SCRIP: Successive Convex Optimization Methods for Risk Parity Portfolio Design

Yiyong Feng and Daniel P. Palomar, *Fellow, IEEE*

***Abstract*—The traditional Markowitz portfolio optimization proposed in the 1950s has not been embraced by practitioners despite its theoretical elegance. Recently, an alternative risk parity portfolio design has been receiving significant attention from both the theoretical and practical sides due to its advantage in diversification of (ex-ante) risk contributions among assets. Such risk contributions can be deemed good predictors for the (ex-post) loss contributions, especially when there exist huge losses. Most of the existing specific problem formulations on risk parity portfolios are highly nonconvex and are solved via standard off-the-shelf numerical optimization methods, e.g., sequential quadratic programming and interior point methods. However, for nonconvex risk parity formulations, such standard numerical approaches may be highly inefficient and may not provide satisfactory solutions. In this paper, we first propose a general risk parity portfolio problem formulation that can fit most of the existing specific risk parity formulations, and then propose a family of simple and efficient successive convex optimization methods for the general formulation. The numerical results show that our proposed methods significantly outperform the existing ones.**

***Index Terms*—Efficient sequential algorithms, risk budgeting, risk parity, successive convex optimization.**

I. Introduction

SINCE the mean-variance portfolio optimization framework was introduced by Markowitz over fifty years ago [1]–[3], it has been well-researched in the academic field. However, practitioners have not embraced such a nice theoretical framework. There are several reasons. The first reason is that this approach relies on the estimates of expected returns and the obtained portfolio is highly unstable. This may be overcome partially by introducing additional constraints on the portfolio. Another severe drawback of the Markowitz portfolio is that this approach tends to provide an excessively concentrated portfolio and risk over a few assets, which goes against the common sense of diversification as a way to reduce the risk. During normal times this may not cause a serious issue, but if a financial crisis were to happen, such a concentrated portfolio would probably incur huge losses.

Manuscript received December 04, 2014; revised May 14, 2015; accepted June 21, 2015. Date of publication July 01, 2015; date of current version September 03, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dmitry Malioutov. This work was supported by the Hong Kong RGC 617312 research grant.

The authors are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (e-mail: yiyong@ust.hk; palomar@ust.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2015.2452219

The new paradigm of risk parity portfolio was precisely introduced to make the portfolio, and hence the risk, truly more diversified. Qian [4], [5] first showed that uniform risk contributions actually lead to a diverse enough portfolio, and the (ex-ante) risk contributions (i.e., the risks computed using historical data) are not only a mathematical measurement of how diverse the risk is, but also good indicators of the (ex-post) loss contributions of the assets (i.e., the observed risks and losses in the future), especially when there exist large losses. According to this observation, the way to avoid a potential huge loss is to distribute the risk contributions. This method has been receiving significant attention recently, especially after the 2008 financial crisis [6], and it has been shown that a risk parity portfolio is also more robust than the Markowitz portfolio [7].

By taking volatility as the risk measurement, Maillard *et al.* [8] first analyzed the properties of the equal risk contribution portfolio and showed that it is actually a trade-off between the minimum variance and equally weighted portfolios. To find the risk parity portfolio, they formulated one logarithmic constrained convex problem for the long-only risk parity portfolio and one nonconvex problem for the general risk parity portfolio. Following [8], there were some works exploring different extensions. A number of empirical experiments were conducted in [9] and it was found that the risk parity portfolio does have several interesting characteristics, e.g., balanced risk allocation and less (ex-post) volatile performance characteristics (e.g., Sharpe ratio) over time. Later, Bai *et al.* [10] considered a slight variation of the problem formulation that appears in [8] by simplifying the objective function. At the same time, the risk parity portfolio was extended to the risk budgeting portfolio [11], and the group risk parity portfolio [10], [12]. Instead of focusing on the assets directly, the risk factor model was introduced into the risk parity portfolio formulation [12], [13], and then the risk was diversified among the underlying risk factors. Apart from focusing on risk only, the expected returns were also incorporated into the problem formulation [14]. The aforementioned works mainly took volatility as a measure of risk, more realistic measures of risk including Value-at-Risk (VaR) and Conditional-VaR (CVaR) were considered in [15]–[17] in the context of the risk parity. Meanwhile, the robustness of the risk parity was studied in [18] and it was found that incorporating some constraints on the risk parity portfolio generates an improvement in the out-of-sample performance [13]. A recent book [7] serves as a good summary on the state-of-the-art.

We need to point out that in the previous reviewed literature most of the formulations for the risk parity portfolio are nonconvex. To compute such risk parity portfolio, usually

traditional off-the-shelf nonlinear optimization methods, like sequential quadratic programming (SQP) [19] and interior point methods (IPM) [20] built in the MATLAB function `fmincon`, are used in practice [8], [10]–[12], [17], [21]. However, for the nonconvex risk parity problem, in general they are time consuming and sometimes may not even converge globally [10], [17], [21]. There exists one improved algorithm based on the alternative linearization method [10], but it is too specific and only works for the problem in [10]. For the special long-only risk parity portfolio, there exist several ad-hoc methods. Chaves *et al.* [22] first proposed a Newton-based efficient algorithm, however, it only works when there are no constraints on the portfolio, which is unrealistic in practice where constraints always exist. Later, two different efficient algorithms, i.e., Newton-Nesterov (NN) method [23] and Cyclical Coordinate Descent (CCD) method [21], were proposed and they work in a similar fashion.

To the best of our knowledge, there does not exist any numerically efficient method for the general risk parity portfolio problem. The main contribution of this paper is first to propose a framework general enough to characterize most of the existing specific risk parity problems taking volatility, Gaussian VaR or Gaussian CVaR as risk measurements and then to provide a family of efficient algorithms having the following good properties: i) provable convergence to a stationary point; ii) able to cope with any kind of portfolio constraints (e.g., both long-only and general long/short constraints); and iii) much better performance compared with benchmarks for the general long/short portfolio and similar performance for the long-only portfolio.

This paper is organized as follows. Section II briefly reviews the concepts of risk contribution and risk parity/budget portfolio. Section III proposes the general risk parity portfolio problem formulation with connections to the existing specific formulations. Then Section IV presents the proposed efficient solving approach. Section V considers the detailed explorations for some specific cases, and Section VI extends the approach by considering more alternative approximations. At last, Section VII provides some numerical results and Section VIII concludes the paper.

Notation: We use a boldface lower letter for vectors $\mathbf{a}$, and upper case letter for matrices $\mathbf{A}$. The notation $\mathbf{1}$ denotes all-one column vector with proper size. The transpose, Moore-Penrose pseudo-inverse, and inverse operators are denoted by the symbols $(\cdot)^T$, $(\cdot)^\dagger$, and $(\cdot)^{-1}$, respectively. $\mathrm{Diag}(\mathbf{A})$ denotes a diagonal matrix with diagonal elements equal to that of $\mathbf{A}$, and its principal square root is denoted by $\mathrm{Diag}^{1/2}(\mathbf{A})$. The scalar c denotes a constant with proper value.

II. Risk Parity/Budgeting Portfolio

A. Risk Contribution

Suppose there are n assets with random returns $\mathbf{r} \in \mathbb{R}^n$, and the mean vector and (positive definite) covariance matrix are denoted as $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$. We use $\mathbf{w} \in \mathbb{R}^n$ to denote the normalized portfolio (e.g., $\mathbf{w}^T\mathbf{1} = 1$), which describes how the total capital budget is to be allocated over the assets. To study the risk parity portfolio, we need some well defined risk measurement $f(\mathbf{w})$ so that the "risk contribution" of each asset to the risk of the whole portfolio can be quantified. The following desired property is always introduced in risk parity literature.

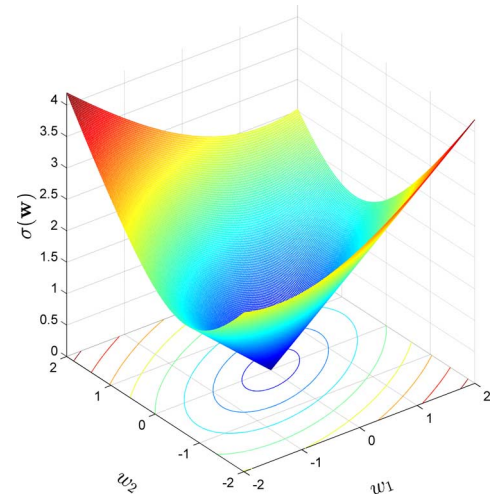


Fig. 1. One example that satisfies Euler property (1).

Theorem 1 (Euler's Theorem): Let a continuous and differentiable function $f: \mathbb{R}^n \mapsto \mathbb{R}$ be a positively homogeneous function of degree one[1]*. Then*

$$f(\mathbf{w}) = \sum_{i=1}^{n} w_i \frac{\partial f}{\partial w_i}. \tag{1}$$

One observation from property (1) is that the component $w_i \frac{\partial f}{\partial w_i}$ can be regarded as the risk contribution from asset i to the total risk $f(\mathbf{w})$.

Interestingly and fortunately, most of the existing risk measurements do satisfy the Euler property (1) either directly (VaR and CVaR) or indirectly (variance) (see the next subsection). Fig. 1 shows an example of $f(\mathbf{w}) = \sigma(\mathbf{w}) = \sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}$ (see (2) later) and we can see that the function is linear along any direction starting from the origin.

B. Risk Measurements Satisfying the Euler Property

1) Volatility: Note that variance $\sigma^2(\mathbf{w}) = \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$ does not satisfy (1) directly. Fortunately, it is easy to check that volatility $\sigma(\mathbf{w}) = \sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}$ does satisfy (1)

$$\begin{aligned}\sum_{i=1}^{n} w_i \frac{\partial \sigma}{\partial w_i} &= \sum_{i=1}^{n} w_i \left(\frac{\boldsymbol{\Sigma}\mathbf{w}}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}}\right)_i \\ &= \frac{1}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}} \sum_{i=1}^{n} w_i (\boldsymbol{\Sigma}\mathbf{w})_i \\ &= \frac{1}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}} \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} = \sigma(\mathbf{w}).\end{aligned} \tag{2}$$

Thus variance fits (1) indirectly via volatility.

2) VaR and CVaR: For any portfolio $\mathbf{w} \in \mathbb{R}^n$, the definition of VaR and CVaR are given as follows:

$$\mathrm{VaR}_{1-\varepsilon}(\mathbf{w}) = \min\left\{\gamma : P\left(\gamma \le -\mathbf{r}^T\mathbf{w}\right) \le \varepsilon\right\}, \tag{3}$$

$$\mathrm{CVaR}_{1-\varepsilon}(\mathbf{w}) = \mathsf{E}\left[-\mathbf{r}^T\mathbf{w} \middle| -\mathbf{r}^T\mathbf{w} \ge \mathrm{VaR}_{1-\varepsilon}(\mathbf{w})\right], \tag{4}$$

where $-\mathbf{r}^T\mathbf{w}$ denotes the (random) loss of portfolio $\mathbf{w}$.

[1]A function $f(\mathbf{w})$ is a positively homogeneous function of degree one if $f(c\mathbf{w}) = cf(\mathbf{w})$ holds for any constant $c > 0$.

It was shown in [24] that $\mathrm{VaR}_{1-\varepsilon}(\mathbf{w})$ is a linear homogeneous function of weight $\mathbf{w}$ and the VaR contribution of the i-th asset is

$$w_i \frac{\partial \mathrm{VaR}_{1-\varepsilon}(\mathbf{w})}{\partial w_i} = \mathsf{E}\left[-r_i w_i | - \mathbf{r}^T\mathbf{w} = \mathrm{VaR}_{1-\varepsilon}(\mathbf{w})\right], \quad (5)$$

thus $\mathrm{VaR}_{1-\varepsilon}(\mathbf{w})$ satisfies property (1).

Similarly, Scaillet [25] showed that the CVaR contribution of the i-th asset is

$$w_i \frac{\partial \mathrm{CVaR}_{1-\varepsilon}(\mathbf{w})}{\partial w_i} = \mathsf{E}\left[-r_i w_i | - \mathbf{r}^T\mathbf{w} \geq \mathrm{VaR}_{1-\varepsilon}(\mathbf{w})\right], \quad (6)$$

thus $\mathrm{CVaR}_{1-\varepsilon}(\mathbf{w})$ satisfies property (1) as well.

In practice, risk contribution expressions (5) and (6) are not used because they are not numerically computable in general. Fortunately, there exist several ways to compute them either exactly or approximately.

Gaussian Case: For the Gaussian distribution, VaR and CVaR can be expressed explicitly as [26]

$$\mathrm{VaR}_{1-\varepsilon}(\mathbf{w}) = -\boldsymbol{\mu}^T\mathbf{w} + \kappa_1(\varepsilon)\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}, \quad (7)$$

$$\mathrm{CVaR}_{1-\varepsilon}(\mathbf{w}) = -\boldsymbol{\mu}^T\mathbf{w} + \kappa_2(\varepsilon)\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}, \quad (8)$$

where $\kappa_1(\varepsilon) \triangleq Q^{-1}(\varepsilon)$ and $\kappa_2(\varepsilon) \triangleq \frac{1}{\sqrt{2\pi}\varepsilon} e^{-\frac{\left(Q^{-1}(\varepsilon)\right)^2}{2}}$. Here, we implicitly assume that ε is small (e.g., $\varepsilon \leq 20\%$) and $\kappa_1(\varepsilon)$ and $\kappa_2(\varepsilon)$ are both positive.

From (7) and (8) we can see that if $\boldsymbol{\mu} \propto \mathbf{1}$, volatility, VaR, and CVaR are equal up to a positive scalar.

More generally, the Gaussian distribution can be extended to elliptical distributions [27] for which VaR and CVaR both are mean and standard deviation trade-off expressions.

Non-Gaussian Case: For the non-Gaussian case, we cannot obtain VaR and CVaR explicitly. The alternative approach is to obtain some tight approximations.

