Recursive Optimization of Convex Risk Measures: Mean-Semideviation Models

Dionysios S. Kalogerias and Warren B. Powell

October 2018

Abstract

We develop recursive, data-driven, stochastic subgradient methods for optimizing a new, versatile, and application-driven class of convex risk measures, termed here as *mean-semideviations*, strictly generalizing the well-known and popular mean-upper-semideviation. We introduce the *MESSAGE^p algorithm*, which is an efficient compositional subgradient procedure for iteratively solving *convex* mean-semideviation risk-averse problems *to optimality*, and constitutes a *parallel* variation of the recently developed, general purpose *T-SCGD algorithm* of Yang, Wang & Fang [Yang et al., 2018]. We analyze the asymptotic behavior of the *MESSAGE^p* algorithm under a flexible and structure-exploiting set of problem assumptions, which reveal a well-defined trade-off between the expansiveness of the random cost and the smoothness of the mean-semideviation risk measure under consideration. In particular:

- Under appropriate stepsize rules, we establish pathwise convergence of the *MESSAGE^p* algorithm in a strong technical sense, confirming its asymptotic consistency.
- Assuming a *strongly convex cost*, we show that, for fixed semideviation order $p > 1$ and for $\epsilon \in [0, 1)$, the *MESSAGE^p* algorithm achieves a squared-$\mathcal{L}_2$ solution suboptimality rate of the order of $\mathcal{O}(n^{-(1-\epsilon)/2})$ iterations, where, for $\epsilon > 0$, pathwise convergence is *simultaneously* guaranteed. This result establishes a rate of order arbitrarily close to $\mathcal{O}(n^{-1/2})$, while ensuring strongly stable pathwise operation. For $p \equiv 1$, the rate order improves to $\mathcal{O}(n^{-2/3})$, which also suffices for pathwise convergence, and matches previous results.
- Likewise, in the general case of a *convex cost*, we show that, for any $\epsilon \in [0, 1)$, the *MESSAGE^p* algorithm *with iterate smoothing* achieves an $\mathcal{L}_1$ objective suboptimality rate of the order of $\mathcal{O}(n^{-(1-\epsilon)/(4\mathbb{1}_{\{p>1\}}+4)})$ iterations. This result provides maximal rates of $\mathcal{O}(n^{-1/4})$, if $p \equiv 1$, and $\mathcal{O}(n^{-1/8})$, if $p > 1$, matching the state of the art, as well.

Finally, we discuss the superiority of the proposed framework for convergence, as compared to that employed earlier in [Yang et al., 2018], within the risk-averse context under consideration. By performing careful analysis and by constructing non-trivial counterexamples, we explicitly demonstrate that the class of mean-semideviation problems supported herein is *strictly larger* than the respective class of problems supported in [Yang et al., 2018]. As a result, this work establishes the applicability of compositional stochastic optimization for a significantly and strictly wider spectrum of convex mean-semideviation risk-averse problems, as compared to the state of the art. This fact justifies the purpose of our work from this perspective, as well.

Keywords. Risk-Averse Optimization, Risk-Aware Learning, Risk Measures, Mean-Upper-Semideviation, Stochastic Optimization, Stochastic Gradient Descent, Compositional Optimization.

The Authors are with the Department of Operations Research & Financial Engineering (ORFE), Princeton University, Sherrerd Hall, Charlton Street, Princeton, NJ 08544, USA. e-mail: {dkalogerias, powell}@princeton.edu.

Contents

1 Introduction

During the last almost twenty years, many significant advances have been made in the now relatively mature area of risk-averse modeling and optimization. These primarily include the fundamental axiomatization and theoretical characterization of risk functionals, also commonly known as risk measures [Kijima and Ohnishi, 1993, Rockafellar and Uryasev, 1997, Artzner et al., 1999, Ogryczak and Ruszczyński, 1999, Ogryczak and Ruszczyński, 2002, Rockafellar and Uryasev, 2002, Rockafellar et al., 2003, 2006, Ruszczyński and Shapiro, 2006b, Shapiro et al., 2014], as well as extensive analysis in the context of risk-averse stochastic programs in both static and sequential decision making problem settings [Rockafellar and Uryasev, 1997, Föllmer and Schied, 2002, Rockafellar et al., 2003, 2006, Ruszczyński and Shapiro, 2006a, Collado et al., 2012, Çavuş and Ruszczyński, 2014a, Asamov and Ruszczyński, 2015, Dentcheva and Ruszczyński, 2017, Grechuk and Zabarankin, 2017, Shapiro, 2017, Fan and Ruszczyński, 2018]. The importance of building a well structured theory of risk is motivated by its natural and intuitive relevance to problems from a large variety of applied domains. Arguably the oldest, archetypical application of risk is in Finance [Kijima and Ohnishi, 1993, Rockafellar and Uryasev, 1997, Andersson et al., 2001, Krokhmal et al., 2001, Chen and Wang, 2008, Shang et al., 2018], which has decisively driven pioneering research in risk-averse modeling and optimization, from its very birth, probably dating back to the work of Markowitz [Markowitz, 1952], to present. Other applications of risk may be found in both classical and contemporary domains such as Energy [Moazeni et al., 2015, Bruno et al., 2016, Jiang and Powell, 2016], Wireless Networks [Ma et al., 2018], Inventory Optimization [Ahmed et al., 2007, Chen et al., 2007, Xinsheng et al., 2015] and Supply Chain Management [Gan et al., 2004, Sawik, 2016], to name a few.

Most recently, the development of effective *computational methods* for applying risk-averse optimization to actual problems has also been attracting considerable attention; see, e.g., [Ruszczyński, 2010, Çavuş and Ruszczyński, 2014b, Moazeni et al., 2017, Tamar et al., 2017, Dentcheva et al., 2017, Huang and Haskell, 2017, Jiang and Powell, 2017, Yu et al., 2018]. This line of work can be divided between sequential settings [Çavuş and Ruszczyński, 2014b, Moazeni et al., 2017, Tamar et al., 2017, Huang and Haskell, 2017, Jiang and Powell, 2017, Yu et al., 2018], and static settings [Tamar et al., 2017, Dentcheva et al., 2017], for a variety of different problem characteristics. Computational recipes also vary. *For instance*, [Ruszczyński, 2010] and [Çavuş and Ruszczyński, 2014b] develop and analyze variations of the well known value and policy iteration algorithms of risk-neutral dynamic programming; [Moazeni et al., 2017] proposes a method for risk-averse nonstationary direct parametric policy search for finite horizon problems; [Tamar et al., 2017], [Dentcheva et al., 2017] and [Yu et al., 2018] rely on the so-called *Sample Average Approximation (SAA)* approach [Shapiro et al., 2014], where an appropriately constructed empirical estimate of the original objective is used as a *surrogate* to that of the original stochastic program, assuming existence of a sufficiently large sample of the processes introducing uncertainty into the corresponding risk-averse objective; [Huang and Haskell, 2017] and [Jiang and Powell, 2017] consider an *Approximate Dynamic Programming (ADP)* [Powell, 2011] approach, where sequential finite state/action risk-averse stochastic programs are tackled via stochastic approximation [Kushner and Yin, 2003].

Following this recent trend, this paper proposes and rigorously analyzes *recursive stochastic subgradient methods* for an important class of *static, convex* risk-averse stochastic programs. In a nutshell, we make the following contributions:

1) Following the Mean-Risk Model paradigm [Shapiro et al., 2014], we introduce a new class of *convex* risk measures, called *mean-semideviations*. These *strictly generalize* the well known mean-upper-semideviation risk measure, and are constructed by replacing the positive part *weighting* function of the latter by another nonlinear map, termed here as a *risk regularizer*, obeying certain properties. Mean-semideviations share the same core analytical structure with the mean-upper-semideviation risk measure; however, they are much more versatile in applications. We study mean-semideviations in terms of their basic properties, and we present a fundamental *constructive* characterization result, demonstrating their generality. Specifically, we show that the class of all mean-semideviation risk measures is *almost in one-to-one correspondence* with the class of cumulative distribution functions (cdfs) of all integrable random variables. This result provides an *analytical device* for constructing mean-semideviations with desirable characteristics, starting from any cdf of the aforementioned type. The flexibility and effectiveness of mean-semideviations are explicitly demonstrated on a classical, chance-constrained newsvendor model, as well.

2) We introduce the *MESSAGE^p (MEan-Semideviation Stochastic compositionAl subGradient dEscent of order p) algorithm*, an efficient, *data-driven* Stochastic Subgradient Descent (SSD) -type procedure for iteratively solving *convex* mean-semideviation risk-averse problems *to optimality*. The *MESSAGE^p* algorithm constitutes a *parallel* variation of general purpose *T-level Stochastic Compositional Gradient Descent (T-SCGD)* algorithm, recently developed in [Yang et al., 2018], under a generic theoretical framework. Although risk-averse optimization is listed in [Yang et al., 2018] as a potential application of stochastic compositional optimization for the mere case of mean-upper-semideviations, this work is the first to propose a general algorithm, applicable to any mean-semideviation model of choice.

3) We analyze the asymptotic behavior of the *MESSAGE^p* algorithm under a new, flexible and *structure-exploiting* set of problem assumptions, which reveal a well-defined trade-off between the expansiveness of the random cost and the smoothness of the mean-semideviation risk measure under consideration. In particular, under our proposed structural framework:

- Under appropriate stepsize rules, we establish pathwise convergence of the *MESSAGE^p* algorithm in a strong technical sense, confirming its asymptotic consistency.
- Assuming a *strongly convex* cost function, the convergence rate of the *MESSAGE^p* algorithm is studied in detail. More specifically, we show that, for fixed semideviation order $p > 1$ and for $\epsilon \in [0, 1)$, the *MESSAGE^p* algorithm achieves a squared-$\mathcal{L}_2$ solution suboptimality rate of the order of $\mathcal{O}(n^{-(1-\epsilon)/2})$ iterations, where, for $\epsilon > 0$, pathwise convergence is *simultaneously* guaranteed. Thus, this new result establishes a rate of order arbitrarily close to $\mathcal{O}(n^{-1/2})$, also ensuring strongly stable pathwise operation of the *MESSAGE^p* algorithm. In the simpler case where the semideviation order is chosen as $p \equiv 1$, the rate order of the proposed algorithm improves to $\mathcal{O}(n^{-2/3})$, which is sufficient for pathwise convergence as well, and matches previous results in the related literature [Wang et al., 2017].
- For the general case of a *convex cost*, we show that, for any $\epsilon \in [0, 1)$, the *MESSAGE^p* algorithm *with iterate smoothing* achieves an $\mathcal{L}_1$ objective suboptimality rate of the order of $\mathcal{O}(n^{-(1-\epsilon)/(4\mathbb{1}_{\{p>1\}}+4)})$. As in the strongly convex case, for $\epsilon > 0$, pathwise convergence is also *simultaneously* guaranteed. For $\epsilon \equiv 0$, this result provides maximal rates of $\mathcal{O}(n^{-1/4})$, if $p \equiv 1$, and $\mathcal{O}(n^{-1/8})$, if $p > 1$, matching the state of the art, as well.

4) We discuss the superiority of the proposed framework for convergence, as compared to that employed earlier in [Yang et al., 2018], within the risk-averse context under consideration.

By performing careful analysis and by constructing non-trivial counterexamples, we explicitly demonstrate that the class of mean-semideviation problems supported herein is *strictly larger* than the respective class of problems supported in [Yang et al., 2018]. As a result, this paper establishes the applicability of compositional stochastic optimization for a significantly and strictly wider spectrum of convex mean-semideviation risk-averse problems, as compared to the state of the art. This fact justifies the purpose of our work from this perspective, as well.

Our contributions, briefly outlined above, are now discussed in greater detail. We also briefly explain how our work relates to and is placed within the existing literature.

1.1 Mean-Semideviation Risk Measures

Mean-semideviation risk measures, as proposed and developed in this work, constitute a new class of risk measures where, given a random cost, the corresponding dispersion measure (the term penalizing the "mean" part of a mean-risk functional) is defined as the $\mathcal{L}_p$-norm of a nonlinear, one-dimensional map of the *centered cost*, or, in other words, *its central deviation*. This map is called a *risk regularizer*, and possesses certain analytical properties: convexity, nonnegativity, monotonicity and nonexpansiveness. Dispersion measures with this structure are suggestively called *generalized semideviations*.

This terminology originates from the presence of the positive part function $(\cdot)_+ \triangleq \max\{\cdot, 0\}$, which is the simplest, prototypical example of a risk regularizer, in the corresponding dispersion measure of the well known mean-upper-semideviation risk measure [Shapiro et al., 2014], i.e., the upper-(central)-semideviation. Mean-semideviations are much more versatile, however, since different choices for the involved risk regularizer correspond to different rules for ranking the relative effect of both riskier (higher than the mean) and less risky (lower than the mean) events, corresponding to specific regions in the range of the (centered) cost. As a result, the choice of the risk regularizer affects the general quality and the roughness/stability of an optimal random cost, in a decision making setting. Consequently, owing to their versatility, mean-semideviations are practically appealing as well, because they are parametrizable and they may incorporate domain specific knowledge more easily than the rigid mean-upper-semideviation.

In this work, after we formulate simple conditions for the existence of mean-semideviation risk measures, we study their basic geometric properties, such as convexity and monotonicity. Contrary to the mean-upper-semideviation alone, mean-semideviations are *not* coherent risk measures, in general (as a class), because they do not satisfy positive homogeneity [Shapiro et al., 2014]. This is due to the potential nonhomogeneity of the risk regularizer involved. They do satisfy convexity, monotonicity and translation equivariance, though and, therefore, they belong to the class of *convex risk measures*, [Föllmer and Schied, 2002, Shapiro et al., 2014], and that of *convex-monotone risk measures*, as well.

Further, we present a fundamental *constructive* characterization result, demonstrating the generality of mean-semideviations. Specifically, on the one hand, this result shows that the class of all mean-semideviation risk measures is *almost in one-to-one correspondence* with the class of cdfs of all integrable random variables (on the line). On the other, it provides an *analytical device* for constructing such risk measures from any cdf of the aforementioned type. Although not studied in this paper, this correspondence between mean-semideviations and cdfs might be of interest in other areas related to stochastically robust optimization such as stochastic dominance; see, for instance, the seminal articles [Ogryczak and Ruszczyński, 1999, Ogryczak and Ruszczyński, 2002] for some interesting connections.

Our discussion on mean-semideviation risk measures is concluded by a demonstration of their practical usefulness and flexibility on a classical, chance-constrained newsvendor model. After we briefly analyze the structure of the problem under consideration, we put risk regularizers -each inducing a mean-semideviation risk measure- *in context*, and we explicitly discuss their construction, so that the resulting mean-semideviation risk measure best reflects problem characteristics, and the objectives of the decision maker. Additionally, we present numerical simulations, experimentally confirming the effectiveness of the proposed risk-averse approach. Our simulations also reveal some interesting features of the resulting risk-averse solutions, which we further discuss.

Relation to the Literature: We are not the first to propose convex risk measures featuring nonlinear weighting functions; see, for instance, [Kijima and Ohnishi, 1993, Chen and Yang, 2011, Fu et al., 2017]. In particular, the recent article [Fu et al., 2017] considers risk measures defined as a nonlinearly weighted, order-1 (lower) semideviation *from a fixed target* (see, for instance, Example 6.25 in [Shapiro et al., 2014]), focusing mainly on their applications on a portfolio selection model. In [Fu et al., 2017], the corresponding weighting function shares the same properties as a risk regularizer (see above), except for nonexpansiveness. However, our proposed mean-semideviation risk measures are substantially different and structurally more complex compared to the risk measures proposed in [Fu et al., 2017]. The main reason is the presence of the *expected cost*, rather than a fixed target, in the definition of mean-semideviations; for more details, compare ([Fu et al., 2017], Definition 1) with Section 3 herein.

1.2 Recursive Optimization of Mean-Semideviations

The main contribution of this work concerns efficient optimization of mean-semideviations, measuring convexly parameterized random cost functions, over a closed and convex set. We introduce and rigorously analyze the *MESSAGE^p (MEan-Semideviation Stochastic compositionAl subGradient dEscent of order p) algorithm* (Algorithm 1 in Section 4.3), which constitutes an efficient Stochastic Subgradient Descent (SSD) -type procedure for iteratively solving our base problem *to optimality*. The *MESSAGE^p* algorithm may be seen as a parameterized (relative to the choice of the risk regularizer), *parallel* variation of the general purpose *T-Level Stochastic Compositional Gradient Descent (T-SCGD) algorithm*, presented and analyzed very recently in [Yang et al., 2018] under generic assumptions. In turn, the *T-SCGD* algorithm is a natural generalization of the *Basic 2-Level SCGD* algorithm, presented and analyzed earlier in [Wang et al., 2017]. A key feature of the aforementioned *compositional* stochastic subgradient schemes is the existence of more than one (T, in general), *pairwise coupled* stochastic approximation updates, or *levels*, *each with a dedicated stepsize*, which are executed *concurrently* through the operation of the algorithm. In the case of the *MESSAGE^p* algorithm, there exist three such levels (that is, $T \equiv 3$), and this results naturally, due our specific problem structure. However, contrary to the *T-SCGD* algorithm, all three stochastic approximation levels of the *MESSAGE^p* algorithm *are executed completely in parallel within every iteration*, presenting additional operational efficiency, potentially important in various applications.

Pathwise convergence and convergence rate analyses of the *T-SCGD* algorithm are presented in [Yang et al., 2018], and [Wang et al., 2017] (where, in the latter, $T \equiv 2$). However, the respective structural framework considered in both [Yang et al., 2018] and [Wang et al., 2017], when applied to the problem class considered in this work, imposes *significant restrictions* in regard to the possible choice of the risk regularizer, partially related to the expansiveness and smoothness (or roughness) of the involved random cost function. This fact significantly limits the type of problems the *T-SCGD* algorithm is provably applicable to, *at least* within the class of risk-averse problems introduced and studied herein. For example, when $p \equiv 1$, arguably the most popular regularizer $(\cdot)_+$, leading to the mean-upper-semideviation risk measure, is *not* supported within the framework of [Wang et al., 2017, Yang et al., 2018]. This is because nonsmooth risk regularizers *exhibiting corner points*, such as $(\cdot)_+$, apparently have *discontinuous subderivatives*, whereas the respective assumptions made in [Wang et al., 2017, Yang et al., 2018] essentially require the respective risk regularizer to be *not only everywhere differentiable*, but to have *Lipschitz derivatives*, as well. This shortcoming of the theoretical framework of [Wang et al., 2017, Yang et al., 2018] naturally carries over to higher values of the semideviation order, p. Naturally, the theoretical narrowness of [Wang et al., 2017, Yang et al., 2018] motivates closer study of any compositional subgradient algorithm whatsoever, one that would exploit the special characteristics of a mean-semideviation risk measure. The ultimate goal is the development of a sufficiently general theoretical framework, which will justify the compositional optimization approach for the whole class of mean-semideviation risk measures, under as weak structural assumptions as possible.

Following this direction, and focusing on optimizing mean-semideviation models, we present a new and flexible set of problem assumptions, *substantially weaker than those employed in* [Wang et al., 2017, Yang et al., 2018], under which we analyze the asymptotic behavior of the $\mathit{MESSAGE}^p$ algorithm, proposed in our work. Our framework carefully exploits the structure of mean-semideviations, and presents a probably fundamental, though practically useful, trade-off between the expansiveness of the random cost function and the smoothness of the chosen risk regularizer, in a very well-defined sense. As previously outlined, our results are restated, as follows.

First, under appropriate stepsize rules, we establish pathwise convergence of the $\mathit{MESSAGE}^p$ algorithm in the same strong sense as in [Wang et al., 2017, Yang et al., 2018], thus confirming its asymptotic consistency.

Second, assuming a *strongly convex* cost function, we study the convergence rate of the $\mathit{MESSAGE}^p$ algorithm, in detail. More specifically, we show that, for fixed semideviation order $p > 1$ and for any choice of $\epsilon \in [0, 1)$, the $\mathit{MESSAGE}^p$ algorithm achieves a squared-$\mathcal{L}_2$ solution suboptimality rate of the order of $\mathcal{O}(n^{-(1-\epsilon)/2})$ iterations. Here, ϵ is a user-specified parameter, which directly affects stepsize selection. If, additionally, ϵ is chosen to be strictly positive, that is, for $\epsilon > 0$, pathwise convergence is *simultaneously* guaranteed. This completely novel result establishes a convergence rate of order arbitrarily close to $\mathcal{O}(n^{-1/2})$ as $\epsilon \to 0$, while ensuring strongly stable pathwise operation of the algorithm. In the structurally simpler case where $p \equiv 1$, the rate order improves to $\mathcal{O}(n^{-2/3})$, which is sufficient for pathwise convergence as well, and matches existing results in compositional stochastic optimization, developed earlier along the lines of [Wang et al., 2017].

Third, for the general case of a convex cost function, we show that, for any $\epsilon \in [0, 1)$, the $\mathit{MESSAGE}^p$ algorithm *with iterate smoothing* achieves an $\mathcal{L}_1$ objective suboptimality rate of the order of $\mathcal{O}(n^{-(1-\epsilon)/(4\mathbb{1}_{\{p>1\}}+4)})$. As in the strongly convex case, for $\epsilon > 0$, pathwise convergence is also *simultaneously* guaranteed. For $\epsilon \equiv 0$, this result provides maximal rates of $\mathcal{O}(n^{-1/4})$, if $p \equiv 1$, and $\mathcal{O}(n^{-1/8})$, if $p > 1$, matching the state of the art, as well [Wang et al., 2017, Yang et al., 2018]. Although those rates may not be particularly satisfying, they quantitatively demonstrate the remarkable speedup achieved by assuming and leveraging strong convexity for the analysis and operation of the $\mathit{MESSAGE}^p$ algorithm.

The proposed structural framework adequately mitigates the aforementioned technical issues of that considered in [Wang et al., 2017, Yang et al., 2018]. For example, we show that, when the random cost function has *bounded* (random) subgradients and its distribution is generally well-behaved, the choice of the risk regularizer can be *completely unconstrained*, *regardless of the value* *of* $p \in [1, \infty)$. As a result, under the new framework, the most popular candidate $(\cdot)_+$, but also every risk regularizer exhibiting corner points, are now valid choices (under appropriate conditions) for any p, contrary to [Wang et al., 2017, Yang et al., 2018].

Finally, in order to show the superiority of our proposed framework compared to that of [Wang et al., 2017, Yang et al., 2018], we present a detailed analytical comparison, which rigorously demonstrates that the class of mean-semideviation programs supported within this work *contains* the respective class of problems supported within [Wang et al., 2017, Yang et al., 2018]; further, *the inclusion is strict*. Such comparison is made possible by performing careful analysis and by constructing non-trivial, *non-cornercase* counterexamples. As a result, the applicability of compositional stochastic optimization is established herein for a significantly and strictly wider spectrum of convex mean-semideviation risk-averse problems, as compared to the state of the art. This fact justifies the purpose of our work from this perspective, in addition to our algorithmic contribution, as well.

Relation to the Literature: Apparently, the results presented in this work are related to those developed in [Wang et al., 2017, Yang et al., 2018], for a generic problem setting. Indeed, as already stated, optimization of mean-upper-semideviation risk measures has been briefly identified in [Wang et al., 2017, Yang et al., 2018] as a potential application of the compositional algorithms proposed therein. However, as mentioned above, the assumptions on problem structure employed in [Wang et al., 2017, Yang et al., 2018] are too restrictive to adequately study the class of mean-semideviation risk measures introduced herein, which include the mean-upper-semideviation as a single member of this class. Except for the aforementioned works, and as also discussed above, there is a significant line of research considering the SAA approach to risk-averse stochastic optimization, both from a fundamental, theoretical perspective [Shapiro, 2013, Guigues et al., 2016, Dentcheva et al., 2017] and from the computational one [Dentcheva et al., 2017, Tamar et al., 2017]. As noted in [Wang et al., 2017, Yang et al., 2018], the compositional, SSD-type optimization algorithms analyzed in this paper present some major natural advantages over the SAA approach. First, the *MESSAGE*^{p} algorithm solves the *original* risk-averse stochastic program *asymptotically to optimality*, whereas, in the SAA approach, the corresponding SAA surrogate to the original program is solved, producing only an approximate solution; as the number of the sample increases the solution to the SAA surrogate approaches that of the original stochastic program, in some well defined sense [Shapiro, 2013, Dentcheva et al., 2017]. Second, because of its nature, the SAAs cannot exploit new information available to the decision maker, so that they can improve their decisions, based on those made so far; in fact, the SAA surrogate needs to be redefined using new available information, and then solved afresh. Of course, the *MESSAGE*^{p} algorithm efficiently exploits new information, due to its *recursive*, sequential nature. Third, as a result of the above, SAAs are not suitable for settings where *information is available sequentially*, and decisions have to be made adaptively over time. Fourth, SAAs might often require a very large number of samples for producing accurate approximations to the optimal decisions corresponding to the original problem, and this might result in optimization problems whose objective is computationally difficult to evaluate. For more details on this, see [Wang et al., 2017]. On the contrary, the *MESSAGE*^{p} algorithm is iterative in nature, and presents *minimal* and *fixed* time and space complexity per iteration.

Organization of the Paper

The rest of the paper is organized as follows. Section 2 establishes the stochastic risk-averse convex programming setting under study, and provides some elementary, albeit necessary preliminaries on the theory of risk measures. In Section 3, we constructively introduce the class of mean-semideviation risk measures, we study their existence and their structural properties, we discuss specific examples, and we develop our above mentioned fundamental characterization result. Section 4 is devoted to the development and analysis of the *MESSAGE*^{p} algorithm, under our proposed theoretical framework for convergence, and includes the rigorous comparison of our results with those presented in [Yang et al., 2018]. Finally, Section 5 concludes the paper.

Note: Some longer proofs of the theoretical results presented in the paper in the form of Theorems, Lemmata and Propositions are excluded from the main body of the paper for clarity in the exposition, and are presented in Section 7 (Appendix).

Notation & Definitions

Matrices and vectors will be denoted by boldface uppercase and boldface lowercase letters, respectively. Calligraphic letters and formal script letters will generally denote sets and σ-algebras, respectively, except for clearly specified exceptions. The operator $(\cdot)^{\boldsymbol{T}}$ will denote vector transposition. The ℓ_p-norm of $\boldsymbol{x} \in \mathbb{R}^n$ is $\|\boldsymbol{x}\|_p \triangleq (\sum_{i=1}^n |x(i)|^p)^{1/p}$, for all $\mathbb{N} \ni p \geq 1$. Similarly, the $\mathcal{L}_p$ norm of an appropriately measurable function $f(\cdot)$ will be $\|f\|_{\mathcal{L}_p} \triangleq \left(\int |f(x)|^p \mathrm{d}\mu(x)\right)^{1/p}$ for $p \in [1, \infty)$, and $\|f\|_{\mathcal{L}_\infty} \triangleq \operatorname{ess\,sup}_x |f(x)|$, where the reference measure μ will be clearly specified by the context. The finite N-dimensional identity operator will be denoted as $\mathbf{I}_N$. Additionally, we define $\mathbb{N}^+ \triangleq \{1, 2, \ldots\}$, $\mathbb{N}_n^+ \triangleq \{1, 2, \ldots, n\}$ and $\mathbb{N}_n \triangleq \{0\} \cup \mathbb{N}_n^+$, for $n \in \mathbb{N}^+$.

If Ω denotes a *base* sample space and $F : \mathbb{R}^N \times \Omega \to \mathbb{R}$ (referring directly to Ω), then, for the sake of clarity, we sometimes drop dependence on $\omega \in \Omega$, and write simply $F(\boldsymbol{x}, \omega) \equiv F(\boldsymbol{x})$ (clear by the context).

For every set $\mathcal{X} \subseteq \mathbb{R}^N$, which is nonempty, closed and convex, the *Euclidean projection onto* $\mathcal{X}$, $\Pi_{\mathcal{X}} : \mathbb{R}^N \to \mathcal{X}$ is defined, as usual, as $\Pi_{\mathcal{X}}(\boldsymbol{x}) \triangleq \arg\min_{\widetilde{\boldsymbol{x}} \in \mathcal{X}} \|\widetilde{\boldsymbol{x}} - \boldsymbol{x}\|_2$, for all $\boldsymbol{x} \in \mathbb{R}^N$. Euclidean projections, as defined above, always exist and are nonexpansive operators.

For every real-valued function $f : \mathbb{R}^N \to \mathbb{R}$, which is differentiable at a point $\boldsymbol{x} \in \mathbb{R}^N$, the vector $\nabla f(\boldsymbol{x}) \in \mathbb{R}^N$ denotes its *gradient at* $\boldsymbol{x}$. If, additionally, f is differentiable on $\mathcal{X} \subseteq \mathbb{R}^N$, the function $\nabla f : \mathcal{X} \to \mathbb{R}^N$ denotes its *gradient function*, mapping each $\boldsymbol{x} \in \mathcal{X}$ to $\nabla f(\boldsymbol{x})$.

If f is nonsmooth and convex, its *subdifferential* is the closed-valued multifunction $\partial f : \mathbb{R}^N \rightrightarrows \mathbb{R}^N$, defined, for every $\boldsymbol{x} \in \mathbb{R}^N$, as the set of all gradients each corresponding to a linear underestimator of f, or, in other words,

$$\partial f(\boldsymbol{x}) \triangleq \left\{ \boldsymbol{y}_{\boldsymbol{x}} \in \mathbb{R}^N \middle| f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \boldsymbol{y}_{\boldsymbol{x}}^{\boldsymbol{T}}(\boldsymbol{z} - \boldsymbol{x}), \quad \forall \boldsymbol{z} \in \mathbb{R}^N \right\}, \quad \forall \boldsymbol{x} \in \mathbb{R}^N. \tag{1}$$

A subgradient (function) of f, suggestively denoted as $\underline{\nabla} f : \mathbb{R}^N \to \mathbb{R}^N$, is defined as *any selection* of the subdifferential multifunction ∂f, that is, for every $\boldsymbol{x} \in \mathbb{R}^N$, it is true that $\underline{\nabla} f(\boldsymbol{x}) \in \partial f(\boldsymbol{x})$; for brevity, we write $\underline{\nabla} f \in \partial f$. For fixed $\boldsymbol{x} \in \mathbb{R}^N$, $\underline{\nabla} f(\boldsymbol{x})$ will be called a *subgradient of* f *at* $\boldsymbol{x}$.

2 Problem Setting & Preliminaries

We now formally introduce the problem of interest in this work. Henceforth, all subsequent probabilistic statements will presume the existence of a *common* probability space $(\Omega, \mathscr{F}, \mathcal{P})$. We refer to $(\Omega, \mathscr{F}, \mathcal{P})$ as the *base space*. We place no topological restrictions on the sample space Ω. However, in order for some mild technicalities to be easily resolved, we conveniently assume that $(\Omega, \mathscr{F}, \mathcal{P})$ constitutes a complete measure space.

Let $F : \mathbb{R}^N \times \mathbb{R}^M \rightarrow \mathbb{R}$ be a bivariate real-valued mapping, such that, for every $\boldsymbol{x} \in \mathbb{R}^N$, the function $F(\boldsymbol{x}, \cdot)$ is $\mathscr{B}\left(\mathbb{R}^M\right)$-measurable and, for every $\boldsymbol{w} \in \mathbb{R}^M$, the path $F(\cdot, \boldsymbol{w})$ is *(real-valued) convex (and subdifferentiable)*. Also, for a given $\mathscr{F}$-measurable (in general) random element $\boldsymbol{W} : \Omega \rightarrow \mathbb{R}^M$, consider the composite function $\widetilde{F} : \mathbb{R}^N \times \Omega \rightarrow \mathbb{R}$, defined as

$$\widetilde{F}(\cdot, \omega) \triangleq F(\cdot, \boldsymbol{W}(\omega)), \quad \forall \omega \in \Omega. \tag{2}$$

It easily follows that, for every $\boldsymbol{x} \in \mathbb{R}^N$, the function $\widetilde{F}(\boldsymbol{x}, \cdot) \equiv F(\boldsymbol{x}, \boldsymbol{W}(\cdot))$ is an $\mathscr{F}$-measurable (in general), real-valued random variable. We additionally assume that, for every $\boldsymbol{x} \in \mathbb{R}^N$, $\widetilde{F}(\boldsymbol{x}, \cdot)$ belongs to the Lebesgue space $\mathcal{L}_q$ for some fixed choice of $q \in [1, \infty]$, relative to the base measure $\mathcal{P}$, that is, $\widetilde{F}(\boldsymbol{x}, \cdot) \in \mathcal{L}_q(\Omega, \mathscr{F}, \mathcal{P}; \mathbb{R}) \triangleq \mathcal{Z}_q$. Of course, if, for every $\boldsymbol{x} \in \mathbb{R}^N$, $F(\boldsymbol{x}, \cdot) \in \mathcal{L}_q\left(\mathbb{R}^M, \mathscr{B}\left(\mathbb{R}^M\right), \mathcal{P}_{\boldsymbol{W}}; \mathbb{R}\right)$, where $\mathcal{P}_{\boldsymbol{W}}$ is the Borel pushforward of $\boldsymbol{W}$, then $\widetilde{F}(\boldsymbol{x}, \cdot) \in \mathcal{Z}_q$, as well. Hereafter, $F(\cdot, \boldsymbol{W})$ will be referred to as a *random cost function*.

With the term *risk measure*, we refer to some fixed and known *real-valued functional* on the Banach space $\mathcal{Z}_q$ [Shapiro et al., 2014]. Among all risk measures on $\mathcal{Z}_q$, we pay special attention to those exhibiting the following basic structural characteristics.

Definition 1. (Convex-Monotone Risk Measures) A real valued functional on $\mathcal{Z}_q$, $\rho : \mathcal{Z}_q \rightarrow \mathbb{R}$, is called a *convex-monotone risk measure*, if and only if it satisfies the following conditions:

R1 *(Convexity)*: For every $Z_1 \in \mathcal{Z}_q$ and $Z_2 \in \mathcal{Z}_q$, it is true that

$$\rho(\alpha Z_1 + (1-\alpha) Z_2) \leq \alpha \rho(Z_1) + (1-\alpha) \rho(Z_2), \tag{3}$$

for all $\alpha \in [0, 1]$.

R2 *(Monotonicity)*: For every $Z_1 \in \mathcal{Z}_q$ and $Z_2 \in \mathcal{Z}_q$, such that $Z_1(\omega) \geq Z_2(\omega)$, for $\mathcal{P}$-almost all $\omega \in \Omega$, it is true that $\rho(Z_1) \geq \rho(Z_2)$.

For a possibly convex-monotone risk measure $\rho : \mathcal{Z}_q \rightarrow \mathbb{R}$ (following Assumption 1), we will be interested in the "static" stochastic program

$$\boxed{\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & \rho\left(\widetilde{F}(\boldsymbol{x}, \cdot)\right) \equiv \rho(F(\boldsymbol{x}, \boldsymbol{W})) \triangleq \phi^{\widetilde{F}}(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{X} \end{array}}, \tag{4}$$

where the set of feasible decisions $\mathcal{X} \subseteq \mathbb{R}^N$ is assumed to be closed and convex.

Under the standard problem setting outlined above, it is straightforward to formulate the following elementary result, provided here without proof, and for completeness.

Proposition 1. (Convexity of Risk-Function Compositions [Shapiro et al., 2014]) *Consider a real-valued random function $f : \mathbb{R}^N \times \Omega \rightarrow \mathbb{R}$, as well as a real-valued risk measure $\rho : \mathcal{Z}_q \rightarrow \mathbb{R}$. Suppose that, for every $\omega \in \Omega$, $f(\cdot, \omega)$ is convex and that ρ is convex-monotone. Then, the real-valued composite function $\phi^f(\cdot) \equiv \rho(f(\cdot, \bullet)) : \mathbb{R}^N \rightarrow \mathbb{R}$ is convex.*

Proposition 1 shows that, under the respective assumptions, (4) constitutes a convex mathematical program in standard form. Thus, application of a subgradient method would require that some selection of the subdifferential multifunction $\partial \phi^{\widetilde{F}}$ can be evaluated at will, at any $\boldsymbol{x} \in \mathcal{X}$.

However, for most choices of the random cost function $F(\cdot, \boldsymbol{W})$ and of the risk measure ρ, even the composition $\phi^{\widetilde{F}}(\cdot) \equiv \rho(F(\cdot, \boldsymbol{W}))$ is impossible to be evaluated exactly, let alone (a selection of) $\partial\phi^{\widetilde{F}}$. Instead, we may be given either realizations of the random exogenous information $\boldsymbol{W}$, or direct evaluations of $F(\cdot, \boldsymbol{W})$ and a subgradient, $\underline{\nabla} F(\cdot, \boldsymbol{W})$, at some *test decision candidate* $\boldsymbol{x}$. It might also be desirable that decision making is performed *sequentially over time*, where decisions are updated adaptively as new information arrives. Such settings motivate the consideration of SSD-type algorithms for solving (4), which are of main interest in this paper.

Some basic assumptions follow, fairly standard in the literature of stochastic approximation [Shapiro et al., 2014, Wang et al., 2017, Yang et al., 2018, Kushner and Yin, 2003]. To this end, let us formally introduce the elementary concept of a *IID process*. Then, our assumptions follow.

Definition 2. (IID Process) A stochastic sequence $\{\boldsymbol{W}^n\}_{n\in\mathbb{N}^+}$ is called *IID* if and only if it consists of statistically independent, $\mathbb{R}^M$-valued random elements, identically distributed according to a fixed Borel measure $\mathcal{P}_{\boldsymbol{W}}$.

Assumption 1. (Availability of Information) ***Either one, or more,*** *mutually independent, IID sequences are available* ***sequentially****, all distributed according to* $\mathcal{P}_{\boldsymbol{W}}$.

Remark 1. Note that in Assumption 1 we *do not* require that the process $\boldsymbol{W}^n$ is actually *observable* to the user, but only *available*, either in the form of a data stream, or by simulation. ■

Assumption 2. (Existence of an $\mathcal{SO}$) *There exists a mechanism, called a* ***Sampling Oracle*** *($\mathcal{SO}$), which, given $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{w} \in \mathbb{R}^M$, returns either $F(\boldsymbol{x}, \boldsymbol{w})$, or $\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{w})$, a subgradient of F relative to $\boldsymbol{x}$, or both. It is further assumed that the $\mathcal{SO}$ has* ***direct access*** *to all available information streams, according to Assumption 1.*

In this work, we propose and analyze efficient algorithms for solving (4) under Assumptions 1 and 2, and *explicitly assuming no prior knowledge* of either the random cost function $F(\cdot, \boldsymbol{W})$, or its respective subgradients. We will be restricting our attention to a new class of convex-monotone risk measures with, however, wide applicability, and whose general structure follows the so-called *Mean-Risk Model* ([Shapiro et al., 2014], Section 6.2). This special class of risk measures is introduced and analyzed, in detail, in Section 3.

3 Mean-Semideviation Models

Under the Mean-Risk Model paradigm [Shapiro et al., 2014], a risk measure $\rho : \mathcal{Z}_q \to \mathbb{R}$ is defined, for each random cost $Z \in \mathcal{Z}_q$, as

$$\rho(Z) \triangleq \mathbb{E}\{Z\} + c\mathbb{D}\{Z\}, \tag{5}$$

where the functional $\mathbb{D} : \mathcal{Z}_q \to \mathbb{R}$ constitutes a *dispersion measure*, and provided that the respective quantities are well defined, for the particular choice of $q \in [1, \infty]$. The dispersion measure $\mathbb{D}$ may be conveniently thought as a *penalty*, weighted by the penalty multiplier $c \geq 0$, *effectively quantifying the uncertainty* of the particular cost Z.

In this section, we introduce a special class of dispersion measures, which constitute natural generalizations of the well-known upper semideviation of order p [Shapiro et al., 2014]. This new class of dispersion measures is termed here as *generalized semideviations*. Reasonably enough, risk measures of the form of (5), where the respective dispersion measure constitutes a generalized semideviation will be called either *mean-semideviation risk measures*, or, interchangeably, *mean-semideviation models*, or, simply, *mean-semideviations.*

This section is structured as follows. First, the simple notion of a *risk regularizer* is introduced; risk regularizers constitute the basic building block of generalized semideviations. The basic properties of risk regularizers are concisely presented, and a formal definition of generalized semideviations is also formulated, along with a brief discussion related to their practical relevance. Mean-semideviation risk measures are then formally introduced, along with their basic properties, and specific examples are discussed, highlighting their versatility. Next, we develop a constructive characterization result, essentially showing that the class of all mean-semideviation risk measures is *almost in one-to-one correspondence* with the class of cumulative distribution functions of all integrable random variables (on the line). This result readily demonstrates an apparent generality of mean-semideviations, as well. Lastly, the usefulness, flexibility and effectiveness of mean-semideviation risk measures are demonstrated on a classical, chance-constrained newsvendor model. In particular, risk regularizers (each inducing a mean-semideviation risk measure) are put *in context*, and their construction is explicitly discussed, reflecting the special characteristics of the specific newsvendor problem under consideration, and the objectives of the decision maker.

3.1 Basic Concepts

We start by introducing the concept of a *risk regularizer.* Risk regularizers are simple, real-valued functions of one variable, which are reasonably structured, so that they, on the one hand, can be used to quantify risk (see below) and, on the other, can result in problems which can be solved efficiently and exactly via convex stochastic optimization.

