# Chapter 3: Some applications of risk measures

## Capital allocation and risk budgeting problems

Noufel Frikha

Université Paris 1 Panthéon-Sorbonne

October 2024

# Contents

# Capital allocation

# Capital allocation

Allocation problem: Investor (financial institution, loan book manager) can invest in $d$ investment possibilities with losses represented by random variables $L_1, \ldots, L_d$.

◦ Objective: Determine appropriate risk capital for each investment opportunity.

1. Compute the overall risk capital $\rho(L)$, where $L = \sum_{i=1}^{d} L_i$ and $\rho$ is a specified risk measure.

2. Allocate the capital $\rho(L)$ to individual investments according to a capital allocation principle such that:

$$\rho(L) = \sum_{i=1}^{d} AC_i,$$

where $AC_i$ is the capital allocated to the investment with potential loss $L_i$.

Set-up: Let $L_1, \ldots, L_d$ be random variables on $(\Omega, \mathcal{F}, P)$. For any $\lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d$, representing portfolio weights of individual investments, we set:

$$L(\lambda) = \sum_{i=1}^{d} \lambda_i L_i \quad \text{(note that } L(\mathbf{1}) = L\text{)}.$$

Fix a risk measure $\rho$ and define the associated risk-measure function:

$$r_\rho(\lambda) = \rho(L(\lambda)),$$

which represents the required risk capital for a position $\lambda$ in the investment possibilities.

Definition: A mapping $\pi_\rho : \mathbb{R}^d \to \mathbb{R}^d$ is called a per-unit capital allocation principle if:

$$\text{For all } \lambda \in \mathbb{R}^d, \quad r_\rho(\lambda) = \sum_{i=1}^{d} \lambda_i \pi_{\rho,i}(\lambda).$$

Interpretation: The $i$-th component $\pi_{\rho,i}(\lambda)$ of the vector $\pi_\rho(\lambda)$ gives the amount of capital allocated to one unit of $L_i$ when the overall position is $L(\lambda)$. The equality means that the overall risk capital $r_\rho(\lambda)$ is fully allocated to the individual portfolio positions.

# Euler's principle and examples

We restrict to risk measures that are positively homogeneous, i.e., $\rho(tL) = t\rho(L)$ for all $t > 0$. This includes coherent risk measures, VaR, and the standard deviation risk measure.

$\rightsquigarrow$ This implies that the associated risk-measure function $r_\rho$ is also positively homogeneous: $r_\rho(t\lambda) = tr_\rho(\lambda)$ for all $t > 0$ and $\lambda \in \mathbb{R}^d$. Assuming that $r_\rho$ is differentiable on $\mathbb{R}^d$, we derive Euler's rule:

$$r_\rho(\lambda) = \sum_{i=1}^{d} \lambda_i \frac{\partial r_\rho}{\partial \lambda_i}(\lambda).$$

Comparing with the definition of the allocation principle, we see that the mapping $\pi_\rho : \mathbb{R}^d \to \mathbb{R}^d$ is given by:

$$\pi_{\rho,i}(\lambda) = \frac{\partial r_\rho}{\partial \lambda_i}(\lambda),$$

which is called allocation by the gradient.

# Standard deviation and covariance principle

Consider the risk measure $\rho_{\mathsf{SD}}(X) = \sqrt{\mathsf{Var}(X)}$. Let $\Sigma$ be the covariance matrix of $L = (L_1, \ldots, L_d)$, and note that:

$$\mathsf{Var}(L(\lambda)) = \mathsf{Var}(\lambda^\top L) = \lambda^\top \Sigma \lambda, \quad \lambda \in \mathbb{R}^d.$$

Therefore,

$$r_{\rho_{\mathsf{SD}}}(\lambda) = \sqrt{\lambda^\top \Sigma \lambda}.$$

We can then derive

$$\nabla r_\rho(\lambda) = \frac{\Sigma \lambda}{r_\rho(\lambda)}.$$

Thus, the allocation by the gradient is given by:

$$\pi_{\rho_{\mathsf{SD}},i}(\lambda) = \frac{(\Sigma\lambda)_i}{r_\rho(\lambda)} = \frac{\sum_{j=1}^d \mathsf{Cov}(L_i, L_j)\lambda_j}{r_\rho(\lambda)} = \frac{\mathsf{Cov}(L_i, L(\lambda))}{\sqrt{\mathsf{Var}(L(\lambda))}}.$$

For the initial portfolio $L = L(\mathbf{1})$, the capital allocated to the $i$-th investment is

$$AC_i = \frac{\mathsf{Cov}(L_i, L)}{\sqrt{\mathsf{Var}(L)}}, \quad i = 1, \ldots, d.$$

This is known as the covariance principle.

