Mirror Descent Algorithms for Risk Budgeting Portfolios

Martin Arnaiz Iglesias [*]   Adil Rengim Cetingoz[†]   Noufel Frikha[‡]

December 24, 2024

Abstract

This paper introduces and examines numerical approximation schemes for computing risk budgeting portfolios associated to positive homogeneous and sub-additive risk measures. We employ Mirror Descent algorithms to determine the optimal risk budgeting weights in both deterministic and stochastic settings, establishing convergence along with an explicit non-asymptotic quantitative rate for the averaged algorithm. A comprehensive numerical analysis follows, illustrating our theoretical findings across various risk measures – including standard deviation, Expected Shortfall, deviation measures, and Variantiles – and comparing the performance with that of the standard stochastic gradient descent method recently proposed in the literature.

1   Introduction

The financial problem of constructing investment portfolios has often been analyzed as a mathematical optimization problem, at least since Markowitz [31] formulated it as maximizing a portfolio's expected return under a constraint on risk, defined by variance. Solving this problem to obtain the optimal portfolio typically requires numerical methods, except in trivial cases where analytical solutions exist, such as when the only constraint is that portfolio weights sum to one. Today, efficient tools exist to compute this portfolio under almost any set of convex constraints, given the expected return vector and the covariance matrix of asset returns (see [32], [46], and [13]). However, this framework is rarely adopted in its classical form in real-life applications, mainly due to the fact that it heavily depends on accurate estimates of expected returns, variances, and covariances. Small errors in these inputs, especially in expected returns, can lead to vastly different and potentially suboptimal portfolio allocations ([12] and [34]).

Alternative portfolio construction approaches that prioritise risk management and diversification over return maximisation have thus been proposed – especially following the dot-com crash and the Great

---

[*]Université Paris 1 Panthéon-Sorbonne, Centre d'Economie de la Sorbonne (CES), 106 Boulevard de l'Hôpital, 75642 Paris Cedex 13 (`Martin.Arnaiz-Iglesias@etu.univ-paris1.fr`).

[†]Université Paris 1 Panthéon-Sorbonne, Centre d'Economie de la Sorbonne (CES), 106 Boulevard de l'Hôpital, 75642 Paris Cedex 13 (`rengimcetingoz@gmail.com`).

[‡]Université Paris 1 Panthéon-Sorbonne, Centre d'Economie de la Sorbonne (CES), 106 Boulevard de l'Hôpital, 75642 Paris Cedex 13 (`noufel.frikha@univ-paris1.fr`). The research of N. Frikha has benefited from the support of the Institut Europlace de Finance.

Financial Crisis – to address investment objectives beyond optimising a mean-variance criterion. Risk budgeting is one such approach, aiming to construct a long-only portfolio in which each asset contributes to the portfolio's total risk according to a risk budget specified by the investor. For a comprehensive introduction, we refer to [44]. This method is thus closely aligned with the principles of diversification, helping to mitigate concentrated risks within the portfolio.

The risk budgeting problem is mathematically formulated as a system of nonlinear equations, which also corresponds to the first-order conditions of a constrained convex minimisation problem. Leveraging this variational characterisation, and under positive homogeneous and sub-additive risk measures, the existence of a unique solution has been extensively studied in the literature; see, for example, [14] and [15]. However, an analytical solution is generally not obtainable, prompting the development of various numerical methods for computing risk budgeting portfolios.

Volatility has been one of the earliest and most extensively studied risk measures in the context of risk budgeting [38]. Initial efforts, including those by [30] and [6], focused on minimizing a least-squares formulation. Subsequently, [17] proposed employing Newton's method, with [45] further advancing this approach by demonstrating the feasibility of Nesterov acceleration, leveraging the self-concordant nature of the objective function to achieve provable convergence. Additionally, [28] advocates for a cyclical coordinate descent algorithm, which proves particularly efficient for high-dimensional portfolios.

Risk measures beyond volatility have also attracted interest in risk budgeting. Expected Shortfall (ES for short) is one of the most extensively studied, as it is considered to be more representative of the true risk of financial investments and satisfies certain desirable mathematical properties, as outlined in [2] and [39]. For example, [33] proposes formulating the problem as a second-order cone program when asset returns are modeled by discrete distributions. For continuous distributions, [18] introduces a cutting planes algorithm, while [4] develops a numerical procedure based on a novel ES estimator that also facilitates statistical inference of ES-based risk budgeting portfolios. Other risk measures have also been considered, such as Mean Absolute Deviation (MAD) [3] and Expectiles [10], highlighting the importance of flexibility in defining risk using a diverse set of risk measures in the risk budgeting framework.

A general stochastic formulation of the risk budgeting problem is discussed in [15], based on a common probabilistic representation shared by many risk measures. Specifically, the risk budgeting problem is framed as a stochastic optimisation problem whenever the risk measure can be expressed as the minimum of a convex function written in expectation form, as is the case, for example, with Bayes risk measures [23]. To compute risk budgeting portfolios across a broad class of risk measures – including entire classes of spectral risk measures [1] and deviation measures as defined in [42] – the authors in [15] propose a standard projected stochastic gradient descent (SGD) algorithm.

Our primary objective in this paper is to propose and study Mirror Descent (MD) algorithms for computing risk budgeting portfolios associated with general positive homogeneous and sub-additive risk measures. Originating from the pioneering work of Nemirovskij and Yudin [36], MD provides a natural framework for addressing optimisation problems, particularly when the mirror mapping (also known as proximal mapping) is explicit. MD has the appealing property of defining a smooth trajectory that remains within the constrained set, thus eliminating the need for additional projection steps, as required in the SGD algorithm proposed in [15]. This effectively pushes the boundaries of the domain to an infinite

distance from any interior point. For a comprehensive introduction and extensions to the stochastic setting and convex-concave stochastic saddle-point problems, we refer to [37] and [35]. Additional almost sure convergence results for convex and non-convex stochastic optimisation problems can be found in [29] and [47]. We also refer to the recent work [19] in which a stochastic MD scheme is proposed for optimal portfolio allocation with ES penalty. Thus, MD appears to be a natural method to implement for solving the constrained convex minimisation problem associated with the risk budgeting problem.

However, a key assumption for deriving $L^1(\mathbb{P})$ convergence (also known as regret bounds) and concentration inequalities for MD schemes – the boundedness of the objective's gradient – is not satisfied in our context. Indeed, the gradient of the objective function diverges on the boundary of the domain, preventing direct application of standard results (see, e.g. [35]). To overcome this major challenge, we instead employ a tailored, tamed version of the gradient, which, like the original gradient, vanishes only at the unique minimiser of the objective function but remains uniformly bounded across the entire domain. This modified gradient enables us to develop both deterministic and stochastic MD schemes and to establish their convergence to the unnormalised risk budgeting portfolio for general risk measures under mild assumptions. In both deterministic and stochastic settings, we further establish a non-asymptotic $a.s.$ convergence rate for the weighted averaged sequence. Our stochastic framework accommodates a wide range of risk measures, including volatility, ES, deviation measures, and Variantiles, among others. To the best of our knowledge, our scheme is the first to ensure $a.s.$ convergence along with a non-asymptotic quantitative convergence rate for computing risk budgeting portfolios across general risk measures.

The paper is organized as follows. In Section 2, we briefly introduce the risk budgeting problem, emphasising its formulation as the solution to a strictly convex optimisation problem, which serves as the cornerstone for our analysis in the subsequent sections. Section 3 begins by presenting our tailored, tamed version of the gradient and some key results essential for establishing the convergence of our schemes. We then introduce our deterministic MD (DMD) algorithm, providing convergence results along with a convergence rate for the weighted averaged sequence. Next, we propose and analyse a stochastic MD (SMD) scheme in a stochastic framework where the risk measure is expressed as the minimum of a convex map. We establish the $a.s.$ convergence of the SMD algorithm, along with a non-asymptotic $a.s.$ convergence rate for the weighted average sequence. In Section 4, we conduct an extensive numerical analysis to assess the practical advantages of the MD approach over standard projected gradient descent methods, particularly in the stochastic setting, for accurately computing risk budgeting portfolios of various sizes under different risk measures. The proofs of the main and auxiliary results are deferred to Appendix A.

Notations: $\mathbb{R}^0$ stands for $\mathbb{R}\backslash\{0\}$ and $\mathbb{R}_+^0$ for $\mathbb{R}_+\backslash\{0\}$. For a vector $x \in \mathbb{R}^d$, we write $\|x\| = \|x\|_1 = \sum_{i=1}^d |x_i|$ for its norm and $\|x\|_\infty = \max_{1 \le i \le d} |x_i| = \sup_{\|y\|_1 \le 1} \langle y, x \rangle$ stands for the dual norm of $\|.\|_1$. The simplex of dimension d is denoted by $\Delta_d = \{u \in \mathbb{R}_+^d : u_1 + \cdots + u_d = 1\}$ and we let let $\Delta_d^{>0} := \{u \in (\mathbb{R}_+^0)^d : u_1 + \cdots + u_d = 1\}$.

2 The Risk Budgeting Problem

We consider a financial market composed of d assets whose returns are given by the $\mathbb{R}^d$-valued random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ which is assumed to be rich enough to accommodate all of the subsequent random variables. A financial portfolio is identified with its corresponding vector of weights $u = (u_1, \cdots, u_d)$ belonging to the simplex Δ_d. If the portfolio is given by the vector $u \in \Delta_d$, then $-\langle u, X \rangle = -\sum_{i=1}^d u_i X_i$ corresponds to its loss, recalling that $\langle ., . \rangle$ stands for the Euclidean scalar product on $\mathbb{R}^d$.

In order to assess the risk of the loss of a financial portfolio, we consider a risk measure ρ that is a function mapping a random variable Z to the real number $\rho(Z) \in \mathbb{R}$. We deliberately omit the space on which ρ is defined but typically one has $Z \in L^0(\mathbb{P})$, $L^1(\mathbb{P})$ or $L^2(\mathbb{P})$.

The risk measure ρ is said to be risk budgeting compatible (RB-compatible for short) if it is positive homogeneous and sub-additive, namely if for any $\lambda \geq 0$,

$$\rho(\lambda Z) = \lambda \rho(Z) \tag{PH}$$

and, for any real-valued random variables Z_1 and Z_2

$$\rho(Z_1 + Z_2) \leq \rho(Z_1) + \rho(Z_2). \tag{SA}$$

To any RB-compatible risk measure ρ, we associate the map

$$r_\rho : \mathbb{R}_+^d \ni y \to \rho(-\langle y, X \rangle) \in \mathbb{R}.$$

In particular, the risk associated to the portfolio u is thus given by $r_\rho(u) = \rho(-\langle u, X \rangle)$.

In what follows, we will always assume that the map r_ρ is continuous on $\mathbb{R}_+^d$ and continuously differentiable on the open set $(\mathbb{R}_+^0)^d$. The positive homogeneity property of the risk measure ρ combined with Euler's homogeneous function theorem implies that the risk linked with a portfolio represented by its vector of weights u can be decomposed as follows

$$r_\rho(u) = \sum_{i=1}^d u_i \partial_{u_i} r_\rho(u), \quad u \in \Delta_d^{>0}. \tag{1}$$

In the risk budgeting literature, the i-th term of the above sum, $u_i \partial_{u_i} r_\rho(u)$, is referred to as the risk contribution of asset i to the overall portfolio risk. The risk budgeting problem, therefore, involves identifying a vector of weights $u^\star$ such that the risk contributions align with predetermined proportions, represented by a vector of risk budgets b, of the total risk.

To be more specific, for a given vector of risk budget $b \in \Delta_d^{>0}$, we say that $u \in \Delta_d^{>0}$ solves the risk budgeting problem RB_b if the following condition holds:

$$u_i \partial_{u_i} r_\rho(u) = b_i r_\rho(u), \quad i = 1, \cdots, d.$$

Throughout the paper, we will assume that the risk of any long-only portfolio is positive, that is, $r_\rho(u) > 0$, for any $u \in \Delta_d$. We refer to [15] and also [8, 26] where a similar assumption is made in the context of ES and shortfall risk minimization.

The existence and uniqueness of a vector of weights that solves RB_b for a given vector of risk budgets $b \in \Delta_d^{>0}$, along with its characterization as the solution to a strictly convex optimization problem, has been investigated in several works, see for instance [4, 15, 14, 28, 45]. Here, we recall the characterization provided in [15].

Theorem 1. *[15, Theorems 1 and 2] Let $b \in \Delta_d^{>0}$. Let $g : \mathbb{R}_+ \to \mathbb{R}$ be a continuously differentiable, convex and increasing function. Let the function $\Gamma_g : (\mathbb{R}_+^0)^d \to \mathbb{R}$ be defined by*

$$\Gamma_g : y \mapsto g(r_\rho(y)) - \sum_{i=1}^d b_i \log y_i.$$

There exists a unique minimizer $y^\star$ of the strictly convex function Γ_g satisfying $\nabla \Gamma_g(y^\star) = 0$ and

$$u^* := \frac{y^\star}{\|y^\star\|_1}$$

solves the risk budgeting problem RB_b. Moreover, if u is a solution to RB_b then

$$u = u^\star.$$

We highlight that this result serves as the cornerstone of our analysis. Indeed, the numerical schemes we propose will be specifically designed to target the unique minimizer $y^\star$ of the aforementioned strictly convex optimization problem. Theorem 1 subsequently demonstrates that the renormalized minimizer yields the unique solution to the risk budgeting problem RB_b.

3   Mirror Descent Algorithm

To solve the minimization problem introduced in Theorem 1, we are naturally led to use gradient descent schemes and their stochastic counterpart when the map ρ writes as an expectation of a random function. However, several specific difficulties appear. The first one is the fact that the gradient of Γ_g diverges on $\partial(\mathbb{R}_+^0)^d$ due to the presence of the term $\sum_{i=1}^d b_i \log(y_i)$, thus requiring a specific treatment. The second difficulty is that the minimization problem is set over $(\mathbb{R}_+^0)^d$ and not $\mathbb{R}^d$.

To address the first challenge, we will employ a tailored, tamed version of the gradient, which, like $\nabla \Gamma_g$, vanishes only at the unique minimizer of Γ_g. The second challenge, stemming from the constrained nature of the minimisation problem, is managed using the MD algorithm-a flexible optimisation technique widely used in convex optimisation and machine learning, as noted in the introduction. This algorithm has the advantageous property of defining a trajectory that remains entirely within the constrained set, thereby eliminating the need for additional projection steps that could be difficult to implement in practice.

3.1   Taming the singularity of the gradient

The target of our MD algorithm is the unique minimizer $y^\star$ of the strictly convex map Γ_g. However, as already mentioned, a significant challenge in efficiently implementing this scheme lies in the divergence of

the gradient at the boundary of its domain $\partial(\mathbb{R}_+^0)^d$. Specifically, since $b_i > 0$, it follows that $\partial_{y_i}\Gamma_g(y) \to -\infty$, as soon as $y_i \downarrow 0$ for some $i \in \{1, \cdots, d\}$. Hence, $\sup_{y \in (\mathbb{R}_+^0)^d} \|\nabla\Gamma_g(y)\|_\infty = \infty$.

As highlighted in [35], the boundedness of the objective function's gradient on the constrained set is a crucial prerequisite for establishing a quantitative convergence rate for this type of recursive procedure.