Definition 3. (Risk Regularizers) A real-valued function $\mathcal{R}:\mathbb{R}\rightarrow\mathbb{R}$ is called a *risk regularizer*, if it satisfies the following conditions:

S1 $\mathcal{R}$ is *convex.*

S2 $\mathcal{R}$ is *nonnegative.*

S3 $\mathcal{R}$ is *nondecreasing.*

S4 For every $\alpha\geq 0$, it is true that $\mathcal{R}(x+\alpha)\leq\mathcal{R}(x)+\alpha$, for all $x\in\mathbb{R}$.

Fig. 3.1 illustrates the shapes of various risk regularizers, other than the arguably most obvious example of the positive part function $(\cdot)_+$. Note that a risk regularizer need not be smooth (a trivial example is $(\cdot)_+$); several of the examples of Fig. 3.1 are indeed nonsmooth, with the respective corner points highlighted by black dots.

Risk regularizers of Definition 3 may be further structurally characterized via the following simple result.

Proposition 2. (Characterization of $\mathcal{R}$) *Consider a real-valued function $\mathcal{R}:\mathbb{R}\rightarrow\mathbb{R}$, satisfying condition* **S3** *of Definition 4. Then, condition* **S4** *holds if and only if $\mathcal{R}$ is nonexpansive.*

Proof of Proposition 2. First, assume that condition **S4** holds. Then, by the fact that $\mathcal{R}$ is nondecreasing (**S3**), it is true that

$$|\mathcal{R}(x)-\mathcal{R}(y)|\equiv(\mathcal{R}(x)-\mathcal{R}(y))\mathbb{1}_{\{x\geq y\}}+(\mathcal{R}(y)-\mathcal{R}(x))\mathbb{1}_{\{x<y\}}$$

$$\equiv (\mathcal{R}(y+(x-y))-\mathcal{R}(y))\mathbb{1}_{\{x\geq y\}}+(\mathcal{R}(x+(y-x))-\mathcal{R}(x))\mathbb{1}_{\{x<y\}}$$
$$\leq (x-y)\mathbb{1}_{\{x\geq y\}}+(y-x)\mathbb{1}_{\{x<y\}}\equiv |x-y|, \tag{6}$$

for all $(x,y)\in\mathbb{R}^2$, showing that $\mathcal{R}$ is a nonexpansive map. *Conversely*, assume that $\mathcal{R}$ is nonexpansive. Then, for any $\alpha\geq 0$, it is true that

$$\begin{aligned} 0\leq \mathcal{R}(x+\alpha)-\mathcal{R}(x) &\equiv |\mathcal{R}(x+\alpha)-\mathcal{R}(x)| \\ &\leq |x+\alpha-x|\equiv\alpha, \end{aligned} \tag{7}$$

for all $x\in\mathbb{R}$, verifying condition **S4**. ■

At this point, let us emphasize the elementary fact that, because of convexity, every (real-valued) risk-regularizer must also be *differentiable almost everywhere*, relative to the Lebesgue measure on the Borel space $(\mathbb{R},\mathscr{B}(\mathbb{R}))$. This also follows either by monotonicity, or due to the fact that a risk regularizer is nonexpansive and, therefore, Lipschitz continuous on $\mathbb{R}$. Further, because of convexity, the set of Lebesgue measure zero of points in $\mathbb{R}$, where a risk regularizer is *nondifferentiable*, is *at most countable*.

The class all possible risk regularizers *induces* that of *generalized semideviations*, which constitute the class of dispersion measures considered in this paper. The definition of a generalized semideviation is presented below.

Definition 4. (Generalized Semideviations) Fix $p\in[1,\infty)$ and choose a risk regularizer $\mathcal{R}:\mathbb{R}\to\mathbb{R}$. A dispersion measure $\mathbb{D}_p^{\mathcal{R}}:\mathcal{Z}_q\to\mathbb{R}$ is called a *generalized semideviation of order p,* if and only if, for $Z\in\mathcal{Z}_q$,

$$\mathbb{D}_p^{\mathcal{R}}\{Z\}\triangleq\left(\mathbb{E}\{(\mathcal{R}(Z-\mathbb{E}\{Z\}))^p\}\right)^{1/p}\equiv\|\mathcal{R}(Z-\mathbb{E}\{Z\})\|_{\mathcal{L}_p}, \tag{8}$$

where it is assumed that all involved quantities are well defined and finite.

The power of generalized semideviations is in the fact that they form a parametric family relative to the choice of the risk regularizer $\mathcal{R}$; different risk regularizers correspond to different rules for ranking the relative effect of both riskier (higher than the mean) and less risky (lower than the mean) events, corresponding to specific regions in the range of the cost. For more details, see Section 3, where we illustrate the versatility of generalized semideviations via additional examples, considering various specific choices for $\mathcal{R}$, with the well known upper-semideviation dispersion measure [Shapiro et al., 2014] being the prototypical representative of this class.

3.2 Mean-Semideviations: Definition, Existence & Structure

Utilizing the concept of generalized semideviations, we may now introduce the class of risk measures of central interest in this work, as follows.

Definition 5. (Mean-Semideviation Risk Measures) Fix $p\in[1,\infty)$ and choose a risk regularizer $\mathcal{R}:\mathbb{R}\to\mathbb{R}$. The *mean-semideviation of order p, induced by $\mathcal{R}$,* or $MS_p^{\mathcal{R}}$, for short, is the real-valued risk measure defined, for $Z\in\mathcal{Z}_q$, as[1]

$$\rho(Z)\equiv\rho_p^{\mathcal{R}}(Z;c)\triangleq\mathbb{E}\{Z\}+c\mathbb{D}_p^{\mathcal{R}}\{Z\}, \tag{9}$$

[1] A mean-semideviation risk measure will be denoted either as $\rho_p^{\mathcal{R}}(Z;c)$, which is proper, or $\rho(Z)$, which is simpler, as long as the choices of $p,\mathcal{R}$ and c are clearly specified.

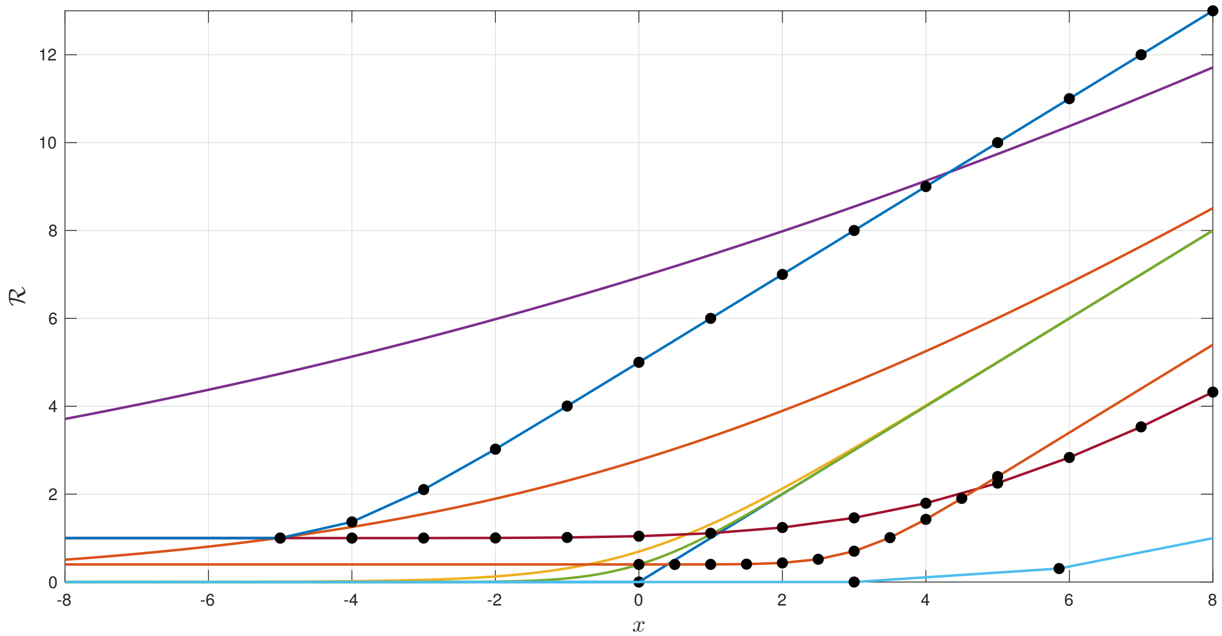

Figure 3.1: Some examples of both smooth and nonsmooth risk regularizers. Black dots highlight the respective corner points of nondifferentiability (some imperceptible).

where $c \geq 0$ constitutes a fixed penalty multiplier, and provided that all involved quantities are well defined and finite.

Next, we state and prove a small number of relatively simple results, related to the existence of mean-semideviation risk measures, introduced in Definition 5, as well as their functional structure. First, as it might be expected, we show that mean-semideviation risk measures of order p may be naturally associated with costs which are also in $\mathcal{L}_p$ (i.e., choosing $p \equiv q$). Recall that, throughout the paper, p is reserved for specifying the order of the mean-semideviation risk measure under consideration, whereas q is related to the integrability of the respective cost.

Proposition 3. (Compatibility of p's and q's) *Fix $p \in [1, \infty)$, $c \geq 0$, and choose any risk regularizer $\mathcal{R} : \mathbb{R} \to \mathbb{R}$. Then, as long as $q \geq p$, the $MS_p^{\mathcal{R}}$ risk measure $\rho_p^{\mathcal{R}}(\cdot; c)$ is well-defined and finite, for every $Z \in \mathcal{Z}_q$.*

Proof of Proposition 3. Since $q \geq p \geq 1$, it is trivial that $Z \in \mathcal{Z}_1$, simply due to the inclusion $\mathcal{Z}_1 \supset \mathcal{Z}_2 \supset \ldots$, for any choice of q. Thus, the expectation of every $Z \in \mathcal{Z}_q$ exists and is finite, and what remains is to prove the result for the dispersion measure $\mathbb{D}_p^{\mathcal{R}}$.

For simplicity, let $q \equiv p$. Using the fact that $\mathbb{E}\{Z\}$ is finite, it is true that, for every $Z \in \mathcal{Z}_p$, the shifted cost $Z - \mathbb{E}\{Z\}$ is in $\mathcal{Z}_p$. It thus suffices to show that, for every $Z - \mathbb{E}\{Z\} \triangleq X \in \mathcal{Z}_p$, $\mathcal{R}(X)$ is in $\mathcal{Z}_p$, as well. Because the risk regularizer $\mathcal{R}$ is nonnegative (condition **S2**), the integral $\mathbb{E}\{(\mathcal{R}(X))^p\}$ exists. Also, due to condition **S4** of Definition 3, it follows that, for every $x \geq 0$, $\mathcal{R}(x) \leq \mathcal{R}(0) + x$, and since $\mathcal{R}$ is nondecreasing (**S3**), it is true that $\mathcal{R}(x) \leq \mathcal{R}(0) + |x|$, for all $x \in \mathbb{R}$. Setting $x \equiv X$, this yields

$$0 \leq \mathcal{R}(X) \leq \mathcal{R}(0) + |X|, \tag{10}$$

and since $X \in \mathcal{Z}_p$, $\mathcal{R}(0) + |X| \in \mathcal{Z}_p$, as well. Consequently, it is true that

$$(\mathbb{E}\{(\mathcal{R}(X))^p\})^{1/p} \leq (\mathbb{E}\{(\mathcal{R}(0) + |X|)^p\})^{1/p} < +\infty, \tag{11}$$

showing that $\mathbb{D}_p^{\mathcal{R}}$ and, therefore, $\rho_p^{\mathcal{R}}(\cdot; c)$, are both well defined and finite, for every $Z \in \mathcal{Z}_p$.

Now, due to the inclusion $\mathcal{Z}_1 \supset \mathcal{Z}_2 \supset \ldots$, we know that, if $Z \in \mathcal{Z}_q$, for some $q \geq p$, then $Z \in \mathcal{Z}_p$, as well. Enough said. ■

Hereafter, for the sake of generality, *we will implicitly assume that p and q are compatible*, so that existence and finiteness of the resulting risk measures considered is ensured. Of course, in actual applications, Proposition 3 may be directly invoked on a case-by-case basis, in order to select the order of the particular dispersion measure of choice, depending on the nature of the random cost, or a family of those, under study.

After characterizing existence and finiteness of mean-semideviation risk measures, as introduced in Definition 5, we focus on their structural properties, from a functional point of view. As the following result suggests, mean-semideviation risk measures are indeed convex-monotone under a standardized assumption on the penalty multiplier c.

Theorem 1. (When are Mean-Semideviations Convex-Monotone?) *Fix $p \in [1, \infty)$ and choose any risk regularizer $\mathcal{R} : \mathbb{R} \to \mathbb{R}$. Then, as long as $c \in [0, 1]$, the $MS_p^{\mathcal{R}}$ risk measure $\rho_p^{\mathcal{R}}(\cdot; c)$ is convex-monotone; that is, it satisfies both conditions* **R1** *and* **R2**.

Proof of Theorem 1. Let us start with verifying convexity (**R1**). Since the expectation term of $\rho_p^{\mathcal{R}}(\cdot; c)$ is a linear functional on $\mathcal{Z}_q$, it will suffice to show that the generalized semideviation term $\mathbb{D}_p^{\mathcal{R}}$ is convex. Indeed, for every $Z_1 \in \mathcal{Z}_q$, $Z_2 \in \mathcal{Z}_q$ and every $\alpha \in [0, 1]$, we may write

$$\begin{aligned}
\mathbb{D}_p^{\mathcal{R}}\{\alpha Z_1 + (1-\alpha) Z_2\} &\equiv \|\mathcal{R}(\alpha Z_1 + (1-\alpha) Z_2 - \mathbb{E}\{\alpha Z_1 + (1-\alpha) Z_2\})\|_{\mathcal{L}_p} \\
&\equiv \|\mathcal{R}(\alpha (Z_1 - \mathbb{E}\{Z_1\}) + (1-\alpha)(Z_2 - \mathbb{E}\{Z_2\}))\|_{\mathcal{L}_p} \\
&\leq \|\alpha \mathcal{R}((Z_1 - \mathbb{E}\{Z_1\})) + (1-\alpha) \mathcal{R}(Z_2 - \mathbb{E}\{Z_2\})\|_{\mathcal{L}_p} \\
&\leq \alpha \|\mathcal{R}((Z_1 - \mathbb{E}\{Z_1\}))\|_{\mathcal{L}_p} + (1-\alpha) \|\mathcal{R}(Z_2 - \mathbb{E}\{Z_2\})\|_{\mathcal{L}_p} \\
&\equiv \alpha \mathbb{D}_p^{\mathcal{R}}\{Z_1\} + (1-\alpha) \mathbb{D}_p^{\mathcal{R}}\{Z_2\},
\end{aligned} \tag{12}$$

where the first inequality is true due to conditions **S1** (convexity) and **S2** (nonnegativity), and the second is due to the triangle (Minkowski) inequality. Thus, $\mathbb{D}_p^{\mathcal{R}}$ is a convex functional, which means that $\rho_p^{\mathcal{R}}(\cdot; c)$ is also convex on $\mathcal{Z}_q$. Note that the value of $c \geq 0$ is not crucial in order to show convexity of $\rho_p^{\mathcal{R}}(\cdot; c)$.

Let us now study monotonicity (**R2**) of the risk measure $\rho_p^{\mathcal{R}}(\cdot; c)$. For every $Z_1 \in \mathcal{Z}_q$ and $Z_2 \in \mathcal{Z}_q$, such that $Z_1(\omega) \geq Z_2(\omega)$, for $\mathcal{P}$-almost all $\omega \in \Omega$, we have

$$\begin{aligned}
\rho_p^{\mathcal{R}}(Z_2; c) &\equiv \mathbb{E}\{Z_2\} + c \|\mathcal{R}(Z_2 - \mathbb{E}\{Z_2\})\|_{\mathcal{L}_p} \\
&\leq \mathbb{E}\{Z_2\} + c \|\mathcal{R}(Z_1 - \mathbb{E}\{Z_2\})\|_{\mathcal{L}_p} \\
&\equiv \mathbb{E}\{Z_2\} + c \|\mathcal{R}(Z_1 - \mathbb{E}\{Z_1\} + \mathbb{E}\{Z_1\} - \mathbb{E}\{Z_2\})\|_{\mathcal{L}_p} \\
&\leq \mathbb{E}\{Z_2\} + c \|\mathcal{R}(Z_1 - \mathbb{E}\{Z_1\}) + \mathbb{E}\{Z_1\} - \mathbb{E}\{Z_2\}\|_{\mathcal{L}_p}
\end{aligned}$$

$$\leq \mathbb{E}\{Z_2\} + c(\mathbb{E}\{Z_1\} - \mathbb{E}\{Z_2\}) + c\|\mathcal{R}(Z_1 - \mathbb{E}\{Z_1\})\|_{\mathcal{L}_p}, \tag{13}$$

where the first inequality is due to conditions **S2** (nonnegativity) and **S3** (monotonicity), the second is due to conditions **S2** (nonnegativity), **S4** (nonexpansiveness), as well as the fact that $\mathbb{E}\{Z_1\} \geq \mathbb{E}\{Z_2\}$, and the third is again due to the triangle inequality. From (13), we readily see that, as long as $c \in [0,1]$, we may further write

$$\rho_p^{\mathcal{R}}(Z_2; c) \leq \mathbb{E}\{Z_1\} + c\|\mathcal{R}(Z_1 - \mathbb{E}\{Z_1\})\|_{\mathcal{L}_p} \equiv \rho_p^{\mathcal{R}}(Z_1; c), \tag{14}$$

completing the proof of the theorem. ■

We may now invoke Proposition 1, presented earlier, to immediately obtain the following key corollary. The proof is trivial and, thus, omitted.

Corollary 1. (When is (4) Convex?) *Fix $p \in [1,\infty)$ and choose any risk regularizer $\mathcal{R} : \mathbb{R} \to \mathbb{R}$. Then, as long as $c \in [0,1]$, the composite function $\phi^{\widetilde{F}}(\cdot) \equiv \rho_p^{\mathcal{R}}(F(\cdot, \boldsymbol{W}); c) \equiv \rho(F(\cdot, \boldsymbol{W}))$ is convex on $\mathbb{R}^N$, and (4) constitutes a convex stochastic program.*

Corollary 1 is an important result, because it shows that, for every mean-semideviation risk measure, or equivalently, for every risk regularizer of choice, problem (4) would be exactly solvable via, for instance, subgradient methods, if the function $\phi^{\widetilde{F}}$ was known in advance. This fact reinforces our hope that it might indeed be possible to solve (4) to optimality, utilizing some carefully designed *stochastic search*, or, more specifically, and based on the assumed subdifferentiability of $F(\cdot, \boldsymbol{W})$, *stochastic subgradient* algorithm. Of course, such an algorithm should be designed to work under Assumptions 1 and 2, *without* the need for explicit knowledge of $F(\cdot, \boldsymbol{W})$, or $\underline{\nabla} F(\cdot, \boldsymbol{W})$.

Remark 2. ***(Coherence?)*** We should mention that mean-semideviations *are not coherent risk measures* ([Shapiro et al., 2014], Section 6.3), since they do not satisfy the axiomatic property of positive homogeneity. This is simply due to the fact that, in general, one may find choices for $\mathcal{R}$ such that

$$\|\mathcal{R}(tZ - \mathbb{E}\{tZ\})\|_{\mathcal{L}_p} \neq t\|\mathcal{R}(Z - \mathbb{E}\{Z\})\|_{\mathcal{L}_p}, \tag{15}$$

for some $t > 0$ and $Z \in \mathcal{Z}_q$. Nevertheless, mean-semideviations may be readily shown to satisfy translation equivariance, although such property is not explicitly required in this work. As a result, except for being convex-monotone, mean-semideviations also belong to the class of *convex risk measures* [Föllmer and Schied, 2002, Shapiro et al., 2014]. ■

3.3 Examples of Mean-Semideviation Models

Before moving on, it would be instructive to discuss some examples of mean-semideviations, highlighting the versatility of this particular class of risk measures. We start from simple, illustrative choices as far as the involved risk regularizer is concerned, and then we generalize.

3.3.1 Mean-*Upper*-Semideviations

The simplest, *prototypical* example of a mean-semideviation risk measure is the *mean-upper-semideviation* of order p ([Shapiro et al., 2014], Sections 6.2.2 & 6.3.2), which is constructed by choosing as risk regularizer the function

$$\mathcal{R}(x) \triangleq (x)_+ \triangleq \max\{x, 0\}, \quad x \in \mathbb{R}, \tag{16}$$

yielding the risk measure

$$\begin{aligned}\rho(Z) &\equiv \mathbb{E}\{Z\} + c\left(\mathbb{E}\left\{\left((Z-\mathbb{E}\{Z\})_+\right)^p\right\}\right)^{1/p} \\ &\equiv \mathbb{E}\{Z\} + c\left\|(Z-\mathbb{E}\{Z\})_+\right\|_{\mathcal{L}_p}, \end{aligned} \tag{17}$$

for $Z \in \mathcal{Z}_q$. Of course, in this case, it is trivial to show that $\mathcal{R}$ satisfies conditions **S1-S4** of Definition 3. Recall that we have assumed that q is appropriately chosen, such that ρ is a well defined, real-valued functional on $\mathcal{Z}_q$.

3.3.2 *Entropic* Mean-Semideviations

Our second example is a generalization of the mean-upper-semideviation risk measure discussed in the previous example. Here, the risk regularizer $\mathcal{R}$ is chosen itself from a parametric family, as

$$\mathcal{R}(x;t) \triangleq \frac{1}{t}\log(1+\exp(tx)), \quad (x,t) \in \mathbb{R}\times\mathbb{R}_{++}, \tag{18}$$

where t is a parameter, regulating the sharpness of the function at zero. It is trivial to verify conditions **S2** (nonnegativity) and **S3** (monotonicity). Also, for fixed t, the first derivative of $\mathcal{R}$ relative to x is the logistic function

$$\frac{\partial\mathcal{R}}{\partial x}(x;t) \equiv \frac{\exp(tx)}{1+\exp(tx)} \in (0,1), \quad \forall x \in \mathbb{R}, \tag{19}$$

showing that $\mathcal{R}$ is a contraction mapping, immediately verifying condition **S4** (nonexpansiveness), via Proposition 2. Likewise, the second derivative of $\mathcal{R}$ is given by

$$\frac{\partial^2\mathcal{R}}{\partial x^2}(x;t) \equiv \frac{t\exp(tx)}{(1+\exp(tx))^2} > 0, \quad \forall x \in \mathbb{R}, \tag{20}$$

and, thus, **S1** (convexity) is readily verified, as well. Hence, $\mathcal{R}$ is a valid risk regularizer. Alternatively and to illustrate the procedure, we may verify condition **S4** directly; for fixed $t > 0$, for every $\alpha \geq 0$ and for every $x \in \mathbb{R}$, we may write

$$\begin{aligned}\mathcal{R}(x+a;t) &\equiv \frac{1}{t}\log(1+\exp(t(x+\alpha))) \\ &\equiv \frac{1}{t}\log\left(\frac{1}{\exp(t\alpha)}+\exp(tx)\right)+\alpha \\ &\leq \frac{1}{t}\log(1+\exp(tx))+\alpha \\ &\equiv \mathcal{R}(x;t)+\alpha, \end{aligned} \tag{21}$$

where the inequality is due to the fact that $t\alpha \geq 0$. It is also easy to see that, for every $x \in \mathbb{R}$, $\mathcal{R}(x;t) \underset{t\to\infty}{\longrightarrow} (x)_+$, showing that $\mathcal{R}(\cdot;t)$ constitutes a smooth approximation to the risk regularizer of the mean-upper-semideviation risk measure discussed previously.

The resulting risk measure is called an *entropic mean-semideviation* of order p, and may be expressed as

$$\rho(Z) \equiv \mathbb{E}\{Z\} + \frac{c}{t}\left\|\log(1+\exp(t(Z-\mathbb{E}\{Z\})))\right\|_{\mathcal{L}_p}, \tag{22}$$

for $Z \in \mathcal{Z}_q$. For obvious reasons, this risk measure may be considered a *soft* version of the mean-upper-semideviation risk measure.

3.3.3 *CDF-Antiderivative (CDFA)* Mean-Semideviations

We now show that, in fact, both previously presented examples are special cases of a much more general approach, which may be utilized for the *construction* of risk regularizers. To this end, let $Y : \Omega \to \mathbb{R}$ be a random variable in $\mathcal{Z}_1$, with cumulative distribution function (cdf) F_Y. Consider the choice

$$\mathcal{R}(x) \triangleq \int_{-\infty}^{x} F_Y(y)\,\mathrm{d}y, \quad x \in \mathbb{R}, \tag{23}$$

where, because F_Y is a nonnegative Borel measurable function, the involved integration is *always* well-defined (might be $+\infty$, though), in the sense of Lebesgue. The particular *antiderivative of the cdf* F_Y, as defined in (23), constitutes a very important quantity in the theory of stochastic dominance; see, for instance, related articles [Ogryczak and Ruszczyński, 1999] and [Ogryczak and Ruszczyński, 2002] for definition and insights. In particular, via Fubini's Theorem (Theorem 2.6.6 in [Ash and Doléans-Dade, 2000]), $\mathcal{R}$ may be easily shown to admit the alternative integral representation

$$\mathcal{R}(x) \equiv \mathbb{E}\left\{(x - Y)_+\right\}, \quad \forall x \in \mathbb{R}. \tag{24}$$

Exploiting the assumption that $Y \in \mathcal{Z}_1$, it follows that $\mathcal{R}(x) < +\infty$, for every $x \in \mathbb{R}$. Also, from (24), it is trivial to see that, because of the structure of the function $(\cdot)_+$, $\mathcal{R}$ is convex (**S1**), nonnegative (**S2**) and nondecreasing (**S3**) on $\mathbb{R}$. Nonexpansiveness (**S4**) may also be readily verified.

Consequently, $\mathcal{R}$ is a valid risk regularizer, and the resulting risk measure, called a *CDF-Antiderivative (CDFA) mean-semideviation*, may be expressed in various forms as

$$\begin{aligned}
\rho(Z) &\equiv \mathbb{E}\{Z\} + c\left\| \int_{-\infty}^{Z-\mathbb{E}\{Z\}} F_Y(y)\,\mathrm{d}y \right\|_{\mathcal{L}_p} \\
&\equiv \mathbb{E}\{Z\} + c\left\| \mathbb{E}\left\{(x - Y)_+\right\}\big|_{x \equiv Z - \mathbb{E}\{Z\}} \right\|_{\mathcal{L}_p} \\
&\equiv \mathbb{E}\{Z\} + c\left\| \int_{\mathbb{R}} ([Z - \mathbb{E}\{Z\}] - y)_+ \,\mathrm{d}\mathcal{P}_Y(y) \right\|_{\mathcal{L}_p} \\
&\equiv \mathbb{E}\{Z\} + c\left\| \mathbb{E}\left\{ ([Z - \mathbb{E}\{Z\}] - Y)_+ \middle| Z \right\} \right\|_{\mathcal{L}_p},
\end{aligned} \tag{25}$$

for $Z \in \mathcal{Z}_q$, where Y can be arbitrarily taken to be *independent of* Z and $\mathcal{P}_Y$ denotes the Borel pushforward of Y.

We may now verify that both mean-upper-semideviation and entropic mean-semideviation risk measures discussed above are special cases of CDFA mean-semideviations. In mean-upper-semideviations, the respective risk regularizer is an antiderivative (taken piecewise) of the cdf corresponding to the Dirac measure at zero. In entropic mean-semideviations, the respective risk regularizer is an antiderivative of (19) (by monotone convergence and via a sequential argument), which is the cdf of a zero-mean element in $\mathcal{Z}_1$. In both cases, the antiderivatives involved are of the form of (23).

3.3.3.1 Special Case: *Gaussian Antiderivative (GA)* Mean-Semideviations An interesting subclass of CDFA mean-semideviations is the one resulting from taking antiderivatives of the cdf of a standard Gaussian random variable $Y \sim \mathcal{N}(0,1)$. In this case, the simplest possible risk regularizer may be constructed as

$$\mathcal{R}(x) \triangleq \int_{-\infty}^{x} \Phi(y)\,\mathrm{d}y \equiv x\Phi(x) + \varphi(x), \quad x \in \mathbb{R}, \tag{26}$$

where $\Phi : \mathbb{R} \to [0,1]$ and $\varphi : \mathbb{R} \to \mathbb{R}$ denote the standard Gaussian cdf and density, respectively. This particular antiderivative of Φ appears naturally in standard treatments of the so-called ranking-&-selection, or best arm identification problem and, more specifically, in lookahead selection policies, such as the Knowledge Gradient and the Expected Improvement [Frazier et al., 2008, Ryzhov, 2016].

The resulting mean-semideviation risk measure is called a *Gaussian Antiderivative (GA) mean-semideviation* of order p, and may be expressed as

$$\rho(Z) \equiv \mathbb{E}\{Z\} + c \left\| (Z - \mathbb{E}\{Z\}) \Phi(Z - \mathbb{E}\{Z\}) + \varphi(Z - \mathbb{E}\{Z\}) \right\|_{\mathcal{L}_p}, \tag{27}$$

for $Z \in \mathcal{Z}_q$. Of course, as it happens for all mean semideviations, the functional ρ, as defined in (27), is a convex risk measure for every $c \geq 0$, and a convex-monotone risk measure, if $c \in [0,1]$.

3.4 A Complete Characterization of Mean-Semideviations

As a result of the discussion in Section 3.3.3 above, it follows that risk regularizers may be formed by taking antiderivatives of the cdf of *any integrable* random variable of choice, resulting in a vast variety of mean-semideviation risk measures, all sharing a common favorable structure.

Here, we show that if we start from a *given* risk regularizer $\mathcal{R}$, *the converse statement is also true.* In this respect, we state and prove the following important result.

Theorem 2. (CDF-Based Representation of Risk Regularizers) *Let $Y : \Omega \to \mathbb{R}$ be a random variable, such that, for every $x \in \mathbb{R}$, $\mathbb{E}\left\{(x-Y)_+\right\} < +\infty$, and let $F_Y : \mathbb{R} \to [0,1]$ denote its cdf. Then, for any fixed $0 \leq C_S \leq 1$ and $C_I \geq 0$, the function $\mathcal{R} : \mathbb{R} \to \mathbb{R}$ defined as*

$$\mathcal{R}(x) \triangleq C_S \int_{-\infty}^{x} F_Y(y)\, dy + C_I, \quad \forall x \in \mathbb{R}, \tag{28}$$

is a valid risk regularizer, where integration may be interpreted either in the improper Riemann sense (for computation), or in the standard sense of Lebesgue (for derivation).

Conversely, let $\mathcal{R} : \mathbb{R} \to \mathbb{R}$ be any risk regularizer. Then, there exist some random variable $Y : \Omega \to \mathbb{R}$, satisfying $\mathbb{E}\left\{(x-Y)_+\right\} < +\infty$, for all $x \in \mathbb{R}$, with cdf $F_Y : \mathbb{R} \to [0,1]$, and constants $0 \leq C_S \leq 1$ and $C_I \geq 0$, such that, for every $x \in \mathbb{R}$, the representation (28) is valid. In particular, if $\mathcal{R}'_+ : \mathbb{R} \to \mathbb{R}$ denotes the right derivative of $\mathcal{R}$, it is always true that $C_S \equiv \sup_{x\in\mathbb{R}} \mathcal{R}'_+(x)$, $C_I \equiv \inf_{x\in\mathbb{R}} \mathcal{R}(x)$, and, as long as $\mathcal{R}$ is nonconstant, it holds that $C_S \neq 0$, and F_Y is given by $F_Y(x) \equiv C_S^{-1} \mathcal{R}'_+(x)$, for all $x \in \mathbb{R}$.

Proof of Theorem 2. See Section 7.1 (Appendix). ∎

Theorem 2 is important for two main reasons, the first being related to the forward statement, and the second to the converse. On the one hand, Theorem 2 provides us with the clean, very versatile and analytically friendly integral formula (28) for constructing risk regularizers of various shapes and types. On the other hand, it informs us that, *necessarily*, any risk regularizer can be expressed in the form of (28) and, as a result, *all possible* risk regularizers may be constructed utilizing (28), each time for some suitably chosen cdf. Therefore, risk regularizers are completely characterized by the cdf-based representation of Theorem 2.

Of course, every risk regularizer induces a unique mean-semideviation risk measure. But also notice that, trivially, every mean-semideviation risk measure corresponds to a uniquely specified risk regularizer (as a functional, or when *all* costs in the corresponding $\mathcal{L}_p$-space $\mathcal{Z}_p$ -the *largest* such space, for the *smallest possible* q- are considered). Therefore, Theorem 2 provides a complete characterization of the whole class of mean-semideviation risk measures. In particular, Theorem 2 implies that the class of all mean-semideviation risk measures is ***almost*** *in one-to-one correspondence* with the class of cdfs of all integrable $\mathbb{R}$-valued random elements. The "almost" in the preceding statement is due to the presence of constants C_S and C_I in Theorem 2, and that actually slightly less is required than (absolute) integrability of the involved random variable Y.

3.5 Practical Illustration of Mean-Semideviation Models

We conclude this section by briefly outlining the relevance of mean-semideviation models in applications, also putting our proposed risk regularizers in context. More specifically, we consider a chance-constrained version of the prototypical, single-product newsvendor problem (see, for instance, Chapter 1 in [Shapiro et al., 2014]), upon which we are based in order to formulate a *doubly risk-averse* newsvendor problem, which jointly controls *both* unmet demand *and* holding costs. We also explicitly demonstrate how the respective risk regularizer may be potentially designed, based on the characteristics of the particular problem under consideration.

Although the single-product newsvendor problem (and its variations) indeed constitutes a one-dimensional, toy example, it provides insights and highlights some important features of the mean-semideviation risk measures advocated herein. Additionally, the simplicity of such a problem facilitates numerical solution, and enables us to present some numerical results, verifying the effectiveness of the proposed mean-semideviation risk measures experimentally, as well.

3.5.1 A Chance-Constrained Single-Product NewsVendor

Suppose that a newsvendor is interested in optimally producing newspapers for an uncertain market, so that they minimize the cost incurred by actual production *and* by *not meeting* market demand, while respecting their holding capacity, or a predefined holding cost target. Let $K^P > 0$, $K^U > 0$ and $K^H > 0$ be known constants, standing for the production, unmet demand and holding costs *per production unit*. Also let $W : \Omega \to \mathbb{R}_+$ be the random market demand, a random variable with cdf $F_W : \mathbb{R} \to [0,1]$, for simplicity assumed to be absolutely continuous relative to the Lebesgue measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Since the market is uncertain, the newsvendor resorts to stochastically deciding their production plan by solving the *chance-constrained program*

$$\begin{array}{ll} \underset{x}{\text{minimize}} & K^P x + \mathbb{E}\left\{K^U (W-x)_+\right\} \\ \text{subject to} & \mathcal{P}\left(K^H (x-W)_+ > h\right) \le \alpha \ , \\ & x \ge 0 \end{array} \tag{29}$$

where, also for simplicity, we assume that the decision variable is real-valued, and where $0 \le \alpha \le 1$ constitutes the newsvendor's tolerance in the event that their holding cost $K^H (x-W)_+$ will exceed a prescribed threshold $h \ge 0$. Both α and h are fixed design parameters decided by the newsvendor beforehand. Chance-constrained newsvendor problems similar to (29) have been previously considered in the literature; see, for instance, the related article [Zhang et al., 2009]. Here, an important detail is that, despite the probabilistic constraint, problem (29) is risk-neutral as far as treatment of unmet demand is concerned. This is because only the expectation of the cost of not meeting the demand, corresponding to $K^U (W-x)_+$, is considered in the objective.

Problem (29) exhibits some interesting features and may be significantly simplified, as follows. First, we may observe that, for every fixed choice of $h \geq 0$,

$$
\begin{aligned}
\mathcal{P}\left(K^{H}(x-W)_{+}>h\right) &= \mathcal{P}\left(K^{H}(x-W)>h\right) \\
&\equiv \mathcal{P}\left(W<x-\frac{h}{K^{H}}\right) \\
&\equiv F_{W}\left(x-\frac{h}{K^{H}}\right), \quad \forall x \in \mathbb{R}.
\end{aligned} \tag{30}
$$

Consequently, it is true that

$$
\begin{aligned}
\mathcal{P}\left(K^{H}(x-W)_{+}>h\right) \leq \alpha \iff & F_{W}\left(x-\frac{h}{K^{H}}\right) \leq \alpha \\
\iff & x \leq F_{W}^{-1}(\alpha)+\frac{h}{K^{H}} && (31) \\
\iff & \mathcal{P}(W<x) \leq F_{W}\left(F_{W}^{-1}(\alpha)+\frac{h}{K^{H}}\right), && (32)
\end{aligned}
$$

where, due to F_W being continuous, the pseudo-inverse or quantile function $F_W^{-1}:[0,1] \rightarrow \mathbb{R}_+$ is defined as

$$
F_{W}^{-1}(\alpha) \triangleq \inf\left\{x \in \mathbb{R} \,|\, F_{W}(x) \geq \alpha\right\} \equiv \sup\left\{x \in \mathbb{R} \,|\, F_{W}(x) \leq \alpha\right\}. \tag{33}
$$

Thus, problem (29) is convex and may be reformulated as

$$
\begin{array}{ll}
\underset{x}{\operatorname{minimize}} & K^{P} x+\mathbb{E}\left\{K^{U}(W-x)_{+}\right\} \\
\text{subject to} & x \in\left[0, F_{W}^{-1}(\alpha)+\dfrac{h}{K^{H}}\right]
\end{array}. \tag{34}
$$

Hereafter, without loss of generality, we may assume that α and h are chosen such that $F_W^{-1}(\alpha)+h/K^H>0$. Otherwise, the problem is trivially solved at $x^* \equiv 0$. To be fully compatible with the generic notation utilized in this paper, we may also define $F(\cdot,\bullet) \triangleq K^P(\cdot)+K^U((\bullet)-(\cdot))_+$, and $\mathcal{X} \triangleq \left[0, F_W^{-1}(\alpha)+h/K^H\right]$.

Next, let us consider the derivative of the objective of (34), relative to x. We have

$$
\nabla \mathbb{E}\{F(x, W)\}=K^{P}-K^{U} \mathcal{P}(W \geq x), \quad \forall x \in \mathcal{X}. \tag{35}
$$

Hence, unless $K^U>K^P$, it readily follows that, for every $x \in \mathcal{X}$, $\nabla \mathbb{E}\{F(x,W)\} \geq 0$, again implying that the choice $x^* \equiv 0$ constitutes a solution of (34); in other words, producing nothing is always optimal whenever $K^U \leq K^P$. On the other hand, it is apparently true that $\nabla \mathbb{E}\{F(x,W)\}<0$, for all $x \in \mathcal{X}$, if and only if

$$
K^{P}-K^{U}(1-\mathcal{P}(W<x))<0, \quad \forall x \in \mathcal{X}, \tag{36}
$$

implying that the condition

$$
K^{P}-K^{U}\left(1-F_{W}\left(F_{W}^{-1}(\alpha)+\frac{h}{K^{H}}\right)\right)<0 \tag{37}
$$

is sufficient to ensure negativity of $\nabla\mathbb{E}\left\{F\left(x,W\right)\right\}$ everywhere within the feasible set $\mathcal{X}$ (where (32) is always satisfied), in which case the choice $x^{*}\equiv F_{W}^{-1}\left(\alpha\right)+h/K^{H}$ constitutes the optimal production level. Putting it altogether, whenever

$$F_{W}\left(F_{W}^{-1}\left(\alpha\right)+\frac{h}{K^{H}}\right)>0, \tag{38}$$

the condition

$$K^{U}\left(1-F_{W}\left(F_{W}^{-1}\left(\alpha\right)+\frac{h}{K^{H}}\right)\right)\leq K^{P}<K^{U} \tag{39}$$

ensures that problem (29) admits nontrivial solutions, thus being of technical interest.

Problem (29) may be solved in closed form. Indeed, either by considering the Karush-Kuhn-Tucker (KKT) conditions for problem (29) (for a constraint qualification, we may observe that Slater's condition is satisfied trivially whenever $F_{W}^{-1}\left(\alpha\right)+h/K^{H}>0$), or by looking at its geometric structure directly, it may be easily shown that its optimal solution may be expressed analytically as

$$x^{*}=\begin{cases}0, & \text{if } K^{U}\leq K^{P}\\ \min\left\{F_{W}^{-1}\left(\dfrac{K^{U}-K^{P}}{K^{U}}\right),F_{W}^{-1}\left(\alpha\right)+\dfrac{h}{K^{H}}\right\}, & \text{if } K^{U}>K^{P}\end{cases}, \tag{40}$$

representing the newsvendor's optimal decision in regard to the quantity of newspapers they would have to plan for, before the random market demand W is revealed.

Remark 3. Problems of the type of (29) are meaningful in various settings; specifically, they are most suitable when holding is operationally more important than unmet demand. For instance, it might be the case that the event where holding exceeds some threshold might have severe economic consequences, while not meeting the demand might be tolerable, although undesirable.

Additionally and perhaps more importantly, we should mention that a chance-constrained approach such as that adopted in (29) allows to efficiently *blend economic with physical quantities* in a single stochastic program. This is simply due to the fact that by defining a quantity $\widetilde{h}\triangleq h/K^{H}\geq0$, the probabilistic constraint of (29) may be written as

$$\mathcal{P}\left(\left(x-W\right)_{+}>\widetilde{h}\right)\leq\alpha, \tag{41}$$

implying that, if we want to, we may directly choose $\widetilde{h}$ as a probabilistic upper bound directly on the excess production $\left(x-W\right)_{+}$.

