly positive value, 0^+ due to the logarithm barrier term in our objective function. Thus,0^+ can be set to some small positive value during implementation.

have some feasible solutions $\boldsymbol{y}^k \in \mathbb{R}^n_+$ and $\boldsymbol{p}^k \in \mathcal{U}_{\boldsymbol{p}}$. The search directions are

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

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

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

The points $\boldsymbol{y}^{k+1}$ and $\boldsymbol{p}^{k+1}$ are the result of linear combinations between two feasible points in each set, respectively. Since the sets are convex, the points $\boldsymbol{y}^{k+1}$ and $\boldsymbol{p}^{k+1}$ are feasible by definition.

We defer to the Barzilai–Borwein method [2] to define the step sizes α_k and γ_k. Specifically, we use the following definition of the Barzilai–Borwein method. For any iteration $k \geq 1$ where $k = 0, 1, \ldots$, we have

$$\alpha_k = \frac{\Big|(\boldsymbol{y}^k - \boldsymbol{y}^{k-1})^\top \big(\nabla_{\boldsymbol{y}} f_{\mathrm{RP}}(\boldsymbol{y}^k, \boldsymbol{p}^k) - \nabla_{\boldsymbol{y}} f_{\mathrm{RP}}(\boldsymbol{y}^{k-1}, \boldsymbol{p}^{k-1})\big)\Big|}{\Big\|\nabla_{\boldsymbol{y}} f_{\mathrm{RP}}(\boldsymbol{y}^k, \boldsymbol{p}^k) - \nabla_{\boldsymbol{y}} f_{\mathrm{RP}}(\boldsymbol{y}^{k-1}, \boldsymbol{p}^{k-1})\Big\|_2^2}, \tag{30}$$

$$\gamma_k = \frac{\Big|(\boldsymbol{p}^k - \boldsymbol{p}^{k-1})^\top \big(\nabla_{\boldsymbol{p}} f_{\mathrm{RP}}(\boldsymbol{y}^k, \boldsymbol{p}^k) - \nabla_{\boldsymbol{p}} f_{\mathrm{RP}}(\boldsymbol{y}^{k-1}, \boldsymbol{p}^{k-1})\big)\Big|}{\Big\|\nabla_{\boldsymbol{p}} f_{\mathrm{RP}}(\boldsymbol{y}^k, \boldsymbol{p}^k) - \nabla_{\boldsymbol{p}} f_{\mathrm{RP}}(\boldsymbol{y}^{k-1}, \boldsymbol{p}^{k-1})\Big\|_2^2}. \tag{31}$$

The Barzilai–Borwein step size is sometimes referred to as the ‘spectral step size’. In the case of projected gradient descent, this class of algorithms is sometimes referred to as ‘spectral projected gradient descent’ [9].

Next, we discuss how to determine the search parameters $\eta_{\boldsymbol{y}}, \eta_{\boldsymbol{p}} \in (0,1]$. Specifically, we favour the non-monotone Grippo–Lampariello–Lucidi (GLL) line search proposed in [26]. 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, 17].

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

$$f_{\mathrm{RP}}(\boldsymbol{y}^k + \eta_{\boldsymbol{y}} \boldsymbol{g}^k, \boldsymbol{p}^k) \leq \max_{j \in \mathcal{J}} f_{\mathrm{RP}}(\boldsymbol{y}^{k-j}, \boldsymbol{p}^{k-j}) + \beta \eta_{\boldsymbol{y}} (\boldsymbol{g}^k)^\top \nabla_{\boldsymbol{y}} f_{\mathrm{RP}}(\boldsymbol{y}^k, \boldsymbol{p}^k) \tag{32}$$

where $\mathcal{J} \triangleq \big\{j \in \mathbb{Z} : 0 \leq j \leq \min\{k, m-1\}\big\}$ and $\beta \in (0,1)$ is some predefined constant. Intuitively, a larger value of $\eta_{\boldsymbol{y}}$ corresponds to a more aggressive descent step. Thus, we can 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 take greedier steps. For example, if we set $m = 1$, then we revert back to a traditional

Armijo-type line search method and the condition in (32) 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 the PGDA algorithm, 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, 17], we purposely avoid claiming that this is true for PGDA. However, we note that for appropriate step sizes, the convergence of constrained convex–concave minimax problems has been previously established [35]. For the purpose of this manuscript, the PGDA algorithm serves as a stepping stone towards the development of the SCP–PGA algorithm.