# Value-at-Risk

For the $\mathrm{VaR}_\alpha$ risk measure, the associated risk-measure function is

$$r_{\mathrm{VaR}_\alpha}(\lambda) = q_\alpha(L(\lambda)),$$

where $q_\alpha$ is the quantile function. The allocation by the gradient is given by

$$\pi_{\mathrm{VaR}_\alpha,i}(\lambda) = \frac{\partial r_{\mathrm{VaR}_\alpha}}{\partial \lambda_i}(\lambda) = \mathbb{E}[L_i \mid L(\lambda) = q_\alpha(L(\lambda))], \quad i = 1, \ldots, d.$$

Let us prove this result for the case where the loss distribution of $L = (L_1, \ldots, L_d)$ has a joint density. Denote by $u \mapsto \varphi(u, L_2, \ldots, L_d)$ the conditional density of $L_1$ given $(L_2, \ldots, L_d)$.

Lemma: For any $\lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d$, with $\lambda_1 \neq 0$, $L(\lambda)$ has density given by

$$f_{L(\lambda)}(t) = \frac{1}{|\lambda_1|}\mathbb{E}\left[\varphi\left(\frac{t - \sum_{j=2}^d \lambda_j L_j}{\lambda_1}, L_2, \ldots, L_d\right)\right].$$

and

$$\mathbb{E}[L_i \mid L(\lambda) = t] = \frac{\mathbb{E}\left[L_i \varphi\left(\frac{t-\sum_{j=2}^d \lambda_j L_j}{\lambda_1}, L_2, \ldots, L_d\right)\right]}{\mathbb{E}\left[\varphi\left(\frac{t-\sum_{j=2}^d \lambda_j L_j}{\lambda_1}, L_2, \ldots, L_d\right)\right]}, \quad i = 2, \ldots, d.$$

Proof. (1) Consider the case $\lambda_1 > 0$ We can write:

$$\begin{aligned}\mathbb{P}[L(\lambda) \leq t] &= \mathbb{E}[\mathbb{P}(L(\lambda) \leq t | L_2, \cdots L_d)] \\ &= \mathbb{E}\left[\mathbb{P}\left[L_1 \leq \lambda_1^{-1}\left(t - \sum_{j=2}^{d} \lambda_j L_j\right) \middle| L_2, \ldots, L_d\right]\right] \\ &= \mathbb{E}\left[\int_{-\infty}^{\lambda_1^{-1}\left(t - \sum_{j=2}^{d} \lambda_j L_j\right)} \varphi(u, L_2, \ldots, L_d) du\right]\end{aligned}$$

The first assertion follows by differentiating under the expectation:

$$f_{L(\lambda)}(t) = \frac{1}{|\lambda_1|}\mathbb{E}\left[\varphi\left(\lambda_1^{-1}\left(t - \sum_{j=2}^{d} \lambda_j L_j\right), L_2, \ldots, L_d\right)\right]$$

By similar arguments, one proves that

$$\frac{\partial}{\partial t}\mathbb{E}[L_i \mathbf{1}_{L(\lambda) \leq t}] = \frac{1}{|\lambda_1|}\mathbb{E}\left[L_i \varphi\left(\lambda_1^{-1}\left(t - \sum_{j=2}^{d} \lambda_j L_j\right), L_2, \ldots, L_d\right)\right]$$

Proof (continued): Observe that we can write:

$$\mathbb{E}[L_i|L(\lambda)=t]=\lim_{h\to 0}\frac{\mathbb{E}[L_i 1_{t<L(\lambda)\leq t+h}]}{\mathbb{P}[t<L(\lambda)\leq t+h]}=\frac{\frac{\partial}{\partial t}\mathbb{E}[L_i 1_{\{L(\lambda)\leq t\}}]}{f_{L(\lambda)}(t)}$$

provided that $f_{L(\lambda)}(t)>0$. The result follows by using the expressions derived in the first assertion. □

Let us now conclude how (2) follows from the above Lemma. Since $L(\lambda)$ has a continuous distribution, we have $\mathbb{P}[L(\lambda) \leq q_\alpha(L(\lambda))] = \alpha$. By setting $k(t) = \lambda_1^{-1}(t - \sum_{j=2}^d \lambda_j L_j)$, we have:

$$\alpha = \mathbb{P}[L(\lambda) \leq r_{\mathsf{VaR}_\alpha}(L(\lambda))] = \mathbb{E}\left[\int_{-\infty}^{k(r_{\mathsf{VaR}_\alpha}(L(\lambda)))} \varphi(u, L_2, \ldots, L_d)du\right].$$

Taking derivatives of this expression with respect to $\lambda_i$, for $i = 2, \ldots, d$, yields:

$$0 = \lambda_1^{-1}\mathbb{E}\left[\left(\frac{\partial r_{\mathsf{VaR}_\alpha}}{\partial \lambda_i}(\lambda) - L_i\right)\varphi(k(r_{\mathsf{VaR}_\alpha}(L(\lambda))), L_2, \ldots, L_d)\right].$$