The modification above can be very useful if we are willing to consider the problem

$$\begin{array}{rl}\underset{x}{\text{minimize}} & K^{P}x+\mathbb{E}\left\{K^{H}\left(x-W\right)_{+}\right\}\\ \text{subject to} & \mathcal{P}\left(K^{U}\left(W-x\right)_{+}>u\right)\leq\alpha\\ & x\geq0\end{array}, \tag{42}$$

which constitutes a dual version of the initial newsvendor problem (29) resulting by interchanging the two respective stochastic costs and where, similarly to (29), $u\geq0$ is a prescribed threshold. In this case, unmet demand is operationally more important than holding, by choice. Of course, problem (42) is structurally very similar to (29), and can be analyzed via almost the same procedure as above. By defining $\widetilde{u} \triangleq u/K^U \geq 0$, problem (42) may be reformulated as

$$
\begin{array}{ll}
\underset{x}{\text{minimize}} & K^P x + \mathbb{E}\left\{K^H (x - W)_+\right\} \\
\text{subject to} & \mathcal{P}\left((W - x)_+ > \widetilde{u}\right) \leq \alpha \\
& x \geq 0
\end{array}
\quad , \tag{43}
$$

where $\widetilde{u}$ can now be preselected directly. We may readily observe that the objective of (43) constitutes an economic quantity (a cost), whereas the probabilistic constraint is placed on the unmet demand itself, which, of course, is a physical quantity. This modification can be extremely useful in a more realistic scenario, since in many practical cases the unit cost of unmet demand, K^U, is either completely unknown, or extremely difficult to estimate based on experience. ■

3.5.2 A Doubly Risk-Averse Single-Product NewsVendor

Suppose now that, due to high variability of the market demand, the newsvendor realizes that minimizing their unmet demand cost in expectation does not constitute a very meaningful objective. Thus, the newsvendor would like to decide on their newspaper production size by explicitly accounting for market variability in their model and, because they are reasonable, they are willing to settle with a potentially slightly higher *expected* monetary penalty for not meeting market demand. In effect, the newsvendor is interested in making their decision by additionally considering the risk incurred due to stochastic variability in the resulting unmet demand, the latter realized when market demand is revealed. In other words, the newsvendor would like to come up with a meaningful *doubly risk-averse* version of the original, chance-constrained problem (29).

The newsvendor may think as follows. For every *fixed and feasible* production decision $x \in \mathcal{X}$, if the *noisy* unmet demand $(W - x)_+$ is *smaller than* $\mathbb{E}\left\{(W - x)_+\right\}$, which is the newsvendor's expectation, then there is no risk incurred, since the newsvendor has been prepared for and has agreed to settle with a cost of unmet demand equal to $\mathbb{E}\left\{K^U (W - x)_+\right\}$. In an actual production scenario, $\mathbb{E}\left\{(W - x)_+\right\}$ might correspond to a small quantity of newspapers which are *not actually produced*, but for which resources have been allocated beforehand, to compensate for the case W is greater than x, but not too much. In other words, we might think about the quantity $\mathbb{E}\left\{(W - x)_+\right\}$ as a *risk-free, first-level safety stock*.

Positive risk is incurred whenever $(W - x)_+ > \mathbb{E}\left\{(W - x)_+\right\}$. However, the newsvendor realizes that *not* all values of the *central deviation*

$$
CD(x, W) \triangleq (W - x)_+ - \mathbb{E}\left\{(W - x)_+\right\} \tag{44}
$$

are of equal importance, or equal severity. In other words, the newsvendor's risk is variable relative to the value of $CD(x, W)$. Under the reasonable assumption that positive risk should be increasing as a function of the deviation $CD(x, W)$, the newsvendor's realization translates naturally into a variable and increasing rate of change of the risk, relative to the values of the deviation. In particular, whenever $CD(x, W) > 0$, the newsvendor identifies the following risk-incurring regions of increasing severity:

1) $CD(x, W) \in (0, t_1 > 0]$. In this case, unmet demand is higher than what the newsvendor expects, but its deviation from their expectation is no higher than a fixed threshold t_1. The value

$\mathbb{E}\left\{(W-x)_+\right\}+t_1$ corresponds to the maximum *partially unplanned or unexpected* production quantity that the newsvendor may be able to produce *today*, potentially using *presently unallocated* resources. We might think about the threshold t_1 as a *risk-incurring, second-level safety stock*.

2) $CD\,(x,W)\in(t_1,t_2>t_1]$. Here, the deviation of the unmet demand from the newsvendor's expectation is exceeds t_1, but is no higher than another fixed threshold t_2. The value t_2-t_1 corresponds to the maximum quantity of newspapers that the newsvendor cannot produce in-house today, but may ask a nearby vendor to produce for them. Of course, such events should incur higher and more severely increasing risk, since the newsvendor essentially *borrows resources* from the nearby vendor. We might call t_2 as the *borrowing threshold*.

3) $CD\,(x,W)\in(t_2,\infty)$. This constitutes an event of "total disaster," in which it is impossible for the newsvendor to compensate for unmet market demand. When $CD\,(x,W)>t_2$, unmet demand is so high that it cannot be met even if the newsvendor borrows the maximum amount of resources from some nearby newsvendor. This might have severe consequences for the newsvendor, since they either might be in debt, or even lose their professional credibility, or both.

Although potentially simplified, a narrative such as the above is reasonable and quite realistic. Of course, what is important for us in the context of this paper, is the fact that the characteristics of the relatively complex risk dynamics discussed above can be succinctly captured by an appropriately shaped risk regularizer, as proposed and analyzed herein. As a simplest example, we may define a *piecewise linear* risk regularizer $\mathcal{R}^{nv}:\mathbb{R}\to\mathbb{R}$ as

$$\mathcal{R}^{nv}\,(x)\triangleq\begin{cases}0, & \text{if } x\le 0\\ \psi_1 x, & \text{if } 0<x\le K^U t_1\\ \psi_2 x+(\psi_1-\psi_2)\,K^U t_1, & \text{if } K^U t_1<x\le K^U t_2\\ x+(\psi_2-1)\,K^U t_2+(\psi_1-\psi_2)\,K^U t_1, & \text{if } x>K^U t_2\end{cases}, \tag{45}$$

where the *risk slopes* $\psi_1\ge 0$ and $\psi_2\ge 0$ are chosen such that $\psi_1\le\psi_2\le 1$. Of course, $\mathcal{R}$ may be rewritten as

$$\begin{aligned}\mathcal{R}^{nv}\,(x)\equiv\psi_1 x\mathbb{1}_{\left[0,K^U t_1\right)}\,(x)&+\Big(\psi_2 x+(\psi_1-\psi_2)\,K^U t_1\Big)\,\mathbb{1}_{\left[K^U t_1,K^U t_2\right)}\,(x)\\ &+\Big(x+(\psi_2-1)\,K^U t_2+(\psi_1-\psi_2)\,K^U t_1\Big)\,\mathbb{1}_{\left[K^U t_2,\infty\right)}\,(x),\end{aligned} \tag{46}$$

for all $x\in\mathbb{R}$, and may be conveniently thought as a generalization of the positive part function of the upper-semideviation dispersion measure. *Equivalently*, the risk regularizer $\mathcal{R}^{nv}$ may be defined as an antiderivative of the cdf $F_Y^{nv}:\mathbb{R}\to[0,1]$ corresponding to some random variable $Y:\Omega\to\mathbb{R}$ in $\mathcal{Z}_\infty$, and defined as

$$F_Y^{nv}\,(x)\triangleq\psi_1\mathbb{1}_{\left[0,K^U t_1\right)}\,(x)+\psi_2\mathbb{1}_{\left[K^U t_1,K^U t_2\right)}\,(x)+\mathbb{1}_{\left[K^U t_2,\infty\right)}\,(x),\quad x\in\mathbb{R}, \tag{47}$$

as suggested by Theorem 2. The cdf F_Y^{nv} expresses precisely the *rate of increase of the risk* incurred at each $x\in\mathbb{R}$, where x may be thought of as the central deviation of the quantity whose risk is assessed by the risk regularizer $\mathcal{R}^{nv}$; in the newsvendor's case, this quantity should be the noisy economic consequence due to unmet demand, i.e., $K^U(W-x)_+$, also justifying the multiplication of thresholds t_1 and t_2 with the unit cost K^U in (45). Essentially, F_Y^{nv} admits an intuitive interpretation, and can be utilized in order to *actually design* $\mathcal{R}^{nv}$, as well; also see Theorem 2.

If the newsvendor chooses the risk slopes ψ_1 and ψ_2 such that $\psi_1 < \psi_2 < 1$ (only *they* know how set specific appropriate values), the quantity $\mathcal{R}^{nv}\left(K^U CD(x,W)\right)$ (for a fixed and feasible $x \in \mathcal{X}$) captures the dynamic behavior of the risk incurred by the central deviation $CD(x,W)$ when taking different values in $\mathbb{R}_+$, as described above. Essentially, $\mathcal{R}^{nv}$ may be regarded as a nonlinear weighting function acting on $K^U CD(x,W)$, whose shape has been carefully designed in order to reflect the newsvendor's particular context.

Now, the newsvendor is interested in utilizing $\mathcal{R}^{nv}\left(K^U CD(\cdot,W)\right)$ for decision making purposes. Since, for each $x \in \mathcal{X}$, $\mathcal{R}^{nv}\left(K^U CD(\cdot,W)\right)$ depends pointwise on the random market demand W, which is unobservable when the newsvendor decides their production plan, it is most reasonable to consider the $\mathcal{L}_p$-norm of $\mathcal{R}^{nv}\left(K^U CD(\cdot,W)\right)$, for some prespecified $p \in [1,\infty)$, *as a measure of magnitude*. Of course, lower values for such deterministic term are preferred. In order to effectively manage their risk during decision making, the newsvendor proceeds by adding the $\mathcal{L}_p$-norm of $\mathcal{R}^{nv}\left(K^U CD(\cdot,W)\right)$ as a penalty term to their original, risk-neutral objective, leading to the risk-averse stochastic program

$$\begin{array}{ll} \underset{x}{\text{minimize}} & K^P x + \mathbb{E}\left\{K^U(W-x)_+\right\} + c\left\|\mathcal{R}^{nv}\left(K^U(W-x)_+ - \mathbb{E}\left\{K^U(W-x)_+\right\}\right)\right\|_{\mathcal{L}_p} \\ \text{subject to} & \mathcal{P}\left(K^H(x-W)_+ > h\right) \le \alpha \\ & x \ge 0 \end{array}, \quad (48)$$

where, in general, $c \ge 0$ denotes the corresponding penalty tradeoff multiplier. It is then easy to see that the objective of problem (48) constitutes a mean-semideviation model. Indeed, by equivalently rewriting $\mathcal{R}^{nv}\left(K^U CD(\cdot,W)\right)$ as

$$\begin{aligned} \mathcal{R}^{nv}\left(K^U CD(x,W)\right) &\equiv \mathcal{R}^{nv}\left(K^P x + K^U(W-x)_+ - \mathbb{E}\left\{K^P x + K^U(W-x)_+\right\}\right) \\ &\equiv \mathcal{R}^{nv}\left(F(x,W) - \mathbb{E}\{F(x,W)\}\right), \quad \forall x \in \mathcal{X}, \end{aligned} \quad (49)$$

problem (48) may be equivalently restated as

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \mathbb{E}\{F(x,W)\} + c\left\|\mathcal{R}^{nv}\left(F(x,W) - \mathbb{E}\{F(x,W)\}\right)\right\|_{\mathcal{L}_p} \\ \text{subject to} & x \in \mathcal{X} \end{array}. \quad (50)$$

Apparently, the objective of problem (50) is a mean-semideviation risk measure. Of course, whenever $c \in [0,1]$, (50) (and thus, (48), as well) constitutes a convex, risk-averse stochastic program, precisely of the form considered in this paper.

To empirically demonstrate the effectiveness of the risk-averse newsvendor problem (50), we have conducted numerical simulations concerning the following three cases: Initial problem (29) (risk neutral), risk-averse problem (50), and risk-averse of the form (50), but with $\mathcal{R}^{nv}$ being replaced by $(\cdot)_+$, resulting to the mean-upper-semideviation risk measure. In all simulations, the random market demand W follows a Rayleigh distribution, and all necessary expectations present in each objective have been approximated utilizing $5 \cdot 10^6$ demand realizations, sampled independently. The precise

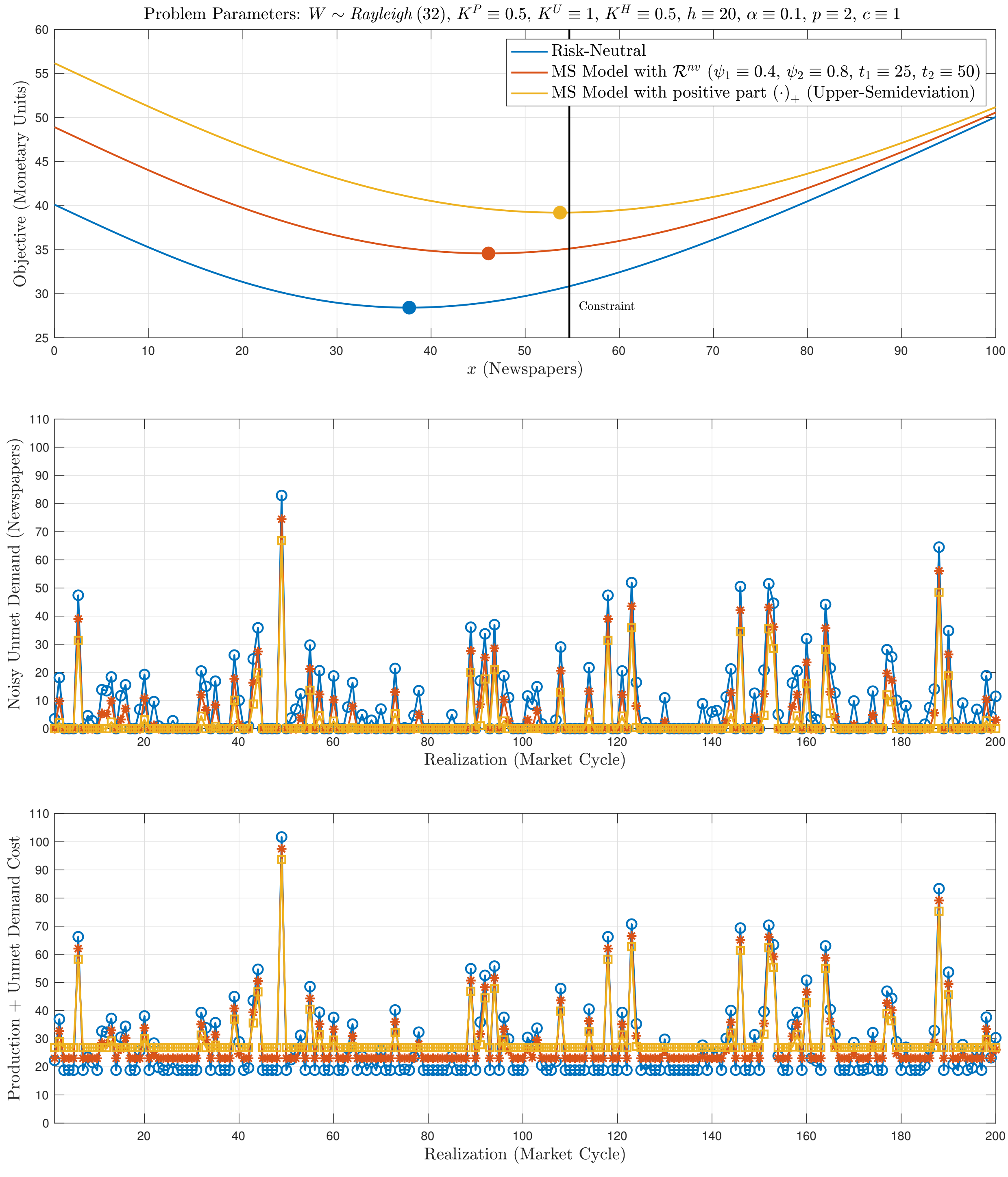

Figure 3.2: Top: Objective as a function of the production decision, for each of the following three cases: Risk neutral (problem (29)), risk-averse of the form (50), and risk-averse of the form (50), but with $\mathcal{R}^{nv}$ being replaced by $(\cdot)_+$. Middle: Unmet demand realizations as a function of the market cycle, for each of the three cases. Bottom: Combined monetary cost of production plus unmet demand realizations as a function of the market cycle, for each of the three cases.

values for the scale of the aforementioned distribution and for all the rest of parameters involved in problem (50) are shown in the title of Fig. 3.2 (top), respectively.

From Fig. 3.2 (top), we observe that the optimal production decision obtained by solving (50) is distinctly different from the respective solutions obtained by solving both the risk neutral problem (29), and the risk-averse problem employing the mean-upper-semideviation risk measure (all optimal solutions are represented by appropriately colored dots in Fig. 3.2 (top)). In particular, the solution of (50) lies somewhere near the midpoint of the respective solutions of the remaining two aforementioned problems. Therefore, we may conclude that the solution of (50) constitutes a *less conservative* risk-averse production decision, compared to the case of the mean-upper-semideviation risk measure, which essentially presumes that all risk-incurring events are of equal operational severity for the newsvendor. Equivalently, the mean-semideviation model utilized in (50) constitutes a less conservative risk-averse objective compared to that involving the mean-upper-semideviation risk measure (which, of course, is itself a mean-semideviation model induced by the trivial risk regularizer $(\cdot)_+$). As it can be readily observed in Fig. 3.2 (middle & bottom), the less conservative character of problem (50) translates directly to the statistical behavior of the realized unmet demand, and that of the combined cost due to production and unmet demand. This is obviously expected in this example, and is due to the simple structure to the original newsvendor problem we started with.

4 The *MESSAGE*^{p} Algorithm

This section is devoted to the introduction and detailed analysis of the *MESSAGE*^{p} algorithm. As also stated in Section 1, the *MESSAGE*^{p} algorithm is a parameterized (relative to the choice of $\mathcal{R}$) parallel version of the general purpose T-*SCGD* algorithm [Yang et al., 2018]. Both the algorithm and analysis presented in this work are new; as compared to [Yang et al., 2018], we propose a significantly milder set of problem assumptions, which, nonetheless, result in asymptotic guarantees of *at least* the same quality, *and more*.

Before proceeding, let us restate the stochastic program under study. Formally, for fixed $p \in [1, \infty)$, we are interested in the convex optimization problem

$$\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & \mathbb{E}\left\{F\left(\boldsymbol{x}, \boldsymbol{W}\right)\right\} + c\left\|\mathcal{R}\left(F\left(\boldsymbol{x}, \boldsymbol{W}\right) - \mathbb{E}\left\{F\left(\boldsymbol{x}, \boldsymbol{W}\right)\right\}\right)\right\|_{\mathcal{L}_p} \\ \text{subject to} & \boldsymbol{x} \in \mathcal{X} \end{array}, \tag{51}$$

where, for every $\boldsymbol{x} \in \mathcal{X}$, the convex real-valued random cost $F\left(\boldsymbol{x}, \boldsymbol{W}\left(\cdot\right)\right) \equiv \widetilde{F}\left(\boldsymbol{x}, \cdot\right)$ is in $\mathcal{Z}_q$, the set of feasible decisions $\mathcal{X} \subseteq \mathbb{R}^N$ is closed and convex, the risk related penalty multiplier is denoted by $c \geq 0$, and where $\mathcal{R} : \mathbb{R} \to \mathbb{R}$ constitutes any risk regularizer of choice. Recall that we implicitly assume that q and p are compatible according to Proposition 3. Also, based on our definitions, the objective is identified as either of the functions $\rho\left(F\left(\cdot, \boldsymbol{W}\right)\right) \equiv \rho_p^{\mathcal{R}}\left(F\left(\cdot, \boldsymbol{W}\right); c\right)$ and $\phi^{\widetilde{F}}$, where the choices of ρ-related quantities p, c, $\mathcal{R}$ are assumed to be *fixed* and made in advance; as such, they will not be explicitly referred to in our notation. Additionally, in the following, we assume that $c \in [0, 1]$, so that, by Corollary 1, (51) constitutes a *convex* problem.

In the following, we first discuss the reformulation of the objective of (51) in a convenient compositional form, key to the development of any compositional algorithm whatsoever. Second, we discuss the technical reasons that motivate the consideration of a compositional SSD-type algorithm for solving (51), as well as differentiability of its objective. Then, we present the *MESSAGE*^{p} algorithm, along with some of its key characteristics.

The section proceeds with the asymptotic analysis of the *MESSAGE*^{p} algorithm. First, our structural assumptions are presented and their main implications are discussed. Second, pathwise convergence of the *MESSAGE*^{p} algorithm is established, in a strong technical sense. Our proof follows the somewhat standard *"almost-supermartingale approach"*, also adopted in [Wang et al., 2017, Yang et al., 2018]. Third, we present a detailed convergence rate analysis of the *MESSAGE*^{p} algorithm, where we systematically develop all the results advertised in Section 1, along with relevant discussions.

Finally, the generality of our structural framework against that utilized in [Yang et al., 2018] is also clearly demonstrated, rigorously showing that the class of mean-semideviation programs supported herein is *strictly larger* than the respective class of problems supported within [Wang et al., 2017, Yang et al., 2018]. This result concludes our discussion related to the consistency of the *MESSAGE*^{p} algorithm, and justifies our effort.

4.1 Mean-Semideviations in Compositional Form

Because we will be interested in determining the structure of the subdifferential of the objective of (51), $\phi^{\widetilde{F}}$, it is convenient to express $\phi^{\widetilde{F}}$ in *compositional form*, similar to the general approach adopted in [Wang et al., 2017, Yang et al., 2018]. To do this, let us define the *expectation functions* $\varrho : \mathbb{R}_+ \to \mathbb{R}$, $g^{\widetilde{F}} : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}_+$ and $\boldsymbol{h}^{\widetilde{F}} : \mathbb{R}^N \to \mathbb{R}^N \times \mathbb{R}$ as

$$\varrho(x) \triangleq x^{1/p}, \quad x > 0, \tag{52}$$

$$g^{\widetilde{F}}(\boldsymbol{x}, y) \triangleq \mathbb{E}\{(\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y))^p\} \quad \text{and} \tag{53}$$

$$\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x}) \triangleq [\boldsymbol{x}\, \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\}], \tag{54}$$

for every admissible choice of F and $\mathcal{P}_{\boldsymbol{W}}$. Then, $\phi^{\widetilde{F}}$ may be alternatively expressed as

$$\phi^{\widetilde{F}}(\boldsymbol{x}) \equiv \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\} + c\varrho\left(g^{\widetilde{F}}\left(\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})\right)\right), \quad \boldsymbol{x} \in \mathcal{X}. \tag{55}$$

We observe that the functional composition term on the RHS of (55) coincides with the dispersion measure in the objective of (51), simply rewritten as a composition of real-valued and, of course, deterministic, functions. In the special case where $p \equiv 1$, (55) becomes

$$\phi^{\widetilde{F}}(\boldsymbol{x}) \equiv \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\} + cg^{\widetilde{F}}\left(\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})\right), \quad \boldsymbol{x} \in \mathcal{X}. \tag{56}$$

Also, it is trivial to see that, if $p \equiv 1$, then, by defining another function $\mathring{g}^{\widetilde{F}} : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ as

$$\mathring{g}^{\widetilde{F}}(\boldsymbol{x}, y) \triangleq y + c\mathbb{E}\{\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y)\}, \tag{57}$$

we may as well write

$$\phi^{\widetilde{F}}(\boldsymbol{x}) \equiv \mathring{g}^{\widetilde{F}}\left(\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})\right), \quad \boldsymbol{x} \in \mathcal{X}, \tag{58}$$

which is exactly the type of problem considered in [Wang et al., 2017], except for the fact that, here, it is formulated under weaker assumptions. See Section 4.3 for details.

The difference between (56) and (55) is subtle. As we will see later on, based on our assumptions, the structure of any SSD-type optimization algorithm suitable for handling objectives of the form of (55) (and, thus, (51)) for $p \in (1, \infty)$ is inherently more complicated, compared to the case of the slightly simpler objective resulting by setting $p \equiv 1$.

Remark 4. Note that, alternatively, we could reexpress $\phi^{\widetilde{F}}$ in the compositional form outlined in ([Wang et al., 2017], Supplemental Materials, Section H.4, or [Yang et al., 2018], Section 4). In particular, if we define the expectation functions $\widehat{\varrho} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$, $\widehat{g}^{\widetilde{F}} : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R} \times \mathbb{R}_+$ and $\widehat{\boldsymbol{h}}^{\widetilde{F}} : \mathbb{R}^N \to \mathbb{R}^N \times \mathbb{R}$ as

$$\widehat{\varrho}\,(x,y) \triangleq x + cy^{1/p}, \quad y > 0, \tag{59}$$

$$\widehat{\boldsymbol{g}}^{\widetilde{F}}\,(\boldsymbol{x},y) \triangleq [y \; \mathbb{E}\left\{(\mathcal{R}\,(F\,(\boldsymbol{x},\boldsymbol{W}) - y))^p\right\}] \quad \text{and} \tag{60}$$

$$\widehat{\boldsymbol{h}}^{\widetilde{F}}\,(\boldsymbol{x}) \triangleq [\boldsymbol{x} \; \mathbb{E}\left\{F\,(\boldsymbol{x},\boldsymbol{W})\right\}], \tag{61}$$

then $\phi^{\widetilde{F}}$ may be written as

$$\phi^{\widetilde{F}}\,(\boldsymbol{x}) = \widehat{\varrho}\left(\widehat{\boldsymbol{g}}^{\widetilde{F}}\left(\widehat{\boldsymbol{h}}^{\widetilde{F}}\,(\boldsymbol{x})\right)\right), \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{62}$$

Of course, the compositional representation (62) is equivalent to (55). For our purposes, though, (55) is perfectly sufficient and, additionally, it is cleaner and somewhat more compact. ■

4.2 Algorithm Motivation and Differentiability of $\phi^{\widetilde{F}}$

As a result of the discussion above, the original problem (4) can be equivalently written as

$$\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & \mathbb{E}\left\{F\,(\boldsymbol{x},\boldsymbol{W})\right\} + c\varrho\left(g^{\widetilde{F}}\left(\boldsymbol{h}^{\widetilde{F}}\,(\boldsymbol{x})\right)\right) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{X} \end{array}. \tag{63}$$

Exploiting the assumed convexity of $\phi^{\widetilde{F}}$ on $\mathcal{X}$, and given that we are interested in solving (63), one would most reasonably hope for a SSD-type algorithm, whose gradient evaluation policy follows a path of the stochastic differential equation

$$\boldsymbol{x}^{n+1} = \Pi_{\mathcal{X}}\left\{\boldsymbol{x}^n - \gamma_n\left[\widetilde{\underline{\nabla}}^{n+1}\phi^{\widetilde{F}}\,(\boldsymbol{x}^n)\right]\right\}, \quad n \in \mathbb{N}, \tag{64}$$

where $\boldsymbol{x}^0 \in \mathcal{X}$ is arbitrarily chosen, $\{\gamma_n > 0\}_{n\in\mathbb{N}}$ is an appropriately chosen stepsize sequence, and $\left\{\widetilde{\underline{\nabla}}^n\phi^{\widetilde{F}}\right\}_{n\in\mathbb{N}^+}$ denotes a sequence of *stochastic subgradients*, that is, a sequence of $\mathbb{R}^N$-valued, *appropriately measurable random functions on* $\mathbb{R}^N \times \Omega$, such that

$$\mathbb{E}\left\{\widetilde{\underline{\nabla}}^n\phi^{\widetilde{F}}\,(\boldsymbol{x})\right\} \equiv \mathbb{E}\left\{\widetilde{\underline{\nabla}}^n\phi^{\widetilde{F}}\,(\boldsymbol{x},\cdot)\right\} \in \partial\phi^{\widetilde{F}}\,(\boldsymbol{x}), \quad \forall\,(n,\boldsymbol{x}) \in \mathbb{N}^+ \times \mathcal{X}, \tag{65}$$

where we recall that the compact-valued multifunction $\partial\phi^{\widetilde{F}} : \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ constitutes the subdifferential associated with the convex function $\phi^{\widetilde{F}}$.

We are now interested in the structural characterization of $\partial\phi^{\widetilde{F}}$. However, even though $\phi^{\widetilde{F}}$ is convex, a *computationally useful* characterization of its subdifferential is highly nontrivial; this is mainly due to the fact that, although any mean-semideviation risk measure is convex-monotone (as long as $c \in [0,1]$), the corresponding dispersion measure can be only guaranteed to be convex. This implies that composition of the latter with a convex random function on $\mathbb{R}^N$ (such as F) does *not* yield a convex function on $\mathbb{R}^N$. Unfortunately, common rules from subdifferential calculus, such as addition and composition, which are essential in order to tractably determine the structure of the multifunction $\partial\phi^{\widetilde{F}}$, are rather complicated for nonconvex functions (see, e.g., Chapter 10 in [Rockafellar and Wets, 2004]), not only conceptually, but most importantly, from a computational point of view, as well. Fortunately, the problem simplifies considerably if we impose some mild regularity requirements on the structure of the random cost function F, thus avoiding unnecessary technical complications.

Assumption 3. *The random function F possesses the following properties:*

P1 *For every $\boldsymbol{x} \in \mathcal{X}$, there exists a measurable set $\mathsf{D}_{\boldsymbol{x}} \subseteq \Omega$, with $\mathcal{P}(\mathsf{D}_{\boldsymbol{x}}) \equiv 1$, such that, for all $\omega \in \mathsf{D}_{\boldsymbol{x}}$, the random function $F(\cdot, \boldsymbol{W}(\omega))$ is differentiable at $\boldsymbol{x}$. In other words, F is differentiable at each $\boldsymbol{x} \in \mathcal{X}$, almost everywhere relative to the base measure $\mathcal{P}$.*

P2 *Let $\mathcal{A}$ be the countable Borel nullset of points where the risk regularizer $\mathcal{R}$ is nondifferentiable. For every $\boldsymbol{x} \in \mathcal{X}$, there exists an event $\mathsf{N}_{\boldsymbol{x}} \subseteq \Omega$, with $\mathcal{P}(\mathsf{N}_{\boldsymbol{x}}) \equiv 1$, such that, for all $\omega \in \mathsf{N}_{\boldsymbol{x}}$, $F(\boldsymbol{x}, \boldsymbol{W}(\omega)) - \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\} \notin \mathcal{A}$. In other words, for every $\boldsymbol{x} \in \mathcal{X}$, it is true that*

$$\mathcal{P}(F(\boldsymbol{x}, \boldsymbol{W}) - \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\} \notin \mathcal{A}) \equiv 1. \tag{66}$$

In addition to Properties **P1** and **P2**, it is technically necessary to make the following basic assumption, concerning the random subdifferential multifunction of F, relative to $\boldsymbol{x}$.

Assumption 4. *There exists a jointly $\mathscr{B}\left(\mathbb{R}^N\right) \otimes \mathscr{B}\left(\mathbb{R}^M\right)$-measurable selection of the closed-valued multifunction $\partial F(\cdot, \bullet) : \mathbb{R}^N \times \mathbb{R}^M \rightrightarrows \mathbb{R}^N$, say $\underline{\nabla} F(\cdot, \bullet) : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^N$; this is provided by the $\mathcal{SO}$, at each $n \in \mathbb{N}$, given current iterate $\boldsymbol{x}^n \in \mathbb{R}^N$ and IID process realization $\boldsymbol{w} \in \mathbb{R}^M$.*

Assumption 4 is important, because it allows us to integrate $\underline{\nabla} F$ on $\mathbb{R}^N \times \mathbb{R}^M$, relative to any qualifying Borel measure, provided such an integral is well defined. This is extremely useful, in case *both arguments* $\boldsymbol{x}$ and $\boldsymbol{W}$ are random elements. Hereafter, Assumption 4 will be considered *implicitly true*; although it has to be verified case-by-case, it is almost always true in practice. Utilizing both properties **P1** and **P2**, the following result may be formulated; it will then be utilized in the design of SSD-type algorithms, specialized for the convex problem (51).

Lemma 1. (Differentiability of $\phi^{\widetilde{F}}$) *Consider the convex function $\phi^{\widetilde{F}}$. Let Assumption 3 be in effect, and suppose that $\mathcal{R}$ is not identically zero everywhere on $\mathbb{R}$. Also, if $p \in (1, \infty)$, and with*

$$\kappa_{\mathcal{R}} \triangleq \sup\{x \in \mathbb{R} \,|\, \mathcal{R}(x) \equiv 0\} \in [-\infty, \infty), \tag{67}$$

suppose that

$$\mathcal{P}(F(\boldsymbol{x}, \boldsymbol{W}) - \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\} \leq \kappa_{\mathcal{R}}) < 1, \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{68}$$

Then $\phi^{\widetilde{F}}$ is differentiable everywhere on $\mathcal{X}$, and its gradient $\nabla\phi^{\widetilde{F}} : \mathbb{R}^N \to \mathbb{R}^N$ may be expressed as

$$\nabla\phi^{\widetilde{F}}(\boldsymbol{x}) \equiv \mathbb{E}\{\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W})\} + c\nabla\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x}) \nabla g^{\widetilde{F}}\left(\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})\right) \nabla\varrho\left(g^{\widetilde{F}}\left(\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})\right)\right), \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{69}$$

where the derivative $\nabla \varrho : \mathbb{R}_+ \to \mathbb{R}$*, Jacobian* $\nabla \boldsymbol{h}^{\widetilde{F}} : \mathbb{R}^N \to \mathbb{R}^{N\times(N+1)}$ *and gradient* $\nabla g^{\widetilde{F}} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ *are given by the expectation functions*

$$\nabla \varrho\,(x) \equiv \begin{cases} 1, \quad x \geq 0, & \textit{if } p \equiv 1 \\ \dfrac{1}{p} x^{(1-p)/p}, \quad x > 0, & \textit{if } p \in (1,\infty) \end{cases}, \tag{70}$$

$$\nabla \boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x}) \equiv \mathbb{E}\left\{\left[\boldsymbol{I}_N \,\middle|\, \underline{\nabla} F\,(\boldsymbol{x},\boldsymbol{W})\right]\right\} \quad \textit{and} \tag{71}$$

$$\nabla g^{\widetilde{F}}(\boldsymbol{x},y) \equiv \mathbb{E}\left\{ p\,(\mathcal{R}\,(F\,(\boldsymbol{x},\boldsymbol{W}) - y))^{p-1}\, \underline{\nabla}\mathcal{R}\,(F\,(\boldsymbol{x},\boldsymbol{W}) - y) \begin{bmatrix} \underline{\nabla} F\,(\boldsymbol{x},\boldsymbol{W}) \\ \hline -1 \end{bmatrix} \right\}, \tag{72}$$

respectively, for every $(\boldsymbol{x},y) \in \{(\boldsymbol{x},y) \in \mathcal{X}\times\mathbb{R} \,|\, y \equiv \mathbb{E}\{F\,(\boldsymbol{x},\boldsymbol{W})\}\} \triangleq \mathrm{Graph}_{\mathcal{X}}\,(\mathbb{E}\{F\,(\cdot,\boldsymbol{W})\})$.

Proof of Lemma 1. See Section 7.2 (Appendix). ■

Remark 5. Note that, in Lemma 1, we have explicitly assumed that the risk regularizer $\mathcal{R}$ is *not* identically zero everywhere on $\mathbb{R}$. If it is, (4) reduces to the standard, risk neutral stochastic program of minimizing the expectation of a convex random function over a closed convex set, well-studied in the literature of stochastic approximation. ■

Remark 6. We would also like to comment on the potential restrictiveness of condition (68). Suppose that $\kappa_{\mathcal{R}} \leq 0$. This essentially means that the risk regularizer $\mathcal{R}$ always *positively* penalizes events for which $F\,(\boldsymbol{x},\boldsymbol{W}) > \mathbb{E}\{F\,(\boldsymbol{x},\boldsymbol{W})\}$. In such a case, (68) will be true if, for every $\boldsymbol{x} \in \mathcal{X}$,

$$\mathcal{P}\,(F\,(\boldsymbol{x},\boldsymbol{W}) - \mathbb{E}\{F\,(\boldsymbol{x},\boldsymbol{W})\} \leq 0) < 1. \tag{73}$$

It is then a standard exercise to show that, for every $\boldsymbol{x} \in \mathcal{X}$, (73) is true if and only if

$$\mathcal{P}\,(F\,(\boldsymbol{x},\boldsymbol{W}) \equiv \mathbb{E}\{F\,(\boldsymbol{x},\boldsymbol{W})\}) < 1. \tag{74}$$

This means that, if $\kappa_{\mathcal{R}} \leq 0$, (68) translates to a truly mild condition on the structure of $F\,(\cdot,\boldsymbol{W})$, namely, that, for every feasible decision $\boldsymbol{x} \in \mathcal{X}$, the cost $F\,(\boldsymbol{x},\boldsymbol{W})$ cannot be equal to a constant, almost everywhere relative to $\mathcal{P}$. In other words, for every $\boldsymbol{x} \in \mathcal{X}$, $F\,(\boldsymbol{x},\boldsymbol{W})$ has to be a nontrivial random variable. Of course, it is not hard to satisfy such a condition in practice. ■

Lemma 1 establishes that, under Assumption 3 (properties **P1** and **P2**) and, *if* $p \in (1,\infty)$, under condition (68), the subdifferential of $\phi^{\widetilde{F}}$ is a singleton, and provides an explicit representation of the gradient vector $\nabla\phi^{\widetilde{F}}$.

Of course, in the original version of the SSD algorithm, one would require the availability of a stochastic subgradient sequence in order to perform the usual SSD update step. However, Lemma 1 reveals that, under our base assumptions, such a stochastic subgradient process is far from obvious to obtain, mainly due to the functional form of $\nabla\phi^{\widetilde{F}}$. In particular, by inspection of (69) in Lemma 1, it is easy to see that $\nabla\phi^{\widetilde{F}}$ exhibits itself *compositional structure*, consisting of products of nested expectation functions. This fact implies that it is generally *not* possible to generate a stochastic gradient in a *single* sampling step, thus leading naturally to the idea of developing a *compositional* *stochastic subgradient algorithm* (see, for instance, [Wang et al., 2017]). In such an algorithm, the respective stochastic gradient step would be implemented in a hierarchical fashion *at every iteration*, starting from the "deepest" *Stochastic Approximation (SA) level*, to the "shallowest" (in most cases, a *biased* procedure); see Section 4.3 for details.

Exploiting Assumptions 1, 2 *and* 3, and to efficiently exploit the special compositional structure of $\nabla\phi^{\widetilde{F}}$, we will be particularly interested in sampled *approximations* of $\nabla\phi^{\widetilde{F}}$, which are constructed using *sampled realizations* of the random cost $F(\cdot,\boldsymbol{W})$, as well as some corresponding subgradient, $\underline{\nabla}F(\cdot,\boldsymbol{W})$. From now on, we will implicitly assume that the risk regularizer $\mathcal{R}$ *is not identically zero everywhere on* $\mathbb{R}$. Otherwise, the problem reduces to its risk-neutral counterpart.