One option is the Cornish-Fisher approximation based on the first two and high-order moments of the distribution of the portfolio return, however, the expression is complicated. Since this is not the main goal of this paper, we do not present it in detail here. The interested reader is referred to [5] and [15] for the Cornish-Fisher approximations of VaR and CVaR, respectively. There are also other popular save convex approximations for VaR and CVaR, see [28].

We need to point out that all the above analytical approximations are differentiable and the solving approach developed in the following content of this paper always applies.

C. Risk Parity/Budgeting Portfolio

The risk parity portfolio is a portfolio such that each asset has the same risk contribution. That is, given the risk measurement $f(\mathbf{w})$, the risk parity portfolio should satisfy

$$w_i \frac{\partial f(\mathbf{w})}{\partial w_i} = w_j \frac{\partial f(\mathbf{w})}{\partial w_j}, \quad \forall i, j. \quad (9)$$

Risk budgeting portfolio is a more general concept. Given a budget vector $\mathbf{b} = [b_1, \ldots, b_n]^T > \mathbf{0}$, and $\mathbf{b}^T\mathbf{1} = 1$, where budget $\mathbf{b}$ is interpreted as a pre-determined percentage risk contribution target for all the assets, the risk budgeting portfolio should satisfy

$$w_i \frac{\partial f(\mathbf{w})}{\partial w_i} = b_i f(\mathbf{w}), \quad \forall i. \quad (10)$$

Obviously, the risk parity portfolio is a special case of the risk budgeting portfolio with $\mathbf{b} = \mathbf{1}/n$.

Due to the popularity of the terminology "risk parity", for clarity of presentation, we mainly refer "risk parity" as a broad portfolio allocation method of risk diversification (e.g., including both risk parity and risk budgeting portfolios) unless specified otherwise in this paper.

III. Problem Formulations

There are many different existing specific formulations on risk parity portfolio due to different risk measurements used or different profiles of investors. In this section, we first propose a general risk parity portfolio problem formulation, and then connect it with the existing specific formulations.

A. General Risk Parity Portfolio Problem Formulation

The general risk parity formulation can be expressed as

$$\begin{aligned} &\underset{\mathbf{w}}{\mathsf{minimize}} \;\; U(\mathbf{w}) \triangleq R(\mathbf{w}) + \lambda F(\mathbf{w}) \\ &\mathsf{subject\ to} \quad \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{aligned} \quad (11)$$

where

- $R(\mathbf{w})$ measures the risk concentration and has the form

$$R(\mathbf{w}) \triangleq \sum_{i=1}^{n} \left(g_i(\mathbf{w})\right)^2 \quad (12)$$

  in which each $g_i(\mathbf{w})$ is a smooth differentiable nonconvex function that measures the risk concentration of the i-th asset. The smaller the quantity $R(\mathbf{w})$ is, the more uniform the risk is distributed among n assets[2] (see Table I later for some specific examples of $g_i(\mathbf{w})$).
- $F(\mathbf{w})$ is a convex function that represents some traditional preferences on the portfolio. For example, it can be the expected portfolio loss (e.g., $F(\mathbf{w}) = -\boldsymbol{\mu}^T\mathbf{w}$), the mean-variance trade-off of the portfolio loss (e.g., $F(\mathbf{w}) = -\boldsymbol{\mu}^T\mathbf{w} + \nu\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$ where $\nu > 0$ is the trade-off parameter), or $F(\mathbf{w}) = 0$ when the goal is to distribute the risk only.
- $\lambda \geq 0$ is some trade-off parameter between the portfolio preference and risk concentration.
- $\mathbf{w}^T\mathbf{1} = 1$ denotes the capital budget constraint.
- $\mathcal{W}$ is a convex set that denotes the investor's profiles, capital limitations, market regulations, etc.

In general, since each function $g_i(\mathbf{w})$ is highly nonconvex, problem (11) is also nonconvex and hard to solve. For technical reasons, we need the following assumptions:

(A1) $\overline{\mathcal{W}}_1 \triangleq \left\{\mathbf{w}^T\mathbf{1} = 1\right\} \cap \mathcal{W}$ is nonempty, closed, and convex;

(A2) R and each g_i are C^1 on an open set containing $\overline{\mathcal{W}}_1$;

[2]In some problem formulations, the definition $\sum_{i,j=1}^{n}\left(g_{ij}(\mathbf{w})\right)^2$ is used where $g_{ij}(\mathbf{w})$ measures the difference between the risk contributions of assets i and j, for which the analytical approach derived in this paper still applies.

TABLE I
LIST OF FUNCTION g. FOR THE NOTATIONS $\mathbf{M}_i$ AND $\mathbf{B}_j$, SEE SECTION III-C

Problem	$g_i(\mathbf{w})$ or $g_{ij}(\mathbf{w})$	$\nabla g_i(\mathbf{w})$ or $\nabla g_{ij}(\mathbf{w})$
(16)	$g_{ij}(\mathbf{w}) = \mathbf{w}^T(\mathbf{M}_i - \mathbf{M}_j)\mathbf{w}$	$\nabla g_{ij}(\mathbf{w}) = \left(\mathbf{M}_i + \mathbf{M}_i^T - \mathbf{M}_j - \mathbf{M}_j^T\right)\mathbf{w}$
(17)	$g_i(\mathbf{w}) = \mathbf{w}^T\mathbf{M}_i\mathbf{w} - \theta$	$\nabla g_i(\mathbf{w}) = \left(\mathbf{M}_i + \mathbf{M}_i^T\right)\mathbf{w}$
(18)	$g_i(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}} - b_i$	$\nabla g_i(\mathbf{w}) = \frac{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})(\mathbf{M}_i+\mathbf{M}_i^T)\mathbf{w} - (\mathbf{w}^T\mathbf{M}_i\mathbf{w})(2\boldsymbol{\Sigma})\mathbf{w}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^2}$
(19)	$g_{ij}(\mathbf{w}) = \mathbf{w}^T\left(\frac{\mathbf{M}_i}{b_i} - \frac{\mathbf{M}_j}{b_j}\right)\mathbf{w}$	$\nabla g_{ij}(\mathbf{w}) = \left(\frac{\mathbf{M}_i}{b_i} + \frac{\mathbf{M}_i^T}{b_i} - \frac{\mathbf{M}_j}{b_j} - \frac{\mathbf{M}_j^T}{b_j}\right)\mathbf{w}$
(20)	$g_i(\mathbf{w}) = \mathbf{w}^T(\mathbf{M}_i - b_i\boldsymbol{\Sigma})\mathbf{w}$	$\nabla g_i(\mathbf{w}) = \left(\mathbf{M}_i + \mathbf{M}_i^T - 2b_i\boldsymbol{\Sigma}\right)\mathbf{w}$
(21)	$g_i(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}} - b_i\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}$	$\nabla g_i(\mathbf{w}) = \frac{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})(\mathbf{M}_i+\mathbf{M}_i^T)\mathbf{w} - (\mathbf{w}^T\mathbf{M}_i\mathbf{w})\boldsymbol{\Sigma}\mathbf{w}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{3/2}} - b_i\frac{\boldsymbol{\Sigma}\mathbf{w}}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}}$
(22)	$g_i(\mathbf{w}) = \frac{\mathbf{w}\mathbf{M}_i\mathbf{w}}{b_i} - \theta$	$\nabla g_i(\mathbf{w}) = \left(\frac{\mathbf{M}_i}{b_i} + \frac{\mathbf{M}_i^T}{b_i}\right)\mathbf{w}$
(24)	$g_i(\mathbf{w}) = \frac{\mathbf{w}\mathbf{M}_i\mathbf{w}}{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}$	$\nabla g_i(\mathbf{w}) = \frac{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})\times(\mathbf{M}_i+\mathbf{M}_i^T)\mathbf{w} - (\mathbf{w}^T\mathbf{M}_i\mathbf{w})\times(2\boldsymbol{\Sigma})\mathbf{w}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^2}$
(26)	$g_k(\mathbf{w}) = \sum_{i\in\mathcal{G}_k}\frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\sqrt{\mathbf{w}\boldsymbol{\Sigma}\mathbf{w}}} - b_k\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}$	$\nabla g_k(\mathbf{w}) = \sum_{i\in\mathcal{G}_k}\frac{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})(\mathbf{M}_i+\mathbf{M}_i^T)\mathbf{w} - (\mathbf{w}^T\mathbf{M}_i\mathbf{w})\boldsymbol{\Sigma}\mathbf{w}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{3/2}} - b_k\frac{\boldsymbol{\Sigma}\mathbf{w}}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}}$
(31)	$g_j(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{B}_j\mathbf{w}}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}} - b_j\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}$	$\nabla g_j(\mathbf{w}) = \frac{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})\left(\mathbf{B}_j+\mathbf{B}_j^T\right)\mathbf{w} - (\mathbf{w}^T\mathbf{B}_j\mathbf{w})\boldsymbol{\Sigma}\mathbf{w}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{3/2}} - b_j\frac{\boldsymbol{\Sigma}\mathbf{w}}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}}$
(32)	$g_i(\mathbf{w}) = -\mu_i w_i + \kappa_2(\varepsilon)\frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}} + b_i\boldsymbol{\mu}^T\mathbf{w} - b_i\kappa_2(\varepsilon)\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}$	$\nabla g_i(\mathbf{w}) = -\mu_i\mathbf{e}_i + \kappa_2(\varepsilon)\frac{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})(\mathbf{M}_i+\mathbf{M}_i^T)\mathbf{w} - (\mathbf{w}^T\mathbf{M}_i\mathbf{w})\boldsymbol{\Sigma}\mathbf{w}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{3/2}} + b_i\boldsymbol{\mu} - b_i\kappa_2(\varepsilon)\frac{\boldsymbol{\Sigma}\mathbf{w}}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}}$

(A3) ∇R is Lipschitz continuous on $\overline{\mathcal{W}}_1$ with constant L_R;
(A4) $F(\mathbf{w})$ is continuous and convex on $\overline{\mathcal{W}}_1$;
(A5) $U(\mathbf{w})$ is coercive with respect to $\overline{\mathcal{W}}_1$.

Note that the above assumptions are standard and are satisfied by a large class of functions. For instance, A3 is satisfied automatically if $\overline{\mathcal{W}}_1$ is bounded, and A4 is satisfied by all the standard F used in portfolio design, including $F(\mathbf{w}) = -\boldsymbol{\mu}^T\mathbf{w}$, $F(\mathbf{w}) = -\boldsymbol{\mu}^T\mathbf{w} + \nu\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$, and $F(\mathbf{w}) = 0$ as listed before. Assumption A5 guarantees that the sequence generated by the solving approach later is bounded, and if $\overline{\mathcal{W}}_1$ is bounded A5 is trivially satisfied. Also, A5 could be dispensed with at the price of a more complex analysis and cumbersome statement of convergence results [29], [31]. Actually for the portfolio design in the real markets, the feasible set will always be bounded due to some practical constraints, e.g., turnover constraints, holding constrains, tracking error constraints, etc. [31]. Next, we move to the existing specific risk parity formulations.

B. Specific Risk Parity Formulations

1) Volatility as Risk Measurement: Recall that the risk contribution of asset i is $\frac{w_i(\boldsymbol{\Sigma}\mathbf{w})_i}{\sqrt{\mathbf{w}\boldsymbol{\Sigma}\mathbf{w}}}$, then risk parity (9) and risk budgeting (10) relationships turn out to be

$$\text{risk parity:} \qquad w_i(\boldsymbol{\Sigma}\mathbf{w})_i = w_j(\boldsymbol{\Sigma}\mathbf{w})_j, \tag{13}$$

$$\text{risk budgeting:} \qquad w_i(\boldsymbol{\Sigma}\mathbf{w})_i = b_i\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}, \tag{14}$$

respectively, where $\mathbf{b} = [b_1, \ldots, b_n]^T > \mathbf{0}$ is the given risk budgeting for n assets and $\mathbf{b}^T\mathbf{1} = 1$. Actually, relationship (13) is a special case of relationship (14) with $b_i = 1/n$ for all i.

Only when $\boldsymbol{\Sigma}$ is diagonal and there exists long-only constraint, the nonlinear equation systems (14) admit an unique solution as follows [7]:

$$w_i = \frac{\sqrt{b_i}/\sqrt{\Sigma_{ii}}}{\sum_{k=1}^{n}\sqrt{b_k}/\sqrt{\Sigma_{kk}}}, \quad i = 1, \ldots, n. \tag{15}$$

However, for non-diagonal $\boldsymbol{\Sigma}$ or when there are some additional constraints, the closed-form solution does not exist anymore and some optimization problems are constructed instead.

Paper [8] is one of the first few papers that focuses on finding the risk parity portfolio via optimization. The proposed problem formulation is to penalize the summation of squared differences among risk contributions:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \sum_{i,j=1}^{n}\left(w_i(\boldsymbol{\Sigma}\mathbf{w})_i - w_j(\boldsymbol{\Sigma}\mathbf{w})_j\right)^2 \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}. \end{aligned} \tag{16}$$

Motivated by problem (16), Bai *et al.* [10] simplified the objective of (16) to solve:

$$\begin{aligned} \underset{\mathbf{w},\theta}{\text{minimize}} \quad & \sum_{i=1}^{n}\left(w_i(\boldsymbol{\Sigma}\mathbf{w})_i - \theta\right)^2 \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}. \end{aligned} \tag{17}$$

To find a portfolio that meets the risk budgeting targets $\mathbf{b}$ as much as possible, Bruder and Roncalli proposed to solve [11]:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \sum_{i=1}^{n}\left(\frac{w_i(\boldsymbol{\Sigma}\mathbf{w})_i}{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}} - b_i\right)^2 \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}. \end{aligned} \tag{18}$$

Similarly, it is easy to have some more alternative (but different) problem formulations, e.g.,

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \sum_{i,j=1}^{n}\left(\frac{w_i(\boldsymbol{\Sigma}\mathbf{w})_i}{b_i} - \frac{\mathbf{w}_j(\boldsymbol{\Sigma}\mathbf{w})_j}{b_j}\right)^2 \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{aligned} \tag{19}$$

and

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \sum_{i=1}^{n}\left(w_i(\boldsymbol{\Sigma}\mathbf{w})_i - b_i\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}\right)^2 \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{aligned} \tag{20}$$

and

$$\begin{array}{ll}\underset{\mathbf{w}}{\text{minimize}} & \displaystyle\sum_{i=1}^{n}\left(\frac{w_i\left(\boldsymbol{\Sigma}\mathbf{w}\right)_i}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}}-b_i\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}\right)^2\\ \text{subject to} & \mathbf{w}^T\mathbf{1}=1,\quad \mathbf{w}\in\mathcal{W},\end{array}\tag{21}$$

and

$$\begin{array}{ll}\underset{\mathbf{w},\theta}{\text{minimize}} & \displaystyle\sum_{i=1}^{n}\left(\frac{w_i\left(\boldsymbol{\Sigma}\mathbf{w}\right)_i}{b_i}-\theta\right)^2\\ \text{subject to} & \mathbf{w}^T\mathbf{1}=1,\quad \mathbf{w}\in\mathcal{W}.\end{array}\tag{22}$$

Note that all the above formulations are nonconvex.