This gives the required result by using Assertion (2) of the Lemma

$$\pi_{\mathsf{VaR}_\alpha, i}(\lambda) = \mathbb{E}[L_i \mid L(\lambda) = q_\alpha(L(\lambda))], \quad i = 1, \ldots, d.$$

In particular, for the initial portfolio $L = L(\mathbf{1})$, we obtain the capital allocated to the $i$-th investment:

$$AC_i = \mathbb{E}[L_i \mid L = \mathsf{VaR}_\alpha(L)], \quad i = 1, \ldots, d.$$

# Expected Shortfall

For the $\mathsf{ES}_\alpha$ risk measure, the associated risk-measure function is:

$$r_{\mathsf{ES}_\alpha}(\lambda) = \frac{1}{1-\alpha}\int_\alpha^1 r_{\mathsf{VaR}_u}(\lambda)\,du,$$

recalling that $r_{\mathsf{VaR}_u}(\lambda) = q_u(L(\lambda))$. By differentiating with respect to $\lambda_i$, and using Assertion (2), we get:

$$\frac{\partial r_{\mathsf{ES}_\alpha}}{\partial \lambda_i}(\lambda) = \frac{1}{1-\alpha}\int_\alpha^1 \frac{\partial r_{\mathsf{VaR}_u}}{\partial \lambda_i}(\lambda)\,du = \frac{1}{1-\alpha}\int_\alpha^1 \mathbb{E}[L_i \mid L(\lambda) = q_u(L(\lambda))]\,du.$$

Assuming that the density $f_{L(\lambda)}$ is strictly positive so that the distribution function $F_{L(\lambda)}$ is invertible, we can make the change of variable $v = q_u(L(\lambda)) = F^{-1}_{L(\lambda)}(u)$ with $du = f_{L(\lambda)}(v)dv$, and get:

$$\begin{aligned}\frac{\partial r_{\mathsf{ES}_\alpha}}{\partial \lambda_i}(\lambda) &= \frac{1}{1-\alpha}\int_{q_\alpha(L(\lambda))}^\infty \mathbb{E}[L_i \mid L(\lambda) = v] f_{L(\lambda)}(v)\,dv \\ &= \frac{1}{1-\alpha}\mathbb{E}[\mathbb{E}[L_i | L(\lambda) \geq q_\alpha(L(\lambda))]] \\ &= \frac{1}{1-\alpha}\mathbb{E}\left[L_i \mid L(\lambda) \geq q_\alpha(L(\lambda))\right].\end{aligned}$$

In particular, for the initial portfolio $L = L(1)$, we obtain the capital allocated to the $i$-th investment possibility:

$$AC_i = \mathbb{E}[L_i \mid L \geq \mathsf{VaR}_\alpha(L)], \quad i = 1, \ldots, d.$$

This is a popular allocation principle in practice, often considered preferable to both the covariance principle and the principle based on VaR.

# Risk budgeting

# Introduction to Portfolio Optimization

- **Classical Optimization (Markowitz, 1952):**
  - **Objective**: Maximize expected return under a risk constraint, typically measured by variance.
  - **Method**: A mathematical formulation based on the mean-variance framework.

$$\text{Maximize} \quad w^t\mu - \lambda w^t\Sigma w$$
$$\text{subject to } w^t\mathbf{1} = 1, \quad w \in \Omega.$$

  - **Outcome**: An "optimal" portfolio $w^\star$, its return $(w^\star)^t\mu$ and risk $(w^\star)^t\Sigma w^\star$; assuming precise parameter estimates (returns $\mu$, variances-covariances $\Sigma$).
- **Challenges and Limitations:**
  - High sensitivity to estimation errors in expected returns and covariances. Michaud, (1989), Chopra, V. K., & Ziemba, W. T. (1993), Meucci, A. (2005).
  - Lack of explicit consideration of individual risk contributions. Maillard, S., Roncalli, T., & Teiletche, J. (2010).
  - Limited applicability in real-world settings (e.g., regulatory constraints, diversification goals).

# Moving Towards a Risk-Centric Approach

- **Shift in Investment Objectives**:
  - Focus on **diversification** and **risk management** following financial crises (dot-com, 2008).
  - Emergence of new paradigms like **Risk Parity** and **Risk Budgeting**.
- **Concept of Risk Budgeting**:
  - **Objective**: Build a portfolio where each asset contributes to total risk according to a pre-defined budget.
  - **Methodology**:
    - Assign a "risk budget" to each asset.
    - Optimize portfolio weights to meet these contributions under long-only constraints.

# Risk Budgeting Problem

◦ Objective: In risk budgeting, we aim to allocate risk capital to different sub-portfolios or asset classes in a way that each component contributes a fixed proportion of the total portfolio risk.

◦ Framework: The financial market is composed of $d$ assets whose returns are given by the $\mathbb{R}^d$-valued random variable $X$. A financial portfolio is given the vector of weights $u = (u_1, \cdots, u_d)$ belonging to the simplex $\Delta_d = \left\{u \in \mathbb{R}^d_+ : u_1 + \cdots + u_d = 1\right\}$, then $-\langle u, X\rangle = -\sum_{i=1}^d u_i X_i$ corresponds to its loss.