4.3 The *MESSAGE*^{p} Algorithm

For any value of $p\in[1,\infty)$, the proposed *MESSAGE*^{p} algorithm consists of *three* SA levels, as naturally suggested by the structure of $\nabla\phi^{\widetilde{F}}$, and assumes the existence of *two* mutually independent, IID information streams, $\boldsymbol{W}_1^n$, $\boldsymbol{W}_2^n$, accessible by the $\mathcal{SO}$ (see Assumption 1), as follows. In the first (shallowest) SA level, at iteration $n\in\mathbb{N}$, and given current, random iterates $\boldsymbol{x}^n\equiv\boldsymbol{x}^n(\omega)\in\mathcal{X}$ and $y^n\equiv y^n(\omega)\in\mathbb{R}$, the $\mathcal{SO}$ provides the samples $F\Big(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\Big)$ and $\underline{\nabla}F\Big(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\Big)$ and the smoothing update

$$y^{n+1}=(1-\beta_n)\,y^n+\beta_nF\Big(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\Big)\tag{75}$$

is performed, where $\{\beta_n>0\}_{n\in\mathbb{N}}$ is an appropriately chosen stepsize sequence. In the second SA level, the $\mathcal{SO}$ provides the samples $F\Big(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\Big)$ and $\underline{\nabla}F\Big(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\Big)$, and another smoothing update

$$z^{n+1}=\begin{cases}1, & \text{if } p=1\\ (1-\gamma_n)\,z^n+\gamma_n\left(\mathcal{R}\left(F\Big(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\Big)-y^n\right)\right)^p, & \text{if } p>1\end{cases}\tag{76}$$

is performed, with $\{\gamma_n>0\}_{n\in\mathbb{N}}$ being another appropriately chosen stepsize sequence. Of course, in the simpler case where $p\equiv1$, no actual update is performed. In the third (deepest) SA level, *with no additional information by the* $\mathcal{SO}$, and by defining the stochastic gradient *approximation* $\widehat{\nabla}^{n+1}\phi^{\widetilde{F}}:\mathbb{R}^{N+2}\times\Omega\to\mathbb{R}$ as

$$\begin{aligned}
&\widehat{\nabla}^{n+1}\phi^{\widetilde{F}}(\boldsymbol{x}^n,y^n,z^n)\equiv\widehat{\nabla}^{n+1}\phi^{\widetilde{F}}(\boldsymbol{x}^n,y^n,z^n,\cdot)\\
&\triangleq\underline{\nabla}F\Big(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\Big)+c\,(z^n)^{(1-p)/p}\left[\boldsymbol{I}_N\,\middle|\,\underline{\nabla}F\Big(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\Big)\right]\\
&\qquad\times\left[\begin{array}{c}\underline{\nabla}F\Big(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\Big)\\ \hline -1\end{array}\right]\underline{\nabla}\mathcal{R}\Big(F\Big(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\Big)-y^n\Big)\Big(\mathcal{R}\Big(F\Big(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\Big)-y^n\Big)\Big)^{p-1}\\
&\equiv\underline{\nabla}F\Big(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\Big)+c\,(z^n)^{(1-p)/p}\Big(\underline{\nabla}F\Big(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\Big)-\underline{\nabla}F\Big(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\Big)\Big)\\
&\qquad\qquad\times\underline{\nabla}\mathcal{R}\Big(F\Big(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\Big)-y^n\Big)\Big(\mathcal{R}\Big(F\Big(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\Big)-y^n\Big)\Big)^{p-1}\\
&\triangleq\underline{\nabla}F\Big(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\Big)+c\Delta^{n+1}(\boldsymbol{x}^n,y^n,z^n),
\end{aligned}\tag{77}$$

Algorithm 1 *MESSAGE*^{p}

Input: Initial points $\boldsymbol{x}^0 \in \mathcal{X}$, $y^0 \in \mathbb{R}$, $z^0 \in \mathbb{R}$, stepsize sequences $\{\alpha_n\}_{n\in\mathbb{N}}$, $\{\beta_n\}_{n\in\mathbb{N}}$, $\{\gamma_n\}_{n\in\mathbb{N}}$, IID sequences $\{\boldsymbol{W}_1^n\}_{n\in\mathbb{N}}$, $\{\boldsymbol{W}_2^n\}_{n\in\mathbb{N}}$ and penalty coefficient $c \in [0,1]$.
Output: Sequence $\{\boldsymbol{x}^n\}_{n\in\mathbb{N}}$.
1: **for** $n = 0, 1, 2, \ldots$ **do**
2: Obtain $F\Big(\boldsymbol{x}^n, \boldsymbol{W}_1^{n+1}\Big)$ and $\underline{\nabla} F\Big(\boldsymbol{x}^n, \boldsymbol{W}_1^{n+1}\Big)$ from the $\mathcal{SO}$.
3: Update (First SA Level):

$$y^{n+1} = (1-\beta_n)\, y^n + \beta_n F\Big(\boldsymbol{x}^n, \boldsymbol{W}_1^{n+1}\Big)$$

4: Obtain $F\Big(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\Big)$ and $\underline{\nabla} F\Big(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\Big)$ from the $\mathcal{SO}$.
5: Update (Second SA Level):

$$z^{n+1} = \begin{cases} 1, & \text{if } p = 1 \\ (1-\gamma_n)\, z^n + \gamma_n \left(\mathcal{R}\left(F\Big(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\Big) - y^n\right)\right)^p, & \text{if } p > 1 \end{cases}$$

6: Define auxiliary variables:

$$\begin{aligned} \delta &= F\left(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\right) - y^n \\ \boldsymbol{\delta}^{\underline{\nabla}} &= \underline{\nabla} F\Big(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\Big) - \underline{\nabla} F\Big(\boldsymbol{x}^n, \boldsymbol{W}_1^{n+1}\Big) \\ \Delta &= \boldsymbol{\delta}^{\underline{\nabla}} \underline{\nabla}\mathcal{R}(\delta)\,(\mathcal{R}(\delta))^{p-1}\,(z^n)^{(1-p)/p} \end{aligned}$$

7: Update (Third SA Level):

$$\boldsymbol{x}^{n+1} = \Pi_{\mathcal{X}}\left\{\boldsymbol{x}^n - \alpha_n \left(\underline{\nabla} F\Big(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\Big) + c\Delta\right)\right\}$$

8: **end for**

we update the current estimate $\boldsymbol{x}^n$ as

$$\begin{aligned} \boldsymbol{x}^{n+1} &\equiv \Pi_{\mathcal{X}}\Big\{\boldsymbol{x}^n - \alpha_n \widehat{\nabla}^{n+1} \phi^{\widetilde{F}}(\boldsymbol{x}^n, y^n, z^n)\Big\} \\ &\equiv \Pi_{\mathcal{X}}\Big\{\boldsymbol{x}^n - \alpha_n \Big(\underline{\nabla} F\Big(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\Big) + c\Delta^{n+1}(\boldsymbol{x}^n, y^n, z^n)\Big)\Big\}, \end{aligned} \tag{78}$$

where $\{\alpha_n \geq 0\}_{n\in\mathbb{N}}$ is another appropriately chosen stepsize sequence. In the above, $\Delta^n : \mathbb{R}^{N+2} \times \Omega \to \mathbb{R}$, $n \in \mathbb{N}^+$ (a random function of the involved quantities) may be viewed as a *risk-averse correction sequence*, weighted by the penalty multiplier $c \in [0,1]$. Again, if $p \equiv 1$, the correction Δ^n is simplified accordingly as

$$\begin{aligned} \Delta^{n+1}(\boldsymbol{x}^n, y^n, z^n) &\equiv \Delta^{n+1}(\boldsymbol{x}^n, y^n, 1) \\ &\equiv \underline{\nabla}\mathcal{R}\Big(F\Big(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\Big) - y^n\Big)\Big(\underline{\nabla} F\Big(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\Big) - \underline{\nabla} F\Big(\boldsymbol{x}^n, \boldsymbol{W}_1^{n+1}\Big)\Big), \end{aligned} \tag{79}$$

for all $n \in \mathbb{N}$. As we will shortly see, whether $p \equiv 1$ or $p > 1$ has nontrivial consequences in regard to the asymptotic performance of the *MESSAGE*^{p} algorithm. The iterative optimization procedure outlined above (the *MESSAGE*^{p} algorithm) is summarized in Algorithm 1.

Exploiting Assumption 4, it is then easy to verify that, for every $(n, \boldsymbol{x}) \in \mathbb{N}^+ \times \mathcal{X}$,

$$\mathbb{E}\left\{\widehat{\nabla}^n \phi^{\widetilde{F}}\left(\boldsymbol{x}, \mathbb{E}\left\{F(\boldsymbol{x},\boldsymbol{W})\right\}, \mathbb{E}\left\{(\mathcal{R}\left(F(\boldsymbol{x},\boldsymbol{W})-\mathbb{E}\left\{F(\boldsymbol{x},\boldsymbol{W})\right\}\right))^p\right\}\right)\right\} \equiv \nabla\phi^{\widetilde{F}}\left(\boldsymbol{x}\right). \tag{80}$$

This implies that while the random function $\widehat{\nabla}^n \phi^{\widetilde{F}}\left(\cdot, y^n, z^n\right)$ *does not* necessarily yield a stochastic gradient of $\phi^{\widetilde{F}}$, for $n \in \mathbb{N}$, $\widehat{\nabla}^n \phi^{\widetilde{F}}\left(\cdot, \mathbb{E}\left\{F(\cdot,\boldsymbol{W})\right\}, \mathbb{E}\left\{(\mathcal{R}\left(F(\cdot,\boldsymbol{W})-\mathbb{E}\left\{F(\cdot,\boldsymbol{W})\right\}\right))^p\right\}\right)$ *does*; this is a key fact in the analysis of general purpose compositional stochastic subgradient algorithms [Yang et al., 2018].

Remark 7. In relation to the brief discussion above, it might be helpful to observe that, by the substitution rule for conditional expectations (again due to Assumption 4), it is also true that

$$\mathbb{E}\Big\{\widehat{\nabla}^{n+1} \phi^{\widetilde{F}}(\boldsymbol{x}^n, \mathbb{E}\{F(\boldsymbol{x},\boldsymbol{W})\}|_{\boldsymbol{x}\equiv\boldsymbol{x}^n}, \mathbb{E}\{(\mathcal{R}\left(F(\boldsymbol{x},\boldsymbol{W})-\mathbb{E}\{F(\boldsymbol{x},\boldsymbol{W})\}\right))^p\}|_{\boldsymbol{x}\equiv\boldsymbol{x}^n}) \Big| \boldsymbol{x}^n\Big\} \equiv \nabla\phi^{\widetilde{F}}(\boldsymbol{x}^n), \tag{81}$$

almost everywhere relative to $\mathcal{P}$ and, apparently, this is *not* the case for the conditional expectation of $\widehat{\nabla}^{n+1} \phi^{\widetilde{F}}\left(\boldsymbol{x}^n, y^n, z^n\right)$ relative to $\sigma\left\{\boldsymbol{x}^n\right\}$, or even $\sigma\left\{\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^n\right\}$. In this sense, we might say that the hierarchical approximate gradient sampling scheme described by (75) and (78) is *conditionally biased*. ■

4.4 Convergence Analysis

Next, we present and discuss the proposed structural framework, explicitly demonstrating its generality and flexibility. Subsequently, we proceed with a detailed presentation of our main technical results, concerning the asymptotic behavior of the *MESSAGE*^{p} algorithm; we study pathwise convergence first, and rate of convergence second.

4.4.1 Structural Assumptions

Hereafter, for some measurable set $\Omega_E \subseteq \Omega$, such that $\mathcal{P}\left(\Omega_E\right) \equiv 1$, let us define the quantities

$$m_l \triangleq \inf_{\boldsymbol{x}\in\mathcal{X}} \inf_{\omega\in\Omega_E} F\left(\boldsymbol{x}, \boldsymbol{W}\left(\omega\right)\right) \quad \text{and} \quad m_h \triangleq \sup_{\boldsymbol{x}\in\mathcal{X}} \sup_{\omega\in\Omega_E} F\left(\boldsymbol{x}, \boldsymbol{W}\left(\omega\right)\right), \tag{82}$$

and let $\mathfrak{R}^{\widetilde{F}} \triangleq \mathrm{cl}\left\{(m_l - m_h, m_h - m_l)\right\}$, where the closure is taken relative to the usual Euclidean topology on $\mathbb{R}$. The structural problem assumptions considered in this paper follow.

Assumption 5. *For $P \in [2, \infty]$ and $Q \in [P/(P-1), \infty]$*[a]*, $F\left(\cdot, \boldsymbol{W}\right)$ and $\mathcal{R}$ satisfy the conditions:*

C1 *For chosen random subgradient $\underline{\nabla} F\left(\cdot, \boldsymbol{W}\right)$, there exists a number $G < \infty$, such that*

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left[\mathbb{E}\left\{\left\|\underline{\nabla} F\left(\boldsymbol{x}, \boldsymbol{W}\right)\right\|_2^P\right\}\right]^{1/P} \triangleq \sup_{\boldsymbol{x}\in\mathcal{X}} \left\|\left\|\underline{\nabla} F\left(\boldsymbol{x}, \boldsymbol{W}\right)\right\|_2\right\|_{\mathcal{L}_P} \leq G.$$

In other words, the ℓ_2-norm of $\underline{\nabla} F\left(\cdot, \boldsymbol{W}\right)$ has bounded $\mathcal{L}_P$-norm, uniformly over $\mathcal{X}$.

C2 *There exists a number $V < \infty$, such that*

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \mathbb{V}\left\{F\left(\boldsymbol{x}, \boldsymbol{W}\right)\right\} \triangleq \sup_{\boldsymbol{x} \in \mathcal{X}} \left[\mathbb{E}\left\{\left(F\left(\boldsymbol{x}, \boldsymbol{W}\right)\right)^2\right\} - \left(\mathbb{E}\left\{F\left(\boldsymbol{x}, \boldsymbol{W}\right)\right\}\right)^2\right] \leq V,$$

that is, uniformly on $\mathcal{X}$, $F\left(\cdot, \boldsymbol{W}\right)$ is of bounded variance.

C3 *For chosen subderivative $\underline{\nabla}\mathcal{R}$, there exists another number $D < \infty$, such that*[b]

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \left|\left|\left| \underline{\nabla}\left(\mathcal{R}\left(z\right)\right)^p\big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_1} - \underline{\nabla}\left(\mathcal{R}\left(z\right)\right)^p\big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_2} \right|\right|\right|_{\mathcal{L}_Q} \leq D\left|y_1 - y_2\right|,$$

for all $(y_1, y_2) \in \left[\mathrm{cl}\left\{(m_l, m_h)\right\}\right]^2$. This is a Lipschitz-in-Expectation type of condition.

C4 *<u>Whenever $p > 1$</u>, it is true that $-\infty < m_l \leq m_h < \infty$, and*

$$0 < \varepsilon \triangleq \mathcal{R}\left(m_l - m_h\right) \leq \mathcal{R}\left(m_h - m_l\right) \triangleq \mathcal{E} < \infty.$$

In other words, the risk regularizer $\mathcal{R}$ is strictly positively uniformly bounded within $\mathfrak{R}^{\widetilde{F}}$.

[a] As usual, the case $P \equiv \infty$ is understood as a limit.

[b] Note that, by convexity, it always is true that $\underline{\nabla}\left(\mathcal{R}\left(z\right)\right)^p \equiv p\left(\mathcal{R}\left(z\right)\right)^{p-1}\underline{\nabla}\mathcal{R}\left(z\right)$.

Let us briefly comment on the various conditions of Assumption 5. First, conditions **C1** and **C3** reveal a probably fundamental trade-off between the size of the ℓ_2-norm of the random subgradient $\underline{\nabla}F\left(\boldsymbol{x}, \boldsymbol{W}\right)$, which may be thought of as a measure of the expansiveness of the random cost function $F\left(\cdot, \boldsymbol{W}\right)$, and the size of the *slope* of the random subgradient $p\left(\mathcal{R}\left(F\left(\cdot, \boldsymbol{W}\right) - \bullet\right)\right)^{p-1} \times \underline{\nabla}\mathcal{R}\left(F\left(\cdot, \boldsymbol{W}\right) - \bullet\right)$, which is directly related to the smoothness of the risk regularizer $\mathcal{R}$, as well as the smoothness of the distribution of $F\left(\cdot, \boldsymbol{W}\right)$. This trade-off is explicitly demonstrated through the following simple result, which presents at least three ways of increasing generality, for ensuring validity of condition **C3**, for certain values of the exponent pair (P, Q).

Proposition 4. (Ensuring Validity of C3) *Assume that, whenever $p > 1$, condition* **C4** *is satisfied. Then, the following statements are true:*

1) *Suppose that the p-th power of $\mathcal{R}$ is differentiable on $\mathfrak{R}^{\widetilde{F}}$, and that there is $D_{\mathcal{R},p} < \infty$, such that*

$$\left|\nabla\left(\mathcal{R}\left(y_1\right)\right)^p - \nabla\left(\mathcal{R}\left(y_1\right)\right)^p\right| \leq D_{\mathcal{R},p}\left|y_1 - y_2\right|, \quad \forall\left(y_1, y_2\right) \in \left[\mathfrak{R}^{\widetilde{F}}\right]^2. \tag{83}$$

Then, condition **C3** *is satisfied for every choice of $Q \in [P/(P-1), \infty]$, for every choice of $P \in [2, \infty]$.*

2) *Choose $\underline{\nabla}\mathcal{R} \equiv \mathcal{R}'_+$, and if $F^{(\cdot)}_{\boldsymbol{W}} : \mathbb{R} \rightarrow [0, 1]$ denotes the cdf of $F\left(\cdot, \boldsymbol{W}\right)$, suppose that there exists $D_{\widetilde{F}} < \infty$, such that*

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \left|F^{\boldsymbol{x}}_{\boldsymbol{W}}\left(y_1\right) - F^{\boldsymbol{x}}_{\boldsymbol{W}}\left(y_2\right)\right| \leq D_{\widetilde{F}}\left|y_1 - y_2\right|, \quad \forall\left(y_1, y_2\right) \in \left[\mathrm{cl}\left\{(m_l, m_h)\right\}\right]^2. \tag{84}$$

Then, condition **C3** *is satisfied for $Q \equiv 1$ (implying that $P \equiv \infty$), for every value of $p \in [1, \infty)$.*

3) *More generally, whenever $p \equiv 1$ and for any choice of $\mathcal{R}$, take $\underline{\nabla}\mathcal{R} \equiv \mathcal{R}'_{+}$, let $Y : \Omega \rightarrow \mathbb{R}$ be the random variable associated with $\mathcal{R}$, as in Theorem 2, and suppose that $F_{\boldsymbol{W}}^{(\cdot)}$ is continuous everywhere on $\mathbb{R}$ (not necessarily Lipschitz). Then, condition* **C3** *is satisfied for $Q \equiv 1$ (implying that $P \equiv \infty$) if and only if there exists some $D_{\mathcal{R}}^{\widetilde{F}} < \infty$, such that the Lipschitz-in-Expectation condition*

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \int \left| F_{\boldsymbol{W}}^{\boldsymbol{x}}(y + y_1) - F_{\boldsymbol{W}}^{\boldsymbol{x}}(y + y_2) \right| \mathrm{d}\mathcal{P}_Y(y) \leq D_{\mathcal{R}}^{\widetilde{F}} |y_1 - y_2|, \tag{85}$$

is satisfied, for all $(y_1, y_2) \in [\mathrm{cl}\{(m_l, m_h)\}]^2$. If $p > 1$, (85) is only sufficient for condition **C3**.

Proof of Proposition 4. See Section 7.4 (Appendix). ■

Proposition 4 demonstrates the versatility of condition **C3**, mainly relative to the choice of Q. First, observe that if $P \equiv 2$, then Q can be *anything* in $[2, \infty]$. If, for instance, we choose $Q \equiv \infty$ (this is most easiest to verify from a technical viewpoint; also see Remark 8 below), the *almost Lipschitz* assumption imposed by condition **C3** on the p-th power of the chosen risk regularizer $\mathcal{R}$ might be *severely restrictive*, depending on the value p. More specifically, a model satisfying (or required to satisfy) Assumption 5 for $Q \equiv \infty$ (in which case **C3** is *almost equivalent* to the respective condition in the first part of Proposition 4) *might not* allow for risk regularizers exhibiting *corner points*. Let us illustrate this by means of an example. Let $p \equiv 1$, choose $\mathcal{R}$ to be the upper-semideviation regularizer, that is, $\mathcal{R} \equiv (\cdot)_{+} \equiv \max\{\cdot, 0\}$, and consider the *linear* objective

$$F(\boldsymbol{x}, \boldsymbol{W}) \triangleq \boldsymbol{W}_1^{\boldsymbol{T}} \boldsymbol{x} + W \in \mathbb{R}, \tag{86}$$

where $\boldsymbol{W} \triangleq \left[\boldsymbol{W}_1^{\boldsymbol{T}}\ W\right]^{\boldsymbol{T}}$, where $\boldsymbol{W}_1 : \Omega \rightarrow \mathbb{R}^N$ constitutes an absolutely continuous random element almost everywhere in $[0,1]^N$, $\mathbb{E}\{\boldsymbol{W}_1\} \equiv \boldsymbol{\mu}$, and $W \sim \mathcal{N}(0,1)$. Then, we are interested in the *nonlinear (convex)*, risk-averse stochastic program

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{\mu}^{\boldsymbol{T}} \boldsymbol{x} + c\mathbb{E}\left\{ \left(\boldsymbol{W}_1^{\boldsymbol{T}} \boldsymbol{x} - \boldsymbol{\mu}^{\boldsymbol{T}} \boldsymbol{x} + W \right)_{+} \right\}, \tag{87}$$

for some closed, convex set $\mathcal{X}$. Note that, for every choice of $\mathcal{X}$ (compact or not), it is true that $m_l \equiv -\infty$ and $m_h \equiv +\infty$, since the random element W is unbounded. Thus, $\mathrm{cl}\{(m_l, m_h)\} \equiv \mathbb{R}$. Consequently, for this problem, condition **C3** (for $Q \equiv \infty$) demands the existence of a number $D < \infty$, such that[2]

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \operatorname*{ess\,sup}_{\omega \in \Omega} \left| \mathbb{1}_{\{F(\boldsymbol{x}, \boldsymbol{W}(\omega)) \geq y_1\}} - \mathbb{1}_{\{F(\boldsymbol{x}, \boldsymbol{W}(\omega)) \geq y_2\}} \right| \leq D |y_1 - y_2|, \tag{88}$$

for all $(y_1, y_2) \in \mathbb{R}^2$, or by definition of the essential supremum ([Bogachev, 2007], p. 250),

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \inf_{\mathscr{F} \ni \Omega' \subseteq \Omega : \mathcal{P}(\Omega') \equiv 1} \sup_{\omega \in \Omega'} \left| \mathbb{1}_{\{F(\boldsymbol{x}, \boldsymbol{W}(\omega)) \geq y_1\}} - \mathbb{1}_{\{F(\boldsymbol{x}, \boldsymbol{W}(\omega)) \geq y_2\}} \right| \leq D |y_1 - y_2|, \tag{89}$$

for all $(y_1, y_2) \in \mathbb{R}^2$. It is relatively easy to show that it is actually impossible for (89) to hold *uniformly in* $(y_1, y_2) \in \mathbb{R}^2$, for any possible choice of $D > 0$. Indeed, for simplicity, consider the symmetric "antidiagonal" case where $y_1 \equiv -y_2 \triangleq z > 0$. Then, for each fixed $\boldsymbol{x} \in \mathcal{X}$, it is true that

$$\left| \mathbb{1}_{\{F(\boldsymbol{x}, \boldsymbol{W}(\omega)) \geq y_1\}} - \mathbb{1}_{\{F(\boldsymbol{x}, \boldsymbol{W}(\omega)) \geq y_2\}} \right| \equiv \left| \mathbb{1}_{\{F(\boldsymbol{x}, \boldsymbol{W}(\omega)) \geq z\}} - \mathbb{1}_{\{F(\boldsymbol{x}, \boldsymbol{W}(\omega)) \geq -z\}} \right|$$

[2] In this case, we simply take $\underline{\nabla}\mathcal{R}(\cdot) \equiv \underline{\nabla}(\cdot)_{+} \equiv \mathbb{1}_{\{(\cdot) \geq 0\}}$

$$
\begin{aligned}
&= \mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W}(\omega))\in[-z,z)\}} \\
&\equiv \mathbb{1}_{\Omega_{\boldsymbol{x}}^z}(\omega), \quad \forall \omega \in \Omega,
\end{aligned} \tag{90}
$$

where the event $\Omega_{\boldsymbol{x}}^z \in \mathscr{F}$ is defined as

$$
\Omega_{\boldsymbol{x}}^z \triangleq \{\omega \in \Omega \,|\, F(\boldsymbol{x}, \boldsymbol{W}(\omega)) \in [-z, z)\}, \quad \forall (\boldsymbol{x}, z) \in \mathcal{X} \times \mathbb{R}_{++}, \tag{91}
$$

and where we emphasize that, due to our assumptions, it holds that $\mathcal{P}(\Omega_{\boldsymbol{x}}^z) > 0$, for every choice of $(\boldsymbol{x}, z) \in \mathcal{X} \times \mathbb{R}_{++}$. Consequently, for an arbitrary event $\Omega' \subseteq \Omega$ such that $\mathcal{P}(\Omega') \equiv 1$, we have

$$
\begin{aligned}
&\sup_{\omega\in\Omega'} \left|\mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W}(\omega))\geq y_1\}} - \mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W}(\omega))\geq y_2\}}\right| \equiv \sup_{\omega\in\Omega'} \mathbb{1}_{\Omega_{\boldsymbol{x}}^z}(\omega) \\
&\quad \equiv \max\left\{\sup_{\omega\in\Omega'\cap\Omega_{\boldsymbol{x}}^z} \mathbb{1}_{\Omega_{\boldsymbol{x}}^z}(\omega), \sup_{\omega\in\Omega'\cap(\Omega_{\boldsymbol{x}}^z)^c} \mathbb{1}_{\Omega_{\boldsymbol{x}}^z}(\omega)\right\} = \max\{1, 0\} \equiv 1, \quad \forall (\boldsymbol{x}, z) \in \mathcal{X} \times \mathbb{R}_{++},
\end{aligned} \tag{92}
$$

which implies in particular that, unless $z \geq (2D)^{-1}$,

$$
1 \equiv \sup_{\boldsymbol{x}\in\mathcal{X}} \operatorname*{ess\,sup}_{\omega\in\Omega} \left|\mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W}(\omega))\geq y_1\}} - \mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W}(\omega))\geq y_2\}}\right| > D|y_1 - y_2| \equiv 2Dz, \tag{93}
$$

for any fixed and finite choice of $D > 0$, thus immediately disproving uniform validity of (89). Of course, the apparent impossibility of (89) may be seen as a consequence of the fact that, for *any fixed* $z \in \mathbb{R}$, the function $\mathbb{1}_{\{z\geq y\}}$ *is discontinuous when* $y \in \mathbb{R}$.

For this example, it is also possible to show that condition **C3** is impossible to hold for $Q \equiv 2$, as well, which corresponds to the smallest choice of Q, when $P \equiv 2$. Indeed, in this case, condition **C3** demands the existence of a number $D < \infty$, such that

$$
\sup_{\boldsymbol{x}\in\mathcal{X}} \left\|\left|\mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W})\geq y_1\}} - \mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W})\geq y_2\}}\right|\right\|_{\mathcal{L}_2} \leq D|y_1 - y_2|, \tag{94}
$$

for all $(y_1, y_2) \in \mathbb{R}^2$. For every $\boldsymbol{x} \in \mathcal{X}$ and for every pair $(y_1, y_2) \in \mathbb{R}^2$, we may write

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W})\geq y_1\}} - \mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W})\geq y_2\}}\right)^2\right\} \\
&\quad \equiv \mathbb{E}\left\{\mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W})\geq y_1\}} + \mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W})\geq y_2\}} - 2\mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W})\geq y_1\}}\mathbb{1}_{\{F(\boldsymbol{x},\boldsymbol{W})\geq y_2\}}\right\} \\
&\quad = \mathcal{P}(F(\boldsymbol{x}, \boldsymbol{W}) \geq y_1) + \mathcal{P}(F(\boldsymbol{x}, \boldsymbol{W}) \geq y_2) - 2\mathcal{P}(F(\boldsymbol{x}, \boldsymbol{W}) \geq \max\{y_1, y_2\}) \\
&\quad = \mathcal{P}(F(\boldsymbol{x}, \boldsymbol{W}) \geq \min\{y_1, y_2\}) - \mathcal{P}(F(\boldsymbol{x}, \boldsymbol{W}) \geq \max\{y_1, y_2\}) \\
&\quad \equiv |\mathcal{P}(F(\boldsymbol{x}, \boldsymbol{W}) \geq y_1) - \mathcal{P}(F(\boldsymbol{x}, \boldsymbol{W}) \geq y_2)| \\
&\quad \equiv \left|F_{\boldsymbol{W}}^{\boldsymbol{x}}(y_1) - F_{\boldsymbol{W}}^{\boldsymbol{x}}(y_2)\right|.
\end{aligned} \tag{95}
$$

Therefore, for (94) to hold, it must be true that, for every $\boldsymbol{x} \in \mathcal{X}$ and for every $(y_1, y_2) \in \mathbb{R}^2$,

$$
\sqrt{\left|F_{\boldsymbol{W}}^{\boldsymbol{x}}(y_1) - F_{\boldsymbol{W}}^{\boldsymbol{x}}(y_2)\right|} \leq D|y_1 - y_2| \iff \left|F_{\boldsymbol{W}}^{\boldsymbol{x}}(y_1) - F_{\boldsymbol{W}}^{\boldsymbol{x}}(y_2)\right| \leq D^2|y_1 - y_2|^2, \tag{96}
$$

implying that $F_{\boldsymbol{W}}^{(\cdot)}$ *must be constant on* $\mathbb{R}$. This is absurd, however, since $F_{\boldsymbol{W}}^{(\cdot)}$ is a proper cdf.

On the other hand, the second and third parts of Proposition 4 show that, if the random variable $\|\underline{\nabla} F(\cdot, \boldsymbol{W})\|_2$ can be afforded to be uniformly in $\mathcal{Z}_\infty$ (for $Q \equiv 1$), the choice of the risk regularizer

$\mathcal{R}$ may be completely unconstrained, as long as the Borel pushforward of $F(\cdot, \boldsymbol{W})$ is uniformly well behaved, in the sense of either (84), or, more generally, (85). Of course, this constitutes a major improvement compared to the case where $Q \equiv \infty$, discussed above, at least in regard to the shape of $\mathcal{R}$. For example, in our previous example, it is also true that

$$\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) = \nabla F(\boldsymbol{x}, \boldsymbol{W}) = \boldsymbol{W}_1 \in [0,1]^N, \quad \mathcal{P}-a.e., \tag{97}$$

and, hence,

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \left\| \|\nabla F(\boldsymbol{x}, \boldsymbol{W})\|_2 \right\|_{\mathcal{L}_\infty} \equiv \sup_{\boldsymbol{x} \in \mathcal{X}} \operatorname*{ess\,sup}_{\omega \in \Omega} \|\boldsymbol{W}_1\|_2 \leq \sqrt{N}. \tag{98}$$

In this case, condition **C3** is loosened to

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \left\| \left| \mathbb{1}_{\{F(\boldsymbol{x}, \boldsymbol{W}) \geq y_1\}} - \mathbb{1}_{\{F(\boldsymbol{x}, \boldsymbol{W}) \geq y_2\}} \right| \right\|_{\mathcal{L}_1} \leq D |y_1 - y_2|, \tag{99}$$

for all $(y_1, y_2) \in \mathbb{R}^2$, whose validity may now be verified for appropriate choices of the distribution $\mathcal{P}_{\boldsymbol{W}}$, as Proposition 4 suggests.

We have seen that the price to be paid for choosing a lower value for the exponent Q is a potentially stronger requirement on the size of the random subgradient $\underline{\nabla} F(\cdot, \boldsymbol{W})$ (condition **C1**). Still, such a requirement is relatively easy to satisfy for many interesting models, other than our particular example discussed above. For example, in the extreme case where $P \equiv \infty$, **C1** will indeed be satisfied in cases involving a compact feasible set $\mathcal{X}$ and a Borel measure $\mathcal{P}_{\boldsymbol{W}}$ with bounded essential support (recall that, by assumption, the domain of $F(\cdot, \boldsymbol{W})$ is the whole Euclidean space $\mathbb{R}^N$).

Remark 8. We would like to emphasize that, for $P \equiv \infty$, a sometimes more easily verifiable sufficient condition for **C1** is the existence of an event $\Omega_{\widetilde{F}} \subseteq \Omega$, with $\mathcal{P}\left(\Omega_{\widetilde{F}}\right) \equiv 1$, as well as a number $G < \infty$ such that

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \sup_{\omega \in \Omega_{\widetilde{F}}} \|\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}(\omega))\|_2 \equiv \sup_{\omega \in \Omega_{\widetilde{F}}} \sup_{\boldsymbol{x} \in \mathcal{X}} \|\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}(\omega))\|_2 \leq G. \tag{100}$$

This follows by definition of the essential supremum ([Bogachev, 2007], p. 250); indeed, we may write

$$\begin{aligned} \sup_{\boldsymbol{x} \in \mathcal{X}} \left\| \|\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W})\|_2 \right\|_{\mathcal{L}_\infty} &\triangleq \sup_{\boldsymbol{x} \in \mathcal{X}} \operatorname*{ess\,sup}_{\omega \in \Omega} \|\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}(\omega))\|_2 \\ &\triangleq \sup_{\boldsymbol{x} \in \mathcal{X}} \inf_{\mathscr{F} \ni \Omega' \subseteq \Omega : \mathcal{P}(\Omega') \equiv 1} \sup_{\omega \in \Omega'} \|\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}(\omega))\|_2 \\ &\leq \sup_{\boldsymbol{x} \in \mathcal{X}} \sup_{\omega \in \widetilde{\Omega}} \|\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}(\omega))\|_2, \end{aligned} \tag{101}$$

where $\widetilde{\Omega} \subseteq \Omega$ is *any* measurable set in $\mathscr{F}$, such that $\mathcal{P}\left(\widetilde{\Omega}\right) \equiv 1$. This technical fact was previously utilized in (98). ■

Let us also comment on the hard boundedness condition **C4**, which is assumed *only when* $p > 1$. Although condition **C4** may not impose significant restrictions on the choice of $\mathcal{R}$, it *does* require that $F(\cdot, \boldsymbol{W})$ is *uniformly bounded* on $\mathcal{X}$, almost everywhere on Ω relative to $\mathcal{P}$. This assumption is explicitly made in condition **C4** mainly for analytical tractability, and is due to the slightly more complicated form of the gradient $\nabla \phi^{\widetilde{F}}$ (see Lemma 1). Without **C4**, asymptotic analysis of the *MESSAGE*^{p} algorithm becomes unnecessarily and uninsightfully complicated, when p is chosen greater than one. Still, uniform boundedness of $F(\cdot, \boldsymbol{W})$ may be verified in many common optimization settings, such as when $\mathcal{X}$ is compact and $\mathcal{P}_{\boldsymbol{W}}$ has bounded essential support (see also the discussion above), or in case F is itself uniformly bounded on $\mathcal{X} \times \mathbb{R}^M$.

Another important structural reason for imposing condition **C4** is that for every choice of $p > 1$, we have implicitly assumed that $F(\cdot, \boldsymbol{W}) \in \mathcal{Z}_q$, for some $q \geq p$, so that problem (51) is well defined. Thus, choosing larger values for p implies that $F(\cdot, \boldsymbol{W})$ behaves more or less like a bounded function (*pointwise on* $\mathcal{X}$). Therefore, condition **C4** may be regarded as an easy way of exploiting this approximate boundedness, compared to the imposition of integral $\mathcal{L}_q$-norm constraints, which are more complicated and harder to handle.

Nevertheless, there are important cases where the choice of $\mathcal{R}$ might make it very difficult to guarantee that $\varepsilon \equiv \mathcal{R}(m_l - m_h) > 0$, even if m_l and m_h are finite. For example, simply take $\mathcal{R}(\cdot) \equiv (\cdot)_+$ (note that $m_l < m_h$, by definition); in this case, $\varepsilon \equiv 0$. Fortunately, there is a simple, cheap-trick remedy to this technical issue. For fixed slack $\eta > 0$, consider a function $\mathcal{R}_\eta : \mathbb{R} \to \mathbb{R}$, defined as

$$\mathcal{R}_\eta(x) \triangleq \mathcal{R}(x) + \eta, \quad \forall x \in \mathbb{R}. \tag{102}$$

It can be readily verified that $\mathcal{R}_\eta$ is a valid risk regularizer and may be seen as a variable, *lower-biased* version of $\mathcal{R}$. Then, problem (51) is replaced by its *slack-adjusted* version

$$\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\} + c\left\|\mathcal{R}_\eta(F(\boldsymbol{x}, \boldsymbol{W}) - \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\})\right\|_{\mathcal{L}_p} \\ \text{subject to} & \boldsymbol{x} \in \mathcal{X} \end{array}, \tag{103}$$

and that it is always true that $\varepsilon \geq \eta$, satisfying the respective requirement of condition **C4**. One then hopes that, as $\eta \to 0$, (103) becomes increasingly equivalent to (51). The size of η should be chosen so that a certain trade-off between algorithmic stability and closeness to the original problem (51) is satisfied.

Lastly, condition **C2** constitutes a common restriction of SSD-type algorithms (either compositional or not) [Kushner and Yin, 2003, Shapiro et al., 2014, Wang et al., 2017, Yang et al., 2018], and will also be made here, without any further comment.

Next, we will exploit Assumption 5, in order to show asymptotic consistency for Algorithm 1, in a strong, pathwise sense.

4.4.2 Pathwise Convergence of the *MESSAGE*^{p} Algorithm

Proving convergence of Algorithm 1 will be based on the so-called *T-Level Almost-Supermartingale Convergence Lemma* by Yang, Wang & Fang [Yang et al., 2018], presented below. This is an inductive generalization of the *Coupled Almost-Supermartingale Convergence Lemma* by Wang & Bertsekas [Wang and Bertsekas, 2016, Wang et al., 2017], which in turn generalizes the well-known *Almost-Supermartingale Convergence Lemma* by Robbins and Siegmund [Robbins and Siegmund, 1971].

Lemma 2. **(T-Level Almost-Supermartingale Convergence Lemma [Yang et al., 2018])** *Let* $\{\xi^n\}_{n\in\mathbb{N}}$, $\{\eta^n\}_{n\in\mathbb{N}}$, $\{\zeta^{n,j}\}_{n\in\mathbb{N}}$, $\{\theta^{n,j}\}_{n\in\mathbb{N}}$, *for* $j \in \mathbb{N}^+_{T-1}$, $\{u^{n,j}\}_{n\in\mathbb{N}}$ *and* $\{\mu^{n,j}\}_{n\in\mathbb{N}}$, *for* $j \in \mathbb{N}^+_T$, *be nonnegative random sequences on* $(\Omega, \mathscr{F})$, *and consider the global filtration* $\{\mathscr{G}^n \subseteq \mathscr{F}\}_{n\in\mathbb{N}}$, *where*

$$\mathscr{G}^n \triangleq \sigma\left\{\xi^i, \eta^i, \zeta^{i,j}, \theta^{i,j}, u^{i,j}, \mu^{i,j}, u^{i,T}, \mu^{i,T}, \ \forall j \in \mathbb{N}^+_{T-1} \text{ and } \forall i \in \mathbb{N}_n\right\}. \tag{104}$$

Let $c_j > 0, j \in \mathbb{N}_T^+$ and suppose that

$$\mathbb{E}\left\{\xi^{n+1}\middle|\mathscr{G}^n\right\} \leq (1+\eta^n)\,\xi^n - u^{n,T} + \sum_{j\in\mathbb{N}_{T-1}^+} c_j\theta^{n,j}\zeta^{n,j} + \mu^{n,T} \quad \textit{and} \tag{105}$$

$$\mathbb{E}\left\{\zeta^{n+1,j}\middle|\mathscr{G}^n\right\} \leq \left(1-\theta^{n,j}\right)\zeta^{n,j} - u^{n,j} + \mu^{n,j}, \quad \forall j \in \mathbb{N}_{T-1}^+, \tag{106}$$

for all $n \in \mathbb{N}$, and that $\sum_{n\in\mathbb{N}}\eta^n < \infty$, $\sum_{n\in\mathbb{N}}\mu^{n,j} < \infty$, for all $j \in \mathbb{N}_T^+$, all almost everywhere relative to $\mathcal{P}$. Then, there exist random variables ξ_* and ζ_*^j, $j \in \mathbb{N}_{T-1}^+$ such that $\xi^n \underset{n\to\infty}{\longrightarrow} \xi_*$ and $\zeta^{n,j} \underset{n\to\infty}{\longrightarrow} \zeta_*^j$, for all $j \in \mathbb{N}_{T-1}^+$ and $\sum_{n\in\mathbb{N}} u^{n,j} < \infty$, for all $j \in \mathbb{N}_T^+$, $\sum_{n\in\mathbb{N}}\theta^{n,j}\zeta^{n,j} < \infty$, for all $j \in \mathbb{N}_{T-1}^+$, all almost everywhere relative to $\mathcal{P}$.

Our proof roadmap is similar to that presented in, say, [Wang et al., 2017, Yang et al., 2018], and is somewhat standard in the literature of stochastic approximation, in general. Before proceeding, let us define the filtration $\{\mathscr{D}^n \subseteq \mathscr{F}\}_{n\in\mathbb{N}}$, generated from *all data observed so far, by both the user and the $\mathcal{SO}$*, with each sub σ-algebra $\mathscr{D}^n$ given by

$$\mathscr{D}^n \triangleq \sigma\left\{\boldsymbol{x}^0,\ldots,\boldsymbol{x}^n, y^0,\ldots,y^n, z^0,\ldots,z^n, \boldsymbol{W}_1^0,\ldots,\boldsymbol{W}_1^n, \boldsymbol{W}_2^0,\ldots,\boldsymbol{W}_2^n\right\}, \quad \forall n \in \mathbb{N}. \tag{107}$$

Also, for the sake of clarity, if $\mathscr{C}$ is some sub σ-algebra of $\mathscr{F}$, we will employ the more compact notation $\mathbb{E}\{\cdot\,|\mathscr{C}\} \equiv \mathbb{E}_{\mathscr{C}}\{\cdot\}$, especially for larger expressions involving conditional expectations.

Our first basic result follows, characterizing the rate of decay of the squared $\mathcal{L}_2$-norm of the inter-iteration error $\boldsymbol{x}^{n+1} - \boldsymbol{x}^n$, relative to $\mathscr{D}^n$.

Lemma 3. (Deep SA Level: Iterate Increment Growth) *Let Assumption 5 be in effect and define a constant*

$$\mathsf{R}_p \equiv \begin{cases} 1, & \textit{if } p \equiv 1 \\ \left(\dfrac{\mathcal{E}}{\varepsilon}\right)^{p-1} & \textit{if } p > 1 \end{cases}. \tag{108}$$

Then, for every $p \in [1,\infty)$, the process $\{\boldsymbol{x}^n\}_{n\in\mathbb{N}}$ generated by the MESSAGE^p algorithm satisfies

$$\mathcal{O}\left(\alpha_n^2\right) \equiv \mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^n\right\|_2^2\right\} \leq \alpha_n^2\left(2c\mathsf{R}_p+1\right)^2 G^2, \quad \forall n \in \mathbb{N}, \tag{109}$$

almost everywhere relative to $\mathcal{P}$.

Proof of Lemma 3. See Section 7.5 (Appendix). ∎

Exploiting Lemma 3, one may also prove the following result, which will be useful later in our analysis. The proof is omitted, since it is essentially provided in ([Wang et al., 2017], Supplementary Material, Section G.1).