In order to circumvent this major difficulty, we build our numerical procedure on a tamed version of $\nabla\Gamma_g$. We let $\underline{y} = \min_{1 \le i \le d} y_i$ and define

$$\kappa(y) = \underline{y} \wedge 1, \quad y \in \mathbb{R}_+^d.$$

We then remark that the map $(\mathbb{R}_+^0)^d \ni y \mapsto \kappa(y)\,\nabla\Gamma_g(y)$ can be extended by continuity on $\mathbb{R}_+^d$.[1] Indeed, note that if $y_i = 0$ for some $i \in \{1, \cdots, d\}$, then, $\underline{y} = 0$ so that

$$(\kappa(y)\,\nabla\Gamma_g(y))_j = \underline{y}\,g'(r_\rho(y))\partial_{y_j}r_\rho(y) - b_j\frac{\underline{y}}{y_j}$$
$$= \begin{cases} -b_j, & \text{if } y_j = \underline{y} = 0 \\ 0, & \text{otherwise} \end{cases}$$

so that $\kappa(y)\,\nabla\Gamma_g(y) \ne 0$ on $\mathbb{R}_+^d \backslash (\mathbb{R}_+^0)^d$. We thus conclude that the unique minimizer $y^\star$ of Γ_g satisfies

$$\{y^*\} = \{y \in (\mathbb{R}_+^0)^d : \nabla\Gamma_g(y) = 0\} = \{y \in \mathbb{R}_+^d : \kappa(y)\,\nabla\Gamma_g(y) = 0\}.$$

It is here important to emphasize that, in contrast to $\nabla\Gamma_g$, the map $y \mapsto \kappa(y)\,\nabla\Gamma_g(y)$ remains bounded on any centred closed ball B_m of radius m and dimension d, since for any $y \in B_m$, the following holds

$$\|\kappa(y)\nabla\Gamma_g(y)\|_\infty$$
$$\le \max_{y \in B_m}\{g'(r_\rho(y))\|\nabla r_\rho(y)\|_\infty\} + \max_{1 \le i \le d} b_i \frac{\underline{y} \wedge 1}{y_i} \tag{2}$$
$$\le M_\star(m, b, d) := \max_{y \in B_m}\{g'(r_\rho(y))\|\nabla r_\rho(y)\|_\infty\} + \max_{1 \le i \le d} b_i.$$

We conclude this section by the following technical lemma that will play a central role to address the convergence of our algorithm.

Lemma 2. *For all $y \in \mathbb{R}_+^d$, $y \ne y^\star$, it holds*

$$\langle y - y^\star, \kappa(y)\nabla\Gamma_g(y)\rangle > 0.$$

Proof. The strict convexity of Γ_g guarantees that the desired property holds on $(\mathbb{R}_+^0)^d$. Now, if $y \in \mathbb{R}_+^d \backslash (\mathbb{R}_+^0)^d$ then $\underline{y} = 0$ so that

$$\langle y - y^\star, \kappa(y)\nabla\Gamma_g(y)\rangle = \sum_{j:y_j=0}(y_j - y_j^\star)(-b_j) = \sum_{j:y_j=0} y_j^\star b_j > 0$$

which concludes the proof. $\square$

---

[1]Throughout the article, we will use the convention $0/0 = 1$.

3.2 Deterministic Mirror Descent Algorithm

Although our primary objective is to address the stochastic framework, where ρ is expressed as an expectation, in certain practical settings the function $y \mapsto r_\rho(y) = \rho(-\langle y, X \rangle)$ is known explicitly or semi-explicitly. In such cases, it is advantageous to rely on a deterministic Mirror Descent (DMD) algorithm to approximate $y^\star$. We introduce the Bregman divergence

$$D_F(y, y') = F(y) - F(y') - \langle \nabla F(y'), y - y' \rangle, \quad y, y' \in (\mathbb{R}_+)^d$$

generated by the negative entropy function $F(y) = \sum_{i=1}^d y_i \log(y_i)$, $y \in \mathbb{R}_+^d$. The map D_F is also known as the Kullback–Leibler divergence inasmuch it satisfies

$$D_F(y, y') = \sum_{i=1}^d y_i \log(y_i/y_i') - \sum_{i=1}^d y_i + \sum_{i=1}^d y_i'.$$

Let $m > 0$. Using the fact that the map $[0, m] \ni y \mapsto y \log(y) - \frac{1}{2m} y^2$ is convex, one deduces

$$D_F(y, y') \geq \frac{1}{2m} \|y - y'\|_1^2, \quad \text{for any } y, y' \in \mathbb{R}_+^d \cap B_m. \tag{3}$$

Let $(\gamma_k)_{k \geq 1}$ be a deterministic positive learning sequence. Starting from $y^0 \in (\mathbb{R}_+^0)^d \cap B_m$, the deterministic MD algorithm constructs the sequence $(y^k)_{k \geq 1}$ inductively where at step k, given $y^k \in (\mathbb{R}_+^0)^d \cap B_m$, y^{k+1} is obtained by solving the optimization problem

$$y^{k+1} = \mathrm{P}_{y^k}^m(\gamma_{k+1} \kappa(y^k) \nabla \Gamma_g(y^k))$$

where, for a given $y \in (\mathbb{R}_+^0)^d \cap B_m$, the proximal mapping $\mathrm{P}_y^m : \mathbb{R}^d \to (\mathbb{R}_+^0)^d \cap B_m$, associated with D_F, is defined by

$$\mathrm{P}_y^m(v) = \underset{w \in (\mathbb{R}_+^0)^d \cap B_m}{\mathrm{argmin}} \{\langle v, w - y \rangle + D_F(y, w)\} = \begin{cases} ye^{-v}, & \text{if } \|ye^{-v}\|_1 \leq m, \\[2mm] \frac{m}{\|ye^{-v}\|_1} ye^{-v}, & \text{otherwise,} \end{cases} \tag{4}$$

where $ye^{-v} = (y_1 e^{-v_1}, \cdots, y_d e^{-v_d})$.

The key point here is that the unique solution to the above constrained strictly convex minimization problem is explicit. Indeed, from (4), it directly follows that $y^{k+1} = (y_1^{k+1}, \cdots, y_d^{k+1})$, for $k \geq 0$, is given by

$$\begin{aligned} y_i^{k+1} &= (\mathrm{P}_{y^k}^m(\gamma_{k+1} \kappa(y^k) \nabla \Gamma_g(y^k)))_i \\[1mm] &= \begin{cases} y_i^k e^{-\gamma_{k+1} \kappa(y^k) \partial_{y_i} \Gamma_g(y^k)}, & \text{if } \|y^k e^{-\gamma_{k+1} \kappa(y^k) \nabla \Gamma_g(y^k)}\|_1 \leq m, \\[2mm] \frac{m}{\|y^k e^{-\gamma_{k+1} \kappa(y^k) \nabla \Gamma_g(y^k)}\|_1} y_i^k e^{-\gamma_{k+1} \kappa(y^k)) \partial_{y_i} \Gamma_g(y^k)}, & \text{otherwise} \end{cases} \quad i = 1, \cdots, d. \end{aligned} \tag{5}$$

The recursive scheme outlined above can be implemented as soon as the gradient $\nabla \Gamma_g$ is available. The following theorem presents the principal result of this section, with its proof deferred to Appendix A.1.

Theorem 3. *Assume that $m \geq \|y^\star\|_1$. Assume additionally that the learning step sequence $(\gamma_n)_{n \geq 1}$ satisfies $\sum_{n \geq 1} \gamma_n = \infty$ and $\sum_{n \geq 1} \gamma_n^2 < \infty$. Then, the sequence $(y^n)_{n \geq 0}$ defined by (5) converges to $y^\star$ as $n \uparrow \infty$. Moreover, the weighted averaged sequence $(\bar{y}^n)_{n \geq 1}$ of $(y^n)_{n \geq 0}$ defined by*

$$\bar{y}^n = \frac{\sum_{k=1}^n \gamma_k y^{k-1}}{\sum_{k=1}^n \gamma_k}, \quad n \geq 1,$$

satisfies

$$\Gamma_g(\bar{y}^n) - \Gamma_g(y^\star) \leq \frac{D_F(y^\star, y^0) + \frac{1}{2} m M_\star^2 \sum_{k=0}^n \gamma_{k+1}^2}{(\min_{0 \leq k \leq n-1} \underline{y}^k \wedge 1) \sum_{k=0}^n \gamma_{k+1}}, \quad n \geq 1, \tag{6}$$

where $M_\star$ is given by (2).

Remark 1. *Concerning the convergence rate given by (6), under the two conditions $\sum_{n \geq 1} \gamma_n = \infty$ and $\sum_{n \geq 1} \gamma_n^2 < \infty$, one may select $\gamma_n = \gamma n^{-\frac{1}{2} - \delta}$, with $\delta \in (0, 1/2]$ and $\gamma > 0$. A straightforward comparison between series and integrals then yields*

$$\Gamma_g(\bar{y}^n) - \Gamma_g(y^\star) \leq K n^{-\frac{1}{2} + \delta}, \quad n \geq 1.$$

Alternatively, choosing $\gamma_n = n^{-1/2} \log(n)^{-(\frac{1}{2} + \delta)}$, for some $\delta > 0$, ensures that the Bertrand series $\sum_{n \geq 1} \gamma_n^2$ converges, yielding

$$\Gamma_g(\bar{y}^n) - \Gamma_g(y^\star) \leq K \frac{(\log n)^{\frac{1}{2} + \delta}}{\sqrt{n}}, \quad n \geq 1,$$

for some constant $K < \infty$ which depends on m, d, $M_\star$ and δ. Thus, we recover the usual convergence rate achieved by deterministic MD schemes, up to a logarithmic factor [35].

Remark 2. *An explicit bound for the initialisation error term $D_F(y^\star, y^0)$ can be established in terms of the parameter m and the dimension d by appropriately choosing y^0. Specifically, if y^0 is selected as $\operatorname{argmin}_{\mathbb{R}_+^d \cap B_m} F = e^{-1}(1, \cdots, 1)$ if $m \geq e^{-1} d$ or $m d^{-1}(1, \cdots, 1)$ otherwise, then $F(y^0) = \min_{\mathbb{R}_+^d \cap B_m} F \geq -d e^{-1}$. Furthermore, if $m \leq 1$ then since $F \leq 0$ on $\mathbb{R}_+^d \cap B_m$, we have $\max_{\mathbb{R}_+^d \cap B_m} F \leq 0$; whereas, if $m \geq 1$, then $F(x) = \|x\|_1 \sum_{i=1}^d \left\{ \frac{x_i}{\|x\|_1} \log(\frac{x_i}{\|x\|_1}) + \frac{x_i}{\|x\|_1} \log(\|x\|_1) \right\} \leq m \log(m)$ for any $x \in \mathbb{R}_+^d \cap B_m$. Hence,*

$$D_F(y^\star, y^0) \leq \max_{\mathbb{R}_+^d \cap B_m} F - \min_{\mathbb{R}_+^d \cap B_m} F \leq (m \log(m))_+ + \frac{d}{e}.$$

This upper bound is particularly valuable as it illuminates the dependence of the convergence rate on the dimension, indicating that this rate grows at most linearly with d.

Remark 3. *The above result indicates that one must select m sufficiently large, ensuring that $m \geq \|y^\star\|_1$. Given that $y^\star$ is unknown, this necessitates a somewhat arbitrary choice for the user.[2] From a numerical standpoint, it is evident that choosing m too small – such that $m < \|y^\star\|_1$ – will prevent the algorithm from converging to $y^\star$. With respect to the convergence rate, although the upper bound is indeed influenced by selecting an excessively large value of m, we do not observe a significant impact on the convergence rate of the weighted averaged sequence $(\bar{y}^n)_{n \geq 1}$ during practical implementations. We refer the reader to Section 4 for a discussion of numerical evidence.*

---

[2]In fact, one can use the identity $r_\rho(y^\star) = \frac{1}{g'(r_\rho(y^\star))}$ (from Theorem 1) to determine a reasonable choice for $m \geq \|y^\star\|_1$. For example, when $g = \mathrm{Id}$, this implies $r_\rho(y^\star) = 1$, indicating that the risk of the portfolio under the weights $y^\star$ equals 1. This equality provides insight into the order of magnitude of the elements of $y^\star$.

3.3 Stochastic Mirror Descent

3.3.1 Stochastic framework

In this section, we propose and analyze a stochastic version of the previous MD algorithm within a stochastic framework. This is achieved in the presence of a convex loss function $\mathbb{R}^2 \ni (\xi, x) \mapsto L(\xi, x)$, such that $L(\xi, -\langle y, X \rangle) \in L^1(\mathbb{P})$ for any $(\xi, y) \in \mathbb{R} \times \mathbb{R}_+^d$, and satisfying the condition:

$$g(r_\rho(y)) = \min_{\xi \in \mathbb{R}} \mathbb{E}[L(\xi, -\langle y, X \rangle)] = \mathbb{E}[L(\xi^\star(y), -\langle y, X \rangle)], \quad y \in (\mathbb{R}_+^0)^d, \tag{7}$$

where the minimizer $\xi^\star(y)$ is assumed to be uniquely defined for every $y \in (\mathbb{R}_+^0)^d$.

As already noticed in [15], when $g = \mathrm{Id}$, the associated risk measure is linked to the notions of Bayes pair and Bayes risk measure mathematically characterized as the minimization of an expected loss function over a set of possible decisions [23]. Such risk measures ensure robust risk minimization and optimal decision-making under uncertainty.

From both practical and theoretical perspectives, it is valuable to consider the general framework where $g \neq \mathrm{Id}$. Perhaps the most significant example is volatility, which attains its minimum under the square map $g(x) = x^2$. While a stochastic algorithm may not be the most efficient numerical method for addressing the problem when volatility is chosen as the risk measure, it is crucial that the stochastic framework accommodates volatility, as it is the most widely used risk measure in the risk budgeting paradigm. Other examples include L^p-deviation risk measures for $p \geq 1$, which similarly are characterized by a minimum under the map $g(x) = x^p$. For a more detailed discussion, we refer to Section 3.3.3.

In this framework, the optimization problem related to the RB problem equivalently writes

$$\min_{y \in (\mathbb{R}_+^0)^d} \Gamma_g(y) = \min_{z := (\xi, y) \in \mathbb{R} \times (\mathbb{R}_+^0)^d} \{h(z) := \mathbb{E}[H(z, X)]\} \tag{8}$$

where

$$H(z, X) := L(\xi, -\langle y, X \rangle) - \sum_{i=1}^{d} b_i \log(y_i), \quad z = (\xi, y) \in \mathbb{R} \times (\mathbb{R}_+^0)^d.$$

The following result demonstrates the existence of a solution to the above optimization problem, which, in turn, provides the unique solution to the risk budgeting problem. The proof of this result is deferred to Appendix A.2.

Proposition 1. *Assume that the function $(\xi, x) \mapsto L(\xi, x)$ satisfies the following regularity and integrability conditions:*

- $\mathbb{P}(X \notin \mathcal{D}_L) = 0$ *where* $\mathcal{D}_L = \{x \in \mathbb{R}^d : \mathbb{R} \times (\mathbb{R}_+^0)^d \ni (\xi, y) \mapsto L(\xi, -\langle y, x \rangle) \text{ is continuously differentiable}\}$.

- *For any centered ball $B_R \subset \mathbb{R} \times (\mathbb{R}_+^0)^d$ of radius $R > 0$,*

$$\sup_{(\xi, y) \in B_R} \{|\partial_\xi L(\xi, -\langle y, X \rangle)| + |X||\partial_x L(\xi, -\langle y, X \rangle)|\} \in L^1(\mathbb{P}).$$

Then, the map h is continuously differentiable on $\mathbb{R} \times (\mathbb{R}_+^0)^d$ with

$$\partial_\xi h(\xi, y) = \mathbb{E}[\partial_\xi L(\xi, -\langle y, X \rangle)] \quad and \quad \partial_{y_i} h(\xi, y) = \mathbb{E}\Big[- X_i \partial_x L(\xi, -\langle y, X \rangle) - \frac{b_i}{y_i} \Big], \quad i = 1, \cdots, d,$$

so that

$$\operatorname{Argmin} h = \big\{ z \in \mathbb{R} \times (\mathbb{R}_+^0)^d : \nabla h(z) = 0 \big\}.$$

Moreover, if $(\xi^\star, y^\star) \in \operatorname{Argmin} h$, then $y^\star$ is the unique minimizer of Γ_g so that $u^\star = \frac{y^\star}{\|y^\star\|_1}$ is the unique solution to the risk budgeting problem RB_b.