2) Herfindahl Index: Now, we use the general risk measurement notation $f\left(\mathbf{w}\right)$. One popular measurement for the risk concentration is called the Herfindahl index, defined as [7]

$$h\left(\mathbf{w}\right)\triangleq\sum_{i=1}^{n}\left(\frac{w_i\frac{\partial f(\mathbf{w})}{\partial w_i}}{f\left(\mathbf{w}\right)}\right)^2.\tag{23}$$

It is easy to check that $\frac{1}{n}\leq h\left(\mathbf{w}\right)\leq 1$, and the extreme case $h\left(\mathbf{w}\right)=1$ means that the risk is concentrated on only one asset. The other extreme case $h\left(\mathbf{w}\right)=\frac{1}{n}$ denotes that the risk is equally distributed among all the assets. Thus, the smaller the Herfindahl index is, the more diversified the risk is.

Then one idea to achieve the risk parity is to minimize the Herfindahl index. For example, when $f\left(\mathbf{w}\right)=\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}$ is chosen, the minimization problem is

$$\begin{array}{ll}\underset{\mathbf{w}}{\text{minimize}} & \displaystyle\sum_{i=1}^{n}\left(\frac{w_i\left(\boldsymbol{\Sigma}\mathbf{w}\right)_i}{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}\right)^2\\ \text{subject to} & \mathbf{w}^T\mathbf{1}=1,\quad \mathbf{w}\in\mathcal{W}.\end{array}\tag{24}$$

If other risk measurements are used, we can easily obtain the corresponding formulations. Also, one can use a weighted Herfindahl index to cover the more general risk budget portfolio as well.

3) Group Risk Parity: The idea of group risk parity is to consider the risk contributions of several assets belonging to the same group (e.g., industry sector) as a whole. For example, suppose there are K ($K<n$) groups, denoted as $\mathcal{G}_1,\ldots,\mathcal{G}_K$ such that they form a partition of n assets, and the risk contribution of the k-th group is

$$RC_{\mathcal{G}_k}\left(\mathbf{w}\right)\triangleq\sum_{i\in\mathcal{G}_k}w_i\frac{\partial f\left(\mathbf{w}\right)}{\partial w_i}.\tag{25}$$

Then we want to find some portfolio so that $RC_{\mathcal{G}_k}\left(\mathbf{w}\right)$, $k=1,\ldots,K$, are less concentrated. For instance, when volatility is used, one formulation can be [12]

$$\begin{array}{ll}\underset{\mathbf{w}}{\text{minimize}} & \displaystyle\sum_{k=1}^{K}\left(\sum_{i\in\mathcal{G}_k}\frac{w_i\left(\boldsymbol{\Sigma}\mathbf{w}\right)_i}{\sqrt{\mathbf{w}\boldsymbol{\Sigma}\mathbf{w}}}-b_k\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}\right)^2\\ \text{subject to} & \mathbf{w}^T\mathbf{1}=1,\quad \mathbf{w}\in\mathcal{W},\end{array}\tag{26}$$

where $\mathbf{b}=\left[b_1,\ldots,b_K\right]^T>\mathbf{0}$ is the given risk budget for the K factors and $\mathbf{b}^T\mathbf{1}=1$.

In fact, we can have more formulations similar to the ones covered before simply by replacing the risk contribution of each asset by that of each group.

4) Risk Parity Portfolio With Risk Factors: Consider the following linear factor model:

$$\mathbf{r}\triangleq\mathbf{A}\mathbf{f}+\boldsymbol{\varepsilon}\tag{27}$$

where $\mathbf{r}\in\mathbb{R}^n$ is the random returns for n assets, $\mathbf{f}\in\mathbb{R}^m$ is the random returns of underlying m factors (probably $m\ll n$) with mean $\bar{\mathbf{f}}\in\mathbb{R}^m$ and covariance $\boldsymbol{\Omega}\in\mathbb{R}^{m\times m}$, $\mathbf{A}\in\mathbb{R}^{n\times m}$ is the factor loading matrix, and $\boldsymbol{\varepsilon}\in\mathbb{R}^n$ is the idiosyncratic component modeled as a noise with zero mean and diagonal covariance matrix $\mathbf{D}\in\mathbb{R}^{n\times n}$. All the mean and covariance parameters are given.

Based on model (27), the random return vector $\mathbf{r}$ has mean and covariance:

$$\boldsymbol{\mu}\triangleq\mathbf{A}\bar{\mathbf{f}},\tag{28}$$

$$\boldsymbol{\Sigma}\triangleq\mathbf{A}\boldsymbol{\Omega}\mathbf{A}^T+\mathbf{D}.\tag{29}$$

Then the original idea of diversifying the risk among n assets may result in a portfolio such that the risk contributions among the m underlying factors are far away from well diversified. Naturally, the idea to overcome such a drawback is to distribute the risk among the risk factors directly. For simplicity, we focus on the case that volatility is used as the risk measurement.

Risk Contributions of Factors: It was shown that when volatility is used the risk contribution of the j-th factor can be defined as [12]

$$RC_{\mathbf{f}_j}\left(\mathbf{w}\right)\triangleq\frac{\left(\mathbf{A}^T\mathbf{w}\right)_j\left(\mathbf{A}^\dagger\boldsymbol{\Sigma}\mathbf{w}\right)_j}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}}.\tag{30}$$

Diversifying the Risk Among Factors: Similar to the previous problem formulations covered in this section, we can have different risk parity formulations simply by replacing the risk contribution of each asset by that of each factor. For example, similar to (21), Roncalli and Weisang studied [12]

$$\begin{array}{ll}\underset{\mathbf{w}}{\text{minimize}} & \displaystyle\sum_{j=1}^{m}\left(RC_{\mathbf{f}_j}\left(\mathbf{w}\right)-b_j\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}\right)^2\\ \text{subject to} & \mathbf{w}^T\mathbf{1}=1,\quad \mathbf{w}\in\mathcal{W},\end{array}\tag{31}$$

where $\mathbf{b}=\left[b_1,\ldots,b_m\right]^T>\mathbf{0}$ is the given risk budget for the m factors and $\mathbf{b}^T\mathbf{1}=1$.

5) VaR and CVaR as Risk Measurements: Apart from volatility, VaR and CVaR have also been introduced into risk parity portfolio formulations [15], [16].

In [16], instead of being formulated into the objective, the risk parity property was added as constraints, e.g., $w_i\frac{\partial f(\mathbf{w})}{\partial w_i}=b_i f\left(\mathbf{w}\right)$, $\forall i$, where $f\left(\mathbf{w}\right)$ can be either VaR or CVaR. Paper [15] focused on CVaR and proposed to minimize the CVaR concentration, e.g., $\max_i\left\{w_i\frac{\partial\mathrm{CVaR}(\mathbf{w})}{\partial w_i}\right\}$. However, this objective, which is a piecewise function of some nonconvex functions, is not differentiable. Due to the intractability of VaR and CVaR expressions, usually VaR and CVaR problems are solved via the

Monte Carlo method [16] or are based on the analytical approximation [15].

In this paper, we can adopt problem formulations similar to the previous ones in Section III-B1 which can be solved numerically. For example, for the Gaussian CVaR (8) we have the risk contribution of asset i expressed in closed-form $w_i \frac{\partial \mathrm{CVaR}(\mathbf{w})}{\partial w_i} = -\mu_i w_i + \kappa_2(\varepsilon) \frac{w_i(\boldsymbol{\Sigma}\mathbf{w})_i}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}}$, then similar to (20), we can formulate

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \sum_{i=1}^{n} \left(-\mu_i w_i + \kappa_2(\varepsilon) \frac{w_i(\boldsymbol{\Sigma}\mathbf{w})_i}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}} \right. \\ & \left. + b_i \boldsymbol{\mu}^T\mathbf{w} - b_i \kappa_2(\varepsilon) \sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}} \right)^2 \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{aligned} \tag{32}$$

If we change $\kappa_2(\varepsilon)$ to $\kappa_1(\varepsilon)$, the problem is then formulating VaR instead.

6) Incorporating Expected Returns Into the Risk Parity Portfolio: Instead of focusing on risk diversification only, paper [14] introduced expected returns into the risk parity portfolio. However, it only focused on the long-only portfolio.

For the general long/short portfolio, the risk parity portfolio with expected returns incorporated can be formulated as

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & R(\mathbf{w}) - \lambda \boldsymbol{\mu}^T\mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{aligned} \tag{33}$$

where $\lambda \geq 0$, $F(\mathbf{w}) = -\boldsymbol{\mu}^T\mathbf{w}$ is the expected loss, and $R(\mathbf{w})$ can be any measurement on the risk concentration.

C. Connections Between the Specific Formulations and (11)

Recall that n, m, and K are the numbers of assets, factors, and groups, respectively. Let us denote $\mathbf{M}_i \in \mathbb{R}^{n\times n}$ the sparse matrix with its i-th row being the same as that of the covariance matrix $\boldsymbol{\Sigma}$ and 0 elsewhere, and denote $\mathbf{E}_j \in \mathbb{R}^{m\times m}$ the sparse matrix with only the j-th diagonal element being 1 and 0 elsewhere, and we define matrices $\mathbf{B}_j \triangleq \mathbf{A}\mathbf{E}_j\mathbf{A}^\dagger\boldsymbol{\Sigma}$, $j = 1, \ldots, m$, where $\mathbf{A}$ and $\boldsymbol{\Sigma}$ are given in (27) and (29).

Then for all the previous specific problems in Section III-B, we can put them in a form of general formulation (11) quite compactly with different specific functions $g_i(\mathbf{w})$, as listed in Table I. Also, we can easily have more possible formulations by combining VaR and CVaR and the different forms of $g_i(\mathbf{w})$. Since the derivation of such formulations is quite similar to the existing specific formulations covered in Section III-B it is thus omitted.

IV. Proposed Solving Approach

As mentioned in the introduction before, the general standard off-the-shelf numerical nonconvex nonlinear optimization methods, like SQP and IPM, are not efficient for the nonconvex problems like (11). In this section, we explore the structure of nonconvex part of $U(\mathbf{w})$, i.e., $R(\mathbf{w}) = \sum_{i=1}^{n} (g_i(\mathbf{w}))^2$, and propose a simple and efficient algorithm with provable global convergence to a stationary point.

A. Solving Procedure

At the k-th iteration, the proposed method aims to solve

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \overbrace{\sum_{i=1}^{n} \left(g_i(\mathbf{w}^k) + (\nabla g_i(\mathbf{w}^k))^T (\mathbf{w} - \mathbf{w}^k) \right)^2}^{P(\mathbf{w};\mathbf{w}^k) \triangleq} \\ & + \frac{\tau}{2} \|\mathbf{w} - \mathbf{w}^k\|_2^2 + \lambda F(\mathbf{w}) \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{aligned} \tag{34}$$

where $\tau > 0$ is the parameter for the regularization term. Here, we convexified the nonconvex term $R(\mathbf{w})$ by linearizing each $g_i(\mathbf{w})$ inside the square operation and added the proximal term $\|\mathbf{w} - \mathbf{w}^k\|_2^2$ for convergence reasons [29].

The beauty of the approximation $P(\mathbf{w}; \mathbf{w}^k)$ is that it is an easily computable quadratic convex function and has the same gradient as $R(\mathbf{w})$ at each iteration point $\mathbf{w}^k$

$$\nabla P(\mathbf{w}; \mathbf{w}^k)|_{\mathbf{w}=\mathbf{w}^k} = \nabla R(\mathbf{w})|_{\mathbf{w}=\mathbf{w}^k} \tag{35}$$

where $\nabla P(\mathbf{w}; \mathbf{w}^k)$ denotes the partial gradient of $P(\mathbf{w}; \mathbf{w}^k)$ with respect to the first argument $\mathbf{w}$.

Because $P(\mathbf{w}; \mathbf{w}^k)$ can be rewritten more compactly as

$$P(\mathbf{w}; \mathbf{w}^k) = \|\mathbf{A}^k(\mathbf{w} - \mathbf{w}^k) + \mathbf{g}(\mathbf{w}^k)\|_2^2 \tag{36}$$

where

$$\mathbf{A}^k \triangleq [\nabla g_1(\mathbf{w}^k), \ldots, \nabla g_n(\mathbf{w}^k)]^T, \tag{37}$$

$$\mathbf{g}(\mathbf{w}^k) \triangleq [g_1(\mathbf{w}^k), \ldots, g_n(\mathbf{w}^k)]^T. \tag{38}$$

Problem (34) can be further rewritten as

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{Q}^k\mathbf{w} + \mathbf{w}^T\mathbf{q}^k + \lambda F(\mathbf{w}) \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{aligned} \tag{39}$$

where

$$\mathbf{Q}^k \triangleq 2(\mathbf{A}^k)^T\mathbf{A}^k + \tau\mathbf{I}, \tag{40}$$

$$\mathbf{q}^k \triangleq 2(\mathbf{A}^k)^T\mathbf{g}(\mathbf{w}^k) - \mathbf{Q}^k\mathbf{w}^k. \tag{41}$$

In general, under the assumption that $F(\mathbf{w})$ is convex, for nonempty convex set $\overline{\mathcal{W}}_1$ (recall that $\overline{\mathcal{W}}_1 = \{\mathbf{w}^T\mathbf{1} = 1\} \cap \mathcal{W}$) and $\tau > 0$, problem (39) is strongly convex and can be solved by the existing efficient solvers (e.g., MOSEK [32], SeDuMi [33], SDPT3 [34], etc.). Moreover, if $F(\mathbf{w})$ is linear or convex quadratic, and $\overline{\mathcal{W}}_1$ only contains linear constraints, problem (39) reduces to a Quadratic Programming (QP).