Lemma 4. (Iterate Increment Summability) *Let Assumption 5 be in effect. Also, consider a sequence $\{\delta_n > 0\}_{n\in\mathbb{N}}$, such that $\sum_{n\in\mathbb{N}}\alpha_n^2\delta_n^{-1} < \infty$. Then, the iterate process $\{\boldsymbol{x}^n\}_{n\in\mathbb{N}}$ generated by the MESSAGE^p algorithm satisfies*

$$\sum_{n\in\mathbb{N}}\delta_n^{-1}\mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^n\right\|_2^2\right\} < \infty, \tag{110}$$

almost everywhere relative to $\mathcal{P}$.

Let us now consider Borel measurable functions $\mathcal{S}^{\widetilde{F}}: \mathcal{X} \to \mathbb{R}$, $\mathcal{D}^{\widetilde{F}}: \mathcal{X} \times \mathbb{R} \to \mathbb{R}_+$ and $\mathcal{D}^{\widetilde{F}}: \mathbb{R} \to \mathbb{R}_+$, defined as[3]

$$\mathcal{S}^{\widetilde{F}}(\boldsymbol{x}) \triangleq \mathbb{E}\left\{F\left(\boldsymbol{x}, \boldsymbol{W}'\right)\right\}, \tag{111}$$

$$\mathcal{D}^{\widetilde{F}}(\boldsymbol{x}, y) \triangleq \mathbb{E}\Big\{\left(\mathcal{R}\left(F(\boldsymbol{x}, \boldsymbol{W}') - y\right)\right)^p\Big\} \quad \text{and} \tag{112}$$

$$\mathcal{D}^{\widetilde{F}}(\boldsymbol{x}) \triangleq \mathbb{E}\Big\{\left(\mathcal{R}\left(F(\boldsymbol{x}, \boldsymbol{W}') - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x})\right)\right)^p\Big\}. \tag{113}$$

where the random element $\boldsymbol{W}': \Omega \to \mathbb{R}^M$ is distributed according to the Borel measure $\mathcal{P}_{\boldsymbol{W}}$, and is arbitrarily taken as independent of the *whole* filtration $\{\mathscr{D}^n\}_{n \in \mathbb{N}}$. For instance, for each $n \in \mathbb{N}$, $\boldsymbol{W}'$ may be substituted by the information stream $\boldsymbol{W}_2^{n+1}$, which is by assumption statistically independent of the sub σ-algebra $\sigma\{\mathscr{D}^n\}$. The main purpose of the auxiliary expectation functions $\mathcal{S}^{\widetilde{F}}$ and $\mathcal{D}^{\widetilde{F}}$ is convenience.

Utilizing $\mathcal{S}^{\widetilde{F}}$, the behavior of the running approximation error $y^{n+1} - \mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}\right)$ may be analyzed in a similar manner as the respective quantity of Lemma 3. The relevant result follows.

Lemma 5. (First SA Level: Error Growth) *Let Assumption 5 be in effect. Also, let $\beta_n \in (0, 1]$, for all $n \in \mathbb{N}$. Then, the composite process $\{(\boldsymbol{x}^n, y^n)\}_{n \in \mathbb{N}}$ generated by the MESSAGE^p algorithm satisfies*

$$\mathbb{E}_{\mathscr{D}^n}\left\{\left|y^{n+1} - \mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}\right)\right|^2\right\} \le (1 - \beta_n)\left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|^2 + \beta_n^{-1} 2G^2 \mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\right\|_2^2\right\} + \beta_n^2 2V, \tag{114}$$

for all $n \in \mathbb{N}$, almost everywhere relative to $\mathcal{P}$.

Proof of Theorem 5. See Section 7.6 (Appendix). ∎

At this point, let us make the following additional assumption, related with the initial conditions of the first and second SA levels of the *MESSAGE^p* algorithm.

Assumption 6. (Initial Values) *Whenever $p > 1$, y^0, β_0 and z^0, γ_0 are chosen such that*

$$\begin{cases} \text{either} \quad y^0 \in [m_l, m_h], & \text{or} \quad \beta_0 \equiv 1 \\ \text{either} \quad z^0 \in [\varepsilon^p, \mathcal{E}^p], & \text{or} \quad \gamma_0 \equiv 1 \end{cases}. \tag{115}$$

It is trivial to see that Assumption 6 *can always be satisfied*, one way or another. However, choosing $\beta_0 \triangleq 1$ and $\gamma_0 \triangleq 1$ might be advantageous in practice, especially for cases where specific values of the constants $m_l, m_h, \varepsilon, \mathcal{E}$ are unknown. In our analysis, Assumption 6 will help us guarantee uniform boundedness of the iterates $\{y^n\}_{n \in \mathbb{N}}$ and $\{z^n\}_{n \in \mathbb{N}}$ of the first and second SA levels of the *MESSAGE^p* algorithm, respectively, whenever the semideviation order is chosen greater than unity, that is, $p > 1$ (see Lemma 11 in Section 7.3 (Appendix)).

Now, similarly to Lemma 5, the growth of the running approximation error $z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)$ may be characterized as follows.

[3] Without any risk of confusion, we use the same name $\mathcal{D}^{\widetilde{F}}$ to refer to the two very similar functions (112) and (113). The two functions will be distinguished by their different number of arguments.

Lemma 6. (Second SA Level: Error Growth) *Let Assumptions 5 and 6 be in effect. Also, choose $p > 1$, and let $\beta_n \in (0,1]$, $\gamma_n \in (0,1]$, for all $n \in \mathbb{N}$. Then, the composite process $\{(\boldsymbol{x}^n, y^n, z^n)\}_{n\in\mathbb{N}}$ generated by the MESSAGE^p algorithm satisfies*

$$\mathbb{E}_{\mathscr{D}^n}\left\{\left|z^{n+1} - \mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}, y^{n+1}\right)\right|^2\right\} \le (1-\gamma_n)\left|z^n - \mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^n, y^n\right)\right|^2$$
$$+ \gamma_n^{-1} 16 G^2 \mathcal{E}^{2p-2} p^2 \mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\right\|_2^2\right\} + \beta_n^2 \gamma_n^{-1} 4\mathcal{E}^{2p-2} p^2 (m_h - m_l)^2 + \gamma_n^2 2\mathcal{E}^{2p}, \quad (116)$$

for all $n \in \mathbb{N}$, almost everywhere relative to $\mathcal{P}$.

Proof of Theorem 6. See Section 7.7 (Appendix). ■

Let $\boldsymbol{x}^* \in \mathcal{X}$ be an optimal solution of problem (51), assuming such solution exists. We now characterize the evolution of optimality error $\boldsymbol{x}^{n+1} - \boldsymbol{x}^*$, showing that the quantity $\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\right\|_2^2$ is an almost-supermartingale nonnegative sequence, of the form (105) in Lemma 2.

Lemma 7. (Third SA Level: Optimality Error Growth) *Let Assumptions 5 and 6 be in effect, let $\beta_n \in (0,1]$, $\gamma_n \in (0,1]$, for all $n \in \mathbb{N}$, and define the constant*

$$\mathsf{B}_p \triangleq \begin{cases} D, & \text{if } p \equiv 1 \\ \left(1 + \mathcal{E}^{p-1} p\right) \max\left\{\left(\dfrac{1}{\varepsilon}\right)^{p-1} D, (p-1)\dfrac{\mathcal{E}^{p-1}}{\varepsilon^{2p-1}}\right\}, & \text{if } p > 1 \end{cases} < \infty. \quad (117)$$

Also, suppose that $\mathcal{X}^* \triangleq \arg\min_{\boldsymbol{x}\in\mathcal{X}} \phi^{\widetilde{F}}(\boldsymbol{x}) \neq \emptyset$ and consider any $\boldsymbol{x}^* \in \mathcal{X}^*$. Then, the composite process $\{(\boldsymbol{x}^n, y^n, z^n)\}_{n\in\mathbb{N}}$ generated by the MESSAGE^p algorithm satisfies

$$\mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\right\|_2^2\right\}$$
$$\le \left(1 + 4\mathsf{B}_p^2 G^2 c^2 \left(\frac{\alpha_n^2}{\beta_n} + \frac{\alpha_n^2}{\gamma_n}\mathbb{1}_{\{p>1\}}\right)\right)\left\|\boldsymbol{x}^n - \boldsymbol{x}^*\right\|_2^2 + \alpha_n^2 \left(2c\mathsf{R}_p + 1\right)^2 G^2 - 2\alpha_n\left(\phi^{\widetilde{F}}(\boldsymbol{x}^n) - \phi_*^{\widetilde{F}}\right)$$
$$+ \beta_n \left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|^2 + \gamma_n \left|z^n - \mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^n, y^n\right)\right|^2 \mathbb{1}_{\{p>1\}}, \quad (118)$$

for all $n \in \mathbb{N}$, almost everywhere relative to $\mathcal{P}$, where $\phi_*^{\widetilde{F}} \in \mathbb{R}$ is the optimal value of problem (51).

Proof of Lemma 7. See Section 7.8 (Appendix). ■

Under the proposed Assumption 5, pathwise convergence of the *MESSAGE*^{p} algorithm is established next, in a rather strong sense. Here, we directly invoke Lemma 2.

Theorem 3. (Pathwise Convergence of the *MESSAGE*^{p} Algorithm) *Let Assumptions 5 and 6 be in effect, and let $\beta_n \in (0,1]$, $\gamma_n \in (0,1]$, for all $n \in \mathbb{N}$. Whenever $p \equiv 1$, suppose that*

$$\sum_{n\in\mathbb{N}} \alpha_n \equiv \infty \quad \textit{and} \quad \sum_{n\in\mathbb{N}} \alpha_n^2 + \beta_n^2 + \frac{\alpha_n^2}{\beta_n} < \infty, \quad (119)$$

whereas, whenever $p > 1$, suppose additionally that

$$\sum_{n\in\mathbb{N}} \gamma_n^2 + \frac{\alpha_n^2}{\gamma_n} + \frac{\beta_n^2}{\gamma_n} < \infty. \tag{120}$$

Then, as long as $\mathcal{X}^* \equiv \arg\min_{\boldsymbol{x}\in\mathcal{X}} \phi^{\widetilde{F}}(\boldsymbol{x}) \neq \emptyset$, the process $\{\boldsymbol{x}^n\}_{n\in\mathbb{N}}$ generated by the MESSAGE^p algorithm satisfies[4]

$$\mathcal{P}\left(\left\{\omega\in\Omega \,\middle|\, \exists\, \boldsymbol{x}^*(\omega)\in\mathcal{X}^* \text{ such that } \boldsymbol{x}^n(\omega) \underset{n\to\infty}{\longrightarrow} \boldsymbol{x}^*(\omega)\right\}\right) \equiv 1. \tag{121}$$

In other words, almost everywhere relative to $\mathcal{P}$, the process $\{\boldsymbol{x}^n\}_{n\in\mathbb{N}}$ converges to a random point in the set of optimal solutions of (63).

Proof of Theorem 3. We present the proof assuming that $p > 1$. If $p \equiv 1$, the proof is almost the same, albeit simpler. Under the assumptions of the theorem, Lemmata 5, 4 and 6 imply that

$$\sum_{n\in\mathbb{N}} \beta_n^{-1} 2G^2 \mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\right\|_2^2\right\} < \infty, \tag{122}$$

and

$$\sum_{n\in\mathbb{N}} \gamma_n^{-1} 16 G^2 \mathcal{E}^{2p-2} p^2 \mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\right\|_2^2\right\} < \infty, \tag{123}$$

whereas it is also true that

$$\sum_{n\in\mathbb{N}} \alpha_n^2 (2c\mathsf{R}_p + 1)^2 G^2 < \infty, \quad \sum_{n\in\mathbb{N}} \beta_n^2 \gamma_n^{-1} 4\mathcal{E}^{2p-2} p^2 (m_h - m_l)^2 < \infty, \tag{124}$$

$$\sum_{n\in\mathbb{N}} \beta_n^2 V < \infty \quad \text{and} \quad \sum_{n\in\mathbb{N}} \gamma_n^2 2\mathcal{E}^{2p} < \infty, \tag{125}$$

as well. Therefore, we may apply Lemma 2 with the identifications

$$\begin{aligned}
\xi^n &\equiv \left\|\boldsymbol{x}^n - \boldsymbol{x}^*\right\|_2^2, & \eta^n &\equiv 4\mathsf{B}_p^2 G^2 c^2 \left(\frac{\alpha_n^2}{\beta_n} + \frac{\alpha_n^2}{\gamma_n}\right), & u^{n,3} &\equiv 2\alpha_n\left(\phi^{\widetilde{F}}(\boldsymbol{x}^n) - \phi_*^{\widetilde{F}}\right) \\
\zeta^{n,1} &\equiv \left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|^2, & \theta^{n,1} &\equiv \beta_n, & u^{n,1} &\equiv 0, \\
\zeta^{n,2} &\equiv \left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|^2, & \theta^{n,2} &\equiv \gamma_n, & u^{n,2} &\equiv 0,
\end{aligned}$$

$$\mu^{n,3} \equiv \alpha_n^2 (2c\mathsf{R}_p + 1)^2 G^2$$

$$\mu^{n,1} \equiv \beta_n^{-1} 2G^2 \mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\right\|_2^2\right\} + \beta_n^2 V,$$

$$\mu^{n,2} \equiv \gamma_n^{-1} 16 G^2 \mathcal{E}^{2p-2} p^2 \mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\right\|_2^2\right\} + \beta_n^2 \gamma_n^{-1} 4\mathcal{E}^{2p-2} p^2 (m_h - m_l)^2 + \gamma_n^2 2\mathcal{E}^{2p},$$

$$\mathscr{G}^n \equiv \mathscr{D}^n,$$

and with $c_1 \equiv c_2 \equiv 1$. The rest of the proof is identical to ([Wang et al., 2017], Proof of Theorem 1 (a)), or ([Yang et al., 2018], Proof of Theorem 2.1 (a) & Proof of Lemma 2.5). ■

[4]Note that (121) is meaningful only if the involved outcome set is an event, that is, $\mathscr{F}$-measurable. In our case, such measurability follows by completeness of the base space $(\Omega, \mathscr{F}, \mathcal{P})$.

Remark 9. Note that, in both ([Wang et al., 2017], Theorem 1) and ([Yang et al., 2018], Theorem 2.1), in addition to the stepsize requirements of Theorem 3, it is assumed that

$$\sum_{n\in\mathbb{N}^+} \beta_n \equiv \infty \quad \text{and} \quad \sum_{n\in\mathbb{N}^+} \gamma_n \equiv \infty, \tag{126}$$

as well. To be best of our knowledge, however, although they do not hurt, *none* of the aforementioned (non)summability conditions are necessary in order to guarantee pathwise convergence of the *MESSAGE*^{p} algorithm, and the same is true for the general purpose *SCGD* algorithm of [Wang et al., 2017] (see statement and proof of Theorem 1 in [Wang et al., 2017]) and *T-SCGD* algorithm of [Yang et al., 2018] (see statement and proof of Theorem 2.1 in [Yang et al., 2018]). ■

Besides the classical Robbins-Monro (RM) conditions [Robbins and Monro, 1951] on the stepsize sequence $\{\alpha_n\}_{n\in\mathbb{N}^+}$ and the square summability conditions on $\{\beta_n\}_{n\in\mathbb{N}^+}$and $\{\gamma_n\}_{n\in\mathbb{N}^+}$, and in agreement with ([Yang et al., 2018], Theorem 2.1), Theorem 3 demands that

$$\sum_{n\in\mathbb{N}^+} \frac{\alpha_n^2}{\beta_n} < \infty, \quad \sum_{n\in\mathbb{N}^+} \frac{\alpha_n^2}{\gamma_n} < \infty \quad \text{and} \quad \sum_{n\in\mathbb{N}^+} \frac{\beta_n^2}{\gamma_n} < \infty. \tag{127}$$

These stepsize requirements imposed by Theorem 3 might seem quite complicated. Nevertheless, there are lots of viable choices for the sequences $\{\alpha_n\}_{n\in\mathbb{N}^+}$, $\{\beta_n\}_{n\in\mathbb{N}^+}$and $\{\gamma_n\}_{n\in\mathbb{N}^+}$, satisfying the conditions in (127).

Let us present a simple, but instructive example. Take, for every $n \in \mathbb{N}^+$,

$$\alpha_n \equiv \frac{1}{n^{\tau_1}}, \quad \beta_n \equiv \frac{1}{n^{\tau_2}} \quad \text{and} \quad \gamma_n \equiv \frac{1}{n^{\tau_3}}, \tag{128}$$

for some $\tau_j \in (0.5, 1]$, for $j \in \{1, 2, 3\}$. In such a case, the RM conditions are automatically satisfied for all three stepsizes since

$$\sum_{n\in\mathbb{N}^+} \frac{1}{n^{\tau_j}} \equiv \infty \quad \text{and} \quad \sum_{n\in\mathbb{N}^+} \frac{1}{n^{2\tau_j}} < \infty, \quad j \in \{1, 2, 3\}. \tag{129}$$

We would like to see how we may choose τ_j, $j \in \{1, 2, 3\}$, such that the summability conditions in (127) are satisfied. First, we demand that

$$\sum_{n\in\mathbb{N}^+} \frac{\alpha_n^2}{\beta_n} \equiv \sum_{n\in\mathbb{N}^+} \frac{1}{n^{2\tau_1-\tau_2}} < \infty, \tag{130}$$

which equivalently yields

$$1 < 2\tau_1 - \tau_2 \quad \iff \quad \tau_2 < 2\tau_1 - 1. \tag{131}$$

Also, for the preceding inequality to yield a feasible lower bound for τ_2, it must be true that

$$2\tau_1 - 1 > \frac{1}{2} \quad \iff \quad \tau_1 > \frac{3}{4}. \tag{132}$$

Consequently, we obtain the conditions

$$\frac{3}{4} < \tau_1 \leq 1 \quad \text{and} \quad \frac{1}{2} < \tau_2 < 2\tau_1 - 1. \tag{133}$$

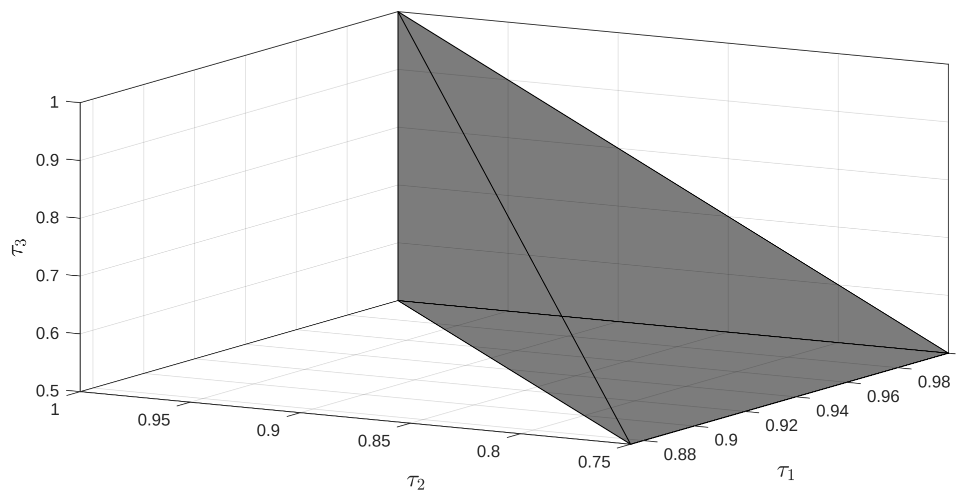

Figure 4.1: A graphical representation of the stepsize constraint set (138)-(140).

Similarly, for the second condition of (127)

$$\sum_{n\in\mathbb{N}^+} \frac{\alpha_n^2}{\gamma_n} \equiv \sum_{n\in\mathbb{N}^+} \frac{1}{n^{2\tau_1-\tau_3}} < \infty, \tag{134}$$

we obtain the constraints

$$\frac{3}{4} < \tau_1 \leq 1 \quad \text{and} \quad \frac{1}{2} < \tau_3 < 2\tau_1 - 1. \tag{135}$$

Now, for the third condition of (127), we demand that

$$\sum_{n\in\mathbb{N}^+} \frac{\beta_n^2}{\gamma_n} \equiv \sum_{n\in\mathbb{N}^+} \frac{1}{n^{2\tau_2-\tau_3}} < \infty, \tag{136}$$

yielding

$$\frac{3}{4} < \tau_2 \leq 1 \quad \text{and} \quad \frac{1}{2} < \tau_3 < 2\tau_2 - 1. \tag{137}$$

Of course, the linear constraints (133), (135) and (137) need to be satisfied simultaneously, yielding the feasible set

$$\frac{7}{8} < \tau_1 \leq 1, \tag{138}$$

$$\frac{3}{4} < \tau_2 < 2\tau_1 - 1 \quad \text{and} \tag{139}$$

$$\frac{1}{2} < \tau_3 < 2\tau_2 - 1. \tag{140}$$

We observe that there are lots of feasible choices for the exponents τ_1, τ_2 and τ_3. For example, one may take $\tau_1 \equiv 1$, $\tau_2 \equiv 0.9$ and $\tau_3 \equiv 0.7$. A graphical representation of the constraint set (138)-(140) is shown in Fig. 4.1.

4.4.3 Convergence Rates of the *MESSAGE^p* Algorithm

We study two standard settings considered in the literature, namely, that involving a convex risk-averse objective, matching all problem assumptions we have made so far, and that involving a *strongly convex objective*, which, as we will shortly see, results naturally by imposing strong convexity *directly* on the random cost function under consideration.

For the convex case, we employ *iterate smoothing*, and we provide detailed bounds on the $\mathcal{L}_1$ *objective suboptimality rate* of the *MESSAGE^p* algorithm. The proof of our result follows directly, by appealing to the respective results developed recently in [Yang et al., 2018].

For the strongly convex case, we develop completely new, detailed and much stronger results on the *squared-$\mathcal{L}_2$ solution suboptimality rate* of the *MESSAGE^p* algorithm, which provide substantial improvement over the convex case, and are much more comparable to rates achievable in risk-neutral stochastic optimization.

The next basic technical result will be useful in our analysis.

Lemma 8. (Approximation Error Boundedness) *Let Assumptions 5, 6 be in effect, and let $\beta_n \in (0,1]$, $\gamma_n \in (0,1]$, for all $n \in \mathbb{N}$. Also, whenever $p \equiv 1$, suppose that $\sup_{n\in\mathbb{N}} \alpha_n^2/\beta_n^2 < \infty$. Then, it is true that*

$$\sup_{n\in\mathbb{N}} \mathbb{E}\left\{\left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|^2\right\} < \infty, \tag{141}$$

for every choice of $p \in [1,\infty)$, and

$$\sup_{n\in\mathbb{N}} \mathbb{E}\left\{\left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|^2\right\} < \infty, \tag{142}$$

for every choice of $p \in (1,\infty)$.

Proof of Lemma 8. If $p > 1$, the conclusion of the lemma is trivial, due to Assumptions 5 and 6, which imply that the involved quantities y^n, $\mathcal{S}^{\widetilde{F}}(\cdot)$, z^n and $\mathcal{D}^{\widetilde{F}}(\cdot,\cdot)$ are uniformly bounded almost everywhere relative to $\mathcal{P}$ (see Lemma 11 in the Appendix (Section 7)).

For the remaining case where $p \equiv 1$, if $\sup_{n\in\mathbb{N}} \alpha_n^2\beta_n^{-2} < \infty$, we may use simple induction exactly as in ([Yang et al., 2018], Appendix, Proof of Lemma 2.3 (c)), exploiting the recursion (114) in Lemma (5), respectively. ∎

4.4.3.1 Convex Random Cost with Iterate Smoothing For the convex case, we consider *iterate smoothing* on top of the *MESSAGE^p* algorithm, by defining averages

$$\widehat{\boldsymbol{x}}^n \triangleq \frac{1}{\lceil n/2\rceil} \sum_{i\in\mathbb{N}_n^{n-\lceil n/2\rceil}} \widehat{\boldsymbol{x}}^i, \quad n \in \mathbb{N}^+, \tag{143}$$

exactly as in [Wang et al., 2017, Yang et al., 2018]. Under this setting, the next result characterizes the $\mathcal{L}_1$ objective suboptimality rate of the *MESSAGE^p* algorithm, when iterate smoothing is employed, for any choice of the semideviation order p.

Theorem 4. (Rate | Convex Case | Subharmonic Stepsizes) *Let Assumptions 5, 6 be in effect, and let the stepsize sequences $\{\alpha_n\}_{n\in\mathbb{N}}$, $\{\beta_n\}_{n\in\mathbb{N}}$ and $\{\gamma_n\}_{n\in\mathbb{N}}$ follow the subharmonic rules*

$$\begin{cases} \alpha_n \triangleq \dfrac{1}{n^{\tau_1}}, \quad \beta_n \triangleq \dfrac{1}{n^{\tau_2}}, & \text{if } p \equiv 1 \text{ with } 1/2 \le \tau_2 < \tau_1 < 1 \\ \alpha_n \triangleq \dfrac{1}{n^{\tau_1}}, \quad \beta_n \triangleq \dfrac{1}{n^{\tau_2}}, \quad \gamma_n \triangleq \dfrac{1}{n^{\tau_3}}, & \text{if } p > 1 \text{ with } 1/2 \le \tau_3 < \tau_2 < \tau_1 < 1 \end{cases}, \quad \forall n \in \mathbb{N}^+, \tag{144}$$

with initial values $\alpha_0 \equiv \beta_0 \equiv \gamma_0 \equiv 1$. *Additionally, suppose that* $\sup_{n\in\mathbb{N}} \mathbb{E}\left\{\left\|\boldsymbol{x}^n - \boldsymbol{x}^*\right\|_2^2\right\} < \infty$, *where* $\boldsymbol{x}^* \in \mathcal{X}^*$. *Then, for every* $n \in \mathbb{N}^+$, *it is true that*

$$\left\|\phi^{\widetilde{F}}(\widehat{\boldsymbol{x}}^n) - \phi_*^{\widetilde{F}}\right\|_{\mathcal{L}_1} \equiv \mathbb{E}\left\{\phi^{\widetilde{F}}(\widehat{\boldsymbol{x}}^n) - \phi_*^{\widetilde{F}}\right\} \leq \begin{cases} \mathcal{K}_1 n^{-\min\{1-\tau_1, \tau_1-\tau_2, 2\tau_2-\tau_1\}}, & \text{if } p \equiv 1 \\ \mathcal{K}_p n^{-\min\{1-\tau_1, \tau_1-\tau_2, 2\tau_3-\tau_1, 2\tau_2-\tau_1-\tau_3\}}, & \text{if } p > 1 \end{cases}, \quad (145)$$

where $0 < \mathcal{K}_p < \infty, p \in [1, \infty)$ *is a problem dependent constant. In particular, if, for some* $\epsilon \in [0, 1)$, $\delta \in (0, 1)$ *and* $\zeta \in (0, 1)$ *such that* $\delta \geq \zeta$,

$$\begin{cases} \tau_1 \equiv \dfrac{3+\epsilon}{4} \quad \text{and} \quad \tau_2 \equiv \dfrac{1+\delta\epsilon}{2}, & \text{if } p \equiv 1 \\ \tau_1 \equiv \dfrac{7+\epsilon}{8}, \quad \tau_2 \equiv \dfrac{3+\delta\epsilon}{4} \quad \text{and} \quad \tau_3 \equiv \dfrac{1+\zeta\epsilon}{2}, & \text{if } p > 1 \end{cases}, \quad (146)$$

then the MESSAGE^p *algorithm satisfies*

$$\mathbb{E}\left\{\phi^{\widetilde{F}}(\widehat{\boldsymbol{x}}^n) - \phi_*^{\widetilde{F}}\right\} \leq \mathcal{K}_p n^{-(1-\epsilon)/\left(4\mathbb{1}_{\{p>1\}}+4\right)}, \quad (147)$$

for every $n \in \mathbb{N}^+$, *for each fixed* ϵ.

Proof of Theorem 4. Although the *MESSAGE*^{p} algorithm is different from the *T-SCGD* algorithm of [Yang et al., 2018], the proof of Theorem 4 shares essentially the same structure with ([Yang et al., 2018], Proof of Theorem 2.2). In a nutshell, except for its native assumptions, the proof exploits Lemma 8 discussed above, the bound of Lemma 3 and the recursions of Lemmata 5, 6, and 7, developed in Section 4.4.2, the convexity of $\phi^{\widetilde{F}}$, as well as the stepsize exponent constraint set (138)-(140). The details of the proof are omitted, and the reader is referred to [Yang et al., 2018], instead. ■

It should be mentioned that, for $\epsilon > 0$, the exponents of the subharmonic stepsizes of Theorem 4 *simultaneously* satisfy the constraints (138)-(140), as discussed in Section 4.4.2, which are sufficient for guaranteeing convergence of the *MESSAGE*^{p} algorithm in the pathwise sense. Therefore, for $\epsilon > 0$, the *MESSAGE*^{p} algorithm attains a $\mathcal{L}_1$ objective suboptimality rate of order *arbitrarily close to* $\mathcal{O}(n^{-1/\left(4\mathbb{1}_{\{p>1\}}+4\right)})$, while provably exhibiting pathwise stability, as well. On the other hand, if $\epsilon \equiv 0$, then the *MESSAGE*^{p} algorithm attains a rate of order precisely $\mathcal{O}(n^{-1/\left(4\mathbb{1}_{\{p>1\}}+4\right)})$, that is, $\mathcal{O}(n^{-1/4})$, if $p \equiv 1$, and $\mathcal{O}(n^{-1/4})$, if $p > 1$, but pathwise convergence is not guaranteed, at least based on the results presented in Section 4.4.2.

Indeed, the conclusions of Theorem 4, which match the respective rate results previously developed for the general purpose *T-SCGD* algorithm in [Yang et al., 2018], are somewhat disappointing, especially when $p > 1$. However, we should note that this result assumes nothing but mere convexity on the random cost $F(\cdot, \boldsymbol{W})$ and, therefore, on $\phi^{\widetilde{F}}$, as well. This means that Theorem 4 is valid for any problematic or pathological choice of the potentially nonsmooth cost $F(\cdot, \boldsymbol{W})$, as long as it is convex (and of course satisfying any additional regularity assumptions made throughout this work).

Nonetheless, the situation changes dramatically if we strengthen our assumptions on the convexity of $\phi^{\widetilde{F}}$, which, as we will see shortly, can be guaranteed very naturally by in turn strengthening the convexity of $F(\cdot, \boldsymbol{W})$, as in classical risk-neutral stochastic optimization. This is the subject of the next paragraph.

4.4.3.2 Strongly Convex Random Cost Here, we assume that the risk-averse objective under consideration, $\phi^{\widetilde{F}}$, is *σ-strongly convex, in the sense that* there exists $\sigma > 0$, such that

$$\phi^{\widetilde{F}}(\boldsymbol{x}) - \phi_*^{\widetilde{F}} \geq \sigma \left\| \boldsymbol{x} - \boldsymbol{x}^* \right\|_2^2, \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{148}$$

where $\boldsymbol{x}^* \in \mathcal{X}^*$, and $\mathcal{X}^*$ is singleton. Although condition (148) will turn out to be very central in our analysis, is not very useful per se, unless we can show that it can be satisfied under reasonable choices of the random cost $F(\cdot, \boldsymbol{W})$, and the risk measure ρ, such that $\rho(F(\cdot, \boldsymbol{W})) \equiv \phi^{\widetilde{F}}(\cdot)$. In other words, it is important to be able to satisfy condition (148) *constructively* within our problem setting, starting from appropriate assumptions on its basic components (bottom-up).

In fact, it turns out that imposing σ-strong convexity on $F(\cdot, \boldsymbol{W})$ *in the usual sense that* there exists $\sigma > 0$, such that

$$F(\cdot, \boldsymbol{w}) - \sigma \left\| \cdot \right\|_2^2 \text{ is convex, for all } \boldsymbol{w} \in \mathbb{R}^M, \tag{149}$$

is *all that is needed* in order to guarantee condition (148) for the objective of our base problem, $\phi^{\widetilde{F}}$. This is a very simple, but important consequence of the fact that mean-semideviations are convex risk measures (that is, convex, monotone and translation equivariant real-valued functionals on $\mathcal{Z}_q$). The relevant results follow.

Proposition 5. (Strong Convexity of Risk-Function Compositions) *Consider a real-valued random function $f : \mathbb{R}^N \times \Omega \to \mathbb{R}$, as well as a real-valued risk measure $\rho : \mathcal{Z}_q \to \mathbb{R}$. Suppose that, for every $\omega \in \Omega$, $f(\cdot, \omega)$ is σ-strongly convex, and that ρ is convex. Then, the real-valued composite function $\phi^f(\cdot) \equiv \rho(f(\cdot, \bullet)) : \mathbb{R}^N \to \mathbb{R}$ is σ-strongly convex, as well.*

Proof of Proposition 5. By σ-strong convexity of $f(\cdot, \omega)$ for all $\omega \in \Omega$, it is true that $f(\cdot, \omega) - \sigma \left\| \cdot \right\|_2^2$ is convex, for all $\omega \in \Omega$. But ρ is a convex-monotone risk measure and, thus, $\rho\left(f(\cdot, \bullet) - \sigma \left\| \cdot \right\|_2^2\right)$ is also convex. Since, additionally, ρ is translation equivariant, it is true that, for every $\boldsymbol{x} \in \mathbb{R}^N$,

$$\rho\left(f(\boldsymbol{x}, \bullet) - \sigma \left\| \boldsymbol{x} \right\|_2^2\right) \equiv \rho(f(\boldsymbol{x}, \bullet)) - \sigma \left\| \boldsymbol{x} \right\|_2^2 \equiv \phi^f(\boldsymbol{x}) - \sigma \left\| \boldsymbol{x} \right\|_2^2, \tag{150}$$

which, of course, implies that the function $\phi^f(\cdot) - \sigma \left\| \cdot \right\|_2^2$ is convex. Enough said. ∎

For the special case of mean-semideviation models, Proposition 5 may be specialized accordingly, as follows. The proof is trivial, and therefore is omitted.

Proposition 6. (Strong Convexity of $\phi^{\widetilde{F}}$) *Fix $p \in [1, \infty)$ and choose any risk regularizer $\mathcal{R} : \mathbb{R} \to \mathbb{R}$. Suppose that, for every $\boldsymbol{w} \in \mathbb{R}^M$, $F(\cdot, \boldsymbol{w})$ is σ-strongly convex on $\mathcal{X}$. Then, as long as $c \in [0, 1]$, the composite function $\phi^{\widetilde{F}}(\cdot) \equiv \rho(F(\cdot, \boldsymbol{W}))$ is σ-strongly convex on $\mathcal{X}$, as well, and satisfies condition (148).*

Consequently, we see that strong convexity of $F(\cdot, \boldsymbol{W})$ suffices for $\phi^{\widetilde{F}}$ being strongly convex, as well. This fact is very important from an operational/practical point of view, because it implies that guaranteeing strong convexity for a risk-averse problem is in principle no harder than guaranteeing strong convexity for the respective risk-neutral problem when mean-semideviations, or, more generally, convex risk measures, are involved. In particular, Proposition 5 holds true for all coherent risk measures, being also convex.

What remains now is to quantitatively characterize the effect of strong convexity on the convergence rates achieved by the *MESSAGE*^{p} algorithm, and how those compare to the more general convex case, briefly analyzed in Section 4.4.3.1.

Our first result is a parametric *"rate generator"*, which provides general stepsize conditions, under which the optimality and approximation errors of all three levels of the *MESSAGE*^{p} algorithm may be jointly combined into a recursive inequality, resembling the respective recursions arising in the rate analysis of standard, risk-neutral SSD algorithms. This result is new, and builds upon some basic ideas found in the earlier work of [Wang et al., 2017], concerning the rate of the general purpose (2-level) *SCGD* algorithm developed therein.

Lemma 9. (Rate Generator | Strongly Convex Case) *Let Assumptions 5, 6 be in effect, and let $\beta_n \in (0,1]$, $\gamma_n \in (0,1]$, for all $n \in \mathbb{N}$. Also, suppose that $\phi^{\widetilde{F}}$ is σ-strongly convex, and that there exists $n_o \in \mathbb{N}^+$, such that, for all $n \in \mathbb{N}^{n_o}$, the following conditions hold simultaneously:*

G1 $\sigma\alpha_n \leq \dfrac{K-1}{K}\min\{\beta_{n-1},\gamma_{n-1}\}$, *for some bounded constant $K > 1$.*

G2 $\alpha_{n+1}\beta_{n-1} \leq \alpha_n\beta_n$ *and, likewise,* $\alpha_{n+1}\gamma_{n-1} \leq \alpha_n\gamma_n$.

For nonnegative sequences $\{\boldsymbol{\Delta}_B^n\}_{n\in\mathbb{N}}$ and $\{\boldsymbol{\Delta}_C^n\}_{n\in\mathbb{N}}$, consider the process

$$
\begin{aligned}
J^n \triangleq \mathbb{E}\left\{\left\|\boldsymbol{x}^n - \boldsymbol{x}^*\right\|_2^2\right\} + \boldsymbol{\Delta}_B^{n-1}\mathbb{E}\left\{\left|y^{n-1} - \mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n-1}\right)\right|^2\right\} \\
+ \boldsymbol{\Delta}_C^{n-1}\mathbb{E}\left\{\left|z^{n-1} - \mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n-1}, y^{n-1}\right)\right|^2\right\}\mathbb{1}_{\{p>1\}}, \quad n \in \mathbb{N}^+.
\end{aligned} \tag{151}
$$

Then, $\{\boldsymbol{\Delta}_B^n\}_{n\in\mathbb{N}}$ and $\{\boldsymbol{\Delta}_C^n\}_{n\in\mathbb{N}}$ may be chosen such that

$$
\begin{aligned}
J^{n+1} \leq (1-\sigma\alpha_n) J^n + \widetilde{\Sigma}\left(\sigma^2\alpha_n^2 + \frac{\sigma^3\alpha_n\alpha_{n-1}^2}{\beta_{n-1}^2} + \sigma\alpha_n\beta_{n-1}\right) \\
+ \widetilde{\Sigma}\left(\frac{\sigma^3\alpha_n\alpha_{n-1}^2}{\gamma_{n-1}^2} + \frac{\sigma\alpha_n\beta_{n-1}^2}{\gamma_{n-1}^2} + \sigma\alpha_n\gamma_{n-1}\right)\mathbb{1}_{\{p>1\}}, \quad \forall n \in \mathbb{N}^{n_o},
\end{aligned} \tag{152}
$$

for some constant $0 < \widetilde{\Sigma} < \infty$. Additionally, under the assumptions of Lemma 8, and for the same choices of $\{\boldsymbol{\Delta}_B^n\}_{n\in\mathbb{N}}$ and $\{\boldsymbol{\Delta}_C^n\}_{n\in\mathbb{N}}$, it is true that $\sup_{n\in\mathbb{N}^+} J^n < \infty$.

Proof of Lemma 9. See Section 7.10 (Appendix). ∎

Leveraging Lemma 9, along with a simple generalization of Chung's Lemma (see Section 7.9 (Appendix)), we may characterize the convergence rates of the *MESSAGE*^{p} algorithm in the strongly convex case, in full and transparent technical detail. We start with the case where $p > 1$.

Theorem 5. (Rate | Strongly Convex Case | Subharmonic Stepsizes | $p > 1$) *Let Assumptions 5 and 6 be in effect. Suppose that $\phi^{\widetilde{F}}$ is σ-strongly convex, and that the stepsize sequences $\{\alpha_n\}_{n\in\mathbb{N}}$, $\{\beta_n\}_{n\in\mathbb{N}}$ and $\{\gamma_n\}_{n\in\mathbb{N}}$ satisfy the subharmonic rules*

$$
\alpha_n \triangleq \frac{1}{\sigma n}, \quad \beta_n \triangleq \frac{1}{n^{\tau_2}} \quad \text{and} \quad \gamma_n \triangleq \frac{1}{n^{\tau_3}}, \quad \forall n \in \mathbb{N}^+, \tag{153}
$$

where $1/2 \le \tau_3 < \tau_2 < 1$, *and with initial values* $\alpha_0 \equiv \beta_0 \equiv \gamma_0 \equiv 1$. *Also, define the quantities*

$$n_o(\tau_2) \equiv \left\lceil \frac{1}{1-\tau_2^{1/(\tau_2+1)}} \right\rceil \in \mathbb{N}^3 \quad \textit{and} \quad \mathsf{R}(\tau_2,\tau_3) \triangleq \frac{1}{1-\max\{2-2\tau_2, 2\tau_2-2\tau_3, \tau_3\}} > 1. \tag{154}$$

Then, for every $n \in \mathbb{N}^{n_o(\tau_2)}$, *it is true that*

$$\mathbb{E}\left\{\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^*\right\|_2^2\right\} \le \widehat{\Sigma}\frac{n_o(\tau_2)}{n} + \widehat{\Sigma}\frac{\mathsf{R}(\tau_2,\tau_3)}{n^{2\min\{1-\tau_2,\tau_2-\tau_3\}}}, \tag{155}$$

for some constant $0 < \widehat{\Sigma} < \infty$. *In particular, if, for some* $\epsilon \in [0,1)$ *and* $\delta \in (0,1)$,

$$\tau_2 \equiv \frac{3+\epsilon}{4} \quad \textit{and} \quad \tau_3 \equiv \frac{1+\delta\epsilon}{2}, \tag{156}$$

then the MESSAGE^p *algorithm satisfies*

$$\mathcal{O}\left(n^{-(1-\epsilon)/2}\right) \equiv \mathbb{E}\left\{\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^*\right\|_2^2\right\} \le \frac{\widehat{\Sigma}\left(n_o(\epsilon)+\dfrac{2}{1-\epsilon}\right)}{n^{(1-\epsilon)/2}}, \tag{157}$$

for every $n \in \mathbb{N}^{n_o(\epsilon)}$, *for each fixed* ϵ.