The above proposition suggests that, to solve the original risk budgeting problem, we must address the stochastic optimization problem (8). To this end, we develop a stochastic Mirror Descent (SMD) algorithm with limiting point in $\{\nabla h = 0\}$.

As it can be noticed, somehow we end up with the same difficulty as before since the gradient of $H(., X)$ is unbounded due to the presence of the term $\sum_{i=1}^d b_i \log(y_i)$. We thus apply the same taming technique as in the deterministic framework, namely, similarly to what we observed in Section 3.1, it holds

$$\left\{ z \in \mathbb{R} \times (\mathbb{R}_+^0)^d : \nabla h(z) = \mathbb{E}\left[\begin{pmatrix} \partial_\xi H(z, X) \\ \nabla_y H(z, X) \end{pmatrix} \right] = 0 \right\} = \left\{ z \in \mathbb{R} \times (\mathbb{R}_+)^d : \mathbb{E}\left[\begin{pmatrix} \partial_\xi H(z, X) \\ \kappa(y) \nabla_y H(z, X) \end{pmatrix} \right] = 0 \right\}.$$

recalling that $\kappa(y) = \underline{y} \wedge 1$ and $\underline{y} = \min_{1 \le i \le d} y_i$.

Indeed, similarly to the deterministic case, if $z \in \mathbb{R} \times (\mathbb{R}_+^d \setminus (\mathbb{R}_+^0)^d)$ then $\underline{y} = 0$ so that $\kappa(y) \partial_{y_i} H(z, X) = -b_i \neq 0$ if $y_i = 0$ and 0 otherwise, which in turn implies

$$\mathbb{E}\left[\begin{pmatrix} \partial_\xi H(z, X) \\ \kappa(y) \nabla_y H(z, X) \end{pmatrix} \right] \neq 0 \text{ for } z \in \mathbb{R} \times (\mathbb{R}_+^d \setminus (\mathbb{R}_+^0)^d).$$

Analogously to the previous approach, we will construct our SMD algorithm using $(\partial_\xi H(z, X), \kappa(y) \nabla_y H(z, X))$ instead of $\nabla H(z, X)$. The following result, akin to Lemma 2 follows from the convexity of h, and its proof is therefore omitted.

Lemma 4. *For all $z \in \mathbb{R} \times \mathbb{R}_+^d$, $z \neq z^\star := (\xi^\star, y^\star)$, it holds*

$$\left\langle z - z^\star, \mathbb{E}\left[\begin{pmatrix} \partial_\xi H(z, X) \\ \kappa(y) \nabla_y H(z, X) \end{pmatrix} \right] \right\rangle > 0.$$

3.3.2 Stochastic Mirror Descent algorithm

In order to deal with the couple $z = (\xi, y)$, we will work with the so-called Bregman divergence over pairs defined by

$$D_G(z, z') = \frac{1}{2}(\xi - \xi')^2 + D_F(y, y'), \quad z, z' \in \mathbb{R} \times (\mathbb{R}_+^0)^d,$$

associated to the map

$$G(z) = \frac{1}{2}\xi^2 + F(y), \quad z = (\xi, y) \in \mathbb{R} \times B_m,$$

recalling that F stands for the negative entropy function. From (3), we readily get

$$D_G(z, z') \geq \frac{1}{2}(\xi - \xi')^2 + \frac{1}{2m}\|y - y'\|_1^2 \geq \frac{1}{4(m \vee 1)}\|z - z'\|_1^2, \quad z, z' \in \mathbb{R} \times (\mathbb{R}_+^d \cap B_m). \tag{9}$$

Let $(\gamma_k)_{k \geq 1}$ be a deterministic positive learning sequence as before. Starting from an initial point $z^0 = (\xi^0, y^0) \in \mathbb{R} \times (\mathbb{R}_+^0)^d$, the SMD algorithm generates the sequence $(z^k)_{k \geq 1}$ inductively. At each iteration k, given $z^k = (\xi^k, y^k) \in \mathbb{R} \times ((\mathbb{R}_+^0)^d \cap B_m)$, the update vector $z^{k+1} = (\xi^{k+1}, y^{k+1})$ is obtained by solving the following optimization problem:

$$z^{k+1} = \mathrm{P}_{z^k}^m \left(\gamma_{k+1} \begin{pmatrix} \partial_\xi H(z^k, X^{k+1}) \\ \kappa(y^k)\nabla_y H(z^k, X^{k+1}) \end{pmatrix} \right)$$

where, for a given $z \in \mathbb{R} \times ((\mathbb{R}_+^0)^d \cap B_m)$, the proximal mapping $P_z^m : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R} \times ((\mathbb{R}_+^0)^d \cap B_m)$, associated to D_G, is defined by

$$\mathrm{P}_z^m(v) = \underset{w \in \mathbb{R} \times ((\mathbb{R}_+^0)^d \cap B_m)}{\operatorname{argmin}} \{\langle v, w - z \rangle + D_G(z, w)\}, \tag{10}$$

and $(X^k)_{k \geq 1}$ is a sequence of i.i.d. copies of X, independent of z^0.

Analogous to the deterministic case, the unique solution to the above strongly convex minimization problem is explicit. Specifically, for $k \geq 0$, the update rule for $z^{k+1} = (\xi^{k+1}, y^{k+1})$ is given by:

$$\begin{cases} \xi^{k+1} = \xi^k - \gamma_{k+1}\partial_\xi H(z^k, X^{k+1}) \\ y^{k+1} = \mathrm{P}_{y^k}^m(\gamma_{k+1}\kappa(y^k)\nabla_y H(z^k, X^{k+1})) \end{cases} \tag{11}$$

where, for $y \in (\mathbb{R}_+^0)^d$, the proximal map P_y^m is defined as in (4). We also introduce the natural filtration $\mathcal{F} = (\mathcal{F}_k)_{k \geq 0}$, $\mathcal{F}_k = \sigma(\xi_0, y_0, X^1, \ldots, X^k)$, $k \geq 0$, associated with the SMD scheme $(z^k)_{k \geq 0}$.

The following theorem presents the central result of this section, establishing the *a.s.* convergence of the sequence $(z^n)_{n \geq 0}$, alongside an *a.s.* convergence rate for the associated weighted averaged sequence $(\bar{z}^n)_{n \geq 1}$ along its trajectory. The proof of this result is deferred to Appendix A.3.

Theorem 5. *Assume that $m \geq \|y^\star\|_1$, that the regularity and integrability conditions on the function L of Proposition 1 are satisfied and that*

$$\sup_{(\xi, y) \in \mathbb{R} \times (\mathbb{R}_+^d \cap B_m)} \{\mathbb{E}[(\partial_\xi L(\xi, -\langle y, X \rangle))^2] + \mathbb{E}[X_i^2(\partial_x L(\xi, -\langle y, X \rangle))^2]\} < \infty, \quad 1 \leq i \leq d. \tag{12}$$

Assume additionally that the learning step sequence $(\gamma_n)_{n \geq 1}$ satisfies $\sum_{n \geq 1} \gamma_n = \infty$ and $\sum_{n \geq 1} \gamma_n^2 < \infty$. Then, the sequence $(z^n)_{n \geq 0}$ defined by (11) converges a.s. to the unique minimum $z^\star$ of h as $n \uparrow \infty$. Moreover, the weighted averaged sequence $(\bar{z}^n)_{n \geq 1}$ of $(z^n)_{n \geq 0}$ defined by

$$\bar{z}^n = \frac{\sum_{k=1}^n \gamma_k z^{k-1}}{\sum_{k=1}^n \gamma_k}, \quad n \geq 1,$$

satisfies the following a.s. upper-bound

$$h(\bar{z}^n) - h(z^\star) \leq \frac{\frac{1}{2}(\xi^0 - \xi^\star)^2 + D_F(y^\star, y^0) - M_n + \frac{1}{2}m \vee 1 \sum_{k=0}^n \gamma_{k+1}^2 Y_{k+1}}{(\min_{0 \leq k \leq n-1} \underline{y}^k \wedge 1) \sum_{k=0}^n \gamma_{k+1}}, \tag{13}$$

where $(M_n)_{n \geq 1}$ is an $\mathcal{F}$-martingale satisfying $\sup_{n \geq 1} \mathbb{E}[M_n^2] < \infty$ (thus converging a.s. as $n \uparrow \infty$ to $M_\infty < \infty$) and $(Y_n)_{n \geq 1}$ is a non-negative sequence satisfying $\sup_{n \geq 0} \mathbb{E}[Y_{n+1}|\mathcal{F}_n] \leq N_\star^2$ a.s. for some constant $N_\star^2 < \infty$.

Remark 4. *Analogous to Remark 2, selecting $y_0 = e^{-1}(1, \cdots, 1)$ if $m \geq e^{-1}d$ or $y_0 = md^{-1}(1, \cdots, 1)$ otherwise, provides explicit control over the initialization error: $D_F(y^\star, y^0) \leq (m \log m)_+ + \frac{d}{e}$.*

Remark 5. *Note that standard convergence rate results for the sequence $(\bar{z}^n)_{n \geq 1}$ are typically expressed as an upper bound on $\mathbb{E}[h(\bar{z}^n) - h(z^\star)]$ of order $(\sum_{k=0}^n \gamma_{k+1})^{-1}$, under the conditions $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$; see, for instance, [36]. In our case, however, the upper-boudn (13) gives*

$$\limsup_n \Big(\sum_{k=0}^n \gamma_{k+1}\Big) \times (h(\bar{z}^n) - h(z^\star)) < \infty \quad a.s.$$

due to the fact that $\sup_{n \geq 1} M_n < \infty$, $\sum_{n \geq 1} \gamma_{n+1}^2 Y_{n+1} < \infty$ and $\min_{n \geq 0} \underline{y}^n > 0$ a.s. While it would be valuable to establish an upper bound for $\mathbb{E}[h(\bar{z}^n) - h(z^\star)]$, this proves to be quite challenging, as it would require an $L^p(\mathbb{P})$-control, for some $p > 1$, on $(\min_{0 \leq k \leq n} \underline{y}^k \wedge 1)^{-1}$ uniformly in n.

Remark 6. *Similarly to the deterministic framework, see Remark 3, it is essential to select m sufficiently large, ensuring that $m \geq \|y^\star\|$. Given that $y^\star$ remains unknown, this necessitates a somewhat arbitrary or blind choice on the part of the user. While an excessively large m may influence the upper bound, we find that, in practical implementations, it does not significantly impact the convergence rate of the weighted averaged sequence $(\bar{z}^n)_{n \geq 1}$. We again refer the reader to Section 4 for a discussion of the numerical evidence regarding the impact of the choice of m.*

3.3.3   Some examples: Volatility, Expected Shortfall, Deviation measures

The most commonly used risk measure for the risk budgeting problem is certainly the volatility. Taking $g(x) = x^2$, $L(\xi, x) = (x - \xi)^2$ and assuming that $X \in L^2(\mathbb{P})$, we observe that

$$g(r_\rho(y)) = y^T \Sigma y = \min_{\xi \in \mathbb{R}} \mathbb{E}[L(\xi, -\langle y, X \rangle)]. \tag{14}$$

Thus, volatility as a risk measure naturally fits within the stochastic framework outlined in Section 3.3.1. In particular, Proposition 1 holds, with $\operatorname{Argmin} h = \{(\xi^\star, y^\star)\}$, $\xi^\star = \mathbb{E}[-\langle y^\star, X \rangle]$ and $y^\star$ is the unique minimizer of Γ_g when volatility is employed as the risk measure. Note, however, that the condition (12) in Theorem 5 is not met, since $\xi \mapsto \mathbb{E}[(\partial_\xi L(\xi, -\langle y, X \rangle))^2]$ and $\mathbb{E}[(\partial_x L(\xi, -\langle y, X \rangle))^2]$ both exhibit quadratic growth in ξ, uniformly in $y \in \mathbb{R}_+^d \cap B_m$. Nevertheless, the proof of Theorem 5 can be readily adapted to overcome this minor difficulty, ensuring *a.s.* convergence of the sequence $(z^n)_{n \geq 1}$, along with a similar convergence rate for $(\bar{z}^n)_{n \geq 1}$. We shall not dwell further on the adaptation of this proof.

Our second example is the Expected Shortfall (ES), calculated at a given confidence level $\alpha \in (0, 1)$. This risk measure is closely related to another well-known measure, the Value-at-Risk (VaR). Both ES and VaR are among the most widely employed risk measures in the finance and insurance industries. As in [24, 2], for a real-valued random vector $Z \in L^1(\mathbb{P})$, these two risk measures are defined by

$$\mathbb{VaR}_\alpha(Z) = \inf\{\xi \in \mathbb{R} : \mathbb{P}(Z \leq \xi) \geq \alpha\}, \qquad \mathbb{ES}_\alpha(Z) = \frac{1}{1 - \alpha} \int_\alpha^1 \mathbb{VaR}_a(Z) \, da. \tag{15}$$

As stated in [40, 11, 9, 7], the VaR and ES are linked by a convex optimization problem. If the cdf F_Z of Z is continuous, then $\mathbb{V}a\mathbb{R}_\alpha(Z)$ is the left-end solution to

$$\underset{\xi \in \mathbb{R}}{\operatorname{argmin}} \left\{ V(\xi) := \mathbb{E}[H(\xi, Z)] \right\}, \quad \text{where} \quad L(\xi, Z) := \xi + \frac{1}{1-\alpha}(Z - \xi)^+, \quad \xi \in \mathbb{R}. \tag{16}$$

Moreover, V is convex and continuously differentiable on $\mathbb{R}$, with $V'(\xi) = \mathbb{E}[\partial_\xi L(\xi, Z)] = \frac{1}{1-\alpha}(F_Z(\xi) - \alpha)$, $\xi \in \mathbb{R}$. If F_Z is additionally increasing, then V is strictly convex and $\mathbb{V}a\mathbb{R}_\alpha(Z)$ is the unique minimizer of V:

$$\mathbb{V}a\mathbb{R}_\alpha(Z) = \operatorname{argmin} V \tag{17}$$

and

$$\mathbb{ES}_\alpha(Z) = \min V.$$

Besides, if Z admits a continuous pdf f_Z, then V is twice continuously differentiable on $\mathbb{R}$, with $V''(\xi) = \frac{1}{1-\alpha} f_Z(\xi)$, $\xi \in \mathbb{R}$.

Hence, assuming that the random vector $X \in L^1(\mathbb{P})$ and that, for any $y \in (\mathbb{R}_+^0)^d$, the cdf of $-\langle y, X \rangle \in L^1(\mathbb{P})$ is continuous and increasing, we see that the ES fits our stochastic framework of Section 3.3.1. In particular, Proposition 1 is valid with $\operatorname{Argmin} h = \{(\xi^\star, y^\star)\}$ where $y^\star$ is the unique minimizer of Γ_g associated to the ES and $\xi^\star = \mathbb{V}a\mathbb{R}_\alpha(-\langle y^\star, X \rangle)$. Moreover, Theorem 5 is also satisfied.