Fix a risk measure $\rho$ and define the associated risk-measure function:

$$r_\rho(y) = \rho(-\langle y, X\rangle), \quad y \in \mathbb{R}^d_+,$$

which represents the required risk capital to hold the position $y$.

Assuming that $\rho$ is positively homogeneous, Euler's rule gives

$$r_\rho(y) = \sum_{i=1}^d y_i \frac{\partial r_\rho}{\partial y_i}(y), \quad y \in \mathbb{R}^d_+$$

The term $y_i \frac{\partial r_\rho}{\partial y_i}(y)$ is referred to as the risk contribution of asset $i$ to the overall portfolio risk.

- Risk budgeting problem: Given a vector of weights $b \in \Delta_d^{>0} := \left\{u \in (\mathbb{R}_+^\star)^d : u_1 + \cdots + u_d = 1\right\}$, called risk budget, find $u \in \Delta_d$ such that the risk contributions align with the predetermined proportions $b$ of the total risk.

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

- Popular Risk Measures: Risk budgeting can be applied to various risk measures:
  - Standard deviation (volatility),
  - Value-at-Risk (VaR),
  - Expected Shortfall (ES).
- References:
  - Course 2023-2024 in Portfolio Allocation and Asset Management, T. Roncalli.
  - Introduction to risk parity and budgeting, T. Roncalli.
  - A. R. Cetingoz & al. "Risk Budgeting portfolios: Existence and computation" (24); T. Griveau-Billion, J.-C. Richard, and T. Roncalli. "A fast algorithm for computing high-dimensional risk parity portfolios" (13), ...

Characterization: Existence and uniqueness of a vector of weights that solves the risk budgeting problem along with its characterization as the solution to a strictly convex optimization problem.

## Theorem

*Let $b \in \Delta_d^{>0}$. Assume that $\rho$ is convex and that $\rho(-\langle u, X\rangle) > 0$ for all $u \in \Delta_d^{>0}$. Let $g : \mathbb{R}_+ \to \mathbb{R}$ be a continuously differentiable, convex and increasing function. Let the function $\Gamma : (\mathbb{R}_+^0)^d \to \mathbb{R}$ 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 $\Gamma_g$ satisfying $\nabla \Gamma_g(y^\star) = 0$ and*

$$u^* := \frac{y^\star}{\sum_{i=1}^d y_i^\star}$$

*solves the risk-budgeting problem. Moreover, if $u$ is a solution to the risk-budgeting problem then*

$$u = u^\star.$$

*Step 1:* Note that $\Gamma_g(y) = g(r_\rho(y)) - \sum_{i=1}^d b_i \log y_i$ is well-defined for all $y \in (\mathbb{R}_+^*)^d$.

Observe also that $\Gamma_g$ is strictly convex because $g$ is convex and increasing, $r_g$ is convex, and the function $y \mapsto -\sum_{i=1}^d b_i \log y_i$ is strictly convex.

To prove the existence of a minimizer of $\Gamma_g$, for any $\theta \in \Delta_{>0}^d$, we introduce the function $\gamma_{g,\theta} : \mathbb{R}_+^* \to \mathbb{R}$ defined as:

$$\gamma_{g,\theta}(\lambda) = \Gamma_g(\lambda\theta) = g(\lambda r_g(\theta)) - \sum_{i=1}^d b_i \log \theta_i - \log \lambda.$$

We note that:

- $\lim_{\lambda \to 0^+} \gamma_{g,\theta}(\lambda) = +\infty$
- $\lim_{\lambda \to +\infty} \gamma_{g,\theta}(\lambda) = +\infty$ since $g(\lambda r_\rho(\theta)) \geq g(c) + g'(c)(\lambda r_\rho(\theta) - c)$ for any $c > 0$ s.t. $g'(c) > 0$.

By continuity, there exists $\lambda^*(\theta) > 0$ such that $\gamma_{g,\theta}(\lambda) \geq \gamma_{g,\theta}(\lambda^*(\theta))$ for every $\lambda > 0$.

*Step 2:* We now show, by contradiction, that $\theta \mapsto \lambda^*(\theta)$ is bounded.
Suppose there exists a sequence $(\theta_n)_n$ in $\Delta^d_{>0}$ such that $\lambda_n := \lambda^*(\theta_n) \to +\infty$.

We can extract$(\theta_{\varphi(n)})_n$ that converges to $\bar{\theta} \in \Delta_d$ and such that $\lambda_{\varphi(n)} \to +\infty$.
For all $n$, $\lambda_n$ satisfies the first-order condition:

$$r_\rho(\theta_{\varphi(n)})g'(\lambda_{\varphi(n)}r_\rho(\theta_{\varphi(n)})) - \frac{1}{\lambda_{\varphi(n)}} = 0 \Leftrightarrow x_n g'(x_n) = 1$$

where $x_n := \lambda_{\varphi(n)}r_\rho(\theta_{\varphi(n)})$.