Proof of Theorem 5. See Section 7.11 (Appendix). ∎

The main conclusion of Theorem 5 is that, for fixed semideviation order $p > 1$ and for any choice of the user-specified parameter $\epsilon \in [0,1)$, the *MESSAGE*^{p} algorithm achieves a squared-$\mathcal{L}_2$ solution suboptimality rate of the order of $\mathcal{O}(n^{-(1-\epsilon)/2})$ iterations. If, additionally, ϵ is chosen to be strictly positive, that is, for $\epsilon > 0$, pathwise convergence is *simultaneously* guaranteed, since the constraints (138)-(140) of Section 4.4.2 are also satisfied. Similarly to the convex case, this completely novel result establishes a convergence rate of order arbitrarily close to $\mathcal{O}(n^{-1/2})$ as $\epsilon \to 0$, while ensuring stable pathwise operation of the algorithm. Of course, when $\epsilon \equiv 0$, the rate of $\mathcal{O}(n^{-1/2})$ iterations is attained, but pathwise convergence of the algorithm is not guaranteed.

Setting aside the fact that rate quantification is different for the convex and strongly convex cases, and by looking at the respective rate exponents, we observe that Theorem 5 provides a rate *strictly four (4) times faster* than that provided by Theorem 4. Of course, this substantial improvement on the rate of convergence of the *MESSAGE*^{p} algorithm is made possible due to imposition of strong convexity on the risk-averse objective $\phi^{\widetilde{F}}$.

When $p \equiv 1$, we also have the following simpler result. As the proof is very similar to that of Theorem 5, it is omitted.

Theorem 6. (Rate | Strongly Convex Case | Subharmonic Stepsizes | $p \equiv 1$**)** *Let Assumptions 5 and 6 be in effect. Suppose that* $\phi^{\widetilde{F}}$ *is* σ*-strongly convex, and that the stepsize sequences* $\{\alpha_n\}_{n\in\mathbb{N}}$, $\{\beta_n\}_{n\in\mathbb{N}}$ *follow the subharmonic rules*

$$\alpha_n \triangleq \frac{1}{\sigma n}, \quad \textit{and} \quad \beta_n \triangleq \frac{1}{n^{\tau_2}} \quad \forall n \in \mathbb{N}^+, \tag{158}$$

where $1/2 < \tau_2 < 1$, and with initial values $\alpha_0 \equiv \beta_0 \equiv 1$. Choose $n_o(\tau_2)$ as in Theorem 5, and define

$$\mathsf{R}(\tau_2) \triangleq \frac{1}{1 - \max\{2 - 2\tau_2, \tau_2\}} > 1 \tag{159}$$

Then, for every $n \in \mathbb{N}^{n_o(\tau_2)}$, it is true that

$$\mathbb{E}\left\{\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\right\|_2^2\right\} \le \frac{\overline{\Sigma}\,(n_o(\tau_2) + \mathsf{R}(\tau_2))}{n^{\min\{2-2\tau_2,\tau_2\}}}, \tag{160}$$

for some constant $0 < \widehat{\Sigma} < \infty$. In particular, the exponent in the denominator is maximized at $\tau_2^* \equiv 2/3$, yielding a rate of the order of $\mathcal{O}(n^{-2/3})$.

In the structurally simpler case where $p \equiv 1$, the rate order improves to $\mathcal{O}(n^{-2/3})$, which is sufficient for pathwise convergence as well, and matches existing results in compositional stochastic optimization, developed earlier along the lines of [Wang et al., 2017]. Compared to the convex case (Theorem 4), Theorem 6 provides a rate which is *at most* $8/3 \approx 2.67$ *times faster*. Note, however, that whereas in the strongly convex case pathwise convergence is always guaranteed for the particular selection of stepsizes, this does not happen in the convex case, which also involves the choice of ϵ. This seems to be a unique feature of mean-semideviation problems of order $p \equiv 1$ (two SA levels), since if $p > 1$ (three SA levels), the trade-off between achieving pathwise convergence and a fast rate of convergence exists in both the convex and strongly convex cases.

4.5 The Choice of $\mathcal{R}$: Comparison with Assumption 2.1 of [Yang et al., 2018]

We now present a detailed comparison between Assumption 5, which is proposed in this paper, and *Assumption 2.1 of [Yang et al., 2018]*, which is utilized for analyzing and proving convergence of the general purpose *T-SCGD* algorithm, formulated therein. In the following, we rigorously show that, as far as problem (51) is concerned, Assumption 5 imposes substantially weaker restrictions on problem structure, compared with ([Yang et al., 2018], Assumption 2.1), therefore providing a much broader structural framework for recursive, compositional SSD-type optimization of mean-semideviation risk measures.

Recall from (58) that the objective of our base problem (51) may be equivalently represented in the form considered in [Wang et al., 2017, Yang et al., 2018] as

$$\phi^{\widetilde{F}}(\boldsymbol{x}) \equiv \widehat{\varrho}\left(\widehat{\boldsymbol{g}}^{\widetilde{F}}\left(\widehat{\boldsymbol{h}}^{\widetilde{F}}(\boldsymbol{x})\right)\right), \quad \forall \boldsymbol{x} \in \mathcal{X}, \quad \text{with} \tag{161}$$

$$\widehat{\varrho}(x, y) \equiv \mathbb{E}\left\{x + cy^{1/p}\right\} \equiv x + cy^{1/p}, \tag{162}$$

$$\begin{aligned}\widehat{\boldsymbol{g}}^{\widetilde{F}}(\boldsymbol{x}, y) &\equiv \mathbb{E}\{[y\ (\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y))^p]\} \\ &\triangleq \mathbb{E}\left\{\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}(\boldsymbol{x}, y)\right\} \quad \text{and}\end{aligned} \tag{163}$$

$$\begin{aligned}\widehat{\boldsymbol{h}}^{\widetilde{F}}(\boldsymbol{x}) &\equiv \mathbb{E}\{[\boldsymbol{x}\ F(\boldsymbol{x}, \boldsymbol{W})]\} \\ &\triangleq \mathbb{E}\left\{\widehat{\boldsymbol{h}}_{\boldsymbol{W}}^{\widetilde{F}}(\boldsymbol{x})\right\}, \quad \boldsymbol{x} \in \mathcal{X}.\end{aligned} \tag{164}$$

Via careful, but relatively straightforward comparison, it follows that, relative to problem (51), ([Yang et al., 2018], Assumption 2.1) translates into the following structural requirements, where the quantities ε and $\mathcal{E}$ are defined precisely as in condition **C4** of Assumption 5. Recall that ε and $\mathcal{E}$ characterize the essential range of $\mathcal{R}(F(\cdot,\boldsymbol{W})-\bullet)$ and the iterate process $\{z^n\}_{n\in\mathbb{N}^+}$, if $z^0\in[\varepsilon^p,\mathcal{E}^p]$ (see Lemma 11). Here, though, we allow the possibility of ε and $\mathcal{E}$ attaining the values zero and infinity, respectively, and we explicitly adopt the generalized definitions

$$\varepsilon \triangleq \lim_{x\to m_l-m_h}\mathcal{R}(x) \quad \text{and} \tag{165}$$

$$\mathcal{E} \triangleq \lim_{x\to m_h-m_l}\mathcal{R}(x), \tag{166}$$

where $m_l\in[-\infty,\infty]$ and $m_h\in[-\infty,\infty]$, respecting the constraint $m_l\le m_h$ (note that, by our assumptions, m_l and m_h cannot be equal and infinite at the same time). Also, $\mathcal{E}$ is finite if and only if *both* m_l and m_h are finite.

W1 The random functions $\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}$ and $\widehat{\boldsymbol{h}}_{\boldsymbol{W}}^{\widetilde{F}}$ are of uniformly bounded variance.

W2 Almost everywhere on Ω, $\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}$ is *differentiable everywhere* on $\mathcal{X}\times\mathrm{cl}\{(m_l,m_h)\}$. In other words, it is true that

$$\mathcal{P}\left(\left\{\omega\in\Omega\,\middle|\,\nabla\widehat{g}_{\boldsymbol{W}(\omega)}^{\widetilde{F}}\text{ exists for all }\mathcal{X}\times\mathrm{cl}\{(m_l,m_h)\}\right\}\right)\equiv 1. \tag{167}$$

W3 The *squared* induced operator ℓ_2-norms of the random subgradient $\underline{\nabla}\widehat{\boldsymbol{h}}_{\boldsymbol{W}}^{\widetilde{F}}$ and the almost everywhere existent random Jacobian function $\nabla\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}$ are uniformly bounded in expectation.

W4 The random Jacobian of $\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}$ is *uniformly Lipschitz* on $\mathcal{X}\times\mathrm{cl}\{(m_l,m_h)\}$, that is, there exists some constant, say $L<\infty$, such that

$$\left\|\nabla\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}(\boldsymbol{x}_1,y_1)-\nabla\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}(\boldsymbol{x}_2,y_2)\right\|_2\le L\sqrt{\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2+|y_1-y_2|^2}, \tag{168}$$

for all $([\boldsymbol{x}_1\,y_1],[\boldsymbol{x}_2\,y_2])\in[\mathcal{X}\times\mathrm{cl}\{(m_l,m_h)\}]^2$, *almost everywhere on* Ω.

W5 The expectation function $\widehat{\boldsymbol{h}}^{\widetilde{F}}$ is Lipschitz on $\mathcal{X}$.

W6 The gradient $\nabla\widehat{\varrho}$ of the outer function $\widehat{\varrho}$ is both uniformly bounded (relative to any ℓ_p-norm) and Lipschitz on $\mathrm{cl}\{(m_l,m_h)\}\times[\varepsilon^p,\mathcal{E}^p]$.

Conditions **W1**$-$**W6** match precisely ([Yang et al., 2018], Assumption 2.1), when the latter is applied to the class of risk-averse problems considered in this paper. In our analysis, we will also impose the following condition *in addition to* **W1** $-$ **W6**, closely resembling condition **C4** of Assumption 5.

$\widetilde{\mathbf{W7}}$ Whenever $p>1$, it is true that $\mathcal{E}<\infty$.

Although condition $\widetilde{\mathbf{W7}}$ is not explicitly considered in [Yang et al., 2018], it is made here in order to simplify and free the comparison with our proposed Assumption 5 from unnecessary technical complications. In effect, considering condition $\widetilde{\mathbf{W7}}$ together with conditions **W1** $-$ **W6** slightly restricts the class of problems supported by the latter. Nonetheless, such restriction is by no means that severe. On the other hand, imposing condition $\widetilde{\mathbf{W7}}$ provides great analytical flexibility; without it, verification of conditions **W1** and **W3** within the framework of [Yang et al., 2018], referring in particular to ([Yang et al., 2018], Assumption 2.1 (iii) & (iv)), becomes rather problematic and uninsightful, for reasons very similar to those justifying condition **C4** as part of Assumption 5. The usefulness of condition $\widetilde{\mathbf{W7}}$ in addition to conditions **W1** – **W6** in the framework of [Yang et al., 2018] is clearly demonstrated in our discussion below.

Of course, condition $\widetilde{\mathbf{W7}}$ is trivially equivalent with *almost half of* condition **C4** of Assumption 5; thus, no further comment is necessary. Amongst all remaining conditions **W1** – **W6**, conditions **W5** and **W6** are the easiest to discuss and may be almost trivially shown to be automatically satisfied by all problems considered in this paper. Between the latter, let us strategically consider condition **W6** first, which requires that the gradient function

$$\nabla\widehat{\varrho}\,(x, y) \equiv \begin{bmatrix} 1 \\ c\dfrac{1}{p}y^{\frac{1-p}{p}} \end{bmatrix} \tag{169}$$

is uniformly bounded and Lipschitz on $\mathrm{cl}\,\{(m_l, m_h)\} \times [\varepsilon^p, \mathcal{E}^p]$. If $p > 1$ (if not, the situation is trivial), in order for $\nabla\widehat{\varrho}$ to be uniformly bounded, we of course need to verify that (any ℓ_p-norm is fine)

$$\begin{aligned} \sup_{(x,y)\in\mathrm{cl}\{(m_l,m_h)\}\times\left[\varepsilon^p,\mathcal{E}^p\right]} \|\nabla\widehat{\varrho}\,(x, y)\|_2 &\equiv \sup_{y\in\left[\varepsilon^p,\mathcal{E}^p\right]} \sqrt{1 + c^2\frac{1}{p^2}y^{\frac{2(1-p)}{p}}} \\ &\equiv \sqrt{1 + c^2\frac{1}{p^2}\sup_{y\in\left[\varepsilon^p,\mathcal{E}^p\right]} y^{\frac{2(1-p)}{p}}} < \infty, \end{aligned} \tag{170}$$

which, due to the fact that $(\cdot)^{\frac{2(1-p)}{p}}$ is a hyperbola, *is only possible if* $\varepsilon > 0$, yielding

$$\sup_{y\in\left[\varepsilon^p,\mathcal{E}^p\right]} y^{\frac{2(1-p)}{p}} = \frac{1}{\varepsilon^{2(p-1)}}. \tag{171}$$

By taking the Jacobian of $\nabla\widehat{\varrho}$, it can be easily shown that strict positivity of ε ensures that $\nabla\widehat{\varrho}$ is Lipschitz on $\mathrm{cl}\,\{(m_l, m_h)\} \times [\varepsilon^p, \mathcal{E}^p]$, as well. Consequently, we see that condition **W6** is implied by condition **C4** of Assumption 5. It is also relatively easy to show that condition **W6** *together with* $\widetilde{\mathbf{W7}}$ are in fact equivalent to **C4**. As far as condition **W5** is concerned, this can be directly verified exploiting Lipschitz continuity of the function $\mathbb{E}\{F(\cdot, \boldsymbol{W})\}$ on $\mathcal{X}$ (see Lemma 10). In the following, we examine the less obvious, remaining conditions **W1** – **W4**, in greater detail.

We start with condition **W1**. In order for $\widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}$ and $\widehat{\boldsymbol{h}}^{\widetilde{F}}_{\boldsymbol{W}}$ to be of uniformly bounded variance, it must be true that

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{y\in\mathrm{cl}\{(m_l,m_h)\}} \mathbb{E}\left\{\left\|\widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}, y) - \widehat{\boldsymbol{g}}^{\widetilde{F}}(\boldsymbol{x}, y)\right\|_2^2\right\} < \infty \quad \text{and} \tag{172}$$

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}\left\{\left\|\widehat{\boldsymbol{h}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}) - \widehat{\boldsymbol{h}}^{\widetilde{F}}(\boldsymbol{x})\right\|_2^2\right\} < \infty, \tag{173}$$

respectively. Let $\boldsymbol{W}' : \Omega \to \mathbb{R}^M$ be an independent copy of the information variable $\boldsymbol{W}$. Then, regarding $\widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}$, we have, for every $\boldsymbol{x} \in \mathcal{X}$ and for every $y \in \mathrm{cl}\,\{(m_l, m_h)\}$,

$$\mathbb{E}\left\{\left\|\widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}, y) - \widehat{\boldsymbol{g}}^{\widetilde{F}}(\boldsymbol{x}, y)\right\|_2^2\right\}$$

$$
\begin{aligned}
&\equiv \mathbb{E}\left\{\|[0\ \ (\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right))^{p}-\mathbb{E}\left\{(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right))^{p}\right\}]\|_{2}^{2}\right\} \\
&\equiv \mathbb{E}\left\{\left((\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right))^{p}-\mathbb{E}\left\{(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right))^{p}\right\}\right)^{2}\right\} \\
&\equiv \mathbb{E}\left\{\left((\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right))^{p}-\mathbb{E}\left\{\left.\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}'\right)-y\right)\right)^{p}\right|\boldsymbol{W}\right\}\right)^{2}\right\} \\
&\equiv \mathbb{E}\left\{\left(\mathbb{E}\left\{(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right))^{p}-\left.\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}'\right)-y\right)\right)^{p}\right|\boldsymbol{W}\right\}\right)^{2}\right\},
\end{aligned}
\tag{174}
$$

which, by Jensen, yields

$$
\mathbb{E}\left\{\left\|\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}\left(\boldsymbol{x},y\right)-\widehat{\boldsymbol{g}}^{\widetilde{F}}\left(\boldsymbol{x},y\right)\right\|_{2}^{2}\right\}\le\mathbb{E}\left\{\left|(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right))^{p}-\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}'\right)-y\right)\right)^{p}\right|^{2}\right\}. \tag{175}
$$

If $p\equiv 1$, nonexpansiveness of $\mathcal{R}$ (condition **S4**) further implies that

$$
\begin{aligned}
\mathbb{E}\left\{\left\|\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}\left(\boldsymbol{x},y\right)-\widehat{\boldsymbol{g}}^{\widetilde{F}}\left(\boldsymbol{x},y\right)\right\|_{2}^{2}\right\}&\le\mathbb{E}\left\{\left|F\left(\boldsymbol{x},\boldsymbol{W}\right)-F\left(\boldsymbol{x},\boldsymbol{W}'\right)\right|^{2}\right\} \\
&\equiv\mathbb{E}\left\{\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)\right)^{2}+\left(F\left(\boldsymbol{x},\boldsymbol{W}'\right)\right)^{2}-2F\left(\boldsymbol{x},\boldsymbol{W}\right)F\left(\boldsymbol{x},\boldsymbol{W}'\right)\right\} \\
&\equiv 2\mathbb{V}\left\{F\left(\boldsymbol{x},\boldsymbol{W}\right)\right\},
\end{aligned}
\tag{176}
$$

for all $\boldsymbol{x}\in\mathcal{X}$ and for all $y\in\mathrm{cl}\left\{(m_l,m_h)\right\}$. For $p>1$, Lemma 12 similarly implies that (recall that we have assumed that condition $\widetilde{\mathbf{W7}}$ is true)

$$
\mathbb{E}\left\{\left\|\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}\left(\boldsymbol{x},y\right)-\widehat{\boldsymbol{g}}^{\widetilde{F}}\left(\boldsymbol{x},y\right)\right\|_{2}^{2}\right\}\le 2\mathcal{E}^{p-1}p\mathbb{V}\left\{F\left(\boldsymbol{x},\boldsymbol{W}\right)\right\}, \tag{177}
$$

for all $\boldsymbol{x}\in\mathcal{X}$ and for all $y\in\mathrm{cl}\left\{(m_l,m_h)\right\}$. Consequently, $F\left(\cdot,\boldsymbol{W}\right)$ being uniformly in $\mathcal{Z}_2$ is *sufficient*, so that $\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}$ is also uniformly in $\mathcal{Z}_2$, as required. Now, note that, for every $\boldsymbol{x}\in\mathcal{X}$,

$$
\mathbb{E}\left\{\left\|\widehat{\boldsymbol{h}}_{\boldsymbol{W}}^{\widetilde{F}}\left(\boldsymbol{x}\right)-\widehat{\boldsymbol{h}}^{\widetilde{F}}\left(\boldsymbol{x}\right)\right\|_{2}^{2}\right\}\equiv\mathbb{E}\left\{\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-\mathbb{E}\left\{F\left(\boldsymbol{x},\boldsymbol{W}\right)\right\}\right)^{2}\right\}\equiv\mathbb{V}\left\{F\left(\boldsymbol{x},\boldsymbol{W}\right)\right\}, \tag{178}
$$

and thus $F\left(\cdot,\boldsymbol{W}\right)$ being uniformly in $\mathcal{Z}_2$ is *equivalent* to $\widehat{\boldsymbol{h}}_{\boldsymbol{W}}^{\widetilde{F}}$ being uniformly in $\mathcal{Z}_2$. Apparently, condition **W1** *implies* condition **C2** of Assumption 5, which directly requires that $F\left(\cdot,\boldsymbol{W}\right)$ is uniformly in $\mathcal{Z}_2$, in turn implying condition **W1**. Therefore, conditions **C2** and **W1** are *equivalent*.

Second, we examine the consequences of assuming everywhere differentiability of $\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}$ primarily on the smoothness on the risk regularizer $\mathcal{R}$, but also that of the random cost function $F\left(\cdot,\boldsymbol{W}\right)$. Suppose that there exists a measurable set $\widehat{\Omega}\subseteq\Omega$, with $\mathcal{P}\left(\widehat{\Omega}\right)\equiv 1$, such that, for all $\omega\in\widehat{\Omega}$, $\widehat{\boldsymbol{g}}_{\boldsymbol{W}(\omega)}^{\widetilde{F}}$ is differentiable everywhere on $\mathcal{X}\times\mathrm{cl}\left\{(m_l,m_h)\right\}$, as in condition **W2**. Without loss of generality, we can take $\widehat{\Omega}\equiv\Omega_E$. Then, for every $\omega\in\widehat{\Omega}$ and for every $(\boldsymbol{x},y)\in\mathcal{X}\times\mathrm{cl}\left\{(m_l,m_h)\right\}$, and due to convexity (see also Proof of Lemma 1 in Section 7.2), the (random) Jacobian of $\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}$ may be expressed as

$$
\nabla\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}\left(\boldsymbol{x},y\right)=\begin{bmatrix}\mathbf{0}_N & \nabla_{\boldsymbol{x}}\left[(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right))^{p}\right]\\ 1 & \nabla_{y}\left[(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right))^{p}\right]\end{bmatrix}
$$

$$
\begin{aligned}
&= \begin{bmatrix} \mathbf{0}_N & p\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right)\right)^{p-1}\underline{\nabla}\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right)\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}\right) \\ 1 & -p\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right)\right)^{p-1}\underline{\nabla}\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right) \end{bmatrix} \\
&\equiv \begin{bmatrix} \mathbf{0}_N & \underline{\nabla}\left(\mathcal{R}\left(z\right)\right)^p|_{z\equiv F(\boldsymbol{x},\boldsymbol{W})-y}\times\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}\right) \\ 1 & -\,\underline{\nabla}\left(\mathcal{R}\left(z\right)\right)^p|_{z\equiv F(\boldsymbol{x},\boldsymbol{W})-y} \end{bmatrix},
\end{aligned}
\tag{179}
$$

where we have assumed that, although $\nabla\widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}}$ exists, the convex functions $F\left(\cdot,\boldsymbol{W}\right)$ and $\mathcal{R}$ may *not* be differentiable *everywhere* on $\mathcal{X}$ and $\mathfrak{R}^{\widetilde{F}}$ (the effective domain of $\mathcal{R}\left(F\left(\cdot,\boldsymbol{W}\right)-(\bullet)\right)$), respectively. Then, the following proposition is true.

Proposition 7. (Masks of Nondifferentiability) *Assume that, for some fixed value of $p\in[1,\infty)$, condition* **W2** *is satisfied. Then, the following statements are necessarily true:*

1) *The p-th power of $\mathcal{R}$ is differentiable everywhere on $\mathfrak{R}^{\widetilde{F}}$.*

2) *Either:*

(a) <u>Everywhere on $\widehat{\Omega}$</u>, the random cost function $F\left(\cdot,\boldsymbol{W}\right)$ is differentiable everywhere on $\mathcal{X}$, Or:

(b) If, <u>for at least one $\omega\in\widehat{\Omega}$</u>, $F\left(\cdot,\boldsymbol{W}\right)$ is nondifferentiable at some $\boldsymbol{x}\in\mathcal{X}$, it must be true that

$$
\mathcal{R}\left(z\right)\equiv C_{ND},\quad \forall z\in\bigcup_{\omega\in\widehat{\Omega}_{ND}}\mathrm{cl}\left\{\left(-\infty,F_{ND}^{*}\left(\omega\right)-m_l\right)\right\},
\tag{180}
$$

where $0\le C_{ND}<\infty$ is some constant, the function $F_{ND}^{*}:\widehat{\Omega}\to[-\infty,\infty]$ is defined as

$$
F_{ND}^{*}\left(\omega\right)\triangleq\sup\{F\left(\boldsymbol{x},\boldsymbol{W}\left(\omega\right)\right)\in\mathbb{R}\,|F\left(\cdot,\boldsymbol{W}\left(\omega\right)\right)\ \textit{is nondifferentiable at } \boldsymbol{x}\},\ \omega\in\widehat{\Omega},
\tag{181}
$$

and the set of elementary events $\widehat{\Omega}_{ND}\subseteq\widehat{\Omega}$ is defined as

$$
\widehat{\Omega}_{ND}\triangleq\left\{\omega\in\widehat{\Omega}\middle|\,F\left(\cdot,\boldsymbol{W}\left(\omega\right)\right)\ \textit{is nondifferentiable at some } \boldsymbol{x}\right\}.
\tag{182}
$$

In other words, $\mathcal{R}$ must be partially constant, as prescribed by (180).

Proof of Proposition 7. First, *existence* of the gradient $\nabla_y\left[\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right)\right)^p\right]$ *for all* $(\boldsymbol{x},y)\in\mathcal{X}\times\mathrm{cl}\left\{(m_l,m_h)\right\}$, which is true by our hypothesis, necessarily implies that the function $\left(\mathcal{R}\left(\cdot\right)\right)^p$ is *differentiable everywhere on* $\mathfrak{R}^{\widetilde{F}}$, for the particular choice of $p\in[1,\infty)$. The case of the gradient $\nabla_{\boldsymbol{x}}\left[\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right)\right)^p\right]$ is slightly more complicated. In order for $\nabla_{\boldsymbol{x}}\left[\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right)\right)^p\right]$ to exist, it must be the case that, for each fixed $\boldsymbol{x}_0\in\mathcal{X}$, either $F\left(\cdot,\boldsymbol{W}\right)$ is differentiable at $\boldsymbol{x}_0$, or whenever $\boldsymbol{x}_0$ is in the restriction of the set of nondifferentiability points of $F\left(\cdot,\boldsymbol{W}\right)$ to $\mathcal{X}$, say $\mathcal{C}_{\boldsymbol{W}(\cdot)}^{F}|_{\mathcal{X}}:\widehat{\Omega}\rightrightarrows\mathbb{R}^N$, defined as

$$
\mathcal{C}_{\boldsymbol{W}(\omega)}^{F}|_{\mathcal{X}}\triangleq\left\{\boldsymbol{x}\in\mathcal{X}|\,F\left(\cdot,\boldsymbol{W}\left(\omega\right)\right)\ \text{is nondifferentiable at } \boldsymbol{x}\right\},\quad \omega\in\widehat{\Omega},
\tag{183}
$$

the stationary point condition

$$
\nabla_y\left[\left(\mathcal{R}\left(F\left(\boldsymbol{x}_0,\boldsymbol{W}\right)-y\right)\right)^p\right]\equiv 0
\tag{184}
$$

is satisfied, *for all* $y \in \mathrm{cl}\left\{(m_l, m_h)\right\}$. Equivalently, we demand that

$$\nabla\left[(\mathcal{R}(z))^p\right] \equiv 0, \tag{185}$$

for all $z \in \mathrm{cl}\left\{(F(\boldsymbol{x}_0, \boldsymbol{W}) - m_h, F(\boldsymbol{x}_0, \boldsymbol{W}) - m_l)\right\}$. Due to convexity, nonnegativity, and monotonicity of $\mathcal{R}$ (conditions **S1**, **S2** and **S3**), it is not hard to see that, for every *qualifying* $\omega \in \widehat{\Omega}$ and for every $\boldsymbol{x}_0 \in \mathcal{C}^F_{\boldsymbol{W}(\omega)}|_{\mathcal{X}}$, $\mathcal{R}$ must partially be of the form

$$\mathcal{R}(z) \equiv C_{ND}, \quad \forall z \in \mathrm{cl}\left\{(-\infty, F(\boldsymbol{x}_0, \boldsymbol{W}(\omega)) - m_l)\right\}, \tag{186}$$

where $0 \le C_{ND} < \infty$ is some constant. Working in the same fashion, utilizing the fact that the multifunction $\mathcal{C}^F_{\boldsymbol{W}(\cdot)}|_{\mathcal{X}}$ is countable-valued and by defining the function $F^*_{ND} : \widehat{\Omega} \to [-\infty, \infty]$ as

$$F^*_{ND}(\omega) \triangleq \sup\left\{F(\boldsymbol{x}, \boldsymbol{W}(\omega)) \in \mathbb{R} \,\middle|\, \boldsymbol{x} \in \mathcal{C}^F_{\boldsymbol{W}(\omega)}|_{\mathcal{X}}\right\}, \quad \omega \in \widehat{\Omega}, \tag{187}$$

we may also obtain the uniform requirement

$$\mathcal{R}(z) \equiv C_{ND}, \quad \forall z \in \bigcup_{\omega \in \widehat{\Omega}_{ND}} \mathrm{cl}\left\{\left(-\infty, F^*_{ND}(\omega) - m_l\right)\right\}, \tag{188}$$

where the set $\widehat{\Omega}_{ND} \subseteq \widehat{\Omega}$ is defined as in (182). Therefore, if, for at least one $\omega \in \widehat{\Omega}$, $F(\cdot, \boldsymbol{W})$ is nondifferentiable at some $\boldsymbol{x} \in \mathcal{X}$, $\mathcal{R}$ must be *partially constant*, as prescribed by (188). ∎

As implied by Proposition 7, condition **W2** *always* requires differentiability of the p-th power of $\mathcal{R}$, everywhere on $\Re^{\widetilde{F}}$. Additionally, any potential nonsmoothness of $F(\cdot, \boldsymbol{W})$ *always* imposes further requirements on the structure of $\mathcal{R}$, significantly restricting the allowable choices in regard to the latter. On the contrary, this is not the case as far as Assumption 5 is concerned, regarding the choice of $\mathcal{R}$. Specifically, there are a lot of cases where, not only $\mathcal{R}$ and/or its powers are allowed to exhibit corner points, but also $F(\cdot, \boldsymbol{W})$ may be nonsmooth, as well. To show that indeed this is the case, let us consider the following, simple example.

Let $p \equiv 1$ (for simplicity), let $\mathcal{R}$ be *any risk regularizer*, and consider the objective function

$$F(\boldsymbol{x}, \boldsymbol{W}) \equiv F(x, W) \triangleq |x - W|, \tag{189}$$

where $W \sim \mathcal{N}(0, 1)$. Although nonsmooth, the random cost function $F(\cdot, W)$ is differentiable almost everywhere relative to $\mathcal{P}$, at each fixed $x \in \mathbb{R}$, thus satisfying condition **P1**. It may also easily argued that condition **P2** is also satisfied, as well. Then, we are interested in the scalar-decision, risk-averse stochastic program

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \mathbb{E}\{|x - W|\} + c\mathbb{E}\left\{\mathcal{R}\left(|x - W| - \mathbb{E}\{|x - W|\}\right)\right\} \\ \text{subject to} & x \in \mathcal{X} \end{array}, \tag{190}$$

for some nonempty, non-singleton, convex and compact set $\mathcal{X}$. Note that, for every choice of $\mathcal{X}$, it is true that

$$m_l \equiv 0 \quad \text{and} \quad m_h \equiv +\infty, \tag{191}$$

since W is unbounded. Thus, $\mathrm{cl}\left\{(m_l, m_h)\right\} \equiv [0, \infty)$. The random subdifferential multifunction of $F(\cdot, W)$ may be expressed as

$$\partial F(x, W) = \begin{cases} \{1\}, & \text{if } x > W \\ [-1, 1], & \text{if } x \equiv W \\ \{-1\}, & \text{if } x < W \end{cases}, \tag{192}$$

and, thus, every subgradient of $F(\cdot, W)$ has the form

$$\underline{\nabla} F(x, W) = \mathbb{1}_{\{x > W\}} - \mathbb{1}_{\{x < W\}} + \delta \mathbb{1}_{\{x \equiv W\}}, \tag{193}$$

where $\delta \in [-1, 1]$. Consequently, it is true that $|\underline{\nabla} F(\cdot, W)| \leq 1$ *uniformly on* $\mathcal{X} \times \Omega$, and condition **C1** of Assumption 5 is satisfied with $P \equiv \infty$. Also, due to W being integrable and $\mathcal{X}$ being compact, it is easy to see that condition **C2** of Assumption 5 is satisfied, as well. Let us now study condition **C3**, related to the choice of $\mathcal{R}$. For each $x \in \mathcal{X}$, the cost $F(x, W)$ follows a *folded normal distribution* with scale x and location 1, since $x - W \sim \mathcal{N}(x, 1)$. On $[0, \infty)$ and for fixed $x \in \mathcal{X}$, the cdf of $F(x, W)$ is given by

$$F_W^x(y) = \Phi(y + x) + \Phi(y - x) - 1. \tag{194}$$

Hence, F_W^x is (uniformly) Lipschitz on $[0, \infty)$, since the Gaussian cdf Φ is Lipschitz on $\mathbb{R}$. Consequently, by choosing $\underline{\nabla} \mathcal{R} \equiv \mathcal{R}'_+$, case (**2**) of Proposition 4 implies that the choice of $\mathcal{R}$ can be *completely unconstrained*.

Now, let us see if problem (190) is supported within the framework set by the necessary conditions of Proposition 7. First, case (**1**) of Proposition 7 directly implies that $\mathcal{R}$ *must be differentiable on* $\mathfrak{R}^{\widetilde{F}}$. Second, case (**2**) of Proposition 7 implies that, since almost everywhere on Ω, $F(\cdot, W)$ is not differentiable everywhere on $\mathcal{X}$, $\mathcal{R}$ *must be partially constant, in addition to the differentiability requirement*. In particular, for every choice of a *certain* event $\widehat{\Omega}$, there exists a pair $(x_0, \omega_0) \in \mathcal{X} \times \widehat{\Omega}$, such that $W(\omega_0) \equiv x_0$, implying the existence of *at least one* point of nondifferentiability of $F(\cdot, W)$ on $\mathcal{X}$ (however happening with $\mathcal{P}$-measure zero, since W is Gaussian). This fact may be shown by the following simple argument. Let $\widehat{\Omega} \subseteq \Omega$ be *any* event such that $\mathcal{P}\left(\widehat{\Omega}\right) \equiv 1$, and consider the preimage

$$W^{-1}(\mathcal{X}) \triangleq \{\omega \in \Omega \,|\, W(\omega) \in \mathcal{X}\} \in \mathscr{F}. \tag{195}$$

Of course, we have $\mathcal{P}\left(W^{-1}(\mathcal{X})\right) > 0$. Suppose that $\widehat{\Omega} \bigcap W^{-1}(\mathcal{X})$ is empty, implying that $\mathcal{P}\left(\widehat{\Omega} \bigcap W^{-1}(\mathcal{X})\right) \equiv 0$. But then it would be true that

$$\begin{aligned} \mathcal{P}\left(\widehat{\Omega} \bigcup W^{-1}(\mathcal{X})\right) &\equiv \mathcal{P}\left(\widehat{\Omega}\right) + \mathcal{P}\left(W^{-1}(\mathcal{X})\right) - \mathcal{P}\left(\widehat{\Omega} \bigcap W^{-1}(\mathcal{X})\right) \\ &\equiv 1 + \mathcal{P}\left(W^{-1}(\mathcal{X})\right) > 1, \end{aligned} \tag{196}$$

which is, of course, absurd. Therefore, the events $\widehat{\Omega}$ and $W^{-1}(\mathcal{X})$ must necessarily have at least one element in common. Call this element ω_0. Since $\omega_0 \in W^{-1}(\mathcal{X})$, there must exist some x_0 in $\mathcal{X}$, such that $W(\omega_0) \equiv x_0$. Now, for every possible choice of $\widehat{\Omega}$, it is trivially true that, if $\omega \in \widehat{\Omega}$ and $F(\cdot, W(\omega))$ is nondifferentiable at some $x \in \mathcal{X}$, then $W(\omega) \equiv x$ and $F(x, W(\omega)) \equiv 0$.

Consequently, with the notation of Proposition 7, it follows that, *regardless of* the choice of the nonempty feasible set $\mathcal{X}$,

$$F^{*}_{ND}(\omega) \equiv 0, \quad \forall \omega \in \widehat{\Omega}, \tag{197}$$

implying that $\mathcal{R}$ *must be constant on* $(-\infty, 0]$. This yields a major limitation of condition **W2**. We should also mention that, for this very simple example, even the choice $\mathcal{R}(\cdot) \equiv (\cdot)_{+}$, which gives the mean-upper-semideviation risk measure, is excluded if condition **W2** is imposed.

Next, let us consider condition **W3**. In this case, the situation is very similar to condition **W1**. Consider the *Frobenius norm* of $\nabla \widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}$ and $\underline{\nabla} \widehat{\boldsymbol{h}}^{\widetilde{F}}_{\boldsymbol{W}}$, respectively. Regarding the Jacobian $\nabla \widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}$, it is true that

$$\begin{aligned}
\left\| \nabla \widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}, y) \right\|_F^2 &\equiv \left\| \begin{bmatrix} \boldsymbol{0}_N & \nabla (\mathcal{R}(z))^p|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y} \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \\ 1 & \nabla (\mathcal{R}(z))^p|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y} \end{bmatrix} \right\|_F^2 \\
&\equiv 1 + \left(\nabla (\mathcal{R}(z))^p|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y} \right)^2 + \left(\nabla (\mathcal{R}(z))^p|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y} \right)^2 \| \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \|_2^2 \\
&\le \begin{cases} 2 + \| \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \|_2^2, & \text{if } p \equiv 1 \\ 1 + p^2 \mathcal{E}^{2(p-1)} \left(1 + \| \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \|_2^2 \right), & \text{if } p > 1 \end{cases},
\end{aligned} \tag{198}$$

for all $\boldsymbol{x} \in \mathcal{X}$ and for all $y \in \operatorname{cl}\{(m_l, m_h)\}$, as a result of Lemma 11, implying that, as long as $\| \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \|_2$ is uniformly in $\mathcal{Z}_2$, $\left\| \nabla \widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}, y) \right\|_F$ must be uniformly in $\mathcal{Z}_2$. Since $\nabla \widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}$ is of rank at most two, we also have

$$\left\| \nabla \widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}, y) \right\|_2^2 \le \left\| \nabla \widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}, y) \right\|_F^2 \le 2 \left\| \nabla \widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}, y) \right\|_2^2, \tag{199}$$

which means that, if $\left\| \nabla \widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}, y) \right\|_F$ is uniformly in $\mathcal{Z}_2$, so is the spectral norm $\left\| \nabla \widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}, y) \right\|_2$ (and conversely). Similarly, the Frobenius norm of $\underline{\nabla} \widehat{\boldsymbol{h}}^{\widetilde{F}}_{\boldsymbol{W}}$ may be explicitly expressed as

$$\left\| \underline{\nabla} \widehat{\boldsymbol{h}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}) \right\|_F^2 \equiv \left\| \left[\boldsymbol{I}_N \,\middle|\, \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \right] \right\|_F^2 = N + \| \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \|_2^2, \tag{200}$$

and because $\underline{\nabla} \widehat{\boldsymbol{h}}^{\widetilde{F}}_{\boldsymbol{W}}$ is of rank at most N, it is true that

$$\left\| \underline{\nabla} \widehat{\boldsymbol{h}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}) \right\|_2^2 \le \left\| \underline{\nabla} \widehat{\boldsymbol{h}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}) \right\|_F^2 \equiv N + \| \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \|_2^2 \le N \left\| \underline{\nabla} \widehat{\boldsymbol{h}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}) \right\|_2^2, \tag{201}$$

for all $\boldsymbol{x} \in \mathcal{X}$. Apparently, we get that the spectral norm of $\underline{\nabla} \widehat{\boldsymbol{h}}^{\widetilde{F}}_{\boldsymbol{W}}$ is uniformly in $\mathcal{Z}_2$ if and only if $\| \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \|_2$ is uniformly in $\mathcal{Z}_2$, as well. This simply means that condition **C1** of Assumption 5 is *equivalent to* condition **W3**, as with the case of condition **W1** and condition **C2** of Assumption 5, discussed above.