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.

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 $\varepsilon_0$; 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 (20–22)
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 $\alpha_k$ as shown in (30)
        Update $\gamma_k$ as shown in (31)
    $\boldsymbol{g}^k = \Pi_{\mathbb{R}^n_+}\Big(\boldsymbol{y}^k - \alpha_k \nabla_{\boldsymbol{y}} f_{\mathrm{RP}}(\boldsymbol{x}^k, \boldsymbol{p}^k)\Big) - \boldsymbol{y}^k$
    $\eta_{\boldsymbol{y}} = 1$
    $\bar{\boldsymbol{y}} = \boldsymbol{y}^k + \eta_{\boldsymbol{y}} \boldsymbol{g}^k$
    while $f_{RP}(\bar{\boldsymbol{y}}, \boldsymbol{p}^k) > \max_{j \in \mathcal{J}} f_{RP}(\boldsymbol{y}^{k-j}, \boldsymbol{p}^{k-j}) + \beta \eta_{\boldsymbol{y}} (\boldsymbol{g}^k)^\top \nabla_{\boldsymbol{y}} f_{RP}(\boldsymbol{x}^k, \boldsymbol{p}^k)$ do
        $\eta_{\boldsymbol{y}} = \eta_{\boldsymbol{y}} \tau$
        $\bar{\boldsymbol{y}} = \boldsymbol{y}^k + \eta_{\boldsymbol{y}} \boldsymbol{g}^k$
    $\boldsymbol{y}^{k+1} = \bar{\boldsymbol{y}}$
    $\boldsymbol{h}^k = \Pi_{\mathcal{U}_{\boldsymbol{p}}}\Big(\boldsymbol{p}^k + \gamma_k \nabla_{\boldsymbol{p}} f_{\mathrm{RP}}(\boldsymbol{x}^k, \boldsymbol{p}^k)\Big) - \boldsymbol{p}^k$
    $\eta_{\boldsymbol{p}} = 1$
    $\bar{\boldsymbol{p}} = \boldsymbol{p}^k + \eta_{\boldsymbol{p}} \boldsymbol{h}^k$
    while $f_{RP}(\boldsymbol{y}^k, \bar{\boldsymbol{p}}) < \min_{j \in \mathcal{J}} f_{RP}(\boldsymbol{y}^{k-j}, \boldsymbol{p}^{k-j}) + \beta \eta_{\boldsymbol{p}} (\boldsymbol{h}^k)^\top \nabla_{\boldsymbol{p}} f_{RP}(\boldsymbol{x}^k, \boldsymbol{p}^k)$ do
        $\eta_{\boldsymbol{p}} = \tau \eta_{\boldsymbol{p}}$
        $\bar{\boldsymbol{p}} = \boldsymbol{p}^k + \eta_{\boldsymbol{p}} \boldsymbol{h}^k$
    $\boldsymbol{p}^{k+1} = \bar{\boldsymbol{p}}$
    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 $\boldsymbol{x}^*$
```

In (25) we stated that the following inequality holds for the saddle point $(\boldsymbol{y}^*, \boldsymbol{p}^*)$ given that $f_{\mathrm{RP}}(\boldsymbol{y}, \boldsymbol{p})$ is convex–concave,

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

As per the risk parity problem in (13), we also have that the saddle point $(\boldsymbol{y}^*, \boldsymbol{p}^*)$ corresponds to $\boldsymbol{y}^* = \boldsymbol{y}^{\mathrm{RP}}(\boldsymbol{p}^*)$. Moreover, projecting $\boldsymbol{y}^*$ onto $\mathcal{X}$ yields the corresponding optimal DRRP portfolio $\boldsymbol{x}^*$, which leads to the following conclusion: $\boldsymbol{x}^* = \boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^*)$. The SCP–PGA algorithm enforces the risk parity condition during every iteration k, i.e., we have $\boldsymbol{x}^k = \boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^k)$. Thus, at convergence, we reach the same conclusion as before: $\boldsymbol{x}^* = \boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^*)$. 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 $f_{\mathrm{RP}}\big(\boldsymbol{x}^{\mathrm{RP}}(\cdot), \cdot\big) : \mathcal{U}_{\boldsymbol{p}} \to \mathbb{R}$ while maintaining the risk parity condition during every iteration.

Since a closed-form solution to $\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p})$ does not exist, we must proceed iteratively by solving $\boldsymbol{x}^k = \boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^k)$ in every iteration k. Although this increases the computational cost per iteration when compared against the PGDA algorithm, we note that convex optimization problems can be efficiently solved by modern optimization algorithms and software packages. Thus, as we will show numerically in Section 4, the additional computational cost per iteration is almost negligible. Moreover, the SCP–PGA algorithm needs less iterations until convergence.

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. Moreover, 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 use of 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 complete this subsection by presenting the SCP–PGA algorithm in Algorithm 2.

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 experiment tests the out-of-sample financial performance of the DRRP portfolio.

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 [23]. Table 1 lists the 30 industry portfolios.

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 script

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 $\varepsilon_0$; 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 (20–22)
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 $\gamma_k$ as shown in (31)
    $\boldsymbol{x}^k = \boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^k)$
    $\boldsymbol{h}^k = \Pi_{\mathcal{U}_{\boldsymbol{p}}}\Big(\boldsymbol{p}^k + \gamma_k \nabla_{\boldsymbol{p}} f_{\mathrm{RP}}(\boldsymbol{x}^k, \boldsymbol{p}^k)\Big) - \boldsymbol{p}^k$
    $\eta_{\boldsymbol{p}} = 1$
    $\bar{\boldsymbol{p}} = \boldsymbol{p}^k + \eta_{\boldsymbol{p}} \boldsymbol{h}^k$
    while $f_{RP}(\boldsymbol{x}^k, \bar{\boldsymbol{p}}) < \min_{j \in \mathcal{J}} f_{RP}(\boldsymbol{x}^{k-j}, \boldsymbol{p}^{k-j}) + \beta \eta_{\boldsymbol{p}} (\boldsymbol{h}^k)^\top \nabla_{\boldsymbol{p}} f_{RP}(\boldsymbol{x}^k, \boldsymbol{p}^k)$ 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 $\boldsymbol{x}^*$
```