Since $\lambda_{\varphi(n)} \to +\infty$ and $\lim_n r_\rho(\theta_{\varphi(n)}) = r_\rho(\bar{\theta}) > 0$, we conclude $x_n \to +\infty$.
This contradicts $x_n g'(x_n) = 1$, as $g$ is convex and increasing.

Conclusion: $\theta \mapsto \lambda^*(\theta)$ is bounded and there exists $M > 0$ such that $\lambda^*(\theta) \leq M$ for all $\theta \in \Delta^d_{>0}$.

*Step 3:* For any $y \in (\mathbb{R}_+^\star)^d$, letting $\theta = y/\sum_{i=1}^d y_i$, we get

$$\Gamma_g(y) = \gamma_{g,\theta}(\sum_{i=1}^d y_i) \geq \gamma_{g,\theta}(\lambda^\star(\theta)) = \Gamma_g\Big(\frac{y}{\sum_{i=1}^d y_i}\lambda^\star(\frac{y}{\sum_{i=1}^d y_i})\Big)$$

Let $C_M := \{y \in (\mathbb{R}_+^*)^d \mid \sum_{i=1}^d y_i \leq M\}$. From the previous inequality, we deduce:

$$\inf_{y\in(\mathbb{R}_+^*)^d} \Gamma_g(y) = \inf_{y\in C_M} \Gamma_g(y).$$

Now, let us consider an arbitrary vector $\bar{y} \in C_M$ and define

$$\epsilon := \min\left(\min_i \bar{y}_i, \min_i \exp\left(\frac{1}{b_i}\left(g(0) - (1-b_i)\log M - \Gamma_g(\bar{y})\right)\right)\right).$$

For any $y \in C_M$, if there exists $j \in \{1,\dots,d\}$ such that $y_j < \epsilon$, then, by definition of $\epsilon$,

$$\Gamma_g(y) = g(r_\rho(y)) - \sum_{i=1}^d b_i \log y_i \geq g(0) - \sum_{i=1}^d b_i \log y_i$$

$$\geq g(0) - \sum_{i\neq j} b_i \log y_i - b_j \log\epsilon \geq g(0) - \log M \sum_{i\neq j} b_i - b_j \log\epsilon$$

$$\geq g(0) - \log M(1-b_j) - b_j \log\epsilon \geq \Gamma_g(\bar{y}).$$

Setting $D_\epsilon := [\epsilon, +\infty)^d$, we therefore have

$$\inf_{y\in C_M} \Gamma_g(y) = \inf_{y\in C_M\cap D_\epsilon} \Gamma_g(y) = \Gamma_g(y^\star).$$

*Step 4:* Finally, the uniqueness of the minimizer follows from the strict convexity of $\Gamma_g$. Since $y^*$ is an interior minimum of $\Gamma_g$, we have:

$$g'(r_\rho(y^*))\partial_i r_\rho(y^*) - \frac{b_i}{y_i^*} = 0 \Leftrightarrow y_i^\star g'(r_\rho(y^*))\partial_i r_\rho(y^*) = b_i \quad \text{for all } i \in \{1,\dots,d\}.$$

Summing over $i$, Euler's theorem on homogeneous functions gives $r_\rho(y^*)g'(r_\rho(y^*)) = 1$, so:

$$y_i^*\partial_i r_\rho(y^*) = b_i r_\rho(y^*), \quad \text{for all } i.$$

Thus, $u^* := \frac{y^*}{\sum_{i=1}^d y_i^*}$ solves the risk-budgeting problem.

# Standard deviation

◦ Volatility is a reasonable choice of risk measure, especially when the probability distributions of asset returns do not exhibit asymmetry and/or heavy tails.

◦ Let $\Sigma$ be the covariance matrix of asset returns, then $r_{\rho_{STD}}(y) := \sqrt{y^T \Sigma y}$. Use a gradient descent algorithm to minimize over $(\mathbb{R}_+^*)^d$:

$$\Gamma_{x^2} : y \mapsto (r_{\rho_{STD}}(y))^2 - \sum_{i=1}^{d} b_i \log y_i = y^T \Sigma y - \sum_{i=1}^{d} b_i \log y_i.$$

◦ Main drawback: Asset and portfolio returns exhibit skewed and heavy-tailed distributions. Numerous studies show that excess returns reward investors for carrying the risk of sudden and significant losses. Therefore, to more efficiently deal with such distributional features in portfolio management, it makes sense to use other risk measures.