We now continue with condition **W4**. Utilizing the fact that, for almost all $\omega \in \Omega$, $\widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}$ is differentiable everywhere on $\mathcal{X} \times \operatorname{cl}\{(m_l, m_h)\}$, condition **W4** demands that, for almost all $\omega \in \Omega$, it is true that

$$\left\| \nabla \widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}_1, y_1) - \nabla \widehat{\boldsymbol{g}}^{\widetilde{F}}_{\boldsymbol{W}}(\boldsymbol{x}_2, y_2) \right\|_2$$

$$
\begin{aligned}
&\equiv \left\| \begin{bmatrix} \mathbf{0}_N & \nabla_{\boldsymbol{x}} \left[\left(\mathcal{R} \left(F \left(\boldsymbol{x}_1, \boldsymbol{W} \right) - y_1 \right) \right)^p \right] \\ 1 & \nabla_y \left[\left(\mathcal{R} \left(F \left(\boldsymbol{x}_1, \boldsymbol{W} \right) - y_1 \right) \right)^p \right] \end{bmatrix} - \begin{bmatrix} \mathbf{0}_N & \nabla_{\boldsymbol{x}} \left[\left(\mathcal{R} \left(F \left(\boldsymbol{x}_2, \boldsymbol{W} \right) - y_2 \right) \right)^p \right] \\ 1 & \nabla_y \left[\left(\mathcal{R} \left(F \left(\boldsymbol{x}_2, \boldsymbol{W} \right) - y_2 \right) \right)^p \right] \end{bmatrix} \right\|_2 \\
&\equiv \left\| \begin{bmatrix} \mathbf{0}_N & \nabla_{\boldsymbol{x}} \left[\left(\mathcal{R} \left(F \left(\boldsymbol{x}_1, \boldsymbol{W} \right) - y_1 \right) \right)^p \right] - \nabla_{\boldsymbol{x}} \left[\left(\mathcal{R} \left(F \left(\boldsymbol{x}_2, \boldsymbol{W} \right) - y_2 \right) \right)^p \right] \\ 0 & \nabla_y \left[\left(\mathcal{R} \left(F \left(\boldsymbol{x}_1, \boldsymbol{W} \right) - y_1 \right) \right)^p \right] - \nabla_y \left[\left(\mathcal{R} \left(F \left(\boldsymbol{x}_2, \boldsymbol{W} \right) - y_2 \right) \right)^p \right] \end{bmatrix} \right\|_2 \\
&= \left\| \begin{bmatrix} \nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}_1, \boldsymbol{W}) - y_1} \times \underline{\nabla} F\left(\boldsymbol{x}_1, \boldsymbol{W}\right) - \nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}_2, \boldsymbol{W}) - y_2} \times \underline{\nabla} F\left(\boldsymbol{x}_2, \boldsymbol{W}\right) \\ - \nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}_1, \boldsymbol{W}) - y_1} + \nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}_2, \boldsymbol{W}) - y_2} \end{bmatrix} \right\|_2 \\
&\le L \sqrt{ \left\| \boldsymbol{x}_1 - \boldsymbol{x}_2 \right\|_2^2 + \left| y_1 - y_2 \right|^2 }, \qquad (202)
\end{aligned}
$$

for all $([\boldsymbol{x}_1 \, y_1], [\boldsymbol{x}_2 \, y_2]) \in [\mathcal{X} \times \mathrm{cl}\{(m_l, m_h)\}]^2$, where $L < \infty$. Without loss of generality, let us call $\widehat{\Omega}$ the certain subset of Ω, such that (202) is true. As with condition **W2**, without loss of generality, we can take $\widehat{\Omega} \equiv \Omega_E$.

We compare (202) with the *strongest variation* of condition **C3**, that is, case (**1**) of Proposition 4. We will see that, *under no additional assumptions*, case (**1**) of Proposition 4 cannot imply (202), by construction. Indeed, *even when* $\boldsymbol{x}_1 \equiv \boldsymbol{x}_2 \triangleq \boldsymbol{x} \in \mathcal{X}$, we may write

$$
\begin{aligned}
&\left\| \nabla \widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}} \left(\boldsymbol{x}, y_1 \right) - \nabla \widehat{\boldsymbol{g}}_{\boldsymbol{W}}^{\widetilde{F}} \left(\boldsymbol{x}, y_2 \right) \right\|_2 \\
&\quad \equiv \left\| \begin{bmatrix} \nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}_1, \boldsymbol{W}) - y_1} \times \underline{\nabla} F\left(\boldsymbol{x}, \boldsymbol{W}\right) - \nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}_2, \boldsymbol{W}) - y_2} \times \underline{\nabla} F\left(\boldsymbol{x}, \boldsymbol{W}\right) \\ - \nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_1} + \nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_2} \end{bmatrix} \right\|_2 \\
&\quad \equiv \left\| \begin{bmatrix} \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \\ \hline -1 \end{bmatrix} \left(\nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_1} - \nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_2} \right) \right\|_2, \qquad (203)
\end{aligned}
$$

for all $(y_1, y_2) \in [\mathrm{cl}\{(m_l, m_h)\}]^2$. It is clear that, in order for (203) to yield a Lipschitz inequality for the involved function, *assuming that case* (**1**) *of Proposition 4 is true*, it would be necessary to impose assumptions on the size of $\nabla F(\cdot, \boldsymbol{W})$. Specifically, it is true that

$$
\begin{aligned}
&\left\| \begin{bmatrix} \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \\ \hline -1 \end{bmatrix} \left(\nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_1} - \nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_2} \right) \right\|_2 \\
&\quad \equiv \sqrt{ \left\| \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \right\|_2^2 + 1 } \left| \nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_1} - \nabla \left(\mathcal{R}\left(z\right) \right)^p \big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_2} \right| \\
&\quad \le \left(\left\| \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \right\|_2 + 1 \right) \left| y_1 - y_2 \right|, \qquad (204)
\end{aligned}
$$

for all $(y_1, y_2) \in [\mathrm{cl}\{(m_l, m_h)\}]^2$, demonstrating need of a bound on $\|\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W})\|_2$, uniform on $\mathcal{X} \times \Omega'$, where $\Omega' \subseteq \Omega$ is a certain event, so that (202) can be verified. Of course, such uniform boundedness assumption is not made directly neither in Assumption 5, nor in ([Yang et al., 2018], Assumption 2.1) (it is made *in expectation*, though). The closest relative to our framework would be to assume that $\|\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W})\|_2$ is in $\mathcal{Z}_\infty$ (that is, with bounded essential supremum), uniformly on $\mathcal{X}$, but this condition is too restrictive if it is imposed *together with* assuming differentiability of the p-th power of $\mathcal{R}$ (case (**1**) of Proposition 4).

On the other hand, suppose that (202) is true. Then, for every $\boldsymbol{x}_1 \equiv \boldsymbol{x}_2 \equiv \boldsymbol{x} \in \mathcal{X}$ and everywhere on $\widehat{\Omega} \equiv \Omega_E$, it is true that

$$\begin{aligned} L\left|y_1 - y_2\right| &\geq \sqrt{\|\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W})\|_2^2 + 1}\left|\nabla\left(\mathcal{R}\left(z\right)\right)^p\right|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_1} - \nabla\left(\mathcal{R}\left(z\right)\right)^p\Big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_2}\Big| \\ &\geq \left|\nabla\left(\mathcal{R}\left(z\right)\right)^p\right|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_1} - \nabla\left(\mathcal{R}\left(z\right)\right)^p\Big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_2}\Big|, \end{aligned} \tag{205}$$

implying that

$$\left|\nabla\left(\mathcal{R}\left(z\right)\right)^p\right|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_1} - \nabla\left(\mathcal{R}\left(z\right)\right)^p\Big|_{z \equiv F(\boldsymbol{x}, \boldsymbol{W}) - y_2}\Big| \leq L\left|F\left(\boldsymbol{x}, \boldsymbol{W}\right) - y_2 - \left(F\left(\boldsymbol{x}, \boldsymbol{W}\right) - y_1\right)\right|, \tag{206}$$

for all $\left(F\left(\boldsymbol{x}, \boldsymbol{W}\right) - y_1, F\left(\boldsymbol{x}, \boldsymbol{W}\right) - y_2\right) \in \left[\mathfrak{R}^{\widetilde{F}}\right]^2$. Therefore, it follows that case (**1**) of Proposition 4 is satisfied with $D_{\mathcal{R},p} \equiv L$. This shows that *condition* **W4** *is in general stronger than the strongest assumption on the smoothness of* $\left(\mathcal{R}\left(z\right)\right)^p$ *considered in this paper* whatsoever.

Driven by the detailed discussion above, let us now formulate the following proposition, which constitutes a precise statement of the fact that the structural framework considered in this work is more general than the one considered in [Wang et al., 2017, Yang et al., 2018]. The proof is based on the above and is omitted.

Proposition 8. (Structural Comparisons) *The class of mean-semideviation programs supported under Assumptions 3 and 5 contains the respective class supported under conditions* $\mathbf{W1} - \mathbf{W6}$ *plus* $\widetilde{\mathbf{W7}}$ *(i.e., Assumption 2.1 of [Yang et al., 2018] +* $\widetilde{\mathbf{W7}}$*). Further, the inclusion is strict.*

5 Conclusion

We have introduced the *MESSAGE^p algorithm*, which is an efficient, data-driven compositional stochastic subgradient procedure for iteratively solving *convex* mean-semideviation risk-averse problems *to optimality*, and constitutes a *parallel* variation of the recently developed, general purpose *T-SCGD algorithm* of Yang, Wang & Fang [Yang et al., 2018]. We have proposed a flexible and structure-exploiting set of problem assumptions, under which we have rigorously analyzed the asymptotic behavior of the *MESSAGE^p* algorithm. Specifically:

- We have established pathwise convergence of the *MESSAGE^p* algorithm in a strong technical sense, confirming its asymptotic consistency.

- In the case of a *strongly convex cost*, we have shown that, for fixed semideviation order $p > 1$, the *MESSAGE^p* algorithm achieves a squared-$\mathcal{L}_2$ solution suboptimality rate of the order of $\mathcal{O}(n^{-(1-\epsilon)/2})$ iterations, where $\epsilon \in [0, 1)$ is a user-specified constant, related to the stepsize selection. In particular, for $\epsilon > 0$, pathwise convergence of the *MESSAGE^p* algorithm is *simultaneously* guaranteed, establishing a rate of order arbitrarily close to $\mathcal{O}(n^{-1/2})$, while ensuring stable pathwise operation. For $p \equiv 1$, the rate order improves to $\mathcal{O}(n^{-2/3})$, which also suffices for pathwise convergence, and matches previous results.

- Likewise, in the general case of a *convex cost*, we have shown that, for any $\epsilon \in [0, 1)$, the *MESSAGE^p* algorithm *with iterate smoothing* achieves an $\mathcal{L}_1$ objective suboptimality rate of the order of $\mathcal{O}(n^{-(1-\epsilon)/\left(4\mathbb{1}_{\{p>1\}}+4\right)})$. This result provides maximal rates $\mathcal{O}(n^{-1/4})$, if $p \equiv 1$, and $\mathcal{O}(n^{-1/8})$, if $p > 1$, matching the state of the art, as well.

We have also discussed the superiority of the proposed framework for convergence, as compared to that employed earlier in [Yang et al., 2018], within the risk-averse context under consideration. First, contrary to [Yang et al., 2018], a unique feature of our framework is that it clearly reveals a well-defined trade-off between the expansiveness of the random cost and the smoothness of the particular mean-semideviation risk measure. This provides great analytical flexibility, which is very important for practical considerations. Additionally, we have rigorously demonstrated that the class of mean-semideviation problems supported herein is *strictly larger* than the respective class of problems supported in [Yang et al., 2018]. As a result, this work establishes the applicability of compositional stochastic optimization for a significantly and strictly wider spectrum of convex mean-semideviation risk-averse problems, as compared to the state of the art. Consequently, the purpose of our work is justified from this perspective, as well.

6 Acknowledgements

We would like to thank our colleague Weidong Han for his useful suggestions.

This material is based upon work supported by the U.S. Navy / SPAWAR Systems Center Pacific under Contract No. N66001-18-C-4031. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Navy / SPAWAR Systems Center Pacific.

7 Appendix: Proofs

7.1 Proof of Theorem 2

The first part of the theorem has essentially already been proved in earlier in Section 3.3.3 (in particular, (28) with $C_S \equiv 1$ and $C_I \equiv 0$, and for some $Y \in \mathcal{Z}_1$, which implies that $\mathbb{E}\left\{(x-Y)_+\right\} < +\infty$, for all $x \in \mathbb{R}$), except for explicitly showing equivalence of interpreting the involved integral in the Lebesgue and improper Riemann senses. Therefore, in addition to this detail, it suffices to prove the second part of the theorem (the converse). The proof, presented below, is technical, but clean and simple.

Consider any *nonconstant* risk regularizer $\mathcal{R} : \mathbb{R} \to \mathbb{R}$. By definition, $\mathcal{R}$ is convex on $\mathbb{R}$ (**S1**) and, thus, it admits both left and right (directional) derivatives, which are *nondecreasing*, *everywhere* on $\mathbb{R}$. Let $\mathcal{R}'_+ : \mathbb{R} \to \mathbb{R}$ be the right derivative of $\mathcal{R}$. Because $\mathcal{R}'_+$ is nondecreasing on $\mathbb{R}$, it exhibits an at most countable number of discontinuities, and of the jump type. By convexity, it follows that $\mathcal{R}'_+$ is right continuous at every such point of discontinuity, as well.

By definition of $\mathcal{R}'_+$, it is true that, for every $x \in \mathbb{R}$, $\mathcal{R}'_+(x) \in \partial\mathcal{R}(x)$, where the compact-valued multifunction $\partial\mathcal{R} : \mathbb{R} \rightrightarrows \mathbb{R}$ denotes the subdifferential of $\mathcal{R}$. Therefore, for every $x \in \mathbb{R}$, the subderivative $\mathcal{R}'_+(x)$ satisfies the defining inequality

$$\mathcal{R}(y) - \mathcal{R}(x) \geq \mathcal{R}'_+(x)(y-x), \tag{207}$$

for every $y \in \mathbb{R}$. Exploiting (207), monotonicity of $\mathcal{R}$ (**S3**) readily implies that $\mathcal{R}'_+(x) \geq 0$, for all $x \in \mathbb{R}$, whereas, from nonexpansiveness of $\mathcal{R}$ (**S4**), it easily follows that $\mathcal{R}'_+(x) \leq 1$, for all $x \in \mathbb{R}$. Additionally, from nonnegativity of $\mathcal{R}$ (**S2**), it is true that, for every $x \in \mathbb{R}$,

$$\mathcal{R}'_+(x) \leq \frac{\mathcal{R}(y) - \mathcal{R}(x)}{y-x} \leq \frac{\mathcal{R}(y)}{y-x}, \quad \forall y \in (x, +\infty), \tag{208}$$

and, using the fact that $\mathcal{R}'_+(x) \geq 0$, for all $x \in \mathbb{R}$, we may pass to the limit as $x \to -\infty$, yielding

$$0 \leq \limsup_{x \to -\infty} \mathcal{R}'_+(x) \leq \limsup_{x \to -\infty} \frac{\mathcal{R}(y)}{y - x} \equiv 0, \tag{209}$$

implying that $\mathcal{R}'_+(x) \underset{x \to -\infty}{\longrightarrow} 0$, as well. On the other hand, since $\mathcal{R}'_+(x) \leq 1$, for all $x \in \mathbb{R}$, it is trivial to see that $0 < \sup_{x \in \mathbb{R}} \mathcal{R}'_+(x) \leq 1$ (for nonconstant $\mathcal{R}$). Consequently, the function $F_Y : \mathbb{R} \to [0,1]$ defined as

$$F_Y(x) \triangleq \frac{\mathcal{R}'_+(x)}{\sup_{x \in \mathbb{R}} \mathcal{R}'_+(x)}, \quad \forall x \in \mathbb{R}, \tag{210}$$

qualifies as the cdf of some random variable $Y : \Omega \to \mathbb{R}$, and we may obviously write

$$\mathcal{R}'_+(x) \equiv F_Y(x) \sup_{x \in \mathbb{R}} \mathcal{R}'_+(x), \quad \forall x \in \mathbb{R}. \tag{211}$$

Now, we know that $\mathcal{R}$ is convex on $\mathbb{R}$ and, if $\mathcal{A}$ denotes the countable set of points where $\mathcal{R}$ is *nondifferentiable*, its derivative exists on $\mathbb{R} \setminus \mathcal{A}$. Let $\mathcal{R}' : \mathbb{R} \to \mathbb{R}$ denote this derivative, defined on the set it exists. Then, by definition, it is true that

$$\mathcal{R}'(x) \equiv \mathcal{R}'_+(x), \quad \forall x \in \mathbb{R} \setminus \mathcal{A}, \tag{212}$$

where, of course, $\mathcal{A}$ is of Lebesgue measure zero. Consequently, for every $(\alpha, x) \in \mathbb{R}^2$, such that $\alpha \leq x$, it follows that $\mathcal{R}' \equiv \mathcal{R}'_+$, almost everywhere relative to the Lebesgue measure on $[\alpha, x]$. Also due to convexity on $\mathbb{R}$ (say), $\mathcal{R}$ is absolutely continuous on $[\alpha, x]$, for every qualifying choice of α and x. Therefore, Lebesgue's Fundamental Theorem of Integral Calculus (Theorems 2.3.4 & 2.3.10 in [Ash and Doléans-Dade, 2000]) implies that

$$\mathcal{R}(x) - \mathcal{R}(\alpha) \equiv \int_\alpha^x \mathcal{R}'(x)\, \mathrm{d}y \equiv \int_\alpha^x \mathcal{R}'_+(y)\, \mathrm{d}y, \tag{213}$$

where integration is interpreted in the sense of Lebesgue, relative to the Lebesgue measure on the Borel space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. By monotone convergence, we may deduce that, since $\mathcal{R}$ is nondecreasing and uniformly bounded from below, its limit at $-\infty$ is finite and, in particular,

$$\mathcal{R}(x) \underset{x \to -\infty}{\longrightarrow} \inf_{x \in \mathbb{R}} \mathcal{R}(x) \geq 0. \tag{214}$$

Also, for every $x \in \mathbb{R}$, (213) is true for every $\mathbb{R} \ni \alpha \leq x$. Therefore, we may pass to the limit in (213) as $\alpha \to -\infty$, to obtain

$$\mathcal{R}(x) - \inf_{x \in \mathbb{R}} \mathcal{R}(x) \equiv \lim_{\alpha \to -\infty} \int_\alpha^x \mathcal{R}'_+(y)\, \mathrm{d}y, \quad \forall x \in \mathbb{R}. \tag{215}$$

Invoking Lebesgue's Monotone Convergence Theorem and via a standard sequential argument, it follows that

$$\begin{aligned} \lim_{\alpha \to -\infty} \int_\alpha^x \mathcal{R}'_+(y)\, \mathrm{d}y &\equiv \lim_{\alpha \to -\infty} \int \mathcal{R}'_+(y) \mathbb{1}_{[\alpha,x]}(y)\, \mathrm{d}y \\ &= \int \lim_{\alpha \to -\infty} \mathcal{R}'_+(y) \mathbb{1}_{[\alpha,x]}(y)\, \mathrm{d}y \end{aligned}$$

$$\equiv \int \mathcal{R}'_{+}(y)\,\mathbb{1}_{[-\infty,x]}(y)\,\mathrm{d}y \equiv \int_{-\infty}^{x} \mathcal{R}'_{+}(y)\,\mathrm{d}y, \quad \forall x \in \mathbb{R}, \tag{216}$$

which, together with (211), further implies that

$$\mathcal{R}(x) \equiv \left(\sup_{x\in\mathbb{R}} \mathcal{R}'_{+}(x)\right) \int_{-\infty}^{x} F_Y(y)\,\mathrm{d}y + \inf_{x\in\mathbb{R}} \mathcal{R}(x), \quad \forall x \in \mathbb{R}. \tag{217}$$

In addition to the above, Fubini's Theorem (Theorem 2.6.6 in [Ash and Doléans-Dade, 2000]) implies that

$$+\infty > \int_{-\infty}^{x} F_Y(y)\,\mathrm{d}y \equiv \mathbb{E}\left\{(x-Y)_{+}\right\}, \tag{218}$$

for all $x \in \mathbb{R}$ and for every random variable $Y : \Omega \to \mathbb{R}$ having F_Y as its cdf.

To show that the integral involved in (217) is well defined in the improper Riemann sense, note first that the nondecreasing function is Riemann integrable. Therefore, the Lebesgue integral in (213) is necessarily equal to the respective Riemann integral. Equivalently, integration in (213) may be interpreted in the Riemann sense, as well. Then, (215) remains true, and the limit on the RHS may be interpreted as an improper Riemann integral, by definition. The validity of (217), where integration is in the improper Riemann sense, follows. Note that, as far as the direct statement of Theorem 2 is concerned, equivalence of the aforementioned Lebesgue and improper Riemann integrals may be shown in exactly the same fashion as above.

Finally, let $\mathcal{R}$ be constant on $\mathbb{R}$. Then, it is trivial to see that $\sup_{x\in\mathbb{R}} \mathcal{R}'_{+}(x) \equiv 0$, and $\inf_{x\in\mathbb{R}} \mathcal{R}(x) \equiv \mathcal{R}(x)$, for all $x \in \mathbb{R}$. Then, for *any* random variable $Y : \Omega \to \mathbb{R}$, such that $\mathbb{E}\left\{(x-Y)_{+}\right\} < +\infty$, for all $x \in \mathbb{R}$, (217) is trivially true, and, apparently, there is at least one such random variable. The result now follows. ■

7.2 Proof of Lemma 1

Certainly, because $\phi^{\widetilde{F}}$ admits the compositional representation

$$\phi^{\widetilde{F}}(\boldsymbol{x}) \equiv \mathbb{E}\left\{F(\boldsymbol{x},\boldsymbol{W})\right\} + c\varrho\left(g^{\widetilde{F}}\left(\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})\right)\right), \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{219}$$

it follows that $\phi^{\widetilde{F}}$ will be differentiable as long as the functions $\mathbb{E}\left\{F(\cdot,\boldsymbol{W})\right\}$, ϱ, $g^{\widetilde{F}}$ and $\boldsymbol{h}^{\widetilde{F}}$ are in the respective effective domains, in which case it must be true that

$$\nabla\phi^{\widetilde{F}}(\boldsymbol{x}) \equiv \nabla\mathbb{E}\left\{F(\boldsymbol{x},\boldsymbol{W})\right\} + c\nabla\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})\,\nabla g^{\widetilde{F}}(\boldsymbol{y})\Big|_{\boldsymbol{y}\equiv\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})} \nabla\varrho(z)\Big|_{z\equiv g^{\widetilde{F}}\left(\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})\right)}, \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{220}$$

where $\nabla\boldsymbol{h}^{\widetilde{F}} : \mathbb{R}^N \to \mathbb{R}^{N\times(N+1)}$ denotes the Jacobian of $\boldsymbol{h}^{\widetilde{F}}$, $\nabla g^{\widetilde{F}} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ denotes the gradient of $g^{\widetilde{F}}$, assumed to exist at least for all $\boldsymbol{y} \in \mathrm{Graph}_{\mathcal{X}}\left(\mathbb{E}\left\{F(\cdot,\boldsymbol{W})\right\}\right)$, and $\nabla\varrho : \mathbb{R} \to \mathbb{R}$ denotes the derivative of ϱ, also assumed to be well defined at least for every z in the range of $g^{\widetilde{F}}$.

Following a bottom-up approach, we first exploit our basic assumption that $F(\cdot,\boldsymbol{W})$ is convex on $\mathcal{X}$ (at least), for every realization $\boldsymbol{W} \equiv \boldsymbol{W}(\omega), \omega \in \Omega$, as well as property **P1**. Under this setting, we may invoke ([Shapiro et al., 2014], Theorem 7.51) for each $\boldsymbol{x} \in \mathcal{X}$, from where it follows that the function $\mathbb{E}\left\{F(\cdot,\boldsymbol{W})\right\}$ is differentiable *everywhere* on $\mathcal{X}$ and that, further, we may interchange differentiation with integration, implying that

$$\nabla\mathbb{E}\left\{F(\boldsymbol{x},\boldsymbol{W})\right\} \equiv \mathbb{E}\left\{\underline{\nabla}F(\boldsymbol{x},\boldsymbol{W})\right\}, \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{221}$$

This result directly yields the existence of the Jacobian of $\boldsymbol{h}^{\widetilde{F}}$, given by

$$\nabla \boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x}) \equiv \left[\boldsymbol{I}_N \,\middle|\, \mathbb{E}\left\{\underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W})\right\}\right], \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{222}$$

which, of course, is the same as (71), in the statement of Lemma 1.

Next, let us discuss differentiability of $g^{\widetilde{F}}$. We know that, due to convexity of $F(\cdot, \boldsymbol{W})$, $\mathcal{R}$ and $(\cdot)^p$, and because of monotonicity of $\mathcal{R}$ and $(\cdot)^p$, the composite function $(\mathcal{R}(F(\cdot, \boldsymbol{W}) - (\bullet)))^p$ is convex in both variables (at least separately). We would also like to show that $(\mathcal{R}(F(\cdot, \boldsymbol{W}) - (\bullet)))^p$ is differentiable at each $(\boldsymbol{x}, y) \in \mathrm{Graph}_{\mathcal{X}}(\mathbb{E}\{F(\cdot, \boldsymbol{W})\})$, almost everywhere relative to $\mathcal{P}$.

Indeed, fix an *arbitrary* point $(\boldsymbol{x}, y_{\boldsymbol{x}}) \equiv (\boldsymbol{x}, \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\})$ in $\mathrm{Graph}_{\mathcal{X}}(\mathbb{E}\{F(\cdot, \boldsymbol{W})\})$. By property **P1**, we know that there is a certain event $\mathsf{D}_{\boldsymbol{x}} \subseteq \Omega$, such that $F(\cdot, \boldsymbol{W}(\omega))$ is differentiable at $\boldsymbol{x} \in \mathcal{X}$, for all $\omega \in \mathsf{D}_{\boldsymbol{x}}$. Consequently, by the fact that the identity $(\bullet): \mathbb{R} \to \mathbb{R}$ is differentiable everywhere on $\mathbb{R}$, it follows that, for every $\omega \in \mathsf{D}_{\boldsymbol{x}}$, the function $F(\cdot, \boldsymbol{W}) - (\bullet)$ is differentiable at $(\boldsymbol{x}, y_{\boldsymbol{x}})$. Utilizing property **P2**, for the same fixed point $(\boldsymbol{x}, y_{\boldsymbol{x}})$, there exists another certain event $\mathsf{N}_{\boldsymbol{x}} \subseteq \Omega$, such that, for every $\omega \in \mathsf{N}_{\boldsymbol{x}}$, $F(\boldsymbol{x}, \boldsymbol{W}(\omega)) - y_{\boldsymbol{x}} \equiv F(\boldsymbol{x}, \boldsymbol{W}(\omega)) - \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\}$ is *not* in $\mathcal{A}$, the countable nullset containing the nondifferentiability points of $\mathcal{R}$. Therefore, for every $\omega \in \mathsf{D}_{\boldsymbol{x}} \bigcap \mathsf{N}_{\boldsymbol{x}}$, with $\mathcal{P}(\mathsf{D}_{\boldsymbol{x}} \bigcap \mathsf{N}_{\boldsymbol{x}}) \equiv 1$, $F(\cdot, \boldsymbol{W}(\omega)) - (\bullet)$ is differentiable at $(\boldsymbol{x}, y_{\boldsymbol{x}})$, and $\mathcal{R}$ is differentiable at $F(\boldsymbol{x}, \boldsymbol{W}(\omega)) - y_{\boldsymbol{x}}$, implying that the composite function $\mathcal{R}(F(\cdot, \boldsymbol{W}(\omega)) - (\bullet))$ is differentiable at $(\boldsymbol{x}, y_{\boldsymbol{x}})$, as well. In other words, we have shown that the function $\mathcal{R}(F(\cdot, \boldsymbol{W}(\omega)) - (\bullet))$ is differentiable at each arbitrary point $(\boldsymbol{x}, y_{\boldsymbol{x}})$ in the set $\mathrm{Graph}_{\mathcal{X}}(\mathbb{E}\{F(\cdot, \boldsymbol{W})\})$, for $\mathcal{P}-$almost every $\omega \in \Omega$. And since the function $(\cdot)^p$ is differentiable everywhere on $\mathbb{R}$, the preceding statement also holds for $(\mathcal{R}(F(\cdot, \boldsymbol{W}) - (\bullet)))^p$.

Further, let us determine the structure of the subdifferential of $(\mathcal{R}(F(\cdot, \boldsymbol{W}) - (\bullet)))^p$. Simply, because the functions $\mathcal{R}(F(\cdot, \boldsymbol{W}) - (\bullet))$ and $(\cdot)^p$ are convex, with the latter being nondecreasing, any subgradient of $(\mathcal{R}(F(\cdot, \boldsymbol{W}) - (\bullet)))^p$ may be expressed as

$$\underline{\nabla}(\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y))^p = p(\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y))^{p-1}\, \underline{\nabla}[\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y)], \tag{223}$$

for all $(\boldsymbol{x}, y) \in \mathrm{Graph}_{\mathcal{X}}(\mathbb{E}\{F(\cdot, \boldsymbol{W})\})$ (at least), where $\underline{\nabla}[\mathcal{R}(\cdot, \boldsymbol{W} - (\bullet))]$ denotes *any* subgradient of $\mathcal{R}(\cdot, \boldsymbol{W} - (\bullet))$. This is a direct application of the composition rule in subgradient calculus. Likewise, another application of the composition rule to the function $\mathcal{R}(F(\cdot, \boldsymbol{W}) - (\bullet))$ yields

$$\underline{\nabla}[\mathcal{R}(F(\cdot, \boldsymbol{W}) - y)] = \underline{\nabla}\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y) \begin{bmatrix} \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \\ \hline -1 \end{bmatrix} \tag{224}$$

and, thus,

$$\underline{\nabla}(\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y))^p \equiv p(\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y))^{p-1}\, \underline{\nabla}\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y) \begin{bmatrix} \underline{\nabla} F(\boldsymbol{x}, \boldsymbol{W}) \\ \hline -1 \end{bmatrix}, \tag{225}$$

for all $(\boldsymbol{x}, y) \in \mathrm{Graph}_{\mathcal{X}}(\mathbb{E}\{F(\cdot, \boldsymbol{W})\})$.

We may now invoke ([Shapiro et al., 2014], Theorem 7.51) for each $(\boldsymbol{x}, y) \in \mathrm{Graph}_{\mathcal{X}}(\mathbb{E}\{F(\cdot, \boldsymbol{W})\})$, from where we obtain that the function $g^{\widetilde{F}}$ is differentiable *everywhere* on $\mathrm{Graph}_{\mathcal{X}}(\mathbb{E}\{F(\cdot, \boldsymbol{W})\})$ and that its gradient is given by

$$\begin{aligned}\nabla g^{\widetilde{F}}(\boldsymbol{x}, y) &\equiv \nabla \mathbb{E}\{(\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y))^p\} \\ &\equiv \mathbb{E}\{\underline{\nabla}(\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y))^p\} \\ &\equiv \mathbb{E}\left\{p(\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y))^{p-1} \underline{\nabla}\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y)\left[\frac{\underline{\nabla}F(\boldsymbol{x}, \boldsymbol{W})}{-1}\right]\right\},\end{aligned} \tag{226}$$

and we are done, since (226) is the same as (72).

As far as the derivative of ϱ is concerned, if $p \in (1, \infty)$ (if not, ϱ is the identity), it exists everywhere on the nonnegative semiaxis, except for the origin, and (70) is obviously true. Thus, from (219), it is clear that we should demand that

$$g^{\widetilde{F}}\left(\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})\right) \equiv \mathbb{E}\{(\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\}))^p\} > 0, \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{227}$$

Fix $\boldsymbol{x} \in \mathcal{X}$. Of course, because $\mathcal{R}$ is nonnegative on $\mathbb{R}$, it is true that

$$\mathbb{E}\{(\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\}))^p\} \equiv 0 \iff \mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\}) \equiv 0,\ \mathcal{P} - a.e. \tag{228}$$

Since, additionally, $\mathcal{R}$ is nondecreasing on $\mathbb{R}$, the RHS statement of (228) is in turn *equivalent to*

$$F(\boldsymbol{x}, \boldsymbol{W}) - \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\} \leq \sup\{x \in \mathbb{R} \,|\, \mathcal{R}(x) \equiv 0\} \triangleq \kappa_{\mathcal{R}} \in \overline{\mathbb{R}}, \quad \mathcal{P} - a.e., \tag{229}$$

where, in general, $\kappa_{\mathcal{R}} \equiv +\infty$ if and only if $\mathcal{R}(x) \equiv 0$, for all $x \in \mathbb{R}$. However, recall that, by assumption, $\mathcal{R}$ is not identically equal to zero everywhere on $\mathbb{R}$; thus, $\kappa_{\mathcal{R}} \in [-\infty, \infty)$. Finally, we have shown that

$$g^{\widetilde{F}}\left(\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})\right) \equiv 0 \iff \mathcal{P}(F(\boldsymbol{x}, \boldsymbol{W}) - \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\} \leq \kappa_{\mathcal{R}}) \equiv 1, \tag{230}$$

which means that, if $\mathcal{P}(F(\boldsymbol{x}, \boldsymbol{W}) - \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\} \leq \kappa_{\mathcal{R}}) < 1$, where $\kappa_{\mathcal{R}}$ is fixed in $[-\infty, \infty)$, then

$$g^{\widetilde{F}}\left(\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})\right) \neq 0 \implies g^{\widetilde{F}}\left(\boldsymbol{h}^{\widetilde{F}}(\boldsymbol{x})\right) > 0. \tag{231}$$

Enough said. ■

7.3 Some Auxiliary Results

In this subsection, let us state some basic, elementary results stemming from Assumption 5, which will be helpful in both the further characterization of condition **C3** of Assumption 5, and the asymptotic analysis of the *MESSAGE*^{p} algorithm. First, a direct, but very useful consequence of condition **C1** is summarized in the next proposition.

Lemma 10. **($\mathbb{E}\{F(\cdot, \boldsymbol{W})\}$ is Lipschitz on $\mathcal{X}$ & More)** *Let condition* **C1** *of Assumption 3 be in effect. Then, the functions* $F(\cdot, \boldsymbol{W})$ *and* $\mathbb{E}\{F(\cdot, \boldsymbol{W})\}$ *satisfy*

$$|\mathbb{E}\{F(\boldsymbol{x}_1, \boldsymbol{W})\} - \mathbb{E}\{F(\boldsymbol{x}_2, \boldsymbol{W})\}| \leq G\,\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \quad \text{and} \tag{232}$$

$$\mathbb{E}\{|F(\boldsymbol{x}_1, \boldsymbol{W}) - F(\boldsymbol{x}_2, \boldsymbol{W})|\} \leq 2G\,\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2, \tag{233}$$

for all $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X} \times \mathcal{X}$.

Proof of Lemma 10. Proof of (232) is straightforward, and omitted. To prove (233), we use the definition of a subgradient and our assumption that the random cost function $F(\cdot, \boldsymbol{W})$ is measurably subdifferentiable on $\mathcal{X}$. For every $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X} \times \mathcal{X}$ and everywhere on Ω, it is true that

$$F(\boldsymbol{x}_1, \boldsymbol{W}) - F(\boldsymbol{x}_2, \boldsymbol{W}) \geq (\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\boldsymbol{T}} \underline{\nabla} F(\boldsymbol{x}_2, \boldsymbol{W}), \tag{234}$$

implying that

$$\begin{aligned} F(\boldsymbol{x}_2, \boldsymbol{W}) - F(\boldsymbol{x}_1, \boldsymbol{W}) &\leq (\boldsymbol{x}_2 - \boldsymbol{x}_1)^{\boldsymbol{T}} \underline{\nabla} F(\boldsymbol{x}_2, \boldsymbol{W}) \\ &\leq \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \|\underline{\nabla} F(\boldsymbol{x}_2, \boldsymbol{W})\|_2 . \end{aligned} \tag{235}$$

Since $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are arbitrary, it follows by symmetry that

$$-(F(\boldsymbol{x}_2, \boldsymbol{W}) - F(\boldsymbol{x}_1, \boldsymbol{W})) \leq \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \|\underline{\nabla} F(\boldsymbol{x}_1, \boldsymbol{W})\|_2 . \tag{236}$$

Consequently, we may write

$$|F(\boldsymbol{x}_2, \boldsymbol{W}) - F(\boldsymbol{x}_1, \boldsymbol{W})| \leq \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 (\|\underline{\nabla} F(\boldsymbol{x}_1, \boldsymbol{W})\|_2 + \|\underline{\nabla} F(\boldsymbol{x}_2, \boldsymbol{W})\|_2), \tag{237}$$

and taking expectations on both sides of (237) yields

$$\begin{aligned} \mathbb{E}\{|F(\boldsymbol{x}_2, \boldsymbol{W}) - F(\boldsymbol{x}_1, \boldsymbol{W})|\} &\leq \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \mathbb{E}\{\|\underline{\nabla} F(\boldsymbol{x}_1, \boldsymbol{W})\|_2 + \|\underline{\nabla} F(\boldsymbol{x}_2, \boldsymbol{W})\|_2\} \\ &\equiv \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 (\mathbb{E}\{\|\underline{\nabla} F(\boldsymbol{x}_1, \boldsymbol{W})\|_2\} + \mathbb{E}\{\|\underline{\nabla} F(\boldsymbol{x}_2, \boldsymbol{W})\|_2\}) \\ &\leq \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \left(\left\| \|\underline{\nabla} F(\boldsymbol{x}_1, \boldsymbol{W})\|_2 \right\|_{\mathcal{L}_P} + \left\| \|\underline{\nabla} F(\boldsymbol{x}_2, \boldsymbol{W})\|_2 \right\|_{\mathcal{L}_P} \right) \\ &\leq \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \, 2G, \end{aligned} \tag{238}$$

where we have exploited condition **C1**. The claim is proved. ■

Second, the next result is on the boundedness of the processes generated by the *MESSAGE*^{p} algorithm, when $p > 1$. It is based on condition **C4** of Assumption 5, as well as Assumption 6.

Lemma 11. (Case $p > 1$: Iterate Boundedness) *Fix $p > 1$, let condition* **C4** *of Assumption 5 be in effect. Also, choose y^0, β_0 and z^0, γ_0 according to Assumption 6 and suppose that $\beta_n \in (0, 1]$, $\gamma_n \in (0, 1]$, for all $n \in \mathbb{N}$. Then, the composite process $\{(\boldsymbol{x}^n, y^n, z^n)\}_{n \in \mathbb{N}}$ generated by the MESSAGE^p algorithm satisfies the uniform pointwise bounds*

$$y^{n+1} \in [m_l, m_h] \tag{239}$$

$$z^{n+1} \in [\varepsilon^p, \mathcal{E}^p] \tag{240}$$

$$\mathcal{R}\left(F\left(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\right) - y^n\right) \in [\varepsilon, \mathcal{E}] \quad \text{and} \tag{241}$$

$$\mathcal{R}\left(F\left(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\right) - \mathbb{E}\left\{F\left(\boldsymbol{x}^n, \boldsymbol{W}'\right)\right\}\right) \in [\varepsilon, \mathcal{E}], \quad \forall n \in \mathbb{N}, \tag{242}$$

almost everywhere relative to $\mathcal{P}$, where $\boldsymbol{W}' \sim \mathcal{P}_{\boldsymbol{W}}$.

Proof of Lemma 11. Let us start with $\left\{\left(y^{n+1}\right)\right\}_{n \in \mathbb{N}}$. It is true that

$$y^1 \equiv (1 - \beta_0) y^0 + \beta_0 F\left(\boldsymbol{x}^0, \boldsymbol{W}_1^1\right)$$

$$\geq \begin{cases} (1-\beta_0)\, m_l + \beta_0 m_l \equiv m_l, & \text{if } y^0 \in [m_l, m_h] \\ (1-1)\, y^0 + m_l \equiv m_l, & \text{if } \beta_0 \equiv 1 \end{cases}, \quad \mathcal{P}-a.e., \tag{243}$$

and the result follows trivially by induction for all $n \in \mathbb{N}^+$, and the fact that $\mathbb{N}$ is countable. The procedure bounding $y^{n+1}, n \in \mathbb{N}$ from above is exactly the same. Since we have shown that $y^{n+1} \in [m_l, m_h]$, for all $n \in \mathbb{N}$, it also readily follows that

$$\varepsilon \equiv \mathcal{R}(m_l - m_h) \leq \mathcal{R}\left(F\left(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\right) - y^n\right) \leq \mathcal{R}(m_h - m_l) \equiv \mathcal{E}, \tag{244}$$

and

$$\varepsilon \leq \mathcal{R}\left(F\left(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\right) - \mathbb{E}\left\{F\left(\boldsymbol{x}^n, \boldsymbol{W}'\right)\right\}\right) \leq \mathcal{E}, \tag{245}$$

as well, almost everywhere relative to $\mathcal{P}$. As far as $\left\{\left(z^{n+1}\right)\right\}_{n\in\mathbb{N}}$ is concerned, we work as above, that is,

$$\begin{aligned} z^1 &\equiv (1-\gamma_0)\, z^0 + \gamma_0 \left(\mathcal{R}\left(F\left(\boldsymbol{x}^0, \boldsymbol{W}_2^1\right) - y^0\right)\right)^p \\ &\geq \begin{cases} (1-\gamma_0)\, \varepsilon^p + \gamma_0 \varepsilon^p \equiv \varepsilon^p, & \text{if } z^0 \in [\varepsilon^p, \mathcal{E}^p] \\ (1-1)\, z^0 + \varepsilon^p \equiv \varepsilon^p, & \text{if } \gamma_0 \equiv 1 \end{cases}, \quad \mathcal{P}-a.e., \end{aligned} \tag{246}$$

and then we use induction, and similarly for the case of the upper bound. ■

Third, another expected, but also useful consequence of condition **C4** is on the expansiveness of the composite function $(\mathcal{R}((\cdot) - \bullet))^p$, as follows.

Lemma 12. ($(\mathcal{R}((\cdot) - \bullet))^p$ is Lipschitz) *Fix $p > 1$ and let condition* **C4** *of Assumption 5 be in effect. Then, it is true that*

$$\begin{aligned} &\left|(\mathcal{R}(F(\boldsymbol{x}_1, \boldsymbol{W}) - y_1))^p - \left(\mathcal{R}\left(F\left(\boldsymbol{x}_2, \boldsymbol{W}'\right) - y_2\right)\right)^p\right| \\ &\qquad\qquad \leq \mathcal{E}^{p-1} p \left(\left|F(\boldsymbol{x}_1, \boldsymbol{W}) - F\left(\boldsymbol{x}_2, \boldsymbol{W}'\right)\right| + |y_1 - y_2|\right), \end{aligned} \tag{247}$$

almost everywhere relative to $\mathcal{P}$, for all $([\boldsymbol{x}_1\, y_1], [\boldsymbol{x}_2\, y_2]) \in [\mathcal{X} \times \mathrm{cl}\{(m_l, m_h)\}]^2$, where $\boldsymbol{W}' : \Omega \to \mathbb{R}^M$ may be taken as any copy of $\boldsymbol{W}$.