Our final example concerns the class of deviation risk measures, of which the standard deviation is undoubtedly the most widely employed instance, see, for example, [42, 43, 41]. Other examples include the mean absolute deviation and semi-deviation, each designed to capture different facets of risk. Formally, a risk measure ρ belongs to this class if it satisfies **(PH)**, **(SA)**, $\rho(Z + c) = \rho(Z)$ for all random variable Z and all constant c and $\rho(Z) > 0$ for any $Z \neq 0$. Building on [15], we consider RB-compatible risk measures that encompass both symmetric and asymmetric deviation measures, to which the stochastic framework outlined in Section 3.3.1 applies. For $a, b > 0$, and $Z \in L^p(\mathbb{P})$, for some $p \geq 1$, we let

$$\rho(Z) = \min_{\xi \in \mathbb{R}} \mathbb{E}[L(\xi, Z)]^{1/p}, \quad \text{with} \quad L(\xi, z) := (a(Z - \xi)_+ + b(Z - \xi)_-)^p. \tag{18}$$

Then, ρ is an RB-compatible risk measure. The aforementioned family encompasses several well-known risk measures for specific choices of a, b and p. When $a = b = 1$, it yields symmetric measures such as the mean absolute deviation around the median (MAD) for $p = 1$ or the standard deviation for $p = 2$. In cases where $a \neq b$, we obtain asymmetric measures. For instance, if $a = \frac{\alpha}{1-\alpha}$, $b = 1$ and $p = 1$, the resulting measure is $\rho(Z) = \mathbb{ES}_\alpha(Z) - \mathbb{E}[Z]$. Similarly, with $a = \sqrt{\alpha}$, $b = \sqrt{1-\alpha}$ and $p = 2$, we derive the square-root of the variantile at level α.

Observe that the map $\mathbb{R} \ni \xi \mapsto \mathbb{E}[L(\xi, Z)]$ is convex. Moreover, if the cdf of Z is continuous then the dominated convergence theorem guarantees that it is continuously differentiable so that

$$\operatorname{Argmin} \mathbb{E}[L(., Z)] = \{\xi \in \mathbb{R} : \mathbb{E}[\partial_\xi L(\xi, Z)] = 0\}$$

where

$$\partial_\xi \mathbb{E}[L(\xi, Z)] = \mathbb{E}[\partial_\xi L(\xi, Z)] = p\mathbb{E}[-a^p(Z - \xi)_+^{p-1} + b^p(Z - \xi)_-^{p-1}]$$

with the convention $(Z - \xi)^0_+ = \mathbf{1}_{Z \geq \xi}$ and $(Z - \xi)^0_- = \mathbf{1}_{Z \leq \xi}$. As with the example of ES, the set $\operatorname{Argmin} \mathbb{E}[L(., Z)]$ does not generally reduce to a singleton. However, if the cdf of Z is increasing then it does. Specifically, for $p = 1$, we have $\operatorname{Argmin} \mathbb{E}[L(., Z)] = \{\mathbb{V}a\mathbb{R}_{a/(a+b)}(Z)\}$.

Now, assuming that the random vector $X \in L^p(\mathbb{P})$ and that, for any $y \in (\mathbb{R}^0_+)^d$, the cdf of $-\langle y, X \rangle$ is continuous and increasing, we observe that ρ aligns with the stochastic framework of Section 3.3.1 with the map $g(x) = x^p$. In particular, Proposition 1 holds with $\operatorname{Argmin} h = \{(\xi^\star, y^\star)\}$, where $y^\star$ is the unique minimizer of Γ_g and $\xi^\star$ is the unique minimizer associated with the optimization problem (18), with $Z = -\langle y^\star, X \rangle$. If $p = 1$, the condition (12) of Theorem 5 is satisfied, ensuring the $a.s.$ convergence of $(z_n)_{n \geq 1}$, along with the $a.s.$ convergence rate of $(\bar{z}_n)_{n \geq 1}$. However, for $p > 1$, as in the first example on the volatility, (12) is violated, since $\xi \mapsto \mathbb{E}[(\partial_\xi L(\xi, -\langle y, X \rangle))^2]$ and $\mathbb{E}[(\partial_x L(\xi, -\langle y, X \rangle))^2]$ exhibit polynomial growth in ξ, uniformly in $y \in \mathbb{R}^d_+ \cap B_m$. Nonetheless, one can readily adapt the proof of Theorem 5 to bypass this minor issue and establish the $a.s.$ convergence of $(z^n)_{n \geq 1}$, along with a similar convergence rate for $(\bar{z}^n)_{n \geq 1}$. Once again, we will refrain from delving into the technical details of this adaptation.

4   Numerical Examples

4.1   Presentation of the model and first results

In this section, we aim to illustrate our theoretical results through a simple example. We begin with a portfolio of three assets, constructing a risk budgeting portfolio under ES at a confidence level of $\alpha = 95\%$. Equal risk budgets are assigned to each asset, i.e., $b = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, which corresponds to the Equal Risk Contribution (ERC) portfolio. Our aim is to showcase the performance of our DMD and SMD algorithms by comparing the outcomes they yield against those from a conventional method.

In the analysis that follows, we assume the joint distribution of asset returns is represented by a mixture of two multivariate Student-t distributions, as this model captures asymmetric and heavy-tailed characteristics commonly observed in asset returns while allowing for precise computation of risk budgeting portfolios under ES using the semi-analytic expressions of VaR and ES [15].

Specifically, we assume that X has the following density with respect to the Lebesgue measure:

$$f_X(x) := p f(x|\mu_1, \Lambda_1, \nu_1) + (1 - p) f(x|\mu_2, \Lambda_2, \nu_2)$$

where p is the probability weight for the first distribution, $f(x|\mu, \Lambda, \nu)$ denotes the density of a multivariate Student-t distribution with mean vector μ, covariance matrix Λ, degrees of freedom ν given by:

$$f(x|\mu_i, \Lambda_i, \nu_i) := \frac{\Gamma\left[\frac{\nu_i + d}{2}\right]}{\Gamma(\nu_i/2)\nu_i^{d/2}\pi^{d/2}\det(\Lambda_i)^{1/2}} \left\{1 + \frac{1}{\nu_i}(x - \mu_i)'\Lambda_i^{-1}(x - \mu_i)\right\}^{-(\nu_i+d)/2}.$$

We work with a realistic model that has been calibrated using daily returns of JPMorgan Chase & Co. (JPM), Pfizer Inc. (PFE) and Exxon Mobil Corporation (XOM) over the period August 2008–April 2022. The model parameters are estimated using the expectation-maximization algorithm. We have obtained

$p = 0.7$, location vectors: $\mu_1 = (0.0001, 0.0002, -0.0003)'$, $\mu_2 = (0.001, 0.0005, 0.0002)'$, scale matrices:

$$
\Lambda_1 = \begin{pmatrix} 9 \cdot 10^{-5} & 3 \cdot 10^{-5} & 5 \cdot 10^{-5} \\ 3 \cdot 10^{-5} & 9 \cdot 10^{-5} & 3 \cdot 10^{-5} \\ 5 \cdot 10^{-5} & 3 \cdot 10^{-5} & 1 \cdot 10^{-4} \end{pmatrix}, \quad \Lambda_2 = \begin{pmatrix} 4 \cdot 10^{-4} & 1 \cdot 10^{-4} & 1 \cdot 10^{-4} \\ 1 \cdot 10^{-4} & 1 \cdot 10^{-4} & 6 \cdot 10^{-5} \\ 1 \cdot 10^{-4} & 6 \cdot 10^{-5} & 1 \cdot 10^{-4} \end{pmatrix}
$$

and degrees of freedom $\nu_1 = 3.4$, $\nu_2 = 2.6$. Under these parameters, we compute the risk budgeting portfolio using the L-BFGS-B algorithm using the semi-analytic expressions of VaR and ES [15]. The resulting portfolio u together with the associated risk contributions $(u_i \partial_{u_i} r_\rho(u))_{1 \leq i \leq d}$ are computed in order to confirm that it is a solution to the risk budgeting problem. The corresponding VaR and ES are also calculated. This portfolio serves as a reliable benchmark for evaluating the convergence and accuracy of our algorithms and will hereafter be referred to as the reference portfolio. The results are presented in Table 1.

Asset	$u_i^\star$	$u_i^\star \partial_i \mathcal{R}(u^\star)$
1	0.2535	0.01096
2	0.3866	0.01096
3	0.3599	0.01096

VaR	0.0193	ES	0.0329

Table 1: Reference portfolio: weights, risk contributions, VaR and ES.

Let us now turn to the application of the DMD and SMD schemes for computing risk budgeting portfolios.

For the implementation of our DMD algorithm, we set $\gamma_n = n^{-0.55}$, $m = 100$, and perform 50,000 iterations. In practice, as will be illustrated later, it suffices to select an m greater than $\|y^\star\|_1$. Here, $\|y^\star\|_1 = 30.4$. The initial point is defined as $y_i^0 = \frac{1}{d\sigma_i^2}$ for $i = 1, \ldots, d$, where σ_i^2 denotes the variance of asset i under the first t-Student distribution, estimated using few samples of X.

To address the stochastic optimization problem outlined in Section 3.3.1 for ES using our SMD algorithm, we generate 10^6 samples of our multivariate Student-t mixture X. The algorithm is run with 10 epochs and a stepsize of $\gamma_n = n^{-0.75}$, with m as in the DMD scheme. We initialize $\xi_0 = 0$ and y^0 as before.

Our results are presented in Table 2. We observe that the DMD algorithm accurately computes the optimal portfolio weights without error. The estimated VaR and ES values are not included, as they correspond to those in Table 1. Additionally, we note that the estimated weights u_i^{SMD} and the estimated VaR from the SMD algorithm closely align with those of the reference portfolio.

---

[1]While implementing a mini-batch version would likely enhance our algorithm's accuracy, we have opted to adhere to our theoretical framework.

Asset	u_i^{SMD}	$\left| u_i^\star - u_i^{\mathrm{SMD}} \right| / u_i^\star$	u_i^{DMD}
1	0.2537	0.08%	0.2535
2	0.3877	0.30%	0.3866
3	0.3586	0.40%	0.3599

$\mathrm{VaR}^{\mathrm{SMD}}$	0.0194	$(\mathrm{VaR}^{\mathrm{SMD}}$ - VaR$)/$ VaR	0.52%

Table 2: Estimated weights and VaR together with relative difference using SMD algorithms and estimated weights using the DMD algorithm.

We now observe the convergence by plotting the three components of the sequence $(u^k)_{k\geq 0}$ generated by the DMD algorithm in Figure 1 with different step sizes: $\gamma_n = 1$, $\gamma_n = n^{-0.55}$, $\gamma_n = n^{-0.75}$, $n \geq 1$.

It is worth noting that the choice of step sequence $(\gamma_n)_{n\geq 1}$ significantly impacts the speed of convergence. Setting $\gamma_n = \gamma = 1$, i.e., a fixed step size of 1 across iterations, appears to be optimal compared to reducing the step size over time, even though our theoretical results do not guarantee convergence in this specific instance. Indeed, convergence is reached in fewer than 1,000 iterations with $\gamma_n = 1$, whereas choosing $\gamma_n = n^{-0.55}$ ensures convergence within 50,000 iterations. By contrast, selecting $\gamma_n = n^{-0.75}$ represents the worst case. This is in line with the upper-bound provided in Theorem 3. These results highlight the significant impact of the selection of the step sequence $(\gamma_n)_{n\geq 1}$ on the convergence rate of the DMD algorithm.
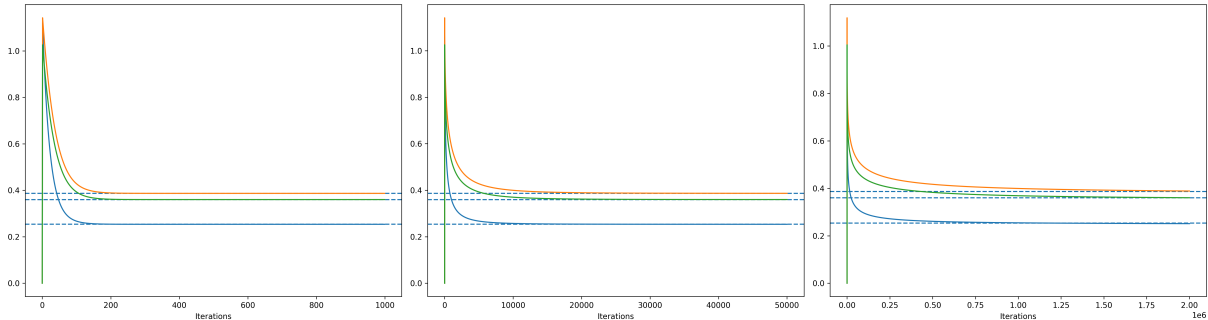

Figure 1: Evolution of the three components of $(u^k)_{k\geq 0}$ of the DMD algorithm for different step sequence $(\gamma_n)_{n\geq 1}$. Left: $\gamma_n \equiv 1$. Center: $\gamma_n = n^{-0.55}$. Right: $\gamma_n = n^{-0.75}$. Dashed lines are the asset weights of the reference portfolio in all subplots.

We now run the SMD algorithm with $N = 250,000$ iterations, comprising 25,000 samples and 10 epochs. We initialize (y^0, ξ^0) and set m as before, with $\gamma_n = n^{-0.55}$ for $n \geq 1$. Convergence is observed by plotting the evolution of the components of both sequences $(y^k)_{0\leq k \leq N}$ and $(u^k := y^k / \|y^k\|_1)_{0\leq k \leq N}$, alongside $(\xi^k)_{0\leq k \leq N}$ in Figure 2. A rapid convergence to the weights and VaR of the reference portfolio is achieved, with the convergence of $(\xi^k)_{k\geq 0}$ appearing faster than that of $(u^k)_{k\geq 0}$.

In Figure 3, we increase the number of iterations in the SMD algorithm, keeping the parameters unchanged except for setting $N = 10^7$ iterations, divided into 10^6 samples and 10 epochs.

Note that the choice of m is not critical, provided it is selected to be larger than $\|y^\star\|_1$. Consequently,
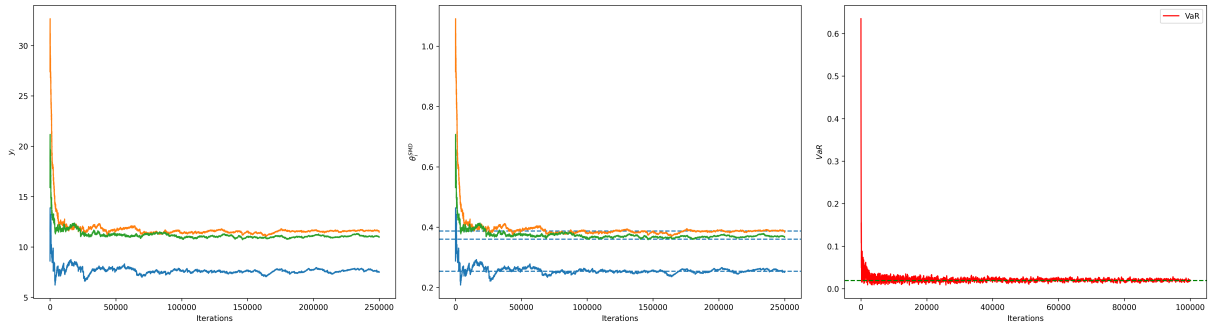
Figure 2: Convergence of the SMD algorithm. Left: evolution of the three components of $(y^k)_{0 \leq k \leq N}$. Center: evolution of the three components of $(u^k)_{0 \leq k \leq N}$ — dashed lines are the weights of the reference portfolio. Right: evolution of $(\xi^k)_{0 \leq k \leq N'}$, with $N' = 100,000$ — dashed line is the VaR of the reference portfolio

we consistently use a large m, as we did before, since selecting an m smaller than $\|y^\star\|_1$ prevents convergence. This effect is illustrated in Figure 4. When $m = 10$, we do not observe convergence of $(u^k)_{k \geq 0}$ towards $u^\star$; however, for values of m larger than $\|y^\star\|_1$, convergence occurs, particularly for $m = 35$, 100, and 1000, with the results for $m = 100$ and $m = 1000$ appearing almost identical due to minimal projections during the algorithm. The DMD algorithm exhibits similar behavior.
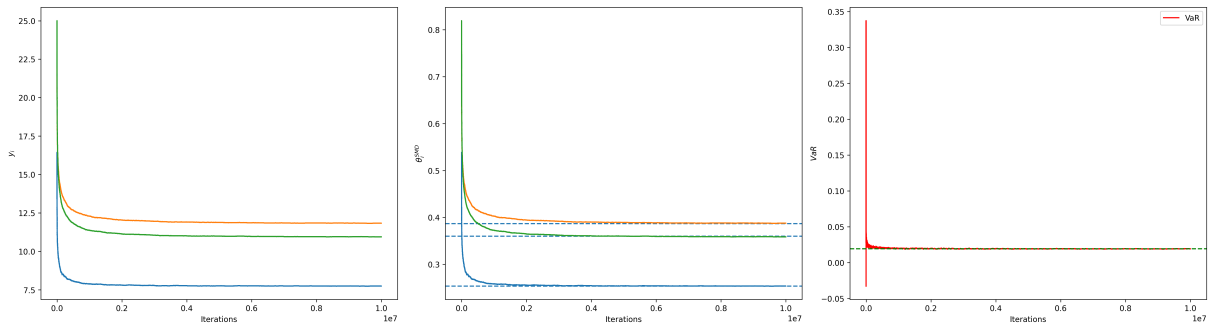

Figure 3: Convergence of the SMD algorithm. Left: evolution of the three components of $(y^k)_{k \geq 0}$. Center: evolution of the three components of $(u^k)_{k \geq 0}$'s — dashed lines are the asset weights of the reference portfolio. Right: evolution of $(\xi^k)_{k \geq 0}$ — dashed line is the VaR of the reference portfolio.