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		

was written using the Julia programming language (version 1.4.0) with the modelling language 'JuMP' [19] and with IPOPT (version 3.12.6) as the optimization solver.

4.1. Numerical performance and tractability

The first part of the numerical performance experiment compares the SCP–PGA algorithm to the PGDA algorithm. The second part extends the results of the SCP–PGA algorithm and compares the results against the nominal risk parity problem.

The experimental setup to compare the SCP–PGA and PGDA algorithms is the following. We randomly generate synthetic datasets with $n = 50$, 200 assets and $T =$ 1,000, 5,000 scenarios, meaning there are 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 optimal 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 $i = 1, \dots, n$.

The SCP–PGA and PGDA algorithms are assessed 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 algorithm appears to converge if the step sizes become ill-conditioned and not enough progress is made in the probability space (i.e., the PGDA algorithm terminates because $\varepsilon \leq \varepsilon_0$ after the step in the $\boldsymbol{p}$-direction becomes negligible). 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 at the DRRP portfolio variance, we must keep in mind that this corresponds to the worst-case estimate of the variance as defined by

[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 $\boldsymbol{x}^{\mathrm{RP}}(\boldsymbol{p}^*)$ after convergence in $\boldsymbol{p}$.

Table 2.: Comparison of numerical performance between the PGDA algorithm (A.1) and the SCP–PGA algorithm (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

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.

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' [43], 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 from a portfolio variance perspective.

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

We begin by 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,

[7] This may exclude high frequency trading environments.

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

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 that of the nominal portfolio, with the DRRP portfolios exhibiting a more even distribution of wealth.