Alg. 1 summarizes the sequential solving approach and we refer to it as SCRIP (Successive Convex optimization for RIsk Parity portfolio) since it is based on a successive convex optimization method.

Algorithm 1: Successive Convex optimization for RIsk Parity portfolio (SCRIP)

Input: $k=0$, $\mathbf{w}^0 \in \overline{\mathcal{W}}_1$, $\tau > 0$, $\{\gamma^k\} > 0$
Output: a stationary point of problem (11)
1: **repeat**
2: Solve (39) to get the optimal solution $\hat{\mathbf{w}}^k$
3: $\mathbf{w}^{k+1} = \mathbf{w}^k + \gamma^k \left(\hat{\mathbf{w}}^k - \mathbf{w}^k\right)$
4: $k \leftarrow k+1$
5: **until** convergence

B. Convergence Analysis

Fortunately, the convergence analysis of Alg. 1 can be obtained based on the theoretical framework developed in [29].

Proposition 2: Under assumptions A1-A5, suppose $\tau > 0$, $\gamma^k \in (0,1]$, $\gamma^k \to 0$, $\sum_k \gamma^k = +\infty$ *and* $\sum_k \left(\gamma^k\right)^2 < +\infty$, *and let* $\left\{\mathbf{w}^k\right\}$ *be the sequence generated by Alg. 1. Then either Alg. 1 converges in a finite number of iterations to a stationary point of* (11) *or every limit of* $\left\{\mathbf{w}^k\right\}$ *(at least one such point exists) is a stationary point of* (11).

Proof: Under assumptions A1–A5 and given $\tau > 0$ and γ^k as above, it is easy to check that the approximated problem (39) is a partial linear approximation of (11) with a quadratic uniformly strongly convex proximal term (since the quadratic coefficient matrix is $\mathbf{Q}^k/2$). That is, [29, Assumptions A1–A4] and [29, condition (b) in Theorem 3] are satisfied, and the proof of Prop. 2 follows directly from [29, Theorem 3]. ■

Remark 3: In the general SQP procedure, usually the derivative vector and the Hessian matrix of the Lagrangian function are used to construct some convex quadratic approximation at each iteration [19]. However, since the general risk parity formulation (11) is highly nonconvex, the Hessian matrix is not necessarily positive semidefinite, and some approximations of the Hessian matrix (e.g., full quasi-Newton, reduced-Hessian approximations [19]) based on the finite difference method are needed instead. Constructing such Hessian approximations may be time consuming and they may not approximate the original problem well since the structure of the objective is not explored at all. If we observe Alg. 1 carefully, we can see that indeed it solves a sequence of strongly convex QPs. That is, it is also a specific SQP method. However, the beauty of Alg. 1 is that it keeps the original convex part $F(\mathbf{w})$ and explores the structure of the nonconvex part $R(\mathbf{w}) = \sum_{i=1}^n \left(g_i(\mathbf{w})\right)^2$, and provides an easily computable convex quadratic approximation $P\left(\mathbf{w};\mathbf{w}^k\right)$, for which we only need to compute some simple closed-form first order derivatives (i.e., ∇g_i). Meanwhile, the convergence of Alg. 1 is perfectly guaranteed by the framework developed in [29], which usually provides a really fast numerical solving procedure for various applications, see [29], [31], [35], [36].

Remark 4: Another observation is that the proposed Alg. 1 may look similar to some numerical optimization methods for a nonlinear least-square problem, e.g., Gauss-Newton and Levenberg-Marquardt methods [19]. This is only because we use the specific quadratic squared loss function (see the next subsection for more general loss functions for which the proposed successive convex approximation based algorithm is still applicable). But even in the quadratic case, the convergence of Gauss-Newton and Levenberg-Marquardt methods in general is not guaranteed in theory [19]. Fortunately, by adding the proximal term $\frac{\tau}{2}\left\|\mathbf{w}-\mathbf{w}^k\right\|_2^2$ in the objective and incorporating the previous iteration point into the update procedure (as stated in step 3 of Alg. 1), Alg. 1 can be guaranteed to converge globally when the step-size parameter γ^k is properly chosen. One practical rule of choosing γ^k is: given $\gamma^0 \in (0,1]$, let

$$\gamma^k = \gamma^{k-1}\left(1-\zeta\gamma^{k-1}\right), \quad k=1,2,\ldots, \tag{42}$$

where $\zeta \in (0,1)$ is a given constant [31], [36]. This rule has been applied in various numerical experiments and in general it enjoys really fast numerical convergence speed (e.g., see the numerical experiments simulated in [29], [31], [35], [36] and Section VII of this paper later).

C. More General Risk Concentration Criteria

In this part, we briefly discuss some risk concentration criteria more general than $R(\mathbf{w}) \triangleq \sum_{i=1}^n \left(g_i(\mathbf{w})\right)^2$ considered before in (12).

Consider the following risk concentration function:

$$\breve{R}(\mathbf{w}) \triangleq \sum_{i=1}^n \rho\left(g_i(\mathbf{w})\right), \tag{43}$$

where $\rho(x): \mathbb{R} \mapsto \mathbb{R}$ is convex and differentiable.

Similar to the above solving procedure, if we define the following convex (not necessarily quadratic) function

$$\breve{P}\left(\mathbf{w};\mathbf{w}^k\right) = \sum_{i=1}^n \rho\left(g_i\left(\mathbf{w}^k\right) + \left(\nabla g_i\left(\mathbf{w}^k\right)\right)^T\left(\mathbf{w}-\mathbf{w}^k\right)\right), \tag{44}$$

it is easy to check that $\breve{R}(\mathbf{w})$ and $\breve{P}\left(\mathbf{w};\mathbf{w}^k\right)$ have the same first order derivative w.r.t. $\mathbf{w}$ at $\mathbf{w}^k$, i.e.,

$$\nabla\breve{P}\left(\mathbf{w};\mathbf{w}^k\right)|_{\mathbf{w}=\mathbf{w}^k} = \nabla\breve{R}(\mathbf{w})|_{\mathbf{w}=\mathbf{w}^k}. \tag{45}$$

Then we can minimize (nonconvex) $\breve{R}(\mathbf{w}) + \lambda F(\mathbf{w})$ via minimizing a sequence of (convex) approximation $\breve{P}\left(\mathbf{w};\mathbf{w}^k\right) + \lambda F(\mathbf{w})$ and the convergence to a stationary point can be guaranteed under similar conditions.

In fact, the previous risk concentration criterion (12) can be regarded as a special case of (43) where

$$\rho(x) = x^2. \tag{46}$$

Moreover, we can have more optional loss functions, e.g., the Huber loss

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & |x| \le \delta \\ \delta\left(|x| - \frac{1}{2}\delta\right), & \text{otherwise} \end{cases} \tag{47}$$

or the log-barrier loss

$$\rho(x) = \begin{cases} -\delta^2\log\left(1-(x/\delta)^2\right), & |x| < \delta \\ +\infty, & \text{otherwise} \end{cases} \tag{48}$$

where δ's in the above Huber and log-barrier losses are predefined parameters.

One step further, we can even have some nondifferentiable loss functions, e.g., the absolute value loss

$$\rho(x) = |x| \tag{49}$$

or the deadzone-linear loss

$$\rho(x) = \max\{0, |x| - \delta\} \tag{50}$$

where $\delta > 0$ is a predefined parameter, since we can always smooth the loss function around the nondifferentiable point with arbitrarily high precision [37].

Interestingly, for any given convex and differentiable loss function $\rho(x)$, the approximations $\breve{P}(\mathbf{w}; \mathbf{w}^k)$ are efficiently computable since all $g_i(\mathbf{w})$ and $\nabla g_i(\mathbf{w})$ listed in Table I are efficiently computable. Thus, the sequential solving procedure based on $\breve{P}(\mathbf{w}; \mathbf{w}^k)$ is really similar to the previously derived one. Because the least-square loss function is used more often in practice and for simplicity, we mainly focus on exploring the risk concentration criterion based on the least-square loss function (i.e., $R(\mathbf{w})$ defined in (12) or equivalently $\rho(x) = x^2$ for $\breve{R}(\mathbf{w})$ defined in (43)) in the following content of the paper.

V. Advanced Solving Approaches

Recall that in step 2 of Alg. 1, problem (39) is convex and we can always use existing efficient solvers to solve it numerically. However, it is still interesting to derive some simple and fast procedures to solve (39) for some cases.

Here, if we furthermore assume $F(\mathbf{w})$ is twice differentiable convex and we use the second order Taylor approximation to approximate $F(\mathbf{w})$, then problem (11) still has a convex quadratic approximation at each iteration and Alg. 1 still converges to a stationary point globally. Without loss of generality, we set $F(\mathbf{w}) \equiv 0$, and the approximated problem (39) becomes

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & Q_k(\mathbf{w}) \triangleq \frac{1}{2}\mathbf{w}^T\mathbf{Q}^k\mathbf{w} + \mathbf{w}^T\mathbf{q}^k \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{aligned} \tag{51}$$

where $\mathbf{Q}^k$ and $\mathbf{q}^k$ are given in (40) and (41).

Specifically, our goal now is to solve (51) efficiently, e.g., either in closed-form or via explicit update equations.

A. Analytical Solution for Linear Equality Constraints

Suppose that the nonempty set $\overline{\mathcal{W}}$ contains only linear equality constraints (including $\mathbf{w}^T\mathbf{1} = 1$) and they can be rewritten as $\mathbf{C}\mathbf{w} = \mathbf{c}$. Here, $\mathbf{C}$ can be assumed to be a non-tall matrix with full row rank, otherwise either it can be reduced to a non-tall matrix with full row rank by eliminating the redundant constraints or the problem is infeasible if there are no redundant constraints.

Then by solving the KKT conditions, the optimal solution of problem (51) can be found in closed-form as

$$\hat{\mathbf{w}}^k = -\left(\mathbf{Q}^k\right)^{-1}\left(\mathbf{q}^k + \mathbf{C}^T\boldsymbol{\lambda}^k\right) \tag{52}$$

where $\boldsymbol{\lambda}^k = -\left(\mathbf{C}\left(\mathbf{Q}^k\right)^{-1}\mathbf{C}^T\right)^{-1}\left(\mathbf{C}\left(\mathbf{Q}^k\right)^{-1}\mathbf{q}^k + \mathbf{c}\right)$. Then the solving approach can be simplified as Alg. 2.

Remark 5: Note that Alg. 2 is a special case of Alg. 1 and inherits the same convergence property as Alg. 1.

Algorithm 2: SCRIP with Linear Equality Constraints (SCRIP-LEC)

Input: $k = 0$, $\mathbf{w}^0 \in \overline{\mathcal{W}}_1$, $\tau > 0$, $\{\gamma^k\} > 0$
Output: a stationary point of problem (11) with LEC
1: **repeat**
2: $\quad \boldsymbol{\lambda}^k = -\left(\mathbf{C}\left(\mathbf{Q}^k\right)^{-1}\mathbf{C}^T\right)^{-1}\left(\mathbf{C}\left(\mathbf{Q}^k\right)^{-1}\mathbf{q}^k + \mathbf{c}\right)$
3: $\quad \hat{\mathbf{w}}^k = -\left(\mathbf{Q}^k\right)^{-1}\left(\mathbf{q}^k + \mathbf{C}^T\boldsymbol{\lambda}^k\right)$
4: $\quad \mathbf{w}^{k+1} = \mathbf{w}^k + \gamma^k\left(\hat{\mathbf{w}}^k - \mathbf{w}^k\right)$
5: $\quad k \leftarrow k + 1$
6: **until** convergence

B. Dual Decomposition for Linear Constraints

First, problem (51) with linear equality and inequality constraints can be written as

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{Q}^k\mathbf{w} + \mathbf{w}^T\mathbf{q}^k \\ \text{subject to} \quad & \mathbf{C}\mathbf{w} = \mathbf{c}, \quad \mathbf{D}\mathbf{w} \le \mathbf{d}, \end{aligned} \tag{53}$$

where $\mathbf{Q}^k$ and $\mathbf{q}^k$ is given in (40) and (41) as before, and $\mathbf{C}$, $\mathbf{c}$, $\mathbf{D}$ and $\mathbf{d}$ are given parameters with proper sizes.

Unfortunately, problem (53) does not admit a closed-form solution [38]. One way to find an optimal solution is via the dual decomposition method. Here, we employ the accelerated dual decomposition method [39, Alg. 1] to solve the approximated problem (53) with rate of convergence being $\mathcal{O}(1/i^2)$, where i is the number of (inner) iterations.

Alg. 3 summarizes the whole solving procedure of the original problem (11) where $\mathbf{B} \triangleq [\,\mathbf{C}^T \quad \mathbf{D}^T\,]^T$ and steps 3–9 denote the accelerated dual decomposition method of solving the inner subproblem (53) at each iteration.