We proceed with our analysis by assessing the robustness of both MD algorithms with respect to the dimension d, corresponding to the number of assets. We consider four scenarios, calibrating our model with 10, 50, 100, and 250 assets (see Section 4.2.1 for details on selecting an arbitrary number of assets d and the calibration process). To demonstrate the convergence of both schemes, we calculate the mean deviation error (MDE) in asset weights compared to those of the reference portfolio, defined as

$$\text{MDE} = \frac{1}{d} \sum_{i=1}^{d} |u_i^N - u_i^\star|.$$

For the DMD algorithm, we use a step sequence $\gamma_n = n^{-0.55}$ and 50,000 iterations. For the SMD algorithm, we set the parameters as follows: $\gamma_n = n^{-0.75}$, with 10^6 samples combined with 10 epochs.
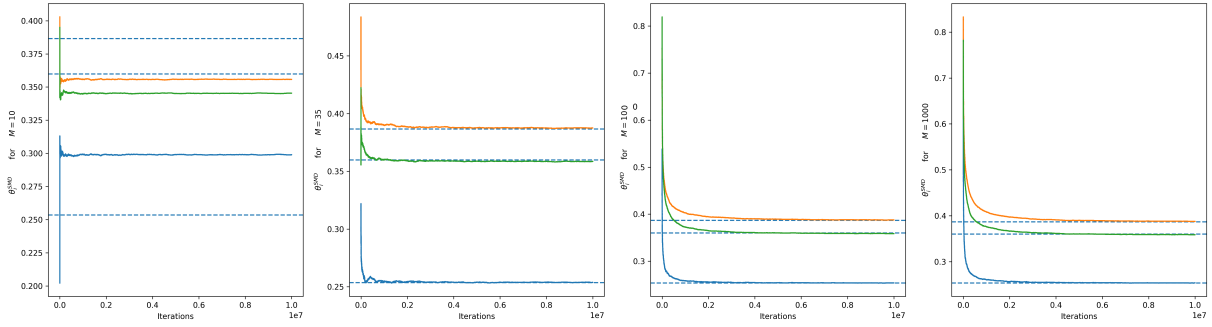
Figure 4: From left to right: evolution of the sequence $(u^k)_{k \geq 0}$ for $m = 10, 35, 100$ and 1000 respectively.

The parameters ξ^0, y^0 and m remain as previously specified. The results are presented in Table 3. Note that values for SMD and DMD are scaled by a factor of 10^3 for clarity. The relative error of VaR is presented as a percentage.

Assets	SMD	$\left|\text{VaR}^{\text{SMD}} - \text{VaR}\right|/\text{VaR}$	DMD
10	1.02	0.005%	0.19
50	0.26	0.980%	0.76
100	0.25	0.075%	1.15
250	0.26	0.977%	0.85

Table 3: MDE for both DMD and SMD algorithms, scaled by 10^3, along with the relative error of VaR with respect to the reference portfolio.

We also performed computations for the DMD algorithm using a constant step size of $\gamma_n = 1$ over 10,000 iterations. These results are not included in Table 3, as no errors were observed in any instance, underscoring the strong robustness of the DMD algorithm when a constant step size is applied. The findings for both SMD and DMD with a decreasing step size similarly demonstrate impressive robustness, as they maintain convergence even with increasing dimensionality. For the variable step size case, the performance remains generally satisfactory, though slightly lower than that achieved with a fixed step size, as convergence is still ongoing and has not yet fully stabilized.

4.2   Error analysis for different portfolio sizes

4.2.1   Taming to the rescue: fixing stability issues

In the traditional approach to solving the risk budgeting problem, the numerical algorithm typically follows the directions defined by the gradient $\nabla \Gamma(y)$, recalling that $g = \text{Id}$ for the case of ES. However, since the gradient diverges on $\partial(\mathbb{R}_+)^d$, as proposed in Section 3.1, we moderate it by applying the factor $\kappa(y)$ to ensure convergence of the algorithm. In this subsection, we aim to demonstrate that, beyond its theoretical benefits, this adjustment also greatly enhances the stability and convergence of the numerical scheme, particularly in stochastic settings.

We begin by comparing the SMD algorithm, following the update rule given in (11) with $m = 100$, to two alternative methods in terms of successful convergence across different scenarios. The first method, referred to as classical-SGD (c-SGD), is the projected stochastic gradient descent algorithm proposed in [15] to solve (8) using the update rule

$$\begin{cases} \xi^{k+1} = \xi^k - \gamma_{k+1} \partial_\xi H(z^k, X^{k+1}) \\ y^{k+1} = \Pi \left(y^k - \gamma_{k+1} \nabla_y H(z^k, X^{k+1}) \right) \end{cases}$$

recalling that $z^k = (\xi^k, y^k)$ and $\Pi : \mathbb{R}^d \to (\mathbb{R}_+^0)^d$ is a projection function designed to ensure that all elements are positive. Specifically, Π is defined to replace any negative elements with a fixed positive value, $\epsilon = 10^{-4}$.

The second approach, referred to as tamed-SGD (t-SGD), incorporates the concept of taming the gradient and involves updating $(z^k)_{k \geq 0}$ as follows:

$$\begin{cases} \xi^{k+1} = \xi^k - \gamma_{k+1} \partial_\xi H(z^k, X^{k+1}) \\ y^{k+1} = \Pi \left(y^k - \gamma_{k+1} \kappa(y^k) \nabla_y H(z^k, X^{k+1}) \right) \end{cases}$$

recalling that $\kappa(y) = \underline{y} \wedge 1$ serves as the taming factor, applied in both DMD and SMD algorithms to adjust the gradient and enhance stability.

At this stage, our primary focus is on assessing the algorithms' ability to approach the true solution across various portfolio sizes, rather than precisely measuring the accuracy of their convergence. This preliminary evaluation offers insight into the algorithms' convergence behavior, setting the groundwork for a subsequent analysis of their accuracy, which will be addressed in the following subsection. In brief, the objective here is to understand the practical significance of gradient taming in mitigating numerical divergence caused by the issue of exploding gradients.

We proceed with computing ERC portfolios for ES. To statistically analyze the convergence rate of different algorithms across various portfolio sizes d, we follow this procedure. We randomly select d stocks from the S&P 500 components (as of Q2 2022) and obtain their daily returns from August 2008 to April 2022. To capture the distributional characteristics of these returns, we fit a two-component Student-t mixture model to the historical data. Based on the estimated parameters, we compute a reference portfolio. We then simulate 10^6 data points from the fitted model and apply the SMD, c-SGD, and t-SGD algorithms to the simulated dataset, analyzing convergence by comparing the results to the reference portfolio.

This procedure is repeated 100 times for each d, with a new random selection of d assets in each iteration. Across these repetitions, we monitor each algorithm's rate of successful convergence, noting instances of divergence due to issues such as exploding gradients. This approach provides insights into the robustness of each algorithm in realistic portfolio construction settings. Specifically, divergence is defined as occurring when a discrepancy in the objective function value is observed at the final iteration $n = 10^6$. That is, if $\Gamma(y^n) - \Gamma(y^\star)$ exceeds a prescribed threshold ε, the algorithm is classified as having diverged in that particular repetition.[3]

---

[3]The error is measured by $\Gamma(y^n) - \Gamma(y^\star)$ – the distance of the objective function value at the last iterate to the minimum – rather than the averaged iterates $\Gamma(\bar{y}^n) - \Gamma(y^\star)$, as discussed in Theorem 6, to mitigate the influence of large values in earlier iterations and to assess convergence within reasonable computational times.

| | | | d | | |
| | 10 | 25 | 50 | 100 | 250 |
ε					
5×10^{-2}	24	11	24	47	35
5×10^{-1}	23	11	23	43	11
5×10^{0}	23	10	13	8	0
5×10^{1}	22	2	0	0	0

Table 4: Number of divergences out of 100 repetitions for the c-SGD algorithm.

Table 4 underscores the inherent stability challenges associated with c-SGD and illustrates the advantages of taming the gradient.[4] While both t-SGD and SMD reliably converge to the reference portfolio (with precision considerations to be discussed in the following section), c-SGD encounters an issue with gradient instability due to certain elements of the vector y approaching zero during iterations. This instability causes divergence in several trials. Figure 5 further illustrates the distribution of error over 100 trials, highlighting outliers that correspond to instances of non-convergence across varying numbers of iterations k. The figure suggests that, although c-SGD can occasionally achieve accurate results, its convergence is highly sensitive to the initial choice of the learning rate γ.[5] In contrast, the tamed methods exhibit notable robustness to hyperparameter selection, consistently avoiding divergence across all tested cases.
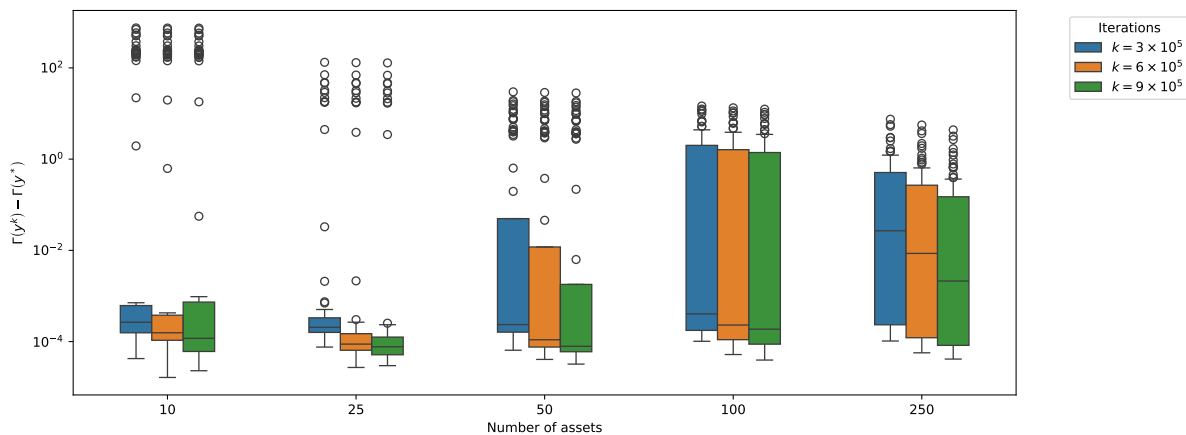

Figure 5: Errors in the objective function values for the unnormalized portfolio weights obtained via c-SGD, calculated across 100 samples for each portfolio size. The error $\Gamma(y^k) - \Gamma(y^*)$ is measured at various iterations $k = 3 \times 10^5$, 6×10^5, and 9×10^5 to illustrate the progression of error throughout the algorithm's execution.

---

[4]Analogous results are not shown for t-SGD and SMD, as neither algorithm exhibited divergence across any of the cases tested.

[5]For all methods, we employ a learning rate scheme defined as $\gamma_n = \gamma n^{-0.65}$. For the tamed methods (t-SGD and SMD), initial learning rates γ are set to 1, 2.5, 5, 10, and 25 for portfolio dimensions d of 10, 25, 50, 100, and 250, respectively. For c-SGD, the corresponding values are set to 5, 1, 0.5, 0.25, and 0.1 for d values of 10, 25, 50, 100, and 250. These learning rates have been selected through iterative tuning to optimize convergence for each algorithm and portfolio size.

4.2.2   Improved accuracy via SMD

We now turn our attention to the SMD and t-SGD algorithms, both of which exhibit stable convergence properties in contrast to c-SGD. Our aim is to evaluate how *closely* the portfolios generated by these methods approximate the reference portfolios when applied to the same dataset. This approach allows for an effective comparison of the efficiency of the update rules associated with t-SGD and SMD, while maintaining all other hyperparameters constant. By isolating the effects of these update mechanisms, we seek to determine whether SMD offers practical advantages in computing risk budgeting portfolios, in addition to its theoretical benefits.
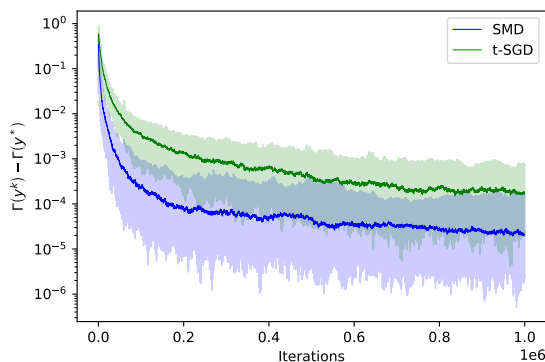

Figure 6: The evolution of error for the SMD and t-SGD algorithms across iterations.

An initial and illustrative assessment can be conducted by tracking the evolution of error throughout the iterations. Specifically, we can simulate data from a fixed model and run the SMD and t-SGD algorithms, recording the error $\Gamma(y^k) - \Gamma(y^*)$ after each iteration. Figure 6 shows the median error along with a 95% confidence interval (computed by running the algorithms on 100 different samples from the same model) for the 3-asset model described in Section 4.1. The results indicate that the SMD algorithm converges to the minimum faster than t-SGD. However, this analysis may be specific to the chosen model and should be extended to general cases and larger portfolio dimensions for a broader conclusion.

For this reason, we apply the same procedure described in Section 4.2.1 to assess the accuracy of the computed portfolios. Table 5 displays the portfolios' accuracy relative to the reference portfolios, where accuracy is evaluated by the difference between the *true* minimum objective function value and the objective function value at the computed y^k. Specifically, the error measure is defined as $(\Gamma(y^k) - \Gamma(y^*)) \times 10^3$. We report the median error and median absolute deviations (in parentheses), computed by repeating the process 100 times with samples drawn from 100 distinct fitted models for each d. This metric serves as a quantitative measure of how closely the computed portfolios approximate the optimal solution.