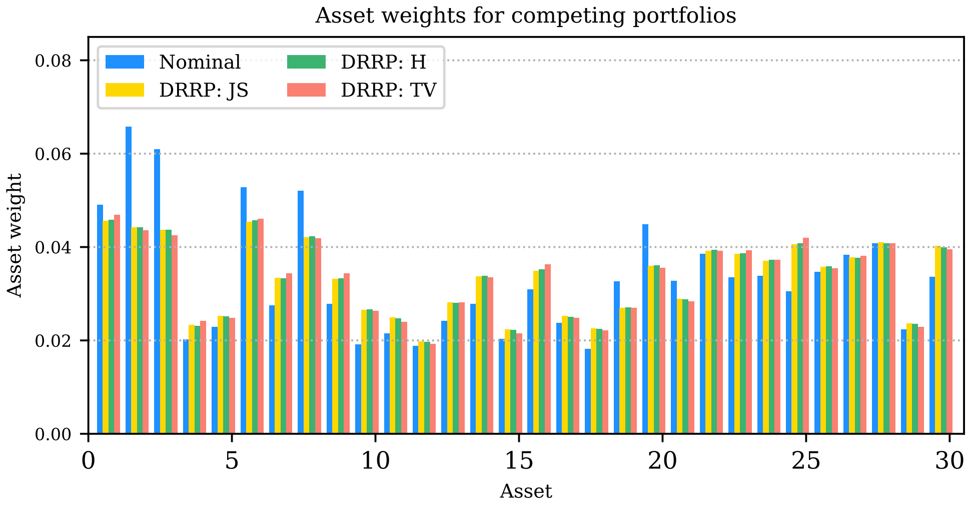

Figure 1.: Asset weights of the nominal and DRRP portfolios with $\omega = 0.3$.

Figure 2 presents a similar analysis, except 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 $R_i \triangleq x_i[\hat{\mathbf{\Sigma}}\boldsymbol{x}]_i$ for some estimate $\hat{\mathbf{\Sigma}}$. 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, $\mathbf{\Sigma}^{\text{nom}} \triangleq \mathbf{\Sigma}(\boldsymbol{q})$. The remaining three plots compares the risk contributions of the nominal portfolio with respect to $\mathbf{\Sigma}^{\text{JS}}$, $\mathbf{\Sigma}^{\text{H}}$ and $\mathbf{\Sigma}^{\text{TV}}$, respectively. These three matrices correspond to the estimated covariance matrix obtained after the 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 from a DRRP portfolio is higher than from the nominal portfolio while simultaneously lower from 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, except we use varying levels of confidence ω. For brevity, these results are summarized 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 corresponding worst-case estimates of the covariance matrix. In addition, we report the level of risk concentration through the coefficient of variation (CV) of the