Algorithm 3: Dual Gradient ascent based SCRIP with Linear Constraints (DualGrid SCRIP-LC)

Input: $k = 0$, $\mathbf{w}^0 \in \overline{\mathcal{W}}_1$, $\tau > 0$, $\{\gamma^k\} > 0$, $\boldsymbol{\lambda}^0 \ge \mathbf{0}$, $\boldsymbol{\mu}^0$
Output: a stationary point of problem (11) with LC
1: **repeat**
2: $\quad i = 0$, $\tilde{\mathbf{w}}^0 = \mathbf{w}^k$, $LC = \left\|\mathbf{B}\left(\mathbf{Q}^k\right)^{-1}\mathbf{B}^T\right\|_2$
3: $\quad$ **repeat**
4: $\quad\quad \tilde{\mathbf{w}}^i = -\left(\mathbf{Q}^k\right)^{-1}\left(\mathbf{q}^k + \mathbf{C}^T\boldsymbol{\lambda}^i + \mathbf{D}^T\boldsymbol{\mu}^i\right)$
5: $\quad\quad \bar{\tilde{\mathbf{w}}}^i = \tilde{\mathbf{w}}^i + \frac{i-1}{i+2}\left(\tilde{\mathbf{w}}^i - \tilde{\mathbf{w}}^{i-1}\right)$
6: $\quad\quad \boldsymbol{\lambda}^{i+1} = \boldsymbol{\lambda}^i + \frac{i-1}{i+2}\left(\boldsymbol{\lambda}^i - \boldsymbol{\lambda}^{i-1}\right) + \frac{1}{LC}\left(\mathbf{C}\bar{\tilde{\mathbf{w}}}^i - \mathbf{c}\right)$
7: $\quad\quad \boldsymbol{\mu}^{i+1} = \left[\boldsymbol{\mu}^i + \frac{i-1}{i+2}\left(\boldsymbol{\mu}^i - \boldsymbol{\mu}^{i-1}\right) + \frac{1}{LC}\left(\mathbf{D}\bar{\tilde{\mathbf{w}}}^i - \mathbf{d}\right)\right]^+$
8: $\quad\quad i \leftarrow i + 1$
9: $\quad$ **until** convergence
10: $\quad \boldsymbol{\lambda}^0 = \boldsymbol{\lambda}^i$, $\boldsymbol{\mu}^0 = \boldsymbol{\mu}^i$
11: $\quad \hat{\mathbf{w}}^k = \tilde{\mathbf{w}}^i$
12: $\quad \mathbf{w}^{k+1} = \mathbf{w}^k + \gamma^k\left(\hat{\mathbf{w}}^k - \mathbf{w}^k\right)$
13: $\quad k \leftarrow k + 1$
14: **until** convergence

Remark 6: Alg. 3 is a special case of Alg. 1 where the approximated problem (53) is solved via the accelerated dual decomposition method, so it converges as Alg. 1.

C. Projected Gradient Method

Since the objective of (51) is simply convex quadratic and its gradient is easily computable, we may simply employ the projected gradient descent method to solve it numerically.

Alg. 4 states the solving procedure, where the approximated problem (51) is solved via the accelerated projected gradient descent method (e.g., steps 3–12) with adaptive restart (e.g., steps 8–10) [40], and step 4 is the projection.

Algorithm 4: Primal Gradient descent based SCRIP (PrimGrad SCRIP)

Input: $k=0$, $\mathbf{w}^0 \in \overline{\mathcal{W}}_1$, $\tau > 0$, $\{\gamma^k\} > 0$
Output: a stationary point of problem (11)
1: **repeat**
2: $i=0$, $\tilde{\mathbf{w}}^0 = \tilde{\bar{\mathbf{w}}}^0 = \mathbf{w}^k$, $\theta^0 = 1$, $\alpha = 1/\lambda_{\max}\left(\mathbf{Q}^k\right)$
3: **repeat**
4: $\tilde{\mathbf{w}}^{i+1} = \left[\tilde{\bar{\mathbf{w}}}^i - \alpha\left(\mathbf{Q}^k\tilde{\bar{\mathbf{w}}}^i + \mathbf{q}^k\right)\right]_{\overline{\mathcal{W}}_1}$
5: $\theta^{i+1} = \theta^i\left(\sqrt{(\theta^i)^2+4} - \theta^i\right)/2$
6: $\beta^{i+1} = \left(1-\theta^i\right)\left(\sqrt{(\theta^i)^2+4} - \theta^i\right)/2$
7: $\tilde{\bar{\mathbf{w}}}^{i+1} = \tilde{\mathbf{w}}^{i+1} + \beta^{i+1}\left(\tilde{\mathbf{w}}^{i+1} - \tilde{\mathbf{w}}^i\right)$
8: **if** $Q_k\left(\tilde{\mathbf{w}}^{i+1}\right) - Q_k\left(\tilde{\mathbf{w}}^i\right) > 0$ **then**
9: $\theta^{i+1} = 1$
10: **end if**
11: $i \leftarrow i+1$
12: **until** convergence
13: $\hat{\mathbf{w}}^k = \tilde{\mathbf{w}}^i$
14: $\mathbf{w}^{k+1} = \mathbf{w}^k + \gamma^k\left(\hat{\mathbf{w}}^k - \mathbf{w}^k\right)$
15: $k \leftarrow k+1$
16: **until** convergence

Remark 7: Suppose the conditions in Prop. 2 hold. Without the adaptive restart (e.g., steps 8–10), the inner loop of Alg. 4, e.g., the accelerated projected gradient descent method, converges at the rate of at least $\mathcal{O}\left(1/i^2\right)$ [41], [42], and Alg. 4 converges as well. When there exists the adaptive restart, even though the analytical convergence analysis is not available, numerically it always converges [40] (and usually it converges much faster than the one without restart, see the example in [40, Sec. 5.3]). For the risk parity problem, Alg. 4 always converges numerically, see numerical examples later in Section VII.

On the Projection: The projection (e.g., step 4 in Alg. 4) is the key step that needs some computational effort. Fortunately, it does admit a closed-form expression for some cases:

- Affine set $\overline{\mathcal{W}}_1 = \{\mathbf{w} \in \mathbb{R}^n | \mathbf{C}\mathbf{w} = \mathbf{c}\}$ [43, Chapter 6]:

$$[\mathbf{w}_0]_{\overline{\mathcal{W}}_1} = \mathbf{w}_0 - \mathbf{C}^\dagger\left(\mathbf{C}\mathbf{w}_0 - \mathbf{c}\right). \tag{54}$$

- Simplex $\overline{\mathcal{W}}_1 = \left\{\mathbf{w} \in \mathbb{R}^n | \mathbf{1}^T\mathbf{w} = 1,\ \mathbf{w} \geq \mathbf{0}\right\}$ [44]:

$$[\mathbf{w}_0]_{\overline{\mathcal{W}}_1} = [\mathbf{w}_0 - \nu\mathbf{1}]^+, \tag{55}$$

for some ν such that $\mathbf{1}^T[\mathbf{w}_0 - \nu\mathbf{1}]^+ = 1$.

Nevertheless, the projection of a given point $\mathbf{w}_0$ onto a general convex set $\overline{\mathcal{W}}_1$ can be always obtained via solving

$$\underset{\mathbf{w}\in\overline{\mathcal{W}}_1}{\text{minimize}} \quad \|\mathbf{w} - \mathbf{w}_0\|_2 \tag{56}$$

or its dual problem, whichever is computationally cheaper [45].

VI. Alternative Approximations

One observation on the previous Algs. 2 and 3 is that, for the full matrix $\mathbf{Q}^k \in \mathbb{R}^{n\times n}$, the computational complexity of $\left(\mathbf{Q}^k\right)^{-1}$ is $\mathcal{O}\left(n^3\right)$ is time consuming especially when n is large. Then one natural upcoming question is can we find another approximation of $R(\mathbf{w})$ at each iteration such that the quadratic coefficient matrix has better structure, say diagonal only, and the computational complexity is reduced?

In the following, we will explore the above question by proposing simpler alternative approximations.

First, notice that $P\left(\mathbf{w};\mathbf{w}^k\right)$ in (36) can be rewritten as

$$P\left(\mathbf{w};\mathbf{w}^k\right) = \left(\mathbf{w}-\mathbf{w}^k\right)\left(\mathbf{A}^k\right)^T\mathbf{A}^k\left(\mathbf{w}-\mathbf{w}^k\right) + 2\left(\mathbf{g}\left(\mathbf{w}^k\right)\right)^T\mathbf{A}^k\left(\mathbf{w}-\mathbf{w}^k\right) + c. \tag{57}$$

Now, we denote another approximation as follows:

$$\tilde{P}\left(\mathbf{w};\mathbf{w}^k\right) \triangleq \left(\mathbf{w}-\mathbf{w}^k\right)\mathrm{Diag}\left(\left(\mathbf{A}^k\right)^T\mathbf{A}^k\right)\left(\mathbf{w}-\mathbf{w}^k\right) + 2\left(\mathbf{g}\left(\mathbf{w}^k\right)\right)^T\mathbf{A}^k\left(\mathbf{w}-\mathbf{w}^k\right). \tag{58}$$

Then it is easy to verify the following property

$$\nabla\tilde{P}\left(\mathbf{w};\mathbf{w}^k\right)|_{\mathbf{w}=\mathbf{w}^k} = \nabla P\left(\mathbf{w};\mathbf{w}^k\right)|_{\mathbf{w}=\mathbf{w}^k} = \nabla R(\mathbf{w})|_{\mathbf{w}=\mathbf{w}^k}. \tag{59}$$

When the approximation $\tilde{P}\left(\mathbf{w};\mathbf{w}^k\right)$ is used, the approximated problem turns out to be

$$\begin{array}{ll}\underset{\mathbf{w}}{\text{minimize}} & \tilde{P}\left(\mathbf{w};\mathbf{w}^k\right) + \dfrac{\tau}{2}\left\|\mathbf{w}-\mathbf{w}^k\right\|_2^2 \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w}\in\mathcal{W},\end{array} \tag{60}$$

which can be further rewritten as

$$\begin{array}{ll}\underset{\mathbf{w}}{\text{minimize}} & \dfrac{1}{2}\mathbf{w}^T\tilde{\mathbf{Q}}^k\mathbf{w} + \mathbf{w}^T\tilde{\mathbf{q}}^k \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w}\in\mathcal{W},\end{array} \tag{61}$$

where

$$\tilde{\mathbf{Q}}^k \triangleq 2\mathrm{Diag}\left(\left(\mathbf{A}^k\right)^T\mathbf{A}^k\right) + \tau\mathbf{I}, \tag{62}$$

$$\tilde{\mathbf{q}}^k \triangleq 2\left(\mathbf{A}^k\right)^T\mathbf{g}\left(\mathbf{w}^k\right) - \tilde{\mathbf{Q}}^k\mathbf{w}^k. \tag{63}$$

Compared with problem (51), problem (61) in fact can be solved more efficiently because matrix $\tilde{\mathbf{Q}}^k$ is diagonal while matrix $\mathbf{Q}^k$ is full.

Alg. 5 summarizes the sequential solving approach when approximation $\tilde{P}\left(\mathbf{w};\mathbf{w}^k\right)$ is used. It is referred to as SCRIP with $d=1$ because it takes 1-by-1 diagonal blocks (e.g., the diagonal

part) of $\left(\mathbf{A}^k\right)^T \mathbf{A}^k$ in the quadratic coefficient matrix (see (62) and Remark 11 later). The convergence analysis is conducted in Prop. 8.

Algorithm 5: SCRIP with $d = 1$

Input: $k = 0$, $\mathbf{w}^0 \in \overline{\mathcal{W}}_1$, $\tau > 0$, $\{\gamma^k\} > 0$
Output: a stationary point of problem (11)
1: **repeat**
2: Solve **(61)** to get the optimal solution $\hat{\mathbf{w}}^k$
3: $\mathbf{w}^{k+1} = \mathbf{w}^k + \gamma^k\left(\hat{\mathbf{w}}^k - \mathbf{w}^k\right)$
4: $k \leftarrow k + 1$
5: **until** convergence

Proposition 8: Under assumptions A1–A5, suppose $\tau > 0$, $\gamma^k \in (0, 1]$, $\gamma^k \to 0$, $\sum_k \gamma^k = +\infty$ *and* $\sum_k \left(\gamma^k\right)^2 < +\infty$, *and let* $\{\mathbf{w}^k\}$ *be the sequence generated by Alg. 5. Then either Alg. 5 converges in a finite number of iterations to a stationary point of* (11) *or every limit of* $\{\mathbf{w}^k\}$ *(at least one such point exists) is a stationary point of* (11).

Proof: The only difference between Props. 8 and 2 is that now $\tilde{P}\left(\mathbf{w}; \mathbf{w}^k\right)$ instead of $P\left(\mathbf{w}; \mathbf{w}^k\right)$ is used as the successive convex approximation of $R(\mathbf{w})$. However, similar to Prop. 2, we can still check that [29, Assumptions A1–A4] and [29, condition (b) in Theorem 3] are satisfied, and the proof of Prop. 2 follows directly from [29, Theorem 3]. ■

Furthermore, comparing Algs. 5 and 1, we have some interesting remarks.

Remark 9: For Alg. 5, we can have corresponding algorithms similar to Algs. 2–4 simply by replacing $\mathbf{Q}^k$ with $\tilde{\mathbf{Q}}^k$.

Remark 10: The computational complexity of $(\mathbf{Q}^k)^{-1}$ is $\mathcal{O}\left(n^3\right)$ while that of $(\tilde{\mathbf{Q}}^k)^{-1}$ is only $\mathcal{O}(n)$. The reduction from $\mathcal{O}\left(n^3\right)$ to $\mathcal{O}(n)$ is actually quite significant. Thus the computational complexity of each iteration of the algorithms based on approximation $\tilde{P}\left(\mathbf{w}; \mathbf{w}^k\right)$ should be much simpler than that of the algorithms based on approximation $P\left(\mathbf{w}; \mathbf{w}^k\right)$, and the sequential solving algorithms based on approximation $\tilde{P}\left(\mathbf{w}; \mathbf{w}^k\right)$ may be faster.