Figure 7 offers further insights into the algorithms' performance. The distribution of the error in the objective function value, $\Gamma(y^k) - \Gamma(y^*)$, is shown for the SMD and t-SGD algorithms, computed over 100 samples across various portfolio sizes: $d = 10, 25, 50, 100$ and 250. The error is calculated for the weights at several iterations $k = 3 \times 10^5$, 6×10^5, and 9×10^5 to illustrate the evolution of error over the course of the algorithms. On average, the SMD algorithm consistently yields more accurate results across tested

d	$k = 3 \times 10^5$		$k = 6 \times 10^5$		$k = 9 \times 10^5$	
	SMD	t-SGD	SMD	t-SGD	SMD	t-SGD
10	0.10 (0.05)	0.22 (0.10)	0.06 (0.02)	0.09 (0.04)	0.05 (0.02)	0.06 (0.03)
25	0.19 (0.06)	0.87 (0.38)	0.11 (0.04)	0.29 (0.15)	0.09 (0.03)	0.17 (0.09)
50	0.24 (0.08)	10.49 (2.31)	0.13 (0.03)	6.13 (1.65)	0.11 (0.03)	4.16 (1.30)
100	0.36 (0.12)	29.52 (3.73)	0.16 (0.05)	23.21 (3.21)	0.11 (0.03)	19.81 (2.95)
250	0.90 (0.27)	33.16 (9.01)	0.43 (0.11)	29.29 (7.93)	0.25 (0.08)	26.73 (7.26)

Table 5: Error in objective function values obtained by SMD and t-SGD across different asset sizes, computed from samples of size 10^6 drawn from fitted models at various iteration steps k.

portfolio sizes and at different stages of iteration. Notably, the advantage of SMD over t-SGD becomes increasingly pronounced as portfolio size grows. A particularly interesting observation is SMD's ability to recover from large errors over the course of iterations, with this robustness clearly seen through outliers, especially for $d = 50$ and $d = 250$. For instance, the outlier at iteration $k = 3 \times 10^5$ is gradually corrected, resulting in more accurate portfolio estimations as the algorithm processes additional data points.
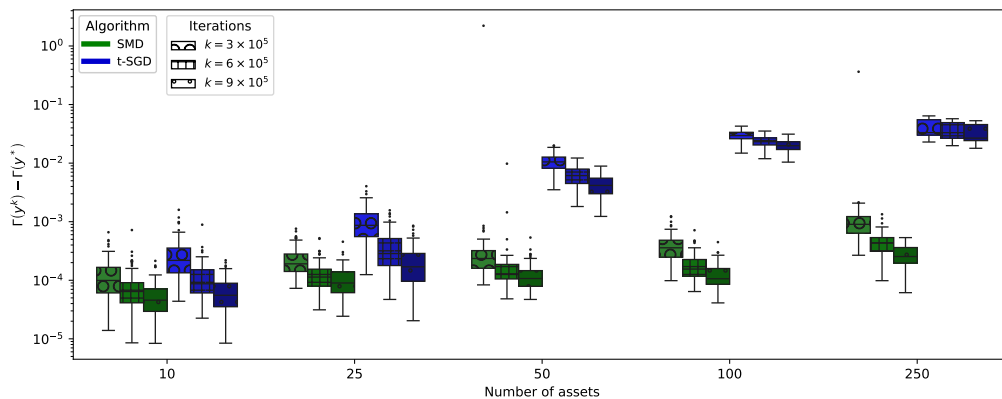

Figure 7: Distribution of errors in the objective function value obtained by using the SMD and t-SGD algorithms, computed over 100 samples for each portfolio size.

Although assessing error and accuracy through differences in objective function values is mathematically sound, this approach lacks practical interpretability in portfolio construction. Therefore, we extend our analysis by evaluating accuracy using the MDE ($\times 10^4$) between the computed portfolio weights and the reference portfolio. In Table 6, we present the median of the MDEs alongside the median absolute deviations of the MDEs (in parentheses), calculated over 100 repetitions with samples drawn from 100 distinct fitted models for each d. This metric offers a more intuitive measure of practical accuracy, as it directly reflects how closely the computed portfolio weights approximate those of the reference portfolio.

In Figure 8, we display the distribution of the MDEs obtained by the SMD and t-SGD algorithms over 100 samples for each portfolio size. The error is calculated at different iterations: $k = 3 \times 10^5$, $k = 6 \times 10^5$ and, $k = 9 \times 10^5$.

d	$k = 3 \times 10^5$		$k = 6 \times 10^5$		$k = 9 \times 10^5$	
	SMD	t-SGD	SMD	t-SGD	SMD	t-SGD
10	7.98 (1.74)	12.09 (3.36)	6.55 (1.32)	8.00 (1.76)	5.43 (1.10)	6.06 (1.41)
25	4.97 (0.69)	9.52 (2.27)	3.84 (0.50)	5.71 (1.61)	3.43 (0.49)	4.74 (1.38)
50	2.57 (0.27)	10.73 (1.03)	2.11 (0.22)	8.97 (0.98)	1.79 (0.18)	7.94 (0.93)
100	1.34 (0.13)	9.19 (0.75)	1.06 (0.11)	8.33 (0.60)	0.95 (0.10)	7.81 (0.54)
250	0.60 (0.08)	4.91 (0.12)	0.47 (0.05)	4.66 (0.12)	0.40 (0.05)	4.49 (0.12)

Table 6: MDEs of the SMD and t-SGD algorithms across different portfolio sizes, computed at iterations $k = 3 \times 10^5$, 6×10^5 and 9×10^5.

Table 6 and Figure 8 further confirm that the SMD algorithm produces more accurate portfolios across all tested portfolio sizes. Additionally, it achieves high accuracy even for portfolios with a large number of assets, highlighting its effectiveness as an optimization method for computing risk budgeting portfolios for ES.
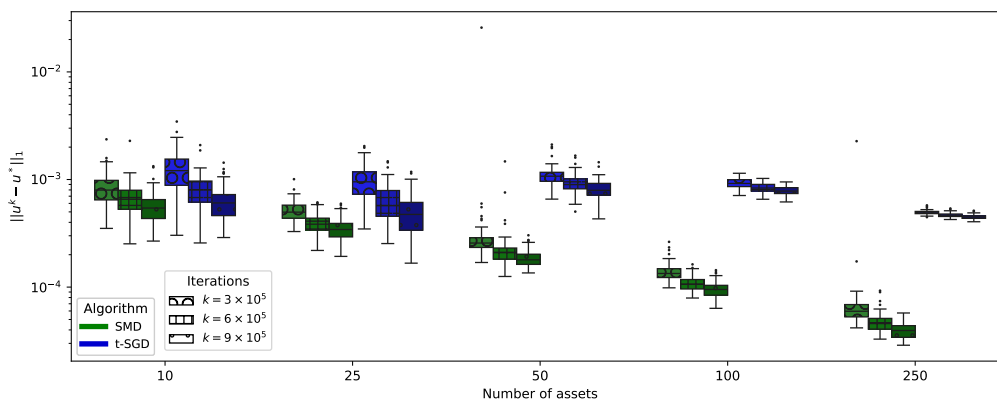

Figure 8: Distribution of the MDEs obtained by the SMD and t-SGD algorithms, computed over 100 samples for each portfolio size and iteration n.

4.3   Convergence using different risk measures

Previous numerical analyses have focused on a specific risk measure – ES – to have a comprehensive understanding of the behavior of SMD and SGD-based algorithms across different portfolio sizes when computing risk budgeting portfolios. However, as detailed in Section 3.3.3, this framework is compatible with a wide variety of risk measures. In this subsection, we aim to understand whether the computational advantage of SMD over t-SGD extends to the broader class of risk measures.

Before proceeding with the comparison of SMD and t-SGD methods using alternative risk measures within the class of deviation measures as formalized by (18) – specifically mean absolute deviation (MAD) ($p = 1$, $a = 1$ and $b = 1$), volatility ($p = 2$, $a = 1$ and $b = 1$), and variantile ($p = 2$, $a = 0.75$ and

$b = 0.25$) – it is essential to verify that the algorithms converge when these measures are used. However, we must avoid modeling X with a Student-t mixture in this context, as it requires a (semi-)analytical representation of the objective function for each risk measure to compute reference portfolios accurately. Therefore, we begin with a model where X follows a centered normal distribution, using the same three assets as in Section 4.1 with the sample covariance matrix of the historical asset returns.

With this particular model, we know how to compute the risk budgeting portfolio for volatility (as defined in Equation (14)) using deterministic algorithms such as DMD, which provides highly precise results and serves as the reference portfolio. Additionally, under the assumption of a centered normal distribution, the aforementioned risk measures should yield the same risk budgeting portfolio, as each can be represented as a function of portfolio volatility. Therefore, we simulate 10^6 data points from this distribution and run SMD algorithm to confirm the algorithm converges to the same reference portfolio.

Asset	Reference	MAD	Volatility	Variantile
1	0.2411	0.2417	0.2417	0.2417
2	0.4150	0.4137	0.4138	0.4140
3	0.3438	0.3446	0.3445	0.3443

Table 7: Reference portfolio weights and computed weights under different portfolio measures.

Table 7 confirms that the update rule we implemented for deviation measures successfully converges to the optimal portfolio in a simple case of three assets. With this verification, we can proceed with our analysis to compare SMD and t-SGD. To make the analysis more realistic, we now model X as a centered multivariate Student-t distribution with fixed degrees of freedom $\nu = 4$, using the maximum-likelihood estimate for the scale matrix Λ. This model has two main advantages: the reference portfolios remain precisely computable, and the stochastic algorithms are exposed to extreme losses, making the case more representative of real-world scenarios.

In line with the procedure outlined in Section 4.2.2, we assess the accuracy of the algorithms across different portfolio sizes. Specifically, we select d assets from our dataset and estimate the corresponding scale matrix Λ. We then compute the reference portfolio – identical across all risk measures – using the centered multivariate Student-t model with the estimated Λ. From this model, we draw $n = 5 \times 10^5$ data points and apply the SMD and t-SGD algorithms. To improve the stability of the approximated portfolios, we employ the Polyak-Ruppert averaging technique. Rather than using the final iterate y^n, we store y^k for the last 20% of iterations (i.e., 10^5 iterations) and calculate their empirical mean, denoted $\bar{y}^n$. The resulting risk budgeting portfolio is then given by $\bar{u}^n = \frac{\bar{y}^n}{\|\bar{y}^n\|_1}$, which provides a more accurate estimate by averaging out the fluctuations in the final iterations.

Table 8 shows the median of the MDEs with median absolute deviations of the MDEs (in parentheses) across different portfolio sizes d and risk measures over 100 repetitions. The initial choice of the learning rate γ is fixed across various portfolio sizes d to avoid tuning the optimal value for each dimension, method, and risk measure. The results indicate that SMD produces more accurate portfolios than t-SGD in most cases, highlighting its suitability for computing risk budgeting portfolios accurately across a range of risk measures. These values, of course, only allow for a comparative analysis, and absolute errors could

d	MAD		Volatility		Variantile	
	SMD	t-SGD	SMD	t-SGD	SMD	t-SGD
10	2.26 (0.38)	2.46 (0.49)	3.91 (0.96)	16.30 (6.99)	4.55 (1.01)	21.76 (9.14)
25	1.04 (0.14)	1.14 (0.17)	2.05 (0.29)	2.75 (0.72)	2.40 (0.36)	3.32 (0.87)
50	0.57 (0.05)	0.61 (0.06)	1.23 (0.17)	1.11 (0.17)	1.40 (0.23)	1.43 (0.25)
100	0.30 (0.02)	0.34 (0.02)	0.69 (0.07)	0.65 (0.08)	0.80 (0.10)	0.79 (0.11)
250	0.13 (0.01)	0.14 (0.01)	0.30 (0.03)	0.31 (0.03)	0.34 (0.04)	0.35 (0.04)

Table 8: MDEs of the portfolios computed using the SMD and t-SGD algorithms across different portfolio sizes.

be further reduced by choosing hyperparameters more carefully.

Conclusion

In this article, leveraging the characterization of risk budgeting portfolios as solutions to a strictly convex optimization problem, we proposed and analyzed the convergence of deterministic and stochastic mirror descent (MD) algorithms for computing these portfolios across general risk measures. Due to the unbounded gradient of the value function, standard convergence results for MD algorithms do not directly apply. By developing a tailored tamed version of the gradient, we introduced both deterministic and stochastic MD schemes, establishing their convergence along with a convergence rate for the weighted averaged sequence. The stochastic framework accommodates several commonly used risk measures, such as volatility, expected shortfall, and deviation measures.

Our numerical results illustrate these theoretical findings, demonstrating that both algorithms remain stable as dimensionality increases and perform more effectively than the standard SGD algorithm recently proposed in the literature.

Potential directions for future research include investigating asymptotic error properties by establishing a central limit theorem for the stochastic MD algorithm presented here. Another promising line of inquiry could be to explore MD algorithms when only biased samples $(X_n^h)_{n \geq 1}$ of X, where h represents the bias parameter, as seen for instance in the two recent works [5] or [19]. In this setting, to further manage complexity, one might employ multi-level or multi-step Richardson-Romberg techniques in stochastic approximation schemes, as originally developed in [25] and [27], and also as applied in recent works on VaR and ES in [20, 22, 21] for the special case of the VaR and ES. It could also be worth investigating the applicability of MD algorithms for calculating risk budgeting portfolios that take into account underlying risk factors, as outlined in [16].

References

[1] C. Acerbi. "Spectral measures of risk: A coherent representation of subjective risk aversion". In: *Journal of Banking & Finance* 26.7 (2002), pp. 1505–1518.

[2] C. Acerbi and D. Tasche. "On the coherence of Expected Shortfall". In: *Journal of Banking & Finance* 26.7 (2002), pp. 1487–1503.

[3] Ç. Ararat et al. "MAD risk parity portfolios". In: *Annals of Operations Research* (2024), pp. 1–26.

[4] V. Asimit et al. *Risk Budgeting under General Risk Measures*. Submitted. URL: https://openaccess.city.ac.uk/id/eprint/33733/.

[5] Y. F. Atchadé, G. Fort, and R. Moulines. "On Perturbed Proximal Gradient Algorithms". In: *Journal of Machine Learning Research* 18.10 (2017), pp. 1–33. URL: http://jmlr.org/papers/v18/15-038.html.

[6] X. Bai, K. Scheinberg, and R. Tutuncu. "Least-squares approach to risk parity in portfolio selection". In: *Quantitative Finance* 16.3 (2016), pp. 357–376.

[7] O. Bardou, N. Frikha, and G. Pagès. "Computing VaR and CVaR using stochastic approximation and adaptive unconstrained importance sampling". In: *Monte Carlo Methods and Applications* 15.3 (2009), pp. 173–210. DOI: doi:10.1515/MCMA.2009.011. URL: https://doi.org/10.1515/MCMA.2009.011.

[8] O. Bardou, N. Frikha, and G. Pagès. "CVaR hedging using quantization-based stochastic approximation algorithm". In: *Mathematical Finance* 26.1 (2016), pp. 184–229.

[9] O. Bardou, N. Frikha, and G. Pagès. "Recursive Computation of Value-at-Risk and Conditional Value-at-Risk using MC and QMC". In: *Monte Carlo and Quasi-Monte Carlo Methods 2008*. Ed. by Pierre L'Ecuyer and Art B. Owen. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 193–208. ISBN: 978-3-642-04107-5.

[10] F. Bellini et al. "Risk parity with expectiles". In: *European journal of operational research* 291.3 (2021), pp. 1149–1163.

[11] A. Ben-Tal and M. Teboulle. "An old-new concept of convex risk measures: the optimized certainty equivalent". In: *Mathematical Finance* 17.3 (2007), pp. 449–476. DOI: https://doi.org/10.1111/j.1467-9965.2007.00311.x.

[12] M. J. Best and R. R. Grauer. "On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results". In: *The Review of Financial Studies* 4.2 (1991), pp. 315–342.

[13] S. Boyd et al. "Markowitz Portfolio Construction at Seventy". In: *arXiv preprint arXiv:2401.05080* (2024).

[14] B. Bruder and T. Roncalli. "Managing Risk Exposures Using the Risk Budgeting Approach". In: *Available at SSRN: https://ssrn.com/abstract=2009778* (2012). DOI: http://dx.doi.org/10.2139/ssrn.2009778.