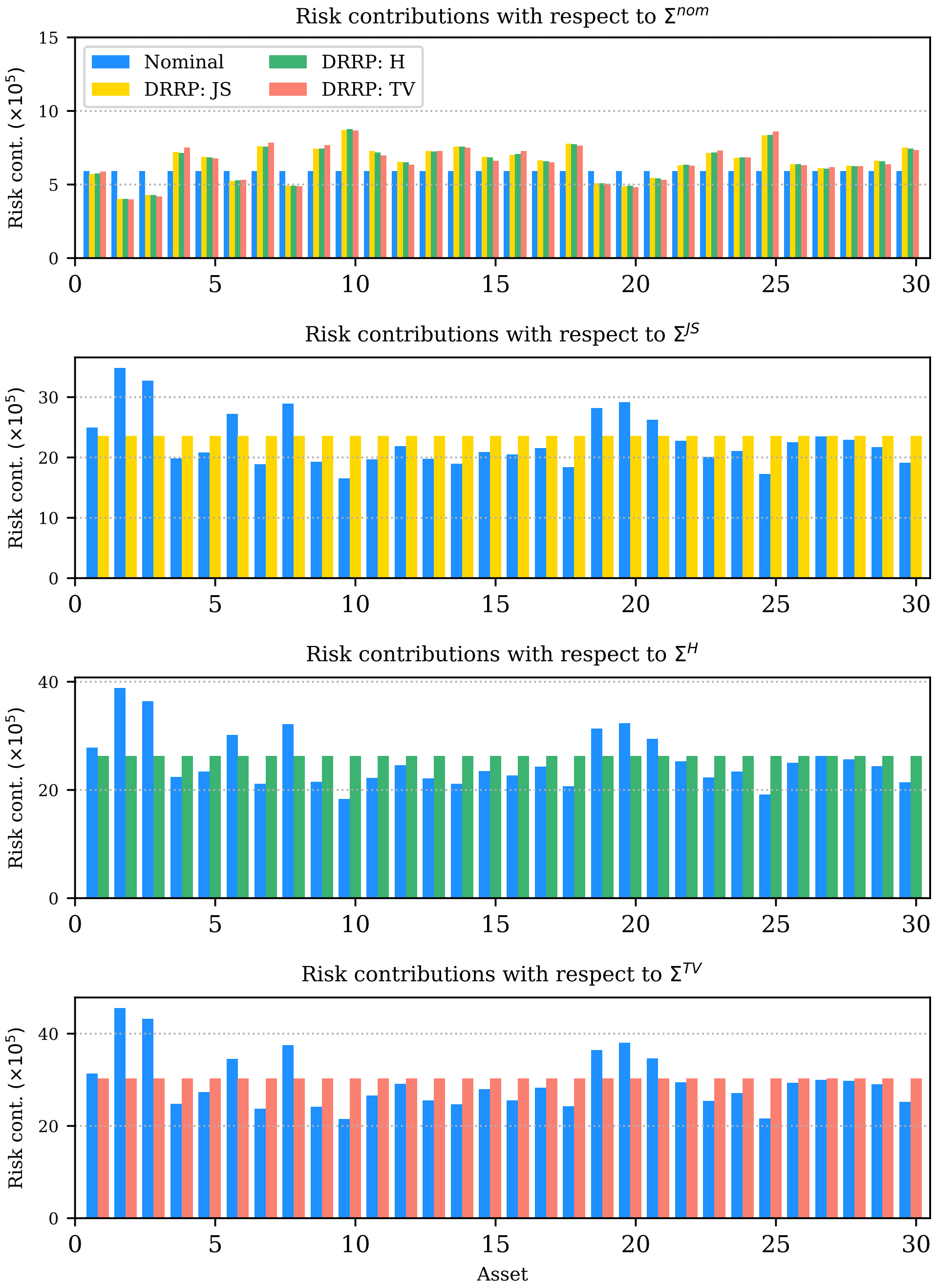

Figure 2.: Risk contributions per asset of the nominal and DRRP portfolios with $\omega = 0.3$. The top plot shows the risk contributions with respect to the nominal estimate of the covariance matrix. The remaining plots show the risk contributions of the nominal and DRRP portfolios based on the robust estimates of the covariance matrix.

risk contributions. The CV is calculated by taking the standard deviation of the risk contributions and dividing them by their average, i.e.,

$$\mathrm{CV} = \frac{\mathrm{SD}\big(\boldsymbol{x} \circ [\hat{\boldsymbol{\Sigma}}\boldsymbol{x}]\big)}{\frac{1}{n}\boldsymbol{x}^{\top}\boldsymbol{\Sigma}\boldsymbol{x}}, \tag{33}$$

where '$\circ$' 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.

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

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

An overview of the out-of-sample experimental setup follows. Our portfolio constituents are the 30 assets listed in Table 1 ($n = 30$). The dataset consists of weekly historical returns from 01–Jan–1998 to 31–Dec–2016, with the data obtained from [23]. 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 them. All estimated parameters and weights are recalibrated every time we rebalance our portfolios. To exemplify 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, having a diverse basket of assets representative of major U.S. industries and a 17-year out-of-sample investment period should 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 shows the total wealth evolution of the nominal portfolio. The remaining three plots show 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, the results suggest 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, the ex post 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 ex ante variance when compared to the nominal portfolio.

We summarize the ex post performance in Table 5, where we show the annualized average excess return, annualized volatility,[8] Sharpe ratio [44] and average turnover rate over the entire investment horizon (2000–2016). We also provide the subset of results corresponding to 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 are becoming increasingly negligible 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 [15]), and our results in 5 are no exception. Thus, the increased transaction costs incurred by