Remark 11: We also should point out that $P\left(\mathbf{w}; \mathbf{w}^k\right)$ with a more complicated structure may approximate the original function $R(\mathbf{w})$ better and thus provide a better solution at each iteration. Then, there may exist a trade-off between simple structure and good approximation. Actually, $\tilde{\mathbf{Q}}^k$ and $\mathbf{Q}^k$ are two extreme cases: $\tilde{\mathbf{Q}}^k$ takes the 1-by-1 diagonal blocks (i.e., the diagonal part) of $\left(\mathbf{A}^k\right)^T \mathbf{A}^k$ as the quadratic coefficient matrix, while $\mathbf{Q}^k$ takes the n-by-n diagonal block (i.e., $\left(\mathbf{A}^k\right)^T \mathbf{A}^k$ as a whole) instead. Similarly, we can even find more alternative approximations by taking the d-by-d diagonal blocks of $\left(\mathbf{A}^k\right)^T \mathbf{A}^k$ as the quadratic coefficient matrix where $1 \le d \le n$, and we could have extended algorithms similar to Algs. 2–4 easily. Such new approximations may explore the trade-off between simple structure and good approximation better.

VII. Numerical Experiments

Thus far we have proposed the general framework for risk parity portfolio with a family of sequential algorithms. In this section we will simulate some numerical experiments to study the performance of the proposed algorithms.

TABLE II
Numerical Methods for Risk Parity Portfolio

Methods	Existing/Proposed	Applicable
Cyclical coordinate descent (CCD)	Existing: [21]	Long-only portfolio
Newton-Nesterov (NN)	Existing: [23]	
fmincon-SQP	Existing	General long/short portfolio
fmincon-IPM	Existing	
SCRIP: Alg. 1 (step 2 solved by MOSEK)	Proposed	
SCRIP-LEC: Alg. 2		
DualGrad SCRIP-LC: Alg. 3		
PrimGrad SCRIP: Alg. 4		
SCRIP with alternative approximations (e.g., different d)		

A. Simulation Setup

The general formulation (11) contains many problems. We first focus on two problems, i.e., problems I and II in the following, in Sections VII-B–VII-D and then study more problems in Section VII-E.

1) Problem I: We take volatility as the risk measurement and set

$$F(\mathbf{w}) = 0, \tag{64}$$

$$R(\mathbf{w}) = \sum_{i=1}^{n}\left(\frac{w_i\left(\mathbf{\Sigma w}\right)_i}{\sqrt{\mathbf{w}^T\mathbf{\Sigma w}}} - b_i\sqrt{\mathbf{w}^T\mathbf{\Sigma w}}\right)^2. \tag{65}$$

Note that $R(\mathbf{w})$ can be further rewritten as

$$R(\mathbf{w}) = \left[\sum_{i=1}^{n}\left(\frac{w_i\left(\mathbf{\Sigma w}\right)_i}{\mathbf{w}^T\mathbf{\Sigma w}} - b_i\right)^2\right]\mathbf{w}^T\mathbf{\Sigma w} \tag{66}$$

where the first summation term measures the risk concentration, and the second term is the total variance. That is, $R(\mathbf{w})$ is a trade-off between the risk concentration and total risk.

2) Problem II: We take Gaussian CVaR as the risk measurement and set

$$F(\mathbf{w}) = -\boldsymbol{\mu}^T\mathbf{w}, \tag{67}$$

$$R(\mathbf{w}) = \sum_{i=1}^{n}\left(-\boldsymbol{\mu}_i w_i + \kappa_2(\varepsilon)\frac{w_i\left(\mathbf{\Sigma w}\right)_i}{\sqrt{\mathbf{w}^T\mathbf{\Sigma w}}} + b_i\boldsymbol{\mu}^T\mathbf{w} - b_i\kappa_2(\varepsilon)\sqrt{\mathbf{w}^T\mathbf{\Sigma w}}\right)^2. \tag{68}$$

3) Compared Methods: Table II lists all the simulated methods and we compare them in terms of numerical convergence speed (e.g., CPU time) and quality of the solution (mainly in Section VII-E). The MATLAB function `fmincon` is used for SQP and IPM methods.

4) Real Data: Here, we consider two types of real market data: Eurostoxx50 and S&P500 constituents. We download the data from Yahoo! Finance from the period 2010-01-01 to 2014-10-30 and use daily adjusted close prices to compute the daily returns. Since the daily returns usually are very small and to avoid algorithms stopping too early, we scale up the daily returns of Eurostoxx50 and S&P500 by positive scalars 10^3 and 10^4 respectively, and then estimate the sample mean and covariance of the scaled daily returns of each data set.

5) Synthetic Data: For the synthetic data, we randomly generate the expected returns as $\boldsymbol{\mu} = \texttt{randn(n, 1)}$ and the covariance matrix as $\mathbf{\Sigma} = \mathbf{V}\mathbf{V}^T$ where $\mathbf{V} = \texttt{rand(n, n)}$.

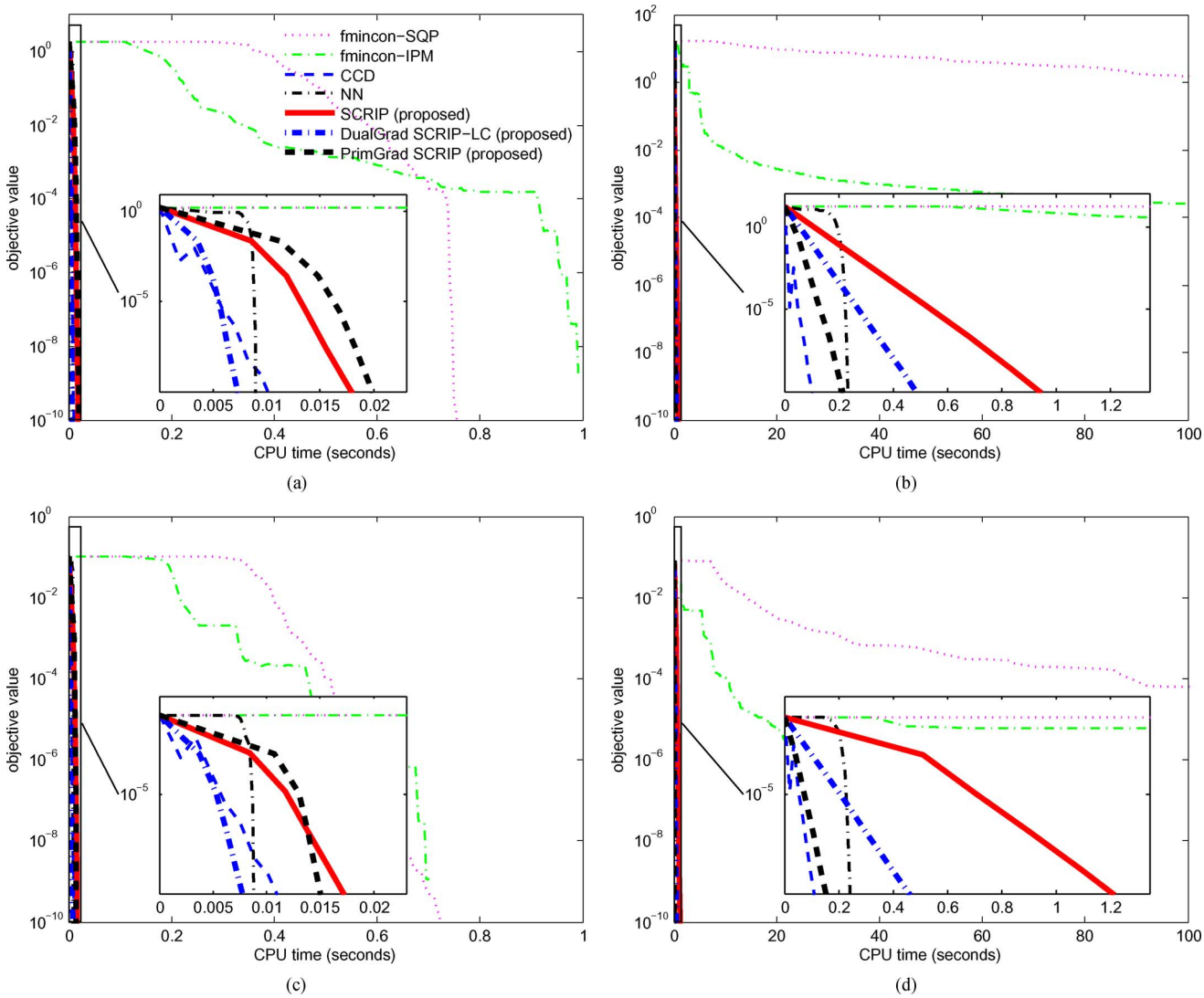


Fig. 2. One realization of problem I with long-only constraints. (a) Eurostocxx50. (b) S&P500. (c) Synthetic data with $n = 50$. (d) Synthetic data with $n = 500$.

6) Practical Implementation: All experiments are implemented in MATLAB on a PC with a 3.20 GHz i5-3470 CPU and 4 GB RAM. For the proposed algorithms, we set $\gamma^k = \gamma^{k-1}\left(1 - 10^{-7}\gamma^{k-1}\right)$ with $\gamma^0 = 0.9$ and $\tau = 0.05\mathrm{Tr}\left(\mathbf{\Sigma}\right)/\left(2n\right)$ (The parameter τ can be much smaller and the proposed algorithms are numerically quite robust w.r.t. τ). With such parameter settings, all the proposed methods converge quite fast and stably in practice.

B. Specific Long-Only Risk Parity Portfolio

For the long-only portfolio, we focus on Problem I

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \sum_{i=1}^{n}\left(\frac{w_i\left(\mathbf{\Sigma w}\right)_i}{\sqrt{\mathbf{w}^T\mathbf{\Sigma w}}} - b_i\sqrt{\mathbf{w}^T\mathbf{\Sigma w}}\right)^2 \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \geq \mathbf{0}. \end{aligned} \tag{69}$$

However, CCD and NN are actually solving other different problems. CCD [21] is proposed to solve the problem

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad \sqrt{\mathbf{x}^T\mathbf{\Sigma x}} - \sum_{i=1}^{n} b_i \ln x_i \tag{70}$$

and NN [23] aims to solve problem

$$\underset{\mathbf{y}\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\mathbf{y}^T\mathbf{C y} - \sum_{i=1}^{n} b_i \ln y_i \tag{71}$$

where $\mathbf{C} = \mathrm{Diag}^{-1/2}\left(\mathbf{\Sigma}\right)\mathbf{\Sigma}\mathrm{Diag}^{-1/2}\left(\mathbf{\Sigma}\right)$ is the correlation matrix.

It is well-known for the special long-only portfolio without any other constraints that the risk budgeting portfolio is unique and thus problem (69) has a unique global minimizer with the optimal objective value being 0 [7], and the relationships among the optimal solutions of the above three problems are [21], [23]

$$\mathbf{w}^\star = \frac{\mathbf{x}^\star}{\mathbf{1}^T\mathbf{x}^\star} = \frac{\mathrm{Diag}^{-1/2}\left(\mathbf{\Sigma}\right)\mathbf{y}^\star}{\mathbf{1}^T\mathrm{Diag}^{-1/2}\left(\mathbf{\Sigma}\right)\mathbf{y}^\star}, \tag{72}$$

where $\mathbf{w}^\star$, $\mathbf{x}^\star$, and $\mathbf{y}^\star$ are the optimal solutions of problems (69), (70), and (71), respectively.

Fig. 2 shows one realization of the objective value of (69) versus the CPU time. Here we set $\mathbf{w}_0 = \mathbf{1}/n$ and randomly generate $\mathbf{b} = \mathbf{b}_0/\mathbf{1}^T\mathbf{b}_0$ where $\mathbf{b}_0 = \mathtt{rand(n,1)}$. We can see that the proposed methods perform better than the SQP and IPM

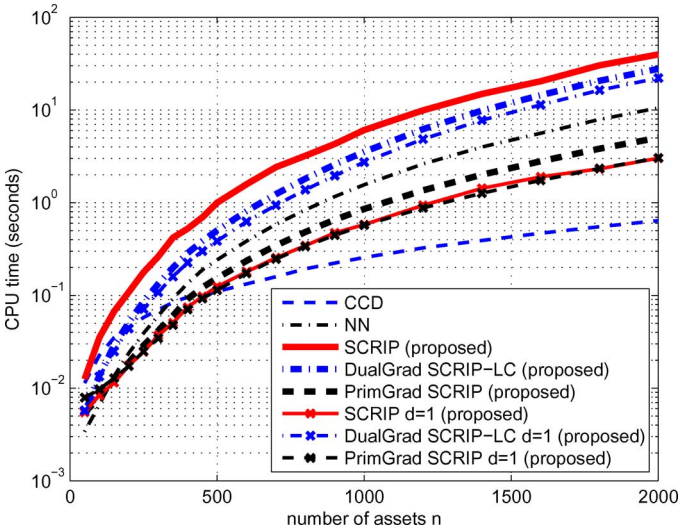


Fig. 3. CPU time of convergence versus the number of assets n of problem I with long-only constraints.

methods, and are comparable with ad-hoc methods, i.e., CCD and NN. Also, the results based on both real and synthetic data look similar, therefore, we can study the effect of the portfolio size n based on synthetic data. Numerical results of the objective value versus the number iterations are quite similar to that in Fig. 2 and thus are omitted.

Fig. 3 shows the average CPU time of convergence over 20 realizations versus the number of assets n based on synthetic data and the convergence criterion is $R(\mathbf{w}) \leq 10^{-10}$. Since the SQP and IMP methods are time consuming and may not even converge when n is large, we only report the results of CCD, NN and the proposed methods. We can see that when $n < 500$, some of the proposed methods converge slightly faster than, or at least similar to, CCD and NN. When $n > 500$, CCD converges fastest, but nerveless, some of the proposed methods beat NN and are still efficient enough since they converge in less than $3s$ even when $n = 2000$ (the proposed methods with $d = 1$ will be explained later in Section VII-D).