[15] A. R. Cetingoz, J.-D. Fermanian, and O. Guéant. "Risk Budgeting portfolios: Existence and computation". In: *Mathematical Finance* (2023).

[16] A. R. Cetingoz and O. Guéant. "Factor Risk Budgeting and Beyond". In: *arXiv preprint arXiv:2312.11132* (2023).

[17] D. Chaves et al. "Efficient algorithms for computing risk parity portfolio weights". In: *Journal of Investing* 21.3 (2012), p. 150.

[18] B. F. P. da Costa, S. M. Pesenti, and R. S. Targino. "Risk budgeting portfolios from simulations". In: *European Journal of Operational Research* 311.3 (2023), pp. 1040–1056.

[19] M. Costa, L. Huang, and S. Gadat. "CV@R penalized portfolio optimization with biased stochastic mirror descent". In: *forthcoming in Finance & Stochastics* (2024).

[20] S. Crépey, N. Frikha, and A. Louzi. "A Multilevel Stochastic Approximation Algorithm for Value-at-Risk and Expected Shortfall Estimation". In: *forthcoming for Finance & Stochastics* (2024).

[21] S. Crépey et al. "Adaptive Multilevel Stochastic Approximation of the Value-at-Risk". In: *arXiv:2408.06531* (2024).

[22] S. Crépey et al. "Asymptotic Error Analysis of Multilevel Stochastic Approximations for the Value-at-Risk and Expected Shortfall". In: *forthcoming for Electronic Journal of Probability* (2024).

[23] P. Embrechts et al. "Bayes risk, elicitability, and the Expected Shortfall". In: *Mathematical Finance* 31.4 (2021), pp. 1190–1217. DOI: https://doi.org/10.1111/mafi.12313. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mafi.12313. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mafi.12313.

[24] H. Föllmer and A. Schied. "Convex risk measures". In: *Encyclopedia of Quantitative Finance*. John Wiley & Sons, Ltd, 2010. ISBN: 9780470061602. DOI: 10.1002/9780470061602.eqf15003.

[25] N. Frikha. "Multi-level stochastic approximation algorithms". In: *The Annals of Applied Probability* 26.2 (2016), pp. 933–985. DOI: 10.1214/15-AAP1109. URL: https://doi.org/10.1214/15-AAP1109.

[26] N. Frikha. "Shortfall risk minimization in discrete time financial market models". In: *SIAM Journal on Financial Mathematics* 5.1 (2014), pp. 384–414. DOI: 10.1137/120903142. URL: https://doi.org/10.1137/120903142.

[27] N. Frikha and L. Huang. "A multi-step Richardson–Romberg extrapolation method for stochastic approximation". In: *Stochastic Processes and their Applications* 125.11 (2015), pp. 4066–4101. ISSN: 0304-4149. DOI: https://doi.org/10.1016/j.spa.2015.05.016. URL: https://www.sciencedirect.com/science/article/pii/S0304414915001386.

[28] T. Griveau-Billion, J.-C. Richard, and T. Roncalli. "A fast algorithm for computing high-dimensional risk parity portfolios". In: (2013). URL: https://arxiv.org/abs/1311.4057.

[29] G. Lan, A. Nemirovski, and A. Shapiro. "Validation analysis of mirror descent stochastic approximation method". In: *Mathematical Programming* 134.2 (2012), pp. 425–458. DOI: 10.1007/s10107-011-0442-6. URL: https://doi.org/10.1007/s10107-011-0442-6.

[30] S. Maillard, T. Roncalli, and J. Teiletche. "The properties of equally weighted risk contribution portfolios". In: *The Journal of Portfolio Management* 36.4 (2010), pp. 60–70.

[31] H. Markowitz. "Portfolio Selection". In: *The Journal of Finance* 7.1 (1952), pp. 77–91.

[32]   H. Markowitz. "The optimization of a quadratic function subject to linear constraints". In: *Naval Research Logistics Quarterly* 3.1-2 (1956), pp. 111–133.

[33]   H. Mausser and O. Romanko. "Long-only equal risk contribution portfolios for CVaR under discrete distributions". In: *Quantitative Finance* 18.11 (2018), pp. 1927–1945.

[34]   R. O. Michaud. "The Markowitz optimization enigma: Is 'optimized'optimal?" In: *Financial Analysts Journal* 45.1 (1989), pp. 31–42.

[35]   A. Nemirovski et al. "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* 19.4 (2009), pp. 1574–1609. DOI: 10.1137/070704277. eprint: https://doi.org/10.1137/070704277. URL: https://doi.org/10.1137/070704277.

[36]   A. S. Nemirovskij and D. B. Yudin. *Problem complexity and method efficiency in optimization.* John Wiley and Sons, 1983.

[37]   Y. Nesterov. "Primal-dual subgradient methods for convex problems". In: *Mathematical Programming* 120.1 (2009), pp. 221–259. DOI: 10.1007/s10107-007-0149-x. URL: https://doi.org/10.1007/s10107-007-0149-x.

[38]   E. Qian et al. "Risk parity portfolios: Efficient portfolios through true diversification". In: *Panagora Asset Management* (2005).

[39]   R. T. Rockafellar and S. Uryasev. "Conditional value-at-risk for general loss distributions". In: *Journal of banking & finance* 26.7 (2002), pp. 1443–1471.

[40]   R. T. Rockafellar and S. Uryasev. "Optimization of Conditional Value-at-Risk". In: *Journal of risk* 2.3 (Apr. 1, 2000), pp. 21–41. DOI: 10.21314/JOR.2000.038. URL: https://doi.org/10.21314/JOR.2000.038.

[41]   R. T. Rockafellar and S. Uryasev. "The fundamental risk quadrangle in risk management, optimization and statistical estimation". In: *Surveys in Operations Research and Management Science* 18.1 (2013), pp. 33–53. ISSN: 1876-7354. DOI: https://doi.org/10.1016/j.sorms.2013.03.001. URL: https://www.sciencedirect.com/science/article/pii/S1876735413000032.

[42]   R. T. Rockafellar, S. Uryasev, and M. Zabarankin. "Generalized deviations in risk analysis". In: *Finance Stochast.* 10 (2006), pp. 51–74. DOI: https://doi.org/10.1007/s00780-005-0165-8.

[43]   R. T. Rockafellar, S. Uryasev, and M. Zabarankin. "Risk Tuning with Generalized Linear Regression". In: *Mathematics of Operations Research* 33 (2008), pp. 712–729. DOI: https://doi.org/10.1287/moor.1080.0313.

[44]   T. Roncalli. *Introduction to Risk Parity and Budgeting.* CRC Press, 2013.

[45]   F. Spinu. *An Algorithm for Computing Risk Parity Weights.* Tech. rep. 2013. URL: https://ssrn.com/abstract=2297383.

[46]   P. Wolfe. "The simplex method for quadratic programming". In: *Econometrica: Journal of the Econometric Society* (1959), pp. 382–398.

[47]   Zhengyuan Zhou et al. "Stochastic Mirror Descent in Variationally Coherent Optimization Problems". In: *Advances in Neural Information Processing Systems.* Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/e6ba70fc093b4ce912d769ede1ceeba8-Paper.pdf.

A Appendix

A.1 Convergence for the DMD algorithm: proof of Theorem 3

Step 1: On the convergence of $(y^n)_{n \geq 1}$.

In order to study the convergence of the sequence $(y^k)_{k \geq 0}$ towards the unique minimizer $y^\star$ of Γ_g, we recall that Lemma 2.1 in [35] (combined with (3)) guarantees that

$$D_F(w, \mathrm{P}_u^m(v)) \leq D_F(w, u) + \langle v, w - u \rangle + m \frac{\|v\|_\infty^2}{2} \tag{19}$$

for any $w, u \in (\mathbb{R}_+^0)^d \cap B_m$, $v \in \mathbb{R}^d$. Selecting $v = \gamma_{k+1} \kappa(y^k) \nabla \Gamma_g(y^k)$, $u = y^k$ and $w = y^\star$ in the above inequality, and taking $m \geq \|y^\star\|_1$, we get

$$\begin{aligned} D_F(y^\star, y^{k+1}) &\leq D_F(y^\star, y^k) - \gamma_{k+1} \langle \kappa(y^k) \nabla \Gamma_g(y^k), y^k - y^\star \rangle \\ &\quad + m \gamma_{k+1}^2 \frac{\|\kappa(y^k) \nabla \Gamma_g(y^k)\|_\infty^2}{2}. \end{aligned} \tag{20}$$

Assuming that the learning rate satisfies $\sum_{k \geq 1} \gamma_k^2 < \infty$ and recalling (2), the Robbins-Siegmund theorem yields

$$D_F(y^\star, y^n) \overset{n \to \infty}{\longrightarrow} D_\infty < \infty \quad \text{and} \quad \sum_{k \geq 0} \gamma_{k+1} \langle \kappa(y^k) \nabla \Gamma_g(y^k), y^k - y^\star \rangle < \infty. \tag{21}$$

Since $\sum_{k \geq 1} \gamma_k = \infty$, we deduce that $\liminf_{n \to \infty} \langle \kappa(y^n) \nabla \Gamma_g(y^n), y^n - y^\star \rangle = 0$. The boundedness of the sequence $(y^n)_{n \geq 0}$ guarantees the existence of a subsequence $(\varphi(n))_{n \geq 0}$ such that

$$y^{\varphi(n)} \to y^\infty \in (\mathbb{R}_+)^d \cap B_m, \quad \text{and} \quad \langle \kappa(y^{\varphi(n)}) \nabla \Gamma_g(y^{\varphi(n)}), y^{\varphi(n)} - y^\star \rangle = 0.$$

By continuity of $\mathbb{R}_+^d \ni y \mapsto \kappa(y) \nabla \Gamma_g(y)$, it follows that $\lim_{n \to \infty} \langle \kappa(y^{\varphi(n)}) \nabla \Gamma_g(y^{\varphi(n)}), y^{\varphi(n)} - y^\star \rangle = \langle \kappa(y^\infty) \nabla \Gamma_g(y^\infty), y^\infty - y^\star \rangle = 0$ which in turn, by Lemma 2, implies that $y^\infty = y^\star$.

Coming back to (21), we eventually deduce that

$$\lim_{n \to \infty} D_F(y^\star, y^n) = \lim_{n \to \infty} D_F(y^\star, y^{\varphi(n)}) = 0$$

which clearly implies that $(y^n)_{n \geq 1}$ converges towards $y^\star$.

Step 2: On the convergence rate of $(\bar{y}^n)_{n \geq 1}$.

In order to derive the corresponding convergence rate, we come back to (20) and write

$$\gamma_{k+1} \langle \kappa(y^k) \nabla \Gamma_g(y^k), y^k - y^\star \rangle \leq D_F(y^\star, y^k) - D_F(y^\star, y^{k+1}) + \frac{1}{2} m \gamma_{k+1}^2 M_\star^2.$$

Summing over $k = 0, \cdots, n-1$ the previous inequality and recalling that $D_F(y^{k+1}, y^\star) \geq 0$, we get

$$\sum_{k=1}^n \gamma_k \langle \kappa(y^{k-1}) \nabla \Gamma_g(y^{k-1}), y^{k-1} - y^\star \rangle \leq D_F(y^\star, y^0) + \frac{1}{2} m M_\star^2 \sum_{k=1}^n \gamma_k^2.$$

I

By convexity of Γ_g and Jensen's inequality, we obtain

$$\sum_{k=1}^{n} \gamma_k \kappa(y^{k-1}) \langle \nabla \Gamma_g(y^{k-1}), y^{k-1} - y^\star \rangle \geq (\min_{0 \leq k \leq n-1} \underline{y}^k \wedge 1) \sum_{k=1}^{n} \gamma_k \langle \nabla \Gamma_g(y^{k-1}), y^{k-1} - y^\star \rangle$$

$$\geq (\min_{0 \leq k \leq n-1} \underline{y}^k \wedge 1) \sum_{k=1}^{n} \gamma_k (\Gamma_g(y^{k-1}) - \Gamma_g(y^\star))$$

$$\geq (\min_{0 \leq k \leq n-1} \underline{y}^k \wedge 1)(\sum_{k=1}^{n} \gamma_k)(\Gamma_g(\bar{y}^n) - \Gamma_g(y^\star))$$

recalling that $(\bar{y}^n)_{n \geq 1}$ is the averaged version of $(y^n)_{n \geq 0}$. Combining the two previous inequalities eventually yields

$$\Gamma_g(\bar{y}^n) - \Gamma_g(y^\star) \leq \frac{D_F(y^\star, y^0) + \frac{1}{2} m M_\star^2 \sum_{k=1}^{n} \gamma_k^2}{(\min_{0 \leq k \leq n-1} \underline{y}^k \wedge 1) \sum_{k=1}^{n} \gamma_k}, \quad n \geq 1,$$

which concludes the proof.

A.2 Proof of Proposition 1

Step 1: The regularity and integrability conditions combined with the dominated convergence theorem guarantee that h is continuously differentiable on $\mathbb{R} \times (\mathbb{R}_+^0)^d$. The expression for the derivatives $\partial_\xi h(\xi, y)$ and $\partial_{y_i} h(\xi, y)$, $i = 1, \cdots, d$, is easily obtained by differentiating under expectation. The convexity of h then guarantees that $\text{Argmin}\, h = \{\nabla h = 0\}$.

Step 2: We here prove that the map $(\mathbb{R}_+^0)^d \ni y \mapsto g(r_\rho(y))$ defined by (7) is continuously differentiable with a derivative given by $\nabla(g(r_\rho(y))) = \mathbb{E}[-X \partial_x L(\xi^\star(y), -\langle y, X \rangle)]$. Let $y, h \in \mathbb{R}^d$. Note that the minimizers $\xi^\star(y+h)$, $\xi^\star(y)$ of $\mathbb{E}[L(., -\langle(y+h), X\rangle)]$ and $\mathbb{E}[L(., -\langle y, X\rangle)]$ respectively, satisfy

$$g(r_\rho(y+h)) - g(r_\rho(y))$$
$$= \mathbb{E}[L(\xi^\star(y+h), -\langle(y+h), X\rangle)] - \mathbb{E}[L(\xi^\star(y), -\langle y, X\rangle)]$$
$$\leq \mathbb{E}[L(\xi^\star(y), -\langle(y+h), X\rangle)] - \mathbb{E}[L(\xi^\star(y), -\langle y, X\rangle)]$$
$$= \langle \mathbb{E}[-X \partial_x L(\xi, -\langle y, X\rangle)]_{\xi=\xi^\star(y)}, h \rangle + o(\|h\|)$$

as $\|h\| \downarrow 0$. Similarly, one has

$$g(r_\rho(y+h)) - g(r_\rho(y))$$
$$\geq \mathbb{E}[L(\xi^\star(y+h), -\langle(y+h), X\rangle)] - \mathbb{E}[L(\xi^\star(y+h), -\langle y, X\rangle)]$$
$$= \langle \mathbb{E}[-X \partial_x L(\xi, -\langle y, X\rangle)]_{\xi=\xi^\star(y+h)}, h \rangle$$
$$+ \int_0^1 \langle \mathbb{E}[-X \partial_x L(\xi, -\langle(y+th), X\rangle)]_{\xi=\xi^\star(y+h)} - \mathbb{E}[-X \partial_x L(\xi, -\langle y, X\rangle)]_{\xi=\xi^\star(y+h)}, h \rangle\, dt$$
$$= \langle \mathbb{E}[-X \partial_x L(\xi, -\langle y, X\rangle)]_{\xi=\xi^\star(y)}, h \rangle$$
$$+ \langle \mathbb{E}[-X \partial_x L(\xi, -\langle y, X\rangle)]_{\xi=\xi^\star(y+h)} - \mathbb{E}[-X \partial_x L(\xi, -\langle y, X\rangle)]_{\xi=\xi^\star(y)}, h \rangle$$
$$+ \int_0^1 \langle \mathbb{E}[-X \partial_x L(\xi, -\langle(y+th), X\rangle)]_{\xi=\xi^\star(y+h)} - \mathbb{E}[-X \partial_x L(\xi, -\langle y, X\rangle)]_{\xi=\xi^\star(y+h)}, h \rangle\, dt.$$

Our aim now is to prove that the two last terms appearing in the right-hand side of the above inequality are $o(\|h\|)$. Clearly, it suffices to prove that

$$\lim_{h\downarrow 0}\mathbb{E}[X\partial_x L(\xi, -\langle y, X\rangle)]_{\xi=\xi^\star(y+h)} = \mathbb{E}[X\partial_x L(\xi, -\langle y, X\rangle)]_{\xi=\xi^\star(y)} \tag{22}$$

and that

$$\lim_{h\downarrow 0}\mathbb{E}[X\partial_x L(\xi, -\langle (y+th), X\rangle)] = \mathbb{E}[X\partial_x L(\xi, -\langle y, X\rangle)] \tag{23}$$

locally uniformly in ξ.