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

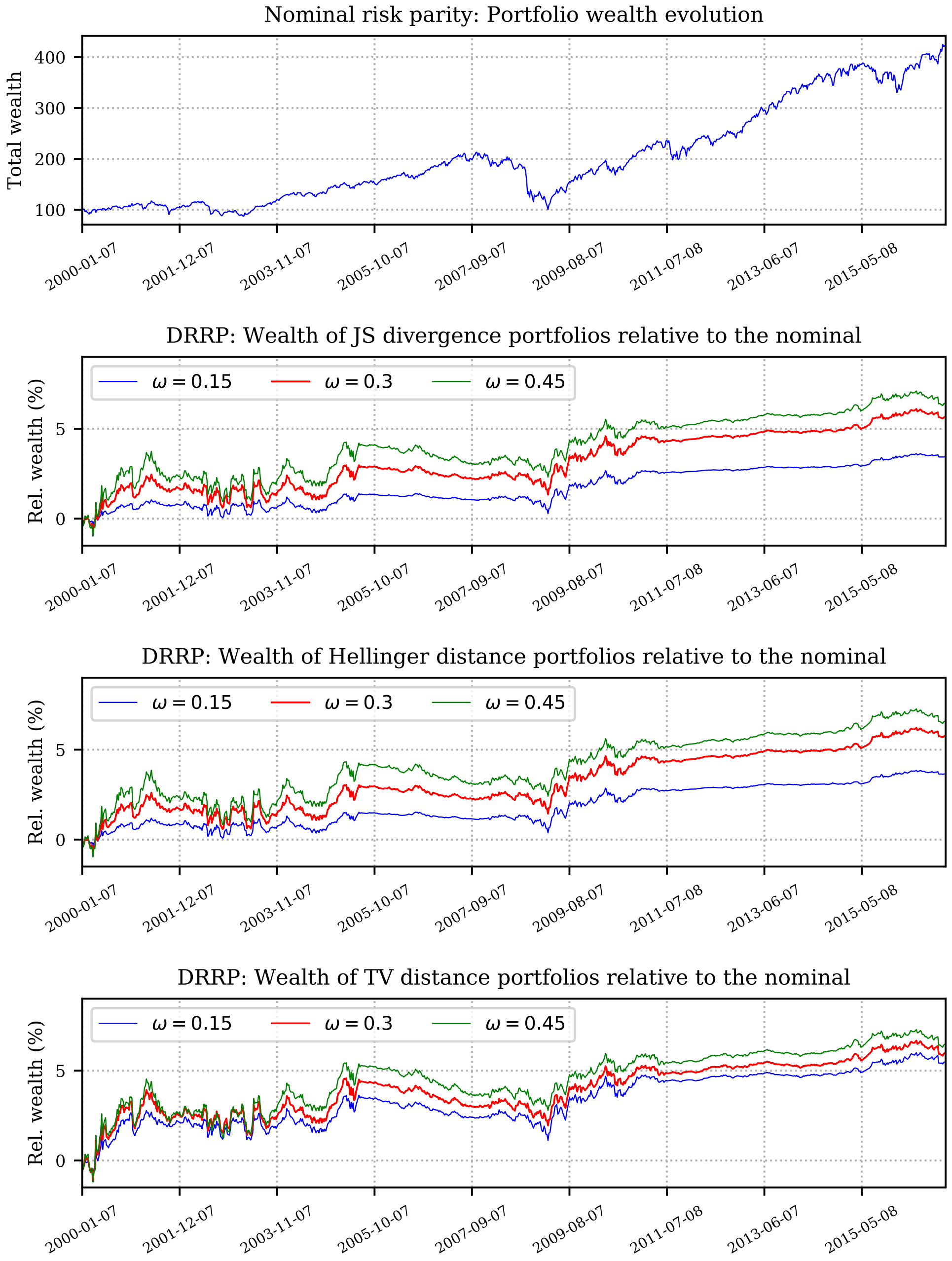

Figure 3.: Wealth evolution for DRRP portfolios. The top plot shows the total wealth evolution of the nominal portfolio. The remaining three plots present the relative wealth evolution of the DRRP portfolios with respect to the nominal portfolio for varying confidence levels.

the DRRP portfolios are somewhat negligible. Finally, we note that our observations are consistent over the 2007–2011, with the robust portfolios having a higher Sharpe ratio over this time period when compared to the nominal.

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

5. Conclusion

This paper introduced a DRO problem specifically designed for risk parity portfolios. Distributional robustness is introduced through a discrete probability distribution that allows us to break away from the assumption that all scenarios in a data-driven parameter estimation process are equally likely. Instead, we can model the probability attached to each as a decision variable, which in turn allows us to formulate a minimax problem that seeks risk parity while simultaneously seeking the most adversarial instance of the discrete distribution such that the portfolio variance is maximized. Our modelling framework allows us to define the probability ambiguity set using any convex function to measure this statistical distance. We exemplify this by implementing three alternative statistical distances: JS, Hellinger, and TV.