C. General Long/Short Risk Parity Portfolio

For the general long/short risk parity portfolio, CCD and NN methods are no longer useful (we can see this easily from formulations (70) and (71)) while our proposed methods are still applicable. To explore more examples, we now consider problem II. In general we now do not know the true optimal value, and we set the convergence criterion to be $\left\|\mathbf{w}^k - \mathbf{w}^{k-1}\right\|_2 / \left\|\mathbf{w}^k\right\|_2 < 10^{-6}$. The initial settings of $\mathbf{w}_0$ and $\mathbf{b}_0$ are the same as problem I before. We set the trade-off parameter to $\lambda = 1$ for real Eurostocxx50 data and synthetic data, and to $\lambda = 0.1$ for real S&P500 data.

First, we consider the linear equality constraints. One example can be that the total investments into different groups should be equal. Since our main focus of this paper is on efficient numerical algorithms, without loss of generality, we randomly separate the stocks into two groups, denoted as $\mathcal{G}_1$ and $\mathcal{G}_2$, and construct the equality constraints $\sum_{i\in\mathcal{G}_1} w_i = \sum_{i\in\mathcal{G}_2} w_i = 0.5$ together with $\mathbf{1}^T\mathbf{w} = 1$.

Fig. 4 shows one realization of the numerical results of different methods. The y axis is the logarithmic of differences between the objective values and the minimum of objective values provided by all the methods (the y axes of the other Figs. 5 and 7 are obtained in the same way). For both Eurostocxx50 and S&P500, we can see that the proposed methods work much better than the existing benchmarks: SQP and IPM.

Next, we consider the case of linear equality and inequality constraints. Again, without loss of generality, we consider inequality constraints $\sum_{i\in\mathcal{G}_1} w_i \leq 0.7$, $\sum_{i\in\mathcal{G}_2} w_i \leq 0.7$, together with the equality constraint $\mathbf{1}^T\mathbf{w} = 1$.

Fig. 5 shows one realization of the numerical results of different methods. Compared with Fig. 4, we have two observations. First, SQP and IPM tend to be worse, and our methods are still efficient. Second, the proposed method PrimGrad SCRIP becomes worse for a small-sized portfolio (see Fig. 4(a) versus Fig. 5(a)). The reason may be that the projection step (i.e., step 4 of Alg. 4) is not in closed-form for problem (53), and we need to call the solver MOSEK to solve the projection problem (56). Calling the solver also takes some time and it may be significant compared with the algorithm update procedure when the problem dimension is small. When the problem dimension becomes larger, the time of calling the solver becomes less significant, and so the method PrimGrad SCRIP still looks quite efficient (see Fig. 5(a) versus Fig. 5(b)). This observation is also verified later in Fig. 8.

D. Do Alternative Approximations Work?

Now we study whether alternative approximations proposed in Section VI can further improve the performance or not. We focus on problem II with linear equality and inequality constraints, e.g., $\sum_{i\in\mathcal{G}_1} w_i \leq 0.7$, $\sum_{i\in\mathcal{G}_2} w_i \leq 0.7$, together with $\mathbf{1}^T\mathbf{w} = 1$. The initial settings and the convergence criterion are the same as problem II simulated before.

Fig. 6 shows the average CPU time of convergence over 20 realizations versus the diagonal block size d based on the synthetic data with $n = 1000$. As mentioned in Remark 11, $d = 1$ and $d = n = 1000$ correspond the two extreme cases of taking the diagonal or full part of matrix $\left(\mathbf{A}^k\right)^T \mathbf{A}^k$ respectively. We can see that, for problem II, smaller d tends to reduce the CPU time for all the proposed algorithms.

Fig. 7 shows a realization of the objective value versus the CPU time with $d = 1$. Clearly, we can see that alternative approximations do reduce the CPU time.

Fig. 8 shows the average CPU time of convergence over 20 realizations versus the number of assets n. We can see that the alternative approximation with $d = 1$ can reduce the CPU time of convergence for a large range of portfolio size n and for all the proposed methods. Also, if we revisit problem I, we can observe similar results in Fig. 3.

E. Studies on Quality of the Solution

The previous numerical experiments focus on comparing the solving speed, this part explores the quality of the solutions given by the different methods for different problems, i.e., (18), (20), (21), and (32).

Note that problem (18) shares similar g_i function as problem (24); problem (20) shares similar g_i function as problems (16), (17) and (22); and problem (21) shares similar g_i function as problems (26) and (31), hence, simulating the proposed algorithms for the four problems (18), (20), (21), and (32) is enough

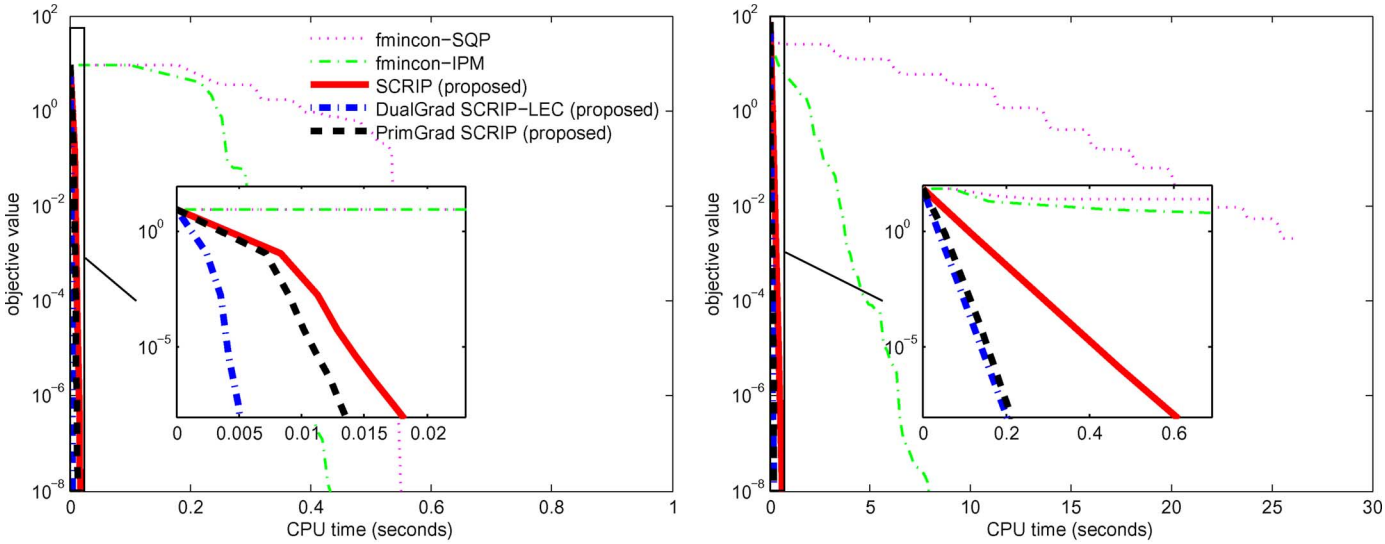


Fig. 4. One realization of problem II with linear equality constraints. (a) Eurostocxx50. (b) S&P500.

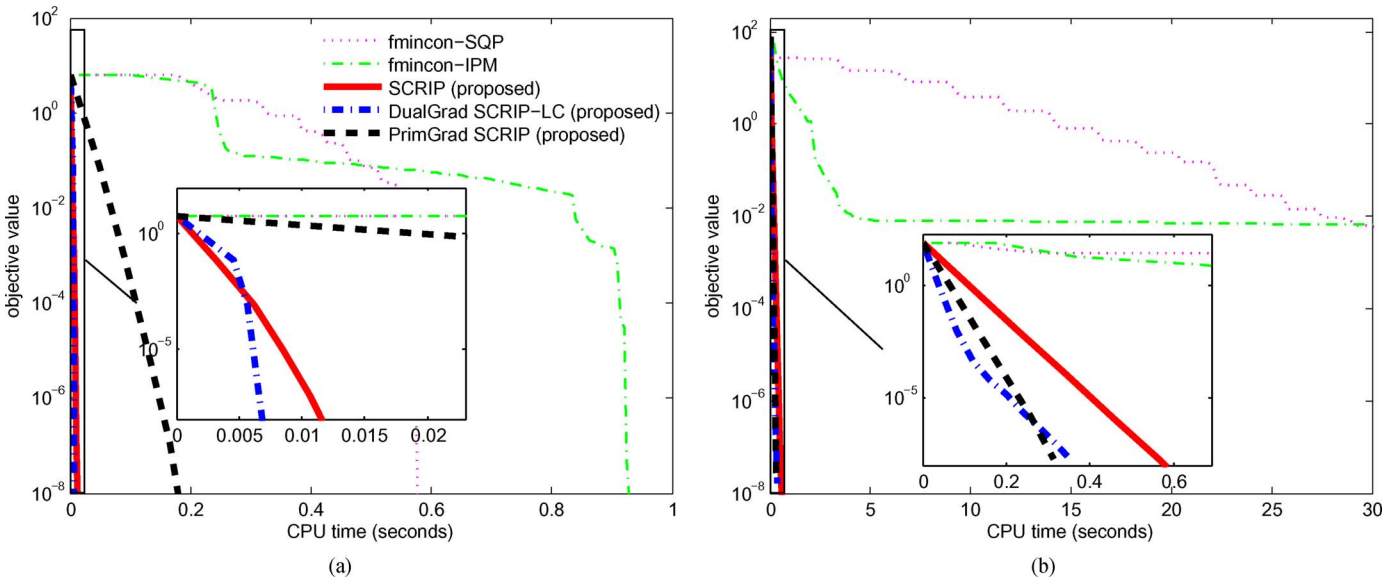


Fig. 5. One realization of problem II with linear equality and inequality constraints. (a) Eurostocxx50. (b) S&P500.

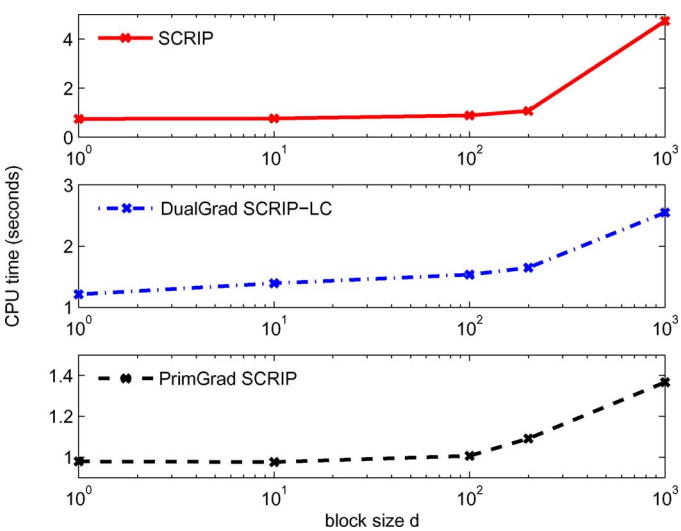


Fig. 6. CPU time of convergence versus diagonal block size d of problem II with linear equality and inequality constraints.

to study the performance of the proposed algorithms for all the problems listed in Table I.

For all the problems, the simulations are based on real data of S&P 500 stocks and we repeat the simulations 35 times. For each simulation realization, the initial settings and the convergence criterion are the same as problem II simulated before.

1) Long-Only Constraints: We first start with the long-only constraints, i.e., $\mathbf{w}^T\mathbf{1} = 1$ and $\mathbf{w} \geq \mathbf{0}$ without any other constraints. Regarding to the proposed methods, for clarity of presentation, we simulate methods SCRIP and PrimGrad SCRIP since the projection onto the long-only constraints admits a closed-form solution.

For problems (18), (20), and (21) that take volatility as the risk measurement, it is known that they have a same unique global minimizer with the optimal value being 0 [7]. The minimizer, denoted as $\mathbf{w}^\star$, satisfies relationship (72) in which $\mathbf{x}^\star$ is the solution given by the CCD algorithm [21] and $\mathbf{y}^\star$ is the solution given by the NN algorithm [23].

TABLE III
QUALITY OF THE SOLUTIONS GIVEN BY DIFFERENT PROBLEMS WITH SPECIFIC LONG-ONLY CONSTRAINTS

	Average CPU time (seconds)				Number of realizations with objective value $\leq 10^{-9}$			
Methods \ Problems	(18)	(20)	(21)	(32)	(18)	(20)	(21)	(32)
fmincon-SQP	11.6790	26.5438	69.9044	38.8989	1	0	0	0
fmincon-IPM	4.8241	49.1794	55.3225	57.9934	0	0	0	0
CCD	0.0737			0.0550	35	35	35	35
NN	0.2293			NA	35	35	35	NA
PrimGrad SCRIP	0.2096	0.2439	0.2297	0.2510	35	35	35	35
SCRIP	1.2958	1.4619	1.0697	1.0251	35	35	35	35

TABLE IV
QUALITY OF THE SOLUTIONS GIVEN BY DIFFERENT PROBLEMS WITH GENERAL LONG/SHORT CONSTRAINTS

	Average CPU time (seconds)				Average ratio between objective values by other methods and SCRIP			
Methods \ Problems	(18)	(20)	(21)	(32)	(18)	(20)	(21)	(32)
fmincon-SQP	10.8048	122.1629	122.0066	122.1541	105.3793	104.9495	104.9582	96.6847
fmincon-IPM	5.8453	50.1504	49.9313	51.8006	102.0960	105.2793	105.2209	96.8853
SCRIP $d = 1$	0.2432	0.2485	0.2416	0.2584	1	1	1	1
SCRIP	1.1400	1.2786	1.1949	1.2327	1	1	1	1

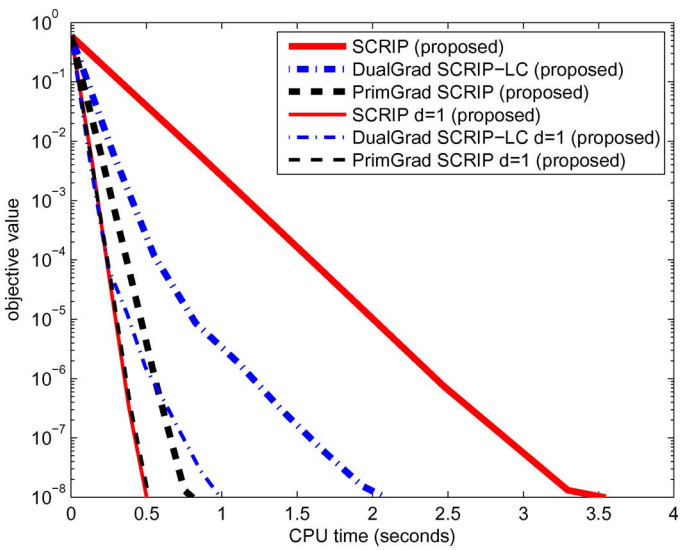


Fig. 7. One realization of problem II with linear equality and inequality constraints based on synthetic data with $n = 1000$.