Note that the convexity of the maps $\xi \mapsto \mathbb{E}[L(\xi, -\langle y, X\rangle)]$, and $\xi \mapsto \mathbb{E}[L(\xi, -\langle (y+h), X\rangle)]$ yields $(\xi - \xi^\star(y))\mathbb{E}[\partial_\xi L(\xi, -\langle y, X\rangle)] > 0$ for any $\xi \neq \xi^\star(y)$ and $(\xi - \xi^\star(y+h))\mathbb{E}[\partial_\xi L(\xi, -\langle (y+h), X\rangle)] > 0$ for any $\xi \neq \xi^\star(y+h)$. Moreover, the continuity of $\mathbb{R} \times (\mathbb{R}_+^0)^d \ni (\xi, y) \mapsto \mathbb{E}[\partial_\xi L(\xi, -\langle y, X\rangle)]$ guarantees that for any fixed y, $\xi \mapsto \mathbb{E}[\partial_\xi L(\xi, -\langle (y+h), X\rangle)]$ converges to $\xi\mathbb{E}[\partial_\xi L(\xi, -\langle y, X\rangle)]$ as $h \downarrow 0$ locally uniformly. Now, [25, Theorem 2.6] guarantees that $\xi^\star(y+h) \to \xi^\star(y)$ as $\|h\| \downarrow 0$ which combined with the continuity of $\xi \mapsto \mathbb{E}[X\partial_z L(\xi, -\langle y, X\rangle)]$ implies (22).

The second point (23) is a consequence of the continuity of $\mathbb{R} \times \mathbb{R}^d \ni (\xi, y) \mapsto \mathbb{E}[X\partial_x L(\xi, -\langle y, X\rangle)]$.

Combining the above arguments, we conclude that $y \mapsto g(r_\rho(y))$ is continuously differentiable and satisfies $\nabla(g(r_\rho(y))) = \mathbb{E}[-X\partial_x L(\xi^\star(y), -\langle y, X\rangle)]$.

Step 3: If $(\xi^\star, y^\star) \in \mathrm{Argmin}\, h$, then $\partial_{y_i} h(\xi^\star, y^\star) = 0$, $i = 1, \cdots, d$, which, according to the conclusion of the previous step, implies that

$$\partial_{y_i}\Gamma_g(y^\star) = \partial_{y_i}(g(r_\rho(y^\star))) - \frac{b_i}{y_i^\star} = 0, \quad i = 1, \cdots, d.$$

Hence, $y^\star$ is the unique minimizer of Γ_g. The proof is now complete.

A.3 Convergence of the SMD algorithm: proof of Theorem 5

Step 1: On the convergence of $(z^n)_{n \geq 1}$

We use the inequality (19) with G instead of F (noting (9)) namely

$$D_G(z, \mathrm{P}_{z'}^m(v)) \leq D_G(z, z') + \langle v, z - z'\rangle + m \vee 1\|v\|_\infty^2.$$

with $v = \gamma_{k+1}(\partial_\xi H(z^k, X^{k+1}), \kappa(y^k)\nabla_y H(z^k, X^{k+1}))^T$, $z' = z^k$ and $z = z^\star$. Hence,

$$\begin{aligned}
D_G(z^\star, z^{k+1}) \leq D_G(z^\star, z^k) &- \gamma_{k+1}\left\langle \mathbb{E}\left[\begin{pmatrix} \partial_\xi H(z, X) \\ \kappa(y)\nabla_y H(z, X) \end{pmatrix}\right]_{|z=z^k}, z^k - z^\star\right\rangle \\
&- \gamma_{k+1}\langle \Delta^{k+1}, z^k - z^\star\rangle + m \vee 1\gamma_{k+1}^2 Y_{k+1},
\end{aligned} \tag{24}$$

where we used the fact that X^{k+1} is independent of $(z^j)_{0 \leq j \leq k}$ and introduced the notations

$$\begin{aligned}
\Delta^{k+1} &= \begin{pmatrix} \partial_\xi H(z^k, X^{k+1}) \\ \kappa(y^k)\nabla_y H(z^k, X^{k+1}) \end{pmatrix} - \mathbb{E}\left[\begin{pmatrix} \partial_\xi H(z, X) \\ \kappa(y)\nabla_y H(z, X) \end{pmatrix}\right]_{|z=z^k}, \\
Y_{k+1} &:= \left\|\begin{pmatrix} \partial_\xi H(z^k, X^{k+1}) \\ \kappa(y^k)\nabla_y H(z^k, X^{k+1}) \end{pmatrix}\right\|_\infty^2.
\end{aligned} \qquad k \geq 0. \tag{25}$$

III

Note that since z^k is $\mathcal{F}_k$-measurable and X^{k+1} is independent of $\mathcal{F}_k$, $(\Delta^k)_{k\geq 1}$ is a sequence of $\mathcal{F}$-martingale increments

$$\mathbb{E}[\Delta^{k+1}|\mathcal{F}_k] = 0, \quad \text{for } k \geq 0.$$

Now, the previous identity together with (24) and observing that

$$\mathbb{E}[Y_{k+1}|\mathcal{F}_k] \leq N_\star^2 := \sup_{(\xi,y)\in\mathbb{R}\times\mathbb{R}_+^d\cap B_m)} \mathbb{E}\left[\left\|\begin{pmatrix}\partial_\xi H(z,X)\\\kappa(y)\nabla_y H(z,X)\end{pmatrix}\right\|_\infty^2\right] < \infty, \quad k \geq 0$$

yields

$$\mathbb{E}[D_G(z^\star, z^{k+1})|\mathcal{F}_k] \leq D_G(z^\star, z^k) - \gamma_{k+1}\left\langle\mathbb{E}\left[\begin{pmatrix}\partial_\xi H(z,X)\\\kappa(y)\nabla_y H(z,X)\end{pmatrix}\right]_{|z=z^k}, z^k - z^\star\right\rangle \tag{26}$$
$$+ m \vee 1 N_\star^2 \gamma_{k+1}^2.$$

The Robbins-Siegmund theorem guarantees that

$$D_G(z^\star, z^n) \overset{n\to\infty}{\longrightarrow} D_\infty \in L^1(\mathbb{P}) \quad a.s. \tag{27}$$

and

$$\sum_{n\geq 0}\gamma_{n+1}\left\langle\mathbb{E}\left[\begin{pmatrix}\partial_\xi H(z,X)\\\kappa(y)\nabla_y H(z,X)\end{pmatrix}\right]_{|z=z^n}, z^n - z^\star\right\rangle < \infty, \quad a.s.$$

The first assertion above implies that the sequence $(z^n)_{n\geq 0}$ is bounded while the second together with Lemma 4 and the condition $\sum_{k\geq 1}\gamma_k = \infty$ guarantees that

$$\liminf_{n\to\infty}\left\langle\mathbb{E}\left[\begin{pmatrix}\partial_\xi H(z,X)\\\kappa(y)\nabla_y H(z,X)\end{pmatrix}\right]_{|z=z^n}, z^n - z^\star\right\rangle = 0.$$

As a consequence, we may assume the existence of a subsequence $(\varphi(n))_{n\geq 0}$ such that

$$z^{\varphi(n)} \to z^\infty \in \mathbb{R} \times (\mathbb{R}_+^d \cap B_m^d)$$

and

$$\lim_{n\to\infty}\left\langle\mathbb{E}\left[\begin{pmatrix}\partial_\xi H(z,X)\\\kappa(y)\nabla_y H(z,X)\end{pmatrix}\right]_{|z=z^{\varphi(n)}}, z^{\varphi(n)} - z^\star\right\rangle = 0.$$

By continuity of the map $z \mapsto (\mathbb{E}[\partial_\xi H(z,X)], \mathbb{E}[\kappa(y)\nabla_y H(z,X)])$, it follows that:

$$\lim_{n\to\infty}\left\langle\mathbb{E}\left[\begin{pmatrix}\partial_\xi H(z,X)\\\kappa(y)\nabla_y H(z,X)\end{pmatrix}\right]_{|z=z^{\varphi(n)}}, z^{\varphi(n)} - z^\star\right\rangle = \left\langle\mathbb{E}\left[\begin{pmatrix}\partial_\xi H(z^\infty,X)\\\kappa(y^\infty)\nabla_y H(z^\infty,X)\end{pmatrix}\right], z^\infty - z^\star\right\rangle = 0$$

and, by Lemma 4, implies that $z^\infty = z^\star$. Subsequently, coming back to (27), we deduce that

$$\lim_{n\to\infty}D_G(z^\star, z^n) = \lim_{n\to\infty}D_G(z^{\varphi(n)}, z^\star) = 0$$

which clearly implies that the sequence $(z^n)_{n\geq 0}$ converges $a.s.$ towards $z^\star$. Note additionally that taking expectation on both sides of (26) and using Lemma 4 together with a direct induction argument give

$$\sup_{n\geq 1}\mathbb{E}[D_G(z^n, z^\star)] \leq \mathbb{E}[D_G(z^\star, z^0)] + m \vee 1 M_\star^2 \sum_{n\geq 1}\gamma_n^2. \tag{28}$$

IV

The above bound will be useful in order to investigate the convergence rate of $(\bar{z}^n)_{n \geq 1}$ in the next step.

Step 2: On the convergence rate of $(\bar{z}^n)_{n \geq 1}$

We now come back to (24) getting

$$
\gamma_{k+1} \left\langle \mathbb{E}\left[\begin{pmatrix} \partial_\xi H(z, X) \\ \kappa(y)\nabla_y H(z, X) \end{pmatrix} \right]_{|z=z^k}, z^k - z^\star \right\rangle \leq D_G(z^\star, z^k) - D_G(z^\star, z^{k+1}) - \gamma_{k+1}\langle \Delta^{k+1}, z^k - z^\star \rangle
$$
$$
+ m \vee 1 \gamma_{k+1}^2 Y_{k+1}
$$

Summing over $k = 0, 1, \cdots, n$ the previous inequality and recalling that $D_G(z^\star, z^{n+1}) \geq 0$, we get

$$
\sum_{k=0}^n \gamma_{k+1} \left\langle \mathbb{E}\left[\begin{pmatrix} \partial_\xi H(z, X) \\ \kappa(y)\nabla_y H(z, X) \end{pmatrix} \right]_{|z=z^k}, z^k - z^\star \right\rangle
$$
$$
\leq D_G(z^\star, z^0) - M_n + m \vee 1 \sum_{k=0}^n \gamma_{k+1}^2 Y_{k+1} \tag{29}
$$

where we introduced the notation

$$
M_n := \sum_{k=0}^n \gamma_{k+1}\langle \Delta^{k+1}, z^k - z^\star \rangle, \quad n \geq 0.
$$

Note that since $(\Delta^k)_{k \geq 1}$ is a sequence of $\mathcal{F}$-martingale increments, $(M_n)_{n \geq 0}$ is an $\mathcal{F}$-martingale.

Combining the convexity of h with Jensen's inequality, we obtain

$$
\sum_{k=0}^n \gamma_{k+1} \left\langle \mathbb{E}\left[\begin{pmatrix} \partial_\xi H(z, X) \\ \kappa(y)\nabla_y H(z, X) \end{pmatrix} \right]_{|z=z^k}, z^k - z^\star \right\rangle \geq \sum_{k=0}^n \gamma_{k+1}\kappa(y^k)\langle \nabla h(z^k), z^k - z^\star \rangle
$$
$$
\geq \sum_{k=0}^n \gamma_{k+1}\kappa(y^k)(h(z^k) - h(z^\star)) \tag{30}
$$
$$
\geq \min_{0 \leq k \leq n} \kappa(y^k)(\sum_{k=0}^n \gamma_{k+1})(h(\bar{z}^n) - h(z^\star)).
$$

Combining (29) with (30) eventually yields

$$
h(\bar{z}^n) - h(z^*) \leq \frac{1}{\min_{0 \leq k \leq n} \kappa(y^k)} \frac{1}{\sum_{k=0}^n \gamma_{k+1}} \left[D_G(z^*, z^0) - M_n + m \vee 1 \sum_{k=0}^n \gamma_{k+1}^2 Y_{k+1} \right], \quad a.s. \tag{31}
$$

Step 3:

To conclude the proof, we provide some uniform controls on the two last terms involved in the right-hand side of the previous upper-bound (31). Recall that $\left\{ M_n := \sum_{k=0}^n \gamma_{k+1}\langle \Delta^{k+1}, z^k - z^\star \rangle, n \geq 1 \right\}$ is an $\mathcal{F}$-martingale. Since

$$
\mathbb{E}[\|\Delta^{k+1}\|_2^2 | \mathcal{F}_k] \leq d\mathbb{E}\left[\left\| \begin{pmatrix} \partial_\xi H(z, X) \\ \kappa(y)\nabla_y H(z, X) \end{pmatrix} \right\|_\infty^2\right] \leq dN_\star^2
$$

V

we get

$$\langle M \rangle_n \le d M_\star^2 \sup_{k \ge 0} \| z^k - z^\star \|_2^2 \sum_{k=0}^n \gamma_{k+1}^2$$

which in turn implies that $\langle M \rangle_\infty = \lim_n \langle M \rangle_n < \infty$ a.s. Hence, $(M_n)_{n \ge 1}$ converges a.s. to $M_\infty < \infty$ a.s. Moreover, from (9) and (28), we deduce

$$\sup_{n \ge 0} \mathbb{E}[\| z^\star - z^n \|_2^2] \le 2(m \vee 1) \sup_{n \ge 0} \mathbb{E}[D_G(z^\star, z^n)] \le \mathbb{E}[D_G(z^\star, z^0)] + \frac{1}{2}(m \vee 1) N_\star^2 \sum_{n \ge 1} \gamma_n^2 < \infty$$

so that

$$\sup_{n \ge 1} \mathbb{E}[M_n^2] \le \sum_{n \ge 0} \gamma_{n+1}^2 \mathbb{E}[\| \Delta^{n+1} \|_2^2 \| z^\star - z^n \|_2^2] \le d N_\star^2 \sup_{n \ge 0} \mathbb{E}[\| z^\star - z^n \|_2^2] \sum_{n \ge 0} \gamma_{n+1}^2 < \infty.$$

Hence, $(M_n)_{\ge 1}$ is bounded in $L^2(\mathbb{P})$.

To conclude, note that the third term in (31) is bounded a.s. and in $L^1(\mathbb{P})$ since

$$\mathbb{E}\Big[\sum_{n \ge 0} \gamma_{n+1}^2 Y_{k+1} \Big] \le N_\star^2 \sum_{n \ge 0} \gamma_{n+1}^2.$$

The proof is now complete.