The DRRP problem is a constrained convex–concave minimax problem over convex sets. We apply projected gradient methods to iterate over the DRRP problem in both descent and ascent directions. The projections ensure that we retain feasibility after each iteration. However, iteratively moving in both directions may lead to instability and slow convergence. Instead, we propose a novel algorithmic framework to solve our DRRP problem. The proposed 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. From a financial perspective, our experiments show that a DRRP portfolio is able to attain a higher risk-adjusted return when compared to the nominal portfolio.

Finally, we note that the DRRP problem can be adapted to solve other portfolio selection problems that 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 future research.

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] 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.

[14] Costa, G. and Kwon, R. H. (2020a). Generalized risk parity portfolio optimization: An ADMM approach. *Journal of Global Optimization*, 78:207–238.

[15] Costa, G. and Kwon, R. H. (2020b). A regime-switching factor model for mean–variance optimization. *Journal of Risk*, 22(4):31–59.

[16] Costa, G. and Kwon, R. H. (2020c). A robust framework for risk parity portfolios. *Journal of Asset Management.*

[17] Dai, Y.-H. and Fletcher, R. (2005). Projected barzilai-borwein methods for large-scale box-constrained quadratic programming. *Numerische Mathematik*, 100(1):21–47.

[18] Delage, E. and Ye, Y. (2010). Distributionally robust optimization under mo-

ment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612.

[19] 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.

[20] 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.

[21] Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860.

[22] Fabozzi, F. J., Kolm, P. N., Pachamanova, D. A., and Focardi, S. M. (2007). Robust portfolio optimization. *Journal of Portfolio Management*, 33(3):40.

[23] French, K. R. (2020). Data library. `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. [Online; accessed 20-May-2020].

[24] 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.

[25] Goldfarb, D. and Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38.

[26] 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.

[27] 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.

[28] 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.

[29] Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

[30] 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.*

[31] Maillard, S., Roncalli, T., and Teiletche, J. (2010). The properties of equally weighted risk contribution portfolios. *Journal of Portfolio Management*, 36(4):60–70.

[32] Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1):77–91.

[33] Mausser, H. and Romanko, O. (2014). Computing equal risk contribution portfolios. *IBM Journal of Research and Development*, 58(4):5–1.

[34] Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of financial economics*, 8(4):323–361.

[35] Nedić, A. and Ozdaglar, A. (2009). Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228.

[36] Neumann, J. v. (1928). Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320.

[37] Rahimian, H. and Mehrotra, S. (2019). Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659.*

[38] Rustem, B. and Howe, M. (2009). *Algorithms for worst-case design and applications to risk management.* Princeton University Press.

[39] Scarf, H. (1958). A min-max solution of an inventory problem. *Studies in the mathematical theory of inventory and production*, pages 201–209.

[40] Shapiro, A. and Ahmed, S. (2004). On a class of minimax stochastic programs. *SIAM Journal on Optimization*, 14(4):1237–1249.

[41] Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). *Lectures on stochastic programming: modeling and theory.* SIAM.

[42] Shapiro, A. and Kleywegt, A. (2002). Minimax analysis of stochastic problems. *Optimization Methods and Software*, 17(3):523–542.

[43] Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442.

[44] Sharpe, W. F. (1994). The sharpe ratio. *Journal of Portfolio Management*, 21(1):49–58.

[45] Tütüncü, R. H. and Koenig, M. (2004). Robust asset allocation. *Annals of Operations Research*, 132(1-4):157–187.

[46] Xiu, N. and Zhang, J. (2003). Some recent advances in projection-type methods for variational inequalities. *Journal of Computational and Applied Mathematics*, 152(1-2):559–585.

[47] 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

Here we describe how to numerically implement the squared Hellinger distance in (17) and the TV distance in (18). 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 (26). However, in their current form, most optimization solvers will reject them.

If we wish to use the squared Hellinger distance in (17), then the projection optimization problem can be implemented as follows.

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

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 (21) 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 (18), then the projection optimization problem can be implemented as follows.

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

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 (22).