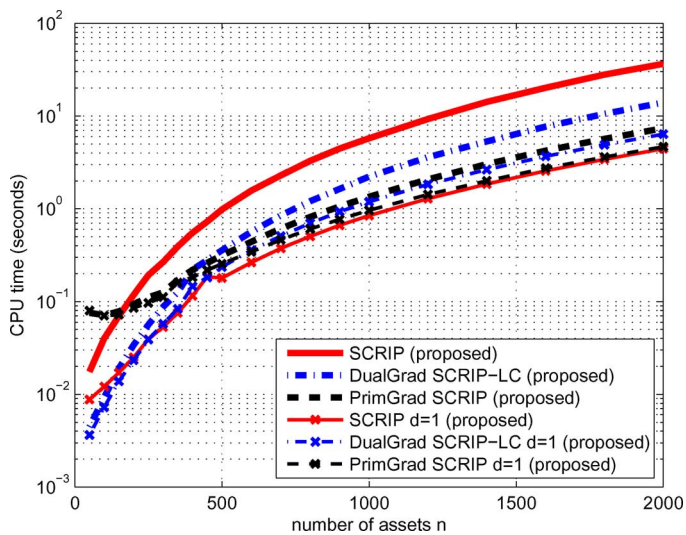


Fig. 8. CPU time of convergence versus the number of assets n of problem II with linear equality and inequality constraints.

For problem (32) that takes Gaussian CVaR as the risk measurement, it is also known that it has a unique global minimizer with the optimal value being 0 [7] and there exists an extension of the CCD algorithm [21] that can find the unique minimizer. However, now the NN algorithm [23] is not applicable (NA) any more.

Table III shows the average CPU time and the number of realizations with objective value equal to or less than 10^{-9} out of the total 35 realizations. We can see that, for all the simulated problems, the proposed methods (especially PrimGrad SCRIP) are very efficient: they outperform the existing SQP and IPM methods or at least are comparable with the ad-hoc CCD and NN methods in terms of both converge speed and solution quality.

2) General Long/Short Constraints: As for the general long/short constraints, we consider the following linear constraints: $\mathbf{w}^T\mathbf{1} = 1$, $\sum_{i=1}^{250} w_i = 0.5$, $-\frac{1}{n} \leq \mathbf{w} \leq \frac{3}{n}$ where n is the number of stocks. Regarding to the proposed methods, for clarity of presentation, we simulate methods SCRIP and SCRIP with $d = 1$.

For the general long/short constrained problems, the different problems have different optimal values strictly larger than 0. Table IV shows the average CPU time and average ratio between objective values by other methods (i.e., fmincon-SQP, fmincon-IPM, and SCRIP with $d = 1$) and SCRIP. We clearly observe that, for all the simulated problems, the proposed methods (i.e., SCRIP with $d = 1$ and SCRIP) cost much shorter CPU time and achieve much smaller objective values than the existing SQP and IPM methods. Also, SCRIP with $d = 1$ achieves the same objective values as SCRIP and even reduces the CPU time.

VIII. Conclusion

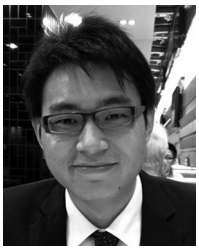
In this paper, we have proposed a general formulation characterizing many existing specific formulations on the risk parity portfolio with a simple and efficient sequential solving approach based on the successive convex approximation method. Furthermore, we have derived various advanced algorithms for different specific cases or based on alternative convex approximations. Theoretically, all the proposed algorithms enjoy global convergence to a stationary point. Extensive numerical experiments based on both synthetic and real data show that, in terms of numerical computation speed (i.e., CPU time) and solution quality (i.e., the objective value of the obtained solution), our proposed methods outperform the existing methods significantly for the general and more practical long/short risk parity portfolio, for which existing methods may not even work, and are comparable with some ad-hoc methods for the specific long-only portfolio.

Acknowledgment

The authors would like to thank Dr. G. W. Peters for early discussions on the topic.

References

[1] H. Markowitz, "Portfolio selection," *J. Finan.*, vol. 7, no. 1, pp. 77–91, 1952.
[2] H. Markowitz, "The optimization of a quadratic function subject to linear constraints," *Nav. Res. Log. Quarter.*, vol. 3, no. 1-2, pp. 111–133, 1956.
[3] H. M. Markowitz, *Portfolio Selection: Efficient Diversification of Investments*. New Haven, CT, USA: Yale Univ. Press, 1968.
[4] E. Qian, "Risk parity portfolios: Efficient portfolios through true diversification," *Panagora Asset Manage.*, Sep. 2005.
[5] E. Qian, "On the financial interpretation of risk contribution: Risk budgets do add up," *J. Invest. Manage.*, vol. 4, no. 4, p. 41, 2006.
[6] R. M. Anderson, S. W. Bianchi, and L. R. Goldberg, "Will my risk parity strategy outperform?," *Finan. Anal. J.*, vol. 68, no. 6, pp. 75–93, 2012.
[7] T. Roncalli, *Introduction to Risk Parity and Budgeting*. Boca Raton, FL, USA: CRC, 2013.
[8] S. Maillard, T. Roncalli, and J. Teïletche, "The properties of equally weighted risk contribution portfolios," *J. Portfolio Manage.*, vol. 36, no. 4, pp. 60–70, 2010.
[9] D. Chaves, J. Hsu, F. Li, and O. Shakernia, "Risk parity portfolio vs. other asset allocation heuristic portfolios," *J. Invest.*, vol. 20, no. 1, pp. 108–118, 2011.
[10] X. Bai, K. Scheinberg, and R. Tutuncu, "Least-squares approach to risk parity in portfolio selection," 2013, SSRN 2343406.
[11] B. Bruder and T. Roncalli, "Managing risk exposures using the risk budgeting approach," Univ. Library of Munich, Munich, Germany, Tech. Rep., 2012.
[12] T. Roncalli and G. Weisang, "Risk parity portfolios with risk factors," SSRN 2155159, 2012.
[13] R. Deguest, L. Martellini, and A. Meucci, "Risk parity and beyond-from asset allocation to risk allocation decisions," SSRN 2355778, 2013.
[14] T. Roncalli, "Introducing expected returns into risk parity portfolios: A new framework for asset allocation," SSRN 2321309, 2013.
[15] K. Boudt, P. Carl, and B. G. Peterson, "Asset allocation with conditional value-at-risk budgets," *J. Risk*, vol. 15, no. 3, pp. 39–68, 2013.
[16] M. Haugh, G. Iyengar, and I. Song, "A generalized risk budgeting approach to portfolio construction," SSRN 2462145, 2014.
[17] I. Song, "New quantitative approaches to asset selection and portfolio construction," Ph.D. dissertation, Columbia Univ., New York, NY, USA, 2014.
[18] S. Darolles, C. Gourieroux, and E. Jay, "Robust portfolio allocation with systematic risk contribution restrictions," 2012, SSRN 2192399.
[19] J. Nocedal and S. J. Wright, *Numerical Optimization*, ser. Springer Ser. Operat. Res., 2nd ed. New York, NY, USA: Springer-Verlag, 2006.
[20] R. H. Byrd, M. E. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *SIAM J. Optimiz.*, vol. 9, no. 4, pp. 877–900, 1999.
[21] T. Griveau-Billion, J.-C. Richard, and T. Roncalli, "A fast algorithm for computing high-dimensional risk parity portfolios," 2013, arXiv:1311.4057 [Online]. Available: http://arxiv.org/abs/1311.4057, to be published
[22] D. B. Chaves, J. C. Hsu, F. Li, and O. Shakernia, "Efficient algorithms for computing risk parity portfolio weights," *J. Investing*, vol. 21, pp. 150–163, 2012.
[23] F. Spinu, "An algorithm for computing risk parity weights," SSRN 2297383, 2013.
[24] W. G. Hallerbach, "Decomposing portfolio value-at-risk: A general analysis," *J. Risk*, vol. 5, no. 2, pp. 1–18, 2003.
[25] O. Scaillet, "Nonparametric estimation and sensitivity analysis of expected shortfall," *Math. Finan.*, vol. 14, no. 1, pp. 115–129, 2004.
[26] A. McNeil, R. Frey, and P. Embrechts, *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton, NJ, USA: Princeton Univ. Press, 2005.
[27] Z. Landsman and E. Valdez, "Tail conditional expectations for elliptical distributions," *N. Amer. Actuar. J.*, vol. 7, no. 4, pp. 55–71, 2003.
[28] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton, NJ, USA: Princeton Univ. Press, 2009.
[29] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 641–656, Feb. 2014.
[30] F. J. Fabozzi, S. M. Focardi, and P. N. Kolm, *Quantitative Equity Investing: Techniques and Strategies*. New York, NY, USA: Wiley, 2010.
[31] F. Facchinei, G. Scutari, and S. Sagratella, "Parallel selective algorithms for nonconvex big data optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1874–1889, Apr. 2015.
[32] "The MOSEK Optimization Toolbox for MATLAB Manual," MOSEK, Tech. Rep., 2013 [Online]. Available: http://www.mosek.com
[33] J. F. Sturm, "Using SeDuMi 1.02, A MATLAB toolbox for optimization over symmetric cones," *Optim. Method Softw.*, vol. 11, no. 1–4, pp. 625–653, 1999.
[34] K.-C. Toh, M. J. Todd, and R. H. Tütüncü, "On the implementation and usage of SDPT3–A MATLAB software package for semidefinite-quadratic-linear programming, version 4.0," in *Handbook on Semidefinite, Conic and Polynomial Optimization*. New York, NY, USA: Springer-Verlag, 2012, pp. 715–754.
[35] A. Daneshmand, F. Facchinei, V. Kungurtsev, and G. Scutari, "Hybrid random/deterministic parallel algorithms for convex and nonconvex big data optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 3914–3929, Aug. 2015.
[36] Y. Yang, G. Scutari, D. P. Palomar, and M. Pesavento, "A parallel stochastic approximation method for nonconvex multi-agent optimization problems," 2014, arXiv:1410.5076 [Online]. Available: http://arxiv.org/abs/1410.5076, to be published
[37] J. Song, P. Babu, and D. P. Palomar, "Sparse generalized eigenvalue problem via smooth optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1627–1642, Apr. 2015.
[38] C. K. Liew, "Inequality constrained least-squares estimation," *J. Amer. Statist. Assoc.*, vol. 71, no. 355, pp. 746–751, 1976.
[39] P. Giselsson, M. D. Doan, T. Keviczky, B. D. Schutter, and A. Rantzer, "Accelerated gradient methods and dual decomposition in distributed model predictive control," *Automatica*, vol. 49, no. 3, pp. 829–833, 2013.
[40] B. O'Donoghue and E. Candès, "Adaptive restart for accelerated gradient schemes," *Found. Comput. Math.*, pp. 1–18, 2012.
[41] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Soviet Math. Doklady*, vol. 27, no. 2, pp. 372–376, 1983.
[42] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[43] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optimiz.*, vol. 1, no. 3, pp. 123–231, 2013.
[44] D. P. Palomar, "Convex Primal decomposition for multicarrier linear MIMO transceivers," *IEEE Trans. Signal Process.*, vol. 53, no. 12, pp. 4661–4674, 2005.
[45] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*. New York, NY, USA: Springer-Verlag, 2003, vol. 1.



Yiyong Feng received the B.E. degree in electronic and information engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2010.

Since then he has been pursuing the Ph.D. degree with the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology (HKUST). From March 2013 to August 2013, he was with the Systematic Market-Making Group at Credit Suisse (Hong Kong). His research interests are in convex optimization, nonlinear programming, and robust optimization, with applications in signal processing, financial engineering, and machine learning.

Daniel P. Palomar (S'99–M'03–SM'08–F'12) received the electrical engineering and Ph.D. degrees from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1998 and 2003, respectively.

He joined the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2006, where he is a Professor. Since 2013, he has been a Fellow of the Institute for Advance Study (IAS) at HKUST. He had previously held several research appointments, namely, at King's College London (KCL), London, UK; Technical University of Catalonia (UPC), Barcelona; Stanford University, Stanford, CA, USA; Telecommunications Technological Center of Catalonia (CTTC), Barcelona; Royal Institute of Technology (KTH), Stockholm, Sweden; University of Rome "La Sapienza," Rome, Italy; and Princeton University, Princeton, NJ, USA. His current research interests include applications of convex optimization theory, game theory, and variational inequality theory to financial systems and communication systems.

Dr. Palomar is a recipient of the 2004–2006 Fulbright Research Fellowship, the 2004 Young Author Best Paper Award by the IEEE Signal Processing Society, the 2002–2003 best Ph.D. prize in information technologies and communications by the Technical University of Catalonia (UPC), the 2002–2003 Rosina Ribalta first prize for the Best Doctoral Thesis in information technologies and communications by the Epson Foundation, and the 2004 prize for the best Doctoral Thesis in Advanced Mobile Communications by the Vodafone Foundation and COIT. He has been an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY and of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, a Guest Editor of the IEEE SIGNAL PROCESSING MAGAZINE 2010 Special Issue on "Convex Optimization for Signal Processing," the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2008 Special Issue on "Game Theory in Communication Systems," and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2007 Special Issue on "Optimization of MIMO Transceivers for Realistic Communication Networks."