# Expected Shortfall

◦ An alternative risk measure is the ES which is known to be coherent. However, except for elliptical distributions, the ES must be computed numerically. One solution is to use the variational characterization of ES known as the Rockafellar-Uryasev formula:

$$\mathsf{ES}_\alpha(Z) = \min_{\xi\in\mathbb{R}} \mathbb{E}[\xi + \frac{1}{1-\alpha}(Z-\xi)_+]$$

so that the function $\Gamma_{Id}$ writes

$$\Gamma_{\mathsf{Id}}(y) = \min_{\xi\in\mathbb{R}} \mathbb{E}\Big[\xi + \frac{1}{1-\alpha}(-\langle y, X\rangle - \xi)_+ - \sum_{i=1}^d b_i \log(y_i).\Big]$$

Hence, solving the risk-budgeting problem boils down to solving the stochastic optimization problem

$$\min_{(y,\xi)\in(\mathbb{R}_+^\star)^d\times\mathbb{R}} \mathbb{E}\Big[\xi + \frac{1}{1-\alpha}(-\langle y, X\rangle - \xi)_+ - \sum_{i=1}^d b_i \log(y_i)\Big].$$

◦ The above stochastic optimization problem suggests to implement a stochastic gradient descent (SGD) algorithm based on the gradient

$$\partial_{y_i}\mathbb{E}\Big[\xi+\frac{1}{1-\alpha}(-\langle y,X\rangle-\xi)_+-\sum_{i=1}^d b_i\log(y_i)\Big]=\mathbb{E}\Big[\frac{-X_i}{1-\alpha}\mathbf{1}_{-\langle y,X\rangle\geq\xi}-b_i\Big]$$

$$\partial_{\xi}\mathbb{E}\Big[\xi+\frac{1}{1-\alpha}(-\langle y,X\rangle-\xi)_+-\sum_{i=1}^d b_i\log(y_i)\Big]=\mathbb{E}\Big[1-\frac{1}{1-\alpha}\mathbf{1}_{-\langle y,X\rangle\geq\xi}\Big]$$

so that the (formal) SGD algorithm writes

$$\begin{cases} y_i^{n+1}=y_i^n-\gamma_{n+1}\Big(\frac{-X_i^{n+1}}{1-\alpha}\mathbf{1}_{-\langle y^n,X^{n+1}\rangle\geq\xi^n}-\frac{b_i}{y_i^n}\Big), & i=1,\cdots d,\\ \xi^{n+1}=\xi^n-\gamma_{n+1}\Big(1-\frac{1}{1-\alpha}\mathbf{1}_{-\langle y^n,X^{n+1}\rangle\geq\xi^n}\Big). \end{cases}$$

with $(y^0,\xi^0)\in(\mathbb{R}_+^\star)^d\times\mathbb{R}$.

◦ Main drawback:

- The sequence $(y^n,\xi^n)_{n\geq0}$ is not guaranteed to live in $(\mathbb{R}_+^\star)^d\times\mathbb{R}$!
- Convergence?

# Projected SGD

- To circumvent the first difficulty, one may use a projected SGD:

$$\begin{cases} y^{n+1} = \Pi\left(y^n - \gamma_{n+1}\left(\dfrac{-X^{n+1}}{1-\alpha}\mathbf{1}_{-\langle y^n, X^{n+1}\rangle \geq \xi^n} - \dfrac{b_i}{y_i^n}\right)\right) \\ \xi^{n+1} = \xi^n - \gamma_{n+1}\left(1 - \dfrac{1}{1-\alpha}\mathbf{1}_{-\langle y^n, X^{n+1}\rangle \geq \xi^n}\right) \end{cases}$$

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

- But still the convergence is difficult to establish...
- Requires fine tuning of the learning step sequence $(\gamma_n)_{n\geq 1}$.

# Convergence of Projected SGD

◦ The classical SA algorithm solves problem

$$\min_{x \in X} g(x) = \mathbb{E}[G(x, Z)] \tag{1}$$

where $X \subset \mathbb{R}^d$ is a nonempty bounded closed convex set and $Z$ a random variable from which we can easily sample, by mimicking the simplest subgradient descent method.

◦ For chosen $x_1 \in X$ and a sequence $\gamma_j > 0$, $j = 1, \ldots$, of stepsizes, it generates the iterates by the formula

$$x_{j+1} := \Pi_X \left( x_j - \gamma_j \nabla G(x_j, Z^j) \right), \tag{2}$$

where $\Pi_X(x) = \arg\min_{x' \in X} \|x - x'\|_2$ and $(Z^j)_{j \geq 1}$ is an i.i.d. sequence with the same law as $Z$.

◦ Of course, the crucial question of that approach is how to choose the stepsizes $\gamma_j$.

◦ Let $\bar{x}$ be an optimal solution of problem (1). Note that since the set $X$ is compact and $g$ is continuous, problem (1) has an optimal solution. Note also that the iterate $x_j = x_j(Z_{[j-1]})$ is a function of the history $Z_{[j-1]} := (Z_1, \ldots, Z_{j-1})$ of the generated random process and hence is random.

◦ Denote $A_j := \frac{1}{2}\|x_j - \bar{x}\|_2^2$ and $a_j := \mathbb{E}[A_j] = \frac{1}{2}\mathbb{E}[\|x_j - \bar{x}\|_2^2]$. By using the fact that $\Pi_X$ is a contraction operator and since $\bar{x} \in X$ and hence $\Pi_X(\bar{x}) = \bar{x}$, we can write

$$\begin{aligned} A_{j+1} &= \frac{1}{2}\left\|\Pi_X\left(x_j - \gamma_j \nabla G(x_j, Z^j)\right) - \bar{x}\right\|_2^2 \\ &\leq A_j + \frac{\gamma_j^2}{2}\|\nabla G(x_j, Z^j)\|_2^2 - \gamma_j (x_j - \bar{x})^\top \nabla G(x_j, Z^j). \end{aligned}$$

We also have

$$\begin{aligned} \mathbb{E}[(x_j - \bar{x})^\top \nabla G(x_j, Z^j)] &= \mathbb{E}\left\{\mathbb{E}\left[(x_j - \bar{x})^\top \nabla G(x_j, Z^j) | \mathcal{F}_{j-1}\right]\right\} \\ &= \mathbb{E}[(x_j - \bar{x})^\top \nabla g(x_j)] \end{aligned}$$

using the fact that $x_j$ is $\mathcal{F}_{j-1}$ measurable and $Z^j$ is independent of $\mathcal{F}_j$.

◦ Therefore, by taking expectation on both sides of the previous inequality, we obtain

$$a_{j+1} \leq a_j - \gamma_j \mathbb{E}[(x_j - \bar{x})^\top \nabla g(x_j)] + \frac{\gamma_j^2}{2} M^2, \tag{3}$$

where

$$M^2 := \sup_{x \in X} \mathbb{E}[\|G(x, \xi)\|_2^2]. \tag{4}$$

$\rightsquigarrow$ We assume that the above constant $M$ is finite.

◦ Suppose, further, that the expectation function $g$ is strongly convex on $X$, i.e., there is a constant $c > 0$ such that

$$g(x') \geq g(x) + (x' - x)^\top \nabla g(x) + \frac{c}{2}\|x' - x\|_2^2, \quad \forall x', x \in X, \tag{5}$$

or equivalently

$$(x' - x)^\top (\nabla g(x') - \nabla g(x)) \geq c\|x' - x\|_2^2, \quad \forall x', x \in X. \tag{6}$$

◦ Strong convexity of $g$ implies that the minimizer $\bar{x}$ is unique. By optimality of $\bar{x}$, we have that

$$(x - \bar{x})^\top \nabla g(\bar{x}) \geq 0, \quad \forall x \in X. \tag{7}$$

Combining (7) with (6), we obtain

$$\begin{aligned} \mathbb{E}[(x_j - \bar{x})^\top \nabla g(x_j)] &\geq \mathbb{E}[(x_j - \bar{x})^\top (\nabla f(x_j) - \nabla f(\bar{x}))] \\ &\geq c\mathbb{E}[\|x_j - \bar{x}\|_2^2] = 2ca_j. \end{aligned} \tag{8}$$

Substituting (8) into (3), we get

$$a_{j+1} \leq (1 - 2c\gamma_j)a_j + \frac{\gamma_j^2}{2} M^2. \tag{9}$$

◦ Let us take stepsizes $\gamma_j = \frac{\gamma}{j}$ for some constant $\gamma > \frac{1}{2c}$. Then by (9), we have

$$a_{j+1} \leq \left(1 - \frac{2c\gamma}{j}\right) a_j + \frac{\gamma^2 M^2}{2j^2}. \tag{10}$$

By induction, it follows that

$$a_j \leq \frac{\kappa}{j}, \tag{11}$$

where

$$\kappa := \max\left\{\frac{\gamma^2 M^2}{2(2c\gamma - 1)}, a_1\right\}. \tag{12}$$

◦ Suppose further that $\bar{x}$ is an interior point of $X$ and $\nabla g(x)$ is Lipschitz continuous, i.e., there exists a constant $L > 0$ such that

$$\|\nabla g(x') - \nabla g(x)\|_2 \leq L\|x' - x\|_2, \quad \forall x', x \in X. \tag{13}$$

Then,

$$g(x) \leq g(\bar{x}) + \frac{L}{2}\|x - \bar{x}\|_2^2, \quad \forall x \in X, \tag{14}$$

and hence

$$\mathbb{E}[g(x_j) - g(\bar{x})] \leq La_j \leq \frac{L\kappa}{j}. \tag{15}$$

◦ Conclusion: Under the specified assumptions, it follows that after $n$ iterations, the expected error of the current solution is of order $O(n^{-1/2})$, and the expected error of the corresponding objective value is of order $O(n^{-1})$, provided that $\gamma > \frac{1}{2c}$.

◦ Caution: However, this result is highly sensitive to the choice of $c$. Overestimating $c$ can lead to suboptimal convergence,

# Some difficulties of GD and SGD

- Implementing a GD or SGD scheme requires fine tuning of learning sequence $(\gamma_k)_{k\geq 1}$.

# Some difficulties of GD and SGD

- Implementing a GD or SGD scheme requires fine tuning of learning sequence $(\gamma_k)_{k\geq 1}$.
- **Basic example:** Consider the optimization problem:

$$\text{Minimize } \left\{ f(x) = \frac{x^2}{10} \right\}, \quad \text{over} \quad X = [-1, 1] \subset \mathbb{R}.$$

# Some difficulties of GD and SGD

- Implementing a GD or SGD scheme requires fine tuning of learning sequence $(\gamma_k)_{k\geq 1}$.
- **Basic example:** Consider the optimization problem:

$$\text{Minimize } \left\{ f(x) = \frac{x^2}{10} \right\}, \quad \text{over} \quad X = [-1, 1] \subset \mathbb{R}.$$

- **Iteration Rule:** of the standard GD scheme with learning rate $\gamma_k = \frac{\gamma}{k}$, $\gamma > 0$.

$$x_{k+1} = x_k - \frac{f'(x_k)}{k} = \left(1 - \frac{\gamma}{5k}\right) x_k.$$

# Some difficulties of GD and SGD

- Implementing a GD or SGD scheme requires fine tuning of learning sequence $(\gamma_k)_{k\geq 1}$.
- **Basic example:** Consider the optimization problem:

$$\text{Minimize } \left\{ f(x) = \frac{x^2}{10} \right\}, \quad \text{over} \quad X = [-1, 1] \subset \mathbb{R}.$$

- **Iteration Rule:** of the standard GD scheme with learning rate $\gamma_k = \frac{\gamma}{k}$, $\gamma > 0$.

$$x_{k+1} = x_k - \frac{f'(x_k)}{k} = \left(1 - \frac{\gamma}{5k}\right) x_k.$$

- **Parameters:** $x_1 = 1$ and $\gamma = 1$. It holds

$$x_n = \prod_{k=1}^{n-1} (1 - \frac{1}{5k}) = e^{-\sum_{k=1}^{n-1} \ln(1+\frac{1}{5k-1})} > e^{-(0.25+\int_1^{n-1} \frac{1}{5t-1}\, dt)} > 0.8 n^{-\frac{1}{5}}$$

# Some difficulties of GD and SGD

- Implementing a GD or SGD scheme requires fine tuning of learning sequence $(\gamma_k)_{k\geq 1}$.

- **Basic example:** Consider the optimization problem:

$$\text{Minimize } \left\{ f(x) = \frac{x^2}{10} \right\}, \quad \text{over} \quad X = [-1, 1] \subset \mathbb{R}.$$

- **Iteration Rule:** of the standard GD scheme with learning rate $\gamma_k = \frac{\gamma}{k}$, $\gamma > 0$.

$$x_{k+1} = x_k - \frac{f'(x_k)}{k} = \left(1 - \frac{\gamma}{5k}\right) x_k.$$

- **Parameters:** $x_1 = 1$ and $\gamma = 1$. It holds

$$x_n = \prod_{k=1}^{n-1} (1 - \frac{1}{5k}) = e^{-\sum_{k=1}^{n-1} \ln(1+\frac{1}{5k-1})} > e^{-(0.25+\int_1^{n-1} \frac{1}{5t-1}\, dt)} > 0.8 n^{-\frac{1}{5}}$$

$\rightsquigarrow$ Slow convergence: For $n = 10^9$, error remains $> 0.013$.

# Some difficulties of GD and SGD

- Implementing a GD or SGD scheme requires fine tuning of learning sequence $(\gamma_k)_{k\geq 1}$.

- **Basic example:** Consider the optimization problem:

$$\text{Minimize } \left\{ f(x) = \frac{x^2}{10} \right\}, \quad \text{over} \quad X = [-1, 1] \subset \mathbb{R}.$$

- **Iteration Rule:** of the standard GD scheme with learning rate $\gamma_k = \frac{\gamma}{k}$, $\gamma > 0$.

$$x_{k+1} = x_k - \frac{f'(x_k)}{k} = \left(1 - \frac{\gamma}{5k}\right) x_k.$$

- **Parameters:** $x_1 = 1$ and $\gamma = 1$. It holds

$$x_n = \prod_{k=1}^{n-1} (1 - \frac{1}{5k}) = e^{-\sum_{k=1}^{n-1} \ln(1+\frac{1}{5k-1})} > e^{-(0.25+\int_1^{n-1} \frac{1}{5t-1}\, dt)} > 0.8 n^{-\frac{1}{5}}$$

$\rightsquigarrow$ Slow convergence: For $n = 10^9$, error remains $> 0.013$.
$\rightsquigarrow$ Optimal choice: $\gamma = 1/c = 5$, yields $x_k = 0$ in a single iteration.