Proof of Lemma 12. Simply, using a telescoping argument and due to the fact that $\mathcal{R}$ is nonexpansive, we proceed directly, also exploiting Lemma 11 (with generic $\boldsymbol{W}$ and $\boldsymbol{W}'$ instead of $\boldsymbol{W}_2^{n+1}$), yielding the inequalities

$$\begin{aligned} &\left|(\mathcal{R}(F(\boldsymbol{x}_1, \boldsymbol{W}) - y_1))^p - \left(\mathcal{R}\left(F\left(\boldsymbol{x}_2, \boldsymbol{W}'\right) - y_2\right)\right)^p\right| \\ &\quad \leq \left|F(\boldsymbol{x}_1, \boldsymbol{W}) - y_1 - F\left(\boldsymbol{x}_2, \boldsymbol{W}'\right) + y_2\right| \sum_{j\in\mathbb{N}_{p-1}} (\mathcal{R}(F(\boldsymbol{x}_1, \boldsymbol{W}) - y_1))^j \left(\mathcal{R}\left(F(\boldsymbol{x}_2, \boldsymbol{W}') - y_2\right)\right)^{p-1-j} \\ &\quad \leq \left(\left|F(\boldsymbol{x}_1, \boldsymbol{W}) - F\left(\boldsymbol{x}_2, \boldsymbol{W}'\right)\right| + |y_1 - y_2|\right) \sum_{j\in\mathbb{N}_{p-1}} \mathcal{E}^j \mathcal{E}^{p-1-j} \\ &\quad \equiv \left(\left|F(\boldsymbol{x}_1, \boldsymbol{W}) - F\left(\boldsymbol{x}_2, \boldsymbol{W}'\right)\right| + |y_1 - y_2|\right) \mathcal{E}^{p-1} p, \quad \mathcal{P}-a.e., \end{aligned} \tag{248}$$

for all $([\boldsymbol{x}_1\, y_1], [\boldsymbol{x}_2\, y_2]) \in [\mathcal{X} \times \mathrm{cl}\{(m_l, m_h)\}]^2$. ■

Remark 10. Observe that, since $\boldsymbol{W}'$ may be taken as *any* copy of $\boldsymbol{W}$ in Lemma 12, the choice $\boldsymbol{W}' \equiv \boldsymbol{W}$ is also perfectly valid. ■

7.4 Proof of Proposition 4

To show case (**1**) of the first part of the result, simply observe that, by assumption, $\underline{\nabla}\mathcal{R} \equiv \nabla\mathcal{R}$. Thus, for every qualifying choice of Q, for every $\boldsymbol{x} \in \mathcal{X}$ and for every $(y_1, y_2) \in [\mathrm{cl}\{(m_l, m_h)\}]^2$, we may write

$$\begin{aligned}\left\|\left|\underline{\nabla}(\mathcal{R}(z))^p|_{z\equiv F(\boldsymbol{x},\boldsymbol{W})-y_1} - \underline{\nabla}(\mathcal{R}(z))^p|_{z\equiv F(\boldsymbol{x},\boldsymbol{W})-y_2}\right|\right\|_{\mathcal{L}_Q} &\le \left\|D_{\mathcal{R},p}|y_1 - y_2|\right\|_{\mathcal{L}_Q} \\ &\equiv D_{\mathcal{R},p}|y_1 - y_2|,\end{aligned} \tag{249}$$

and we are done.

Cases (**2**) and (**3**) of the result will be based on the cdf-based representation of risk regularizers (Theorem 2). Without loss of generality, assume that $\mathcal{R}$ is nonconstant. If it is, the problem is trivial. For nonconstant $\mathcal{R}$, Theorem 2 implies the existence of a random variable $Y : \Omega \to \mathbb{R}$, with $\mathbb{E}\left\{(x - Y)_+\right\} < \infty$, with cdf $F_Y : \mathbb{R} \to [0, 1]$, and of a constant $C_S \in (0, 1]$, such that

$$\mathcal{R}'_+(x) \equiv C_S F_Y(x), \quad \forall x \in \mathbb{R}. \tag{250}$$

Of course, the random variable Y may be taken as independent of $F(\boldsymbol{x}, \boldsymbol{W})$, for all $\boldsymbol{x} \in \mathcal{X}$. First, *whenever* $p > 2$, we have, for every $\boldsymbol{x} \in \mathcal{X}$ and for every $(y_1, y_2) \in [\mathrm{cl}\{(m_l, m_h)\}]^2$,

$$\begin{aligned}&\left\|\left|\underline{\nabla}(\mathcal{R}(z))^p|_{z\equiv F(\boldsymbol{x},\boldsymbol{W})-y_1} - \underline{\nabla}(\mathcal{R}(z))^p|_{z\equiv F(\boldsymbol{x},\boldsymbol{W})-y_2}\right|\right\|_{\mathcal{L}_1} \\ &\le p\left\|\left|(\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_1))^{p-1}\underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_1) - (\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_2))^{p-1}\underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_2)\right|\right\|_{\mathcal{L}_1} \\ &\le p\left\|(\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_1))^{p-1}|\underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_1) - \underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_2)|\right\|_{\mathcal{L}_1} \\ &\qquad + p\left\|\underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_2)\left|(\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_1))^{p-1} - (\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_2))^{p-1}\right|\right\|_{\mathcal{L}_1} && (251) \\ &\le p\mathcal{E}^{p-1}\left\||\underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_1) - \underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_2)|\right\|_{\mathcal{L}_1} + p\mathcal{E}^{p-2}(p-1)|y_1 - y_2|, && (252)\end{aligned}$$

where (251) follows by the triangle inequality and (252) follows from Lemma 12. Similarly, for $p \equiv 2$, we get

$$\begin{aligned}&\left\|\left|\underline{\nabla}(\mathcal{R}(z))^2\Big|_{z\equiv F(\boldsymbol{x},\boldsymbol{W})-y_1} - \underline{\nabla}(\mathcal{R}(z))^2\Big|_{z\equiv F(\boldsymbol{x},\boldsymbol{W})-y_2}\right|\right\|_{\mathcal{L}_1} \\ &\quad \le 2\mathcal{E}\left\||\underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_1) - \underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_2)|\right\|_{\mathcal{L}_1} + 2|y_1 - y_2|,\end{aligned} \tag{253}$$

whereas, for $p \equiv 1$, no further derivation is needed. As far the involved $\mathcal{L}_1$-norm is concerned, since $\underline{\nabla}\mathcal{R} \equiv \mathcal{R}'_+$ by assumption, we may write, for every $\boldsymbol{x} \in \mathcal{X}$ and for every $(y_1, y_2) \in [\mathrm{cl}\{(m_l, m_h)\}]^2$,

$$\begin{aligned}&\left\||\underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_1) - \underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_2)|\right\|_{\mathcal{L}_1} \\ &\quad \equiv \mathbb{E}\{|\underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_1) - \underline{\nabla}\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y_2)|\} \\ &\quad = C_S\mathbb{E}\{|F_Y(F(\boldsymbol{x},\boldsymbol{W})-y_1) - F_Y(F(\boldsymbol{x},\boldsymbol{W})-y_2)|\} && (254) \\ &\quad = C_S\mathbb{E}\{(F_Y(F(\boldsymbol{x},\boldsymbol{W}) - \min\{y_1,y_2\}) - F_Y(F(\boldsymbol{x},\boldsymbol{W}) - \max\{y_1,y_2\}))\} \\ &\quad \equiv C_S\mathbb{E}\{\mathcal{P}(F(\boldsymbol{x},\boldsymbol{W}) - \max\{y_1,y_2\} < Y \le F(\boldsymbol{x},\boldsymbol{W}) - \min\{y_1,y_2\}|\,\boldsymbol{W})\}\end{aligned}$$

$$
\begin{aligned}
&\equiv C_S \mathbb{E}\left\{\int \mathbb{1}_{(F(\boldsymbol{x},\boldsymbol{W})-\max\{y_1,y_2\},F(\boldsymbol{x},\boldsymbol{W})-\min\{y_1,y_2\}]}\left(y\right) \mathrm{d}\mathcal{P}_Y\left(y\right)\right\}\\
&= C_S \int \mathbb{E}\left\{\mathbb{1}_{(F(\boldsymbol{x},\boldsymbol{W})-\max\{y_1,y_2\},F(\boldsymbol{x},\boldsymbol{W})-\min\{y_1,y_2\}]}\left(y\right)\right\} \mathrm{d}\mathcal{P}_Y\left(y\right) \qquad (255)\\
&\equiv C_S \int \mathbb{E}\left\{\mathbb{1}_{[y+\min\{y_1,y_2\},y+\max\{y_1,y_2\})}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)\right)\right\} \mathrm{d}\mathcal{P}_Y\left(y\right)\\
&\equiv C_S \int \mathcal{P}\left(y+\min\left\{y_1,y_2\right\} \le F\left(\boldsymbol{x},\boldsymbol{W}\right) < y+\max\left\{y_1,y_2\right\}\right) \mathrm{d}\mathcal{P}_Y\left(y\right), \qquad (256)
\end{aligned}
$$

where (255) follows from Fubini's Theorem (the involved double integral is always finite) on the product measure space $\left(\mathbb{R}\times\mathbb{R},\mathscr{B}\left(\mathbb{R}\right)\otimes\mathscr{B}\left(\mathbb{R}\right),\mathcal{P}_Y\times\mathcal{P}_{\boldsymbol{W}}^{\boldsymbol{x}}\right)$, with $\mathcal{P}_{\boldsymbol{W}}^{\boldsymbol{x}}$ denoting the Borel measure inducing $F_{\boldsymbol{W}}^{\boldsymbol{x}}$. Exploiting the assumed *continuity* of $F_{\boldsymbol{W}}^{\boldsymbol{x}}$ (Lipschitz or not), we also have, for every $\boldsymbol{x}\in\mathcal{X}$,

$$
\begin{aligned}
&\left\|\left|\underline{\nabla}\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y_1\right)-\underline{\nabla}\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y_2\right)\right|\right\|_{\mathcal{L}_1}\\
&\quad\equiv C_S \int \mathcal{P}\left(y+\min\left\{y_1,y_2\right\} < F\left(\boldsymbol{x},\boldsymbol{W}\right) \le y+\max\left\{y_1,y_2\right\}\right) \mathrm{d}\mathcal{P}_Y\left(y\right)\\
&\quad\equiv C_S \int F_{\boldsymbol{W}}^{\boldsymbol{x}}\left(y+\max\left\{y_1,y_2\right\}\right)-F_{\boldsymbol{W}}^{\boldsymbol{x}}\left(y+\min\left\{y_1,y_2\right\}\right) \mathrm{d}\mathcal{P}_Y\left(y\right)\\
&\quad= C_S \int \left|F_{\boldsymbol{W}}^{\boldsymbol{x}}\left(y+y_1\right)-F_{\boldsymbol{W}}^{\boldsymbol{x}}\left(y+y_2\right)\right| \mathrm{d}\mathcal{P}_Y\left(y\right), \qquad (257)
\end{aligned}
$$

for all $(y_1,y_2)\in\left[\mathrm{cl}\left\{(m_l,m_h)\right\}\right]^2$. If the Lipschitz condition of case (**2**) is true, we further have

$$
\begin{aligned}
\left\|\left|\underline{\nabla}\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y_1\right)-\underline{\nabla}\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y_2\right)\right|\right\|_{\mathcal{L}_1} &\le C_S D_{\widetilde{F}} \int \left|y_1-y_2\right| \mathrm{d}\mathcal{P}_Y\left(y\right)\\
&\equiv C_S D_{\widetilde{F}} \left|y_1-y_2\right|, \qquad (258)
\end{aligned}
$$

for all $(y_1,y_2)\in\left[\mathrm{cl}\left\{(m_l,m_h)\right\}\right]^2$, showing that condition **C3** is satisfied with

$$
D \triangleq \begin{cases} p\mathcal{E}^{p-1}C_S D_{\widetilde{F}} + p\mathcal{E}^{p-2}\left(p-1\right), & \text{if } p>2\\ 2\mathcal{E}C_S D_{\widetilde{F}}+2, & \text{if } p\equiv 2\\ C_S D_{\widetilde{F}}, & \text{if } p\equiv 1 \end{cases}. \qquad (259)
$$

by taking the supremum of (258) relative to $\boldsymbol{x}$ over $\mathcal{X}$. If the Lipschitz-in-Expectation condition of case (**3**) is true, we obtain the desired result of Proposition 4 in exactly the same fashion. In particular, when $p\equiv 1$, the equivalence in case (**3**) of Proposition 4 follows directly by (257), and the fact that $C_S \neq 0$. Enough said. ■

7.5 Proof of Lemma 3

The proof is simple, though somewhat tedious; essentially, it is an exercise on using the triangle and Cauchy-Schwarz inequalities. First, observe that, under Assumption 5, it is true that

$$
\sup_{\boldsymbol{x}\in\mathcal{X}}\left\|\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}\right)\right\|_2\right\|_{\mathcal{L}_2} \le \sup_{\boldsymbol{x}\in\mathcal{X}}\left\|\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}\right)\right\|_2\right\|_{\mathcal{L}_P} \le G < \infty, \qquad (260)
$$

for $P \in [2,\infty]$, due to condition **C1**.

Fix $n \in \mathbb{N}$ and let $p > 1$. Under Assumption 5, by nonexpansiveness of the projection operator onto the closed and convex set $\mathcal{X}$, and by the triangle inequality, we have

$$\begin{aligned}
\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^{n}\right\|_2 &\equiv \left\|\Pi_{\mathcal{X}}\left\{\boldsymbol{x}^{n}-\alpha_n\widehat{\nabla}^{n+1}\phi^{\widetilde{F}}\left(\boldsymbol{x}^{n},y^{n},z^{n}\right)\right\}-\Pi_{\mathcal{X}}\{\boldsymbol{x}^{n}\}\right\|_2 \\
&\le \left\|\alpha_n\widehat{\nabla}^{n+1}\phi^{\widetilde{F}}\left(\boldsymbol{x}^{n},y^{n},z^{n}\right)\right\|_2 \\
&\equiv \alpha_n\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_2^{n+1}\right)+c\Delta^{n+1}\left(\boldsymbol{x}^{n},y^{n},z^{n}\right)\right\|_2 \\
&\le \alpha_n\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_2^{n+1}\right)\right\|_2+\alpha_n c\left\|\Delta^{n+1}\left(\boldsymbol{x}^{n},y^{n},z^{n}\right)\right\|_2,
\end{aligned} \tag{261}$$

where, by Lemmata 11 and 12,

$$\begin{aligned}
\left\|\Delta^{n+1}\left(\boldsymbol{x}^{n},y^{n},z^{n}\right)\right\|_2 &\equiv \Bigg\|\left(z^{n}\right)^{(1-p)/p}\left(\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_2^{n+1}\right)-\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_1^{n+1}\right)\right) \\
&\qquad\times\underline{\nabla}\mathcal{R}\left(F\left(\boldsymbol{x}^{n},\boldsymbol{W}_2^{n+1}\right)-y^{n}\right)\left(\mathcal{R}\left(F\left(\boldsymbol{x}^{n},\boldsymbol{W}_2^{n+1}\right)-y^{n}\right)\right)^{p-1}\Bigg\|_2 \\
&\equiv \left(\frac{\mathcal{E}}{\varepsilon}\right)^{p-1}\left\|\left(\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_2^{n+1}\right)-\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_1^{n+1}\right)\right)\right\|_2 \\
&\le \mathsf{R}_p\left(\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_2^{n+1}\right)\right\|_2+\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_1^{n+1}\right)\right\|_2\right),
\end{aligned} \tag{262}$$

almost everywhere relative to $\mathcal{P}$, where we have defined $\mathsf{R}_p \triangleq \left(\mathcal{E}\varepsilon^{-1}\right)^{p-1}$. Consequently, we may bound the ℓ_2-norm of $\boldsymbol{x}^{n+1}-\boldsymbol{x}^{n}$ from above as

$$\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^{n}\right\|_2 \le \alpha_n\left(1+c\mathsf{R}_p\right)\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_2^{n+1}\right)\right\|_2+\alpha_n c\left(\frac{\mathcal{E}}{\varepsilon}\right)^{p-1}\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_1^{n+1}\right)\right\|_2, \tag{263}$$

almost everywhere relative to $\mathcal{P}$. This, of course, implies that

$$\begin{aligned}
&\mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^{n}\right\|_2^2\right\} \\
&\quad\le \alpha_n^2\mathbb{E}_{\mathscr{D}^n}\left\{\left(\left(1+c\mathsf{R}_p\right)\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_2^{n+1}\right)\right\|_2+c\mathsf{R}_p\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_1^{n+1}\right)\right\|_2\right)^2\right\},
\end{aligned} \tag{264}$$

almost everywhere relative to $\mathcal{P}$, as well. Let us focus more closely on the conditional expectation on the RHS of (264). First, by the substitution rule for conditional expectations, which is guaranteed to be valid in all our discussions in this paper, due to the existence of regular conditional distributions on Borel spaces (see, for instance, [Durrett, 2010]), it is true that

$$\begin{aligned}
&\mathbb{E}_{\mathscr{D}^n}\left\{\left(\left(1+c\mathsf{R}_p\right)\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_2^{n+1}\right)\right\|_2+c\mathsf{R}_p\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_1^{n+1}\right)\right\|_2\right)^2\right\} \\
&\quad\equiv \mathbb{E}\left\{\left(\left(1+c\mathsf{R}_p\right)\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_2^{n+1}\right)\right\|_2+c\mathsf{R}_p\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_1^{n+1}\right)\right\|_2\right)^2\right\}\Bigg|_{\boldsymbol{x}\equiv\boldsymbol{x}^{n}},
\end{aligned} \tag{265}$$

almost everywhere relative to $\mathcal{P}$, where, in the RHS of (265), expectation is with respect to the product measure $\mathcal{P}_{\boldsymbol{W}}\times\mathcal{P}_{\boldsymbol{W}}$ on the Borel measurable space $\left(\mathbb{R}^M\times\mathbb{R}^M,\mathscr{B}\left(\mathbb{R}^M\right)\otimes\mathscr{B}\left(\mathbb{R}^M\right)\right)$. This due to mutual independence of $\boldsymbol{W}_1^{n+1}$ and $\boldsymbol{W}_2^{n+1}$, and also their independence relative to $\mathscr{D}^n$. Then, by the triangle inequality of the $\mathcal{L}_2$-norm on the aforementioned product probability space, we may write, for $\boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned}
&\sqrt{\mathbb{E}\left\{\left((1+c\mathsf{R}_p)\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_2^{n+1}\right)\right\|_2 + c\mathsf{R}_p\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_1^{n+1}\right)\right\|_2\right)^2\right\}} \\
&\qquad\equiv\left\|(1+c\mathsf{R}_p)\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_2^{n+1}\right)\right\|_2 + c\mathsf{R}_p\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_1^{n+1}\right)\right\|_2\right\|_{\mathcal{L}_2} \\
&\qquad\le(1+c\mathsf{R}_p)\left\|\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_2^{n+1}\right)\right\|_2\right\|_{\mathcal{L}_2} + c\mathsf{R}_p\left\|\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_1^{n+1}\right)\right\|_2\right\|_{\mathcal{L}_2} \\
&\qquad\le\left(2c\mathsf{R}_p+1\right)G,
\end{aligned} \tag{266}$$

or, by taking squares on both sides,

$$\mathbb{E}\left\{\left((1+c\mathsf{R}_p)\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_2^{n+1}\right)\right\|_2 + c\mathsf{R}_p\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_1^{n+1}\right)\right\|_2\right)^2\right\} \le (2c+1)^2 G^2, \tag{267}$$

almost everywhere relative to $\mathcal{P}$. Thus, it follows that

$$\mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^n\right\|_2^2\right\} \le \alpha_n^2\left(2c\mathsf{R}_p+1\right)^2 G^2, \tag{268}$$

almost everywhere relative to $\mathcal{P}$. In case $p \equiv 1$, it may be easily shown that the respective bound may be recovered by setting $p \equiv 1$ in (268) (pretending that $\mathcal{E}$ and ε are finite). Finally, note that, for every value of p, (268) holds for each $n \in \mathbb{N}$, and $\mathbb{N}$ is, of course, countable. Enough said. ■

7.6 Proof of Lemma 5

Fix $n \in \mathbb{N}$, and let $y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n) \triangleq E_{\mathcal{S}}^n$, for brevity. Then, we may write

$$\begin{aligned}
\left|E_{\mathcal{S}}^{n+1}\right|^2 &\equiv\left|(1-\beta_n)\,y^n + \beta_n F\left(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\right) - \mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}\right)\right|^2 \\
&\equiv\left|(1-\beta_n)\left(y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right) + \beta_n\left(F\left(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\right) - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right) + \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n) - \mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}\right)\right|^2 \\
&\equiv\left|(1-\beta_n)\,E_{\mathcal{S}}^n + \beta_n\left(F\left(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\right) - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right) + \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n) - \mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}\right)\right|^2 \\
&\le(1+\beta_n)\left|(1-\beta_n)\,E_{\mathcal{S}}^n + \beta_n\left(F\left(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\right) - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right)\right|^2 \\
&\quad+\left(1+\beta_n^{-1}\right)\left|\mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n) - \mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}\right)\right|^2 \\
&\equiv(1+\beta_n)(1-\beta_n)^2\,|E_{\mathcal{S}}^n|^2 + (1+\beta_n)\,\beta_n^2\left|F\left(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\right) - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|^2 \\
&\quad+2\left(1-\beta_n^2\right)\beta_n E_{\mathcal{S}}^n\left(F\left(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\right) - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right) + \left(1+\beta_n^{-1}\right)\left|\mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n) - \mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}\right)\right|^2 \\
&\le(1-\beta_n)\,|E_{\mathcal{S}}^n|^2 + 2\beta_n^2\left|F\left(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\right) - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|^2 \\
&\quad+2\,(1+\beta_n)(1-\beta_n)\,\beta_n E_{\mathcal{S}}^n\left(F\left(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\right) - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right) + 2\beta_n^{-1}G^2\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^n\right\|_2^2,
\end{aligned} \tag{269}$$

where we have used our assumption that $\beta_n \leq 1$. Taking expectations relative to $\mathscr{D}^n$ on both sides, we have

$$\mathbb{E}_{\mathscr{D}^n}\left\{\left|E_{\mathcal{S}}^{n+1}\right|^2\right\} \leq (1-\beta_n)\left|E_{\mathcal{S}}^{n}\right|^2 + \beta_n^2 2V + 0 + \beta_n^{-1}2G^2\mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^{n}\right\|_2^2\right\}, \tag{270}$$

almost everywhere relative to $\mathcal{P}$. The fact that $\mathbb{N}$ is countable completes the proof. ∎

7.7 Proof of Lemma 6

Fix $n \in \mathbb{N}$. By adding and subtracting appropriate terms as in the proof of Lemma 5 above, it is then easy to show that the difference $z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n) \triangleq E_{\mathcal{D}}^n$ may be expressed as

$$\begin{aligned} E_{\mathcal{D}}^{n+1} \equiv (1-\gamma_n) E_{\mathcal{D}}^n &+ \left(\mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n) - \mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}, y^{n+1}\right)\right) \\ &+ \gamma_n\left(\left(\mathcal{R}\left(F\left(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\right) - y^n\right)\right)^p - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right). \end{aligned} \tag{271}$$

Let us consider the quantity $\left|E_{\mathcal{D}}^n\right|^2$. We may expand the square one time, yielding

$$\begin{aligned} \left|E_{\mathcal{D}}^{n+1}\right|^2 &\leq (1+\gamma_n)\left|(1-\gamma_n) E_{\mathcal{D}}^n + \gamma_n\left(\left(\mathcal{R}\left(F\left(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\right) - y^n\right)\right)^p - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right)\right|^2 \\ &\quad + \left(1+\gamma_n^{-1}\right)\left|\mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n) - \mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}, y^{n+1}\right)\right|^2 \\ &\equiv (1+\gamma_n)(1-\gamma_n)^2\left|E_{\mathcal{D}}^n\right|^2 + (1+\gamma_n)\gamma_n^2\left|\left(\mathcal{R}\left(F\left(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\right) - y^n\right)\right)^p - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|^2 \\ &\quad + 2\left(1-\gamma_n^2\right)\gamma_n E_{\mathcal{D}}^n\left(\left(\mathcal{R}\left(F\left(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\right) - y^n\right)\right)^p - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right) \\ &\qquad + \left(1+\gamma_n^{-1}\right)\left|\mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n) - \mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}, y^{n+1}\right)\right|^2 \end{aligned} \tag{272}$$

As in the proof of Lemma 5, taking conditional expectations relative to $\mathscr{D}^n$ on both sides and since $\gamma_n \leq 1$, we get

$$\begin{aligned} \mathbb{E}_{\mathscr{D}^n}\left\{\left|E_{\mathcal{D}}^{n+1}\right|^2\right\} &\leq (1-\gamma_n)\left|E_{\mathcal{D}}^n\right|^2 + 2\gamma_n^2\mathbb{E}_{\mathscr{D}^n}\left\{\left|\left(\mathcal{R}\left(F\left(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\right) - y^n\right)\right)^p - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|^2\right\} \\ &\quad + 2\gamma_n^{-1}\mathbb{E}_{\mathscr{D}^n}\left\{\left|\mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n) - \mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}, y^{n+1}\right)\right|^2\right\} + 0. \end{aligned} \tag{273}$$

Next, we consider the last two nonzero terms of the RHS of (273) separately. First, we may write

$$\begin{aligned} &\mathbb{E}_{\mathscr{D}^n}\left\{\left|\left(\mathcal{R}\left(F\left(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\right) - y^n\right)\right)^p - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|^2\right\} \\ &\quad \leq \mathbb{E}_{\mathscr{D}^n}\left\{\left(\mathcal{R}\left(F\left(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\right) - y^n\right)\right)^{2p}\right\} + \mathbb{E}_{\mathscr{D}^n}\left\{\left(\mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right)^2\right\} \\ &\quad \leq 2\mathcal{E}^{2p}, \quad \mathcal{P}-a.e.. \end{aligned} \tag{274}$$

Second, observe that, by Lemmata 12 and 10, we get

$$\left|\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1}, y^{n+1}\right) - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|$$

$$
\begin{aligned}
&\equiv \left|\mathbb{E}_{\mathscr{D}^{n+1}}\left\{\left(\mathcal{R}\left(F\left(\boldsymbol{x}^{n+1},\boldsymbol{W}'\right)-y^{n+1}\right)\right)^{p}-\left(\mathcal{R}\left(F\left(\boldsymbol{x}^{n},\boldsymbol{W}'\right)-y^{n}\right)\right)^{p}\right\}\right| \\
&\le \mathbb{E}_{\mathscr{D}^{n+1}}\left\{\left|\left(\mathcal{R}\left(F\left(\boldsymbol{x}^{n+1},\boldsymbol{W}'\right)-y^{n+1}\right)\right)^{p}-\left(\mathcal{R}\left(F\left(\boldsymbol{x}^{n},\boldsymbol{W}'\right)-y^{n}\right)\right)^{p}\right|\right\} \\
&\le \mathcal{E}^{p-1}p\left(\mathbb{E}_{\mathscr{D}^{n+1}}\left\{\left|F\left(\boldsymbol{x}^{n+1},\boldsymbol{W}'\right)-F\left(\boldsymbol{x}^{n},\boldsymbol{W}'\right)\right|\right\}+\left|y^{n+1}-y^{n}\right|\right) \\
&\le \mathcal{E}^{p-1}p\left(2G\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^{n}\right\|_{2}+\beta_{n}\left|F\left(\boldsymbol{x}^{n},\boldsymbol{W}_{1}^{n+1}\right)-y^{n}\right|\right) \\
&\le \mathcal{E}^{p-1}p\left(2G\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^{n}\right\|_{2}+\beta_{n}\left(m_{h}-m_{l}\right)\right) \\
&\equiv 2G\mathcal{E}^{p-1}p\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^{n}\right\|_{2}+\beta_{n}\mathcal{E}^{p-1}p\left(m_{h}-m_{l}\right), \quad \mathcal{P}-a.e. \qquad (275)
\end{aligned}
$$

Additionally, it is true that

$$
\left|\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n+1},y^{n+1}\right)-\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n},y^{n}\right)\right|^{2}\le 8G^{2}\mathcal{E}^{2p-2}p^{2}\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^{n}\right\|_{2}^{2}+\beta_{n}^{2}2\mathcal{E}^{2p-2}p^{2}\left(m_{h}-m_{l}\right)^{2}. \qquad (276)
$$

Combining (276), (274) and (273), we end up with the inequality

$$
\begin{aligned}
\mathbb{E}_{\mathscr{D}^{n}}&\left\{\left|E_{\mathcal{D}}^{n+1}\right|^{2}\right\}=\left(1-\gamma_{n}\right)\left|E_{\mathcal{D}}^{n}\right|^{2} \\
&+\gamma_{n}^{-1}16G^{2}\mathcal{E}^{2p-2}p^{2}\mathbb{E}_{\mathscr{D}^{n}}\left\{\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^{n}\right\|_{2}^{2}\right\}+\beta_{n}^{2}\gamma_{n}^{-1}4\mathcal{E}^{2p-2}p^{2}\left(m_{h}-m_{l}\right)^{2}+\gamma_{n}^{2}4\mathcal{E}^{2p}, \qquad (277)
\end{aligned}
$$

being valid almost everywhere relative to $\mathcal{P}$. But $\mathbb{N}$ is countable. ∎

7.8 Proof of Lemma 7

As usual, fix $n\in\mathbb{N}^{+}$, and let $p>1$. Nonexpansiveness of the projection operator onto $\mathcal{X}$ yields

$$
\begin{aligned}
\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^{*}\right\|_{2}^{2}&\equiv\left\|\Pi_{\mathcal{X}}\left\{\boldsymbol{x}^{n}-\alpha_{n}\widehat{\nabla}^{n+1}\phi^{\widetilde{F}}\left(\boldsymbol{x}^{n},y^{n},z^{n}\right)\right\}-\Pi_{\mathcal{X}}\left\{\boldsymbol{x}^{*}\right\}\right\|_{2}^{2} \\
&\le\left\|\boldsymbol{x}^{n}-\boldsymbol{x}^{*}-\alpha_{n}\widehat{\nabla}^{n+1}\phi^{\widetilde{F}}\left(\boldsymbol{x}^{n},y^{n},z^{n}\right)\right\|_{2}^{2} \\
&=\left\|\boldsymbol{x}^{n}-\boldsymbol{x}^{*}\right\|_{2}^{2}+\alpha_{n}^{2}\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_{2}^{n+1}\right)+c\Delta^{n+1}\left(\boldsymbol{x}^{n},y^{n},z^{n}\right)\right\|_{2}^{2} \\
&\quad-2\alpha_{n}\left(\boldsymbol{x}^{n}-\boldsymbol{x}^{*}\right)^{\boldsymbol{T}}\left(\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_{2}^{n+1}\right)+c\Delta^{n+1}\left(\boldsymbol{x}^{n},\mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right),\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right)\right)\right)+\boldsymbol{U}^{n+1} \\
&\equiv\left\|\boldsymbol{x}^{n}-\boldsymbol{x}^{*}\right\|_{2}^{2}+\alpha_{n}^{2}\left\|\widehat{\nabla}^{n+1}\phi^{\widetilde{F}}\left(\boldsymbol{x}^{n},y^{n},z^{n}\right)\right\|_{2}^{2} \\
&\quad-2\alpha_{n}\left(\boldsymbol{x}^{n}-\boldsymbol{x}^{*}\right)^{\boldsymbol{T}}\widehat{\nabla}^{n+1}\phi^{\widetilde{F}}\left(\boldsymbol{x}^{n},\mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right),\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right)\right)+\boldsymbol{U}^{n+1}, \qquad (278)
\end{aligned}
$$

everywhere on Ω, where the function $\boldsymbol{U}^{n+1}:\Omega\to\mathbb{R}$ is defined as

$$
\begin{aligned}
&\boldsymbol{U}^{n+1} \qquad (279) \\
&\triangleq 2c\alpha_{n}\left(\boldsymbol{x}^{n}-\boldsymbol{x}^{*}\right)^{\boldsymbol{T}}\left(\Delta^{n+1}\left(\boldsymbol{x}^{n},\mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right),\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right)\right)-\Delta^{n+1}\left(\boldsymbol{x}^{n},y^{n},z^{n}\right)\right) \\
&\equiv 2c\alpha_{n}\left(\boldsymbol{x}^{n}-\boldsymbol{x}^{*}\right)^{\boldsymbol{T}}\left(\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_{2}^{n+1}\right)-\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_{1}^{n+1}\right)\right)
\end{aligned}
$$

$$\times\left(\underbrace{\underline{\nabla}\left(\mathcal{R}\left(r\right)\right)^{p}\big|_{r\equiv F\left(\boldsymbol{x}^{n},\boldsymbol{W}_{2}^{n+1}\right)-\mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right)}}_{\triangleq\mathsf{A}_{1}^{n}\left(\boldsymbol{x}^{n}\right)}\underbrace{\left(\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right)\right)^{(1-p)/p}}_{\triangleq\mathsf{C}_{1}^{n}\left(\boldsymbol{x}^{n}\right)}-\underbrace{\underline{\nabla}\left(\mathcal{R}\left(r\right)\right)^{p}\big|_{r\equiv F\left(\boldsymbol{x}^{n},\boldsymbol{W}_{2}^{n+1}\right)-y^{n}}}_{\triangleq\mathsf{A}_{2}^{n}\left(\boldsymbol{x}^{n},y^{n}\right)}\underbrace{\left(z^{n}\right)^{(1-p)/p}}_{\triangleq\mathsf{C}_{2}^{n}\left(z^{n}\right)}\right).$$

From the proof of Lemma 3 (Section 7.5), it readily follows that

$$\mathbb{E}_{\mathscr{D}^{n}}\left\{\left\|\widehat{\nabla}^{n+1}\phi^{\widetilde{F}}\left(\boldsymbol{x}^{n},y^{n},z^{n}\right)\right\|_{2}^{2}\right\}\le\left(2c\mathsf{R}_{p}+1\right)^{2}G^{2},\quad\mathcal{P}-a.e. \tag{280}$$

Hence, taking conditional expectations on both sides of (278) relative to $\mathscr{D}^{n}$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathscr{D}^{n}}\left\{\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^{*}\right\|_{2}^{2}\right\}\\
&\quad\le\left\|\boldsymbol{x}^{n}-\boldsymbol{x}^{*}\right\|_{2}^{2}+\alpha_{n}^{2}\left(2c\mathsf{R}_{p}+1\right)^{2}G^{2}\\
&\qquad-2\alpha_{n}\left(\boldsymbol{x}^{n}-\boldsymbol{x}^{*}\right)^{\boldsymbol{T}}\mathbb{E}_{\mathscr{D}^{n}}\left\{\widehat{\nabla}^{n+1}\phi^{\widetilde{F}}\left(\boldsymbol{x}^{n},\mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right),\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right)\right)\right\}+\mathbb{E}_{\mathscr{D}^{n}}\left\{\boldsymbol{U}^{n+1}\right\}\\
&\quad\equiv\left\|\boldsymbol{x}^{n}-\boldsymbol{x}^{*}\right\|_{2}^{2}+\alpha_{n}^{2}\left(2c\mathsf{R}_{p}+1\right)^{2}G^{2}-2\alpha_{n}\left(\boldsymbol{x}^{n}-\boldsymbol{x}^{*}\right)^{\boldsymbol{T}}\nabla\phi^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right)+\mathbb{E}_{\mathscr{D}^{n}}\left\{\boldsymbol{U}^{n+1}\right\}\\
&\quad\le\left\|\boldsymbol{x}^{n}-\boldsymbol{x}^{*}\right\|_{2}^{2}+\alpha_{n}^{2}\left(2c\mathsf{R}_{p}+1\right)^{2}G^{2}-2\alpha_{n}\left(\phi^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right)-\phi_{*}^{\widetilde{F}}\right)+\mathbb{E}_{\mathscr{D}^{n}}\left\{\boldsymbol{U}^{n+1}\right\},
\end{aligned} \tag{281}$$

almost everywhere relative to $\mathcal{P}$, where in the last inequality, we have exploited our assumption that the objective function $\phi^{\widetilde{F}}$ is convex. Therefore, our main concern now is properly bounding $\mathbb{E}_{\mathscr{D}^{n}}\left\{\boldsymbol{U}^{n+1}\right\}$. By Cauchy-Schwarz, $\boldsymbol{U}^{n+1}$ may be bounded from above as

$$\begin{aligned}
\boldsymbol{U}^{n+1}\le 2c\alpha_{n}\left\|\boldsymbol{x}^{n}-\boldsymbol{x}^{*}\right\|_{2}&\left(\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_{1}^{n+1}\right)\right\|_{2}+\left\|\underline{\nabla}F\left(\boldsymbol{x}^{n},\boldsymbol{W}_{2}^{n+1}\right)\right\|_{2}\right)\\
&\times\left|\mathsf{A}_{1}^{n}\left(\boldsymbol{x}^{n}\right)\mathsf{C}_{1}^{n}\left(\boldsymbol{x}^{n}\right)-\mathsf{A}_{2}^{n}\left(\boldsymbol{x}^{n},y^{n}\right)\mathsf{C}_{2}^{n}\left(z^{n}\right)\right|,
\end{aligned} \tag{282}$$

everywhere on Ω, as well.

Now, *let Assumption 5 be in effect.* Making use of the fact that $\underline{\nabla}\mathcal{R}$ is uniformly upper bounded by unity and of Lemmata 11 and 12, and the resulting inequality

$$\begin{aligned}
\mathsf{A}_{1}^{n}\left(\boldsymbol{x}^{n}\right)&\equiv\underline{\nabla}\left(\mathcal{R}\left(r\right)\right)^{p}\big|_{r\equiv F\left(\boldsymbol{x}^{n},\boldsymbol{W}_{2}^{n+1}\right)-\mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right)}\\
&\equiv p\left(\mathcal{R}\left(F\left(\boldsymbol{x}^{n},\boldsymbol{W}_{2}^{n+1}\right)-\mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right)\right)\right)^{p-1}\underline{\nabla}\mathcal{R}\left(F\left(\boldsymbol{x}^{n},\boldsymbol{W}_{2}^{n+1}\right)-\mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^{n}\right)\right)\\
&\le p\mathcal{E}^{p-1},
\end{aligned} \tag{283}$$

we may further bound the absolute difference on the RHS of (282) from above as

$$\begin{aligned}
&\left|\mathsf{A}_{1}^{n}\left(\boldsymbol{x}^{n}\right)\mathsf{C}_{1}^{n}\left(\boldsymbol{x}^{n}\right)-\mathsf{A}_{2}^{n}\left(\boldsymbol{x}^{n},y^{n}\right)\mathsf{C}_{2}^{n}\left(z^{n}\right)\right|\\
&\quad\le\mathsf{C}_{2}^{n}\left(z^{n}\right)\left|\mathsf{A}_{1}^{n}(\boldsymbol{x}^{n})-\mathsf{A}_{2}^{n}(\boldsymbol{x}^{n},y^{n})\right|+\mathsf{A}_{1}^{n}\left(\boldsymbol{x}^{n}\right)\left|\mathsf{C}_{1}^{n}(\boldsymbol{x}^{n})-\mathsf{C}_{2}^{n}(z^{n})\right|\\
&\quad\le\left(\frac{1}{\varepsilon}\right)^{p-1}\left|\mathsf{A}_{1}^{n}(\boldsymbol{x}^{n})-\mathsf{A}_{2}^{n}(\boldsymbol{x}^{n},y^{n})\right|+p\mathcal{E}^{p-1}\left|\mathsf{C}_{1}^{n}(\boldsymbol{x}^{n})-\mathsf{C}_{2}^{n}(z^{n})\right|\\
&\quad\le\left(\frac{1}{\varepsilon}\right)^{p-1}\left|\mathsf{A}_{1}^{n}(\boldsymbol{x}^{n})-\mathsf{A}_{2}^{n}(\boldsymbol{x}^{n},y^{n})\right|+(p-1)\frac{\mathcal{E}^{p-1}}{\varepsilon^{2p-1}}\left|z^{n}-\mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^{n})\right|,
\end{aligned} \tag{284}$$

almost everywhere relative to $\mathcal{P}$. Utilizing (284) and taking conditional expectations relative to $\mathscr{D}^n$ on both sides of (282), we have

$$\begin{aligned}
&\mathbb{E}_{\mathscr{D}^n}\left\{\boldsymbol{U}^{n+1}\right\} \le 2c\alpha_n \left\|\boldsymbol{x}^n - \boldsymbol{x}^*\right\|_2 \\
&\times\left(\left(\frac{1}{\varepsilon}\right)^{p-1}\mathbb{E}_{\mathscr{D}^n}\left\{\left(\left\|\underline{\nabla}F\left(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\right)\right\|_2+\left\|\underline{\nabla}F\left(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\right)\right\|_2\right)\left|\mathsf{A}_1^n\left(\boldsymbol{x}^n\right)-\mathsf{A}_2^n\left(\boldsymbol{x}^n,y^n\right)\right|\right\}\right. \\
&\left.+\left(p-1\right)\frac{\mathcal{E}^{p-1}}{\varepsilon^{2p-1}}\mathbb{E}_{\mathscr{D}^n}\left\{\left(\left\|\underline{\nabla}F\left(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\right)\right\|_2+\left\|\underline{\nabla}F\left(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\right)\right\|_2\right)\left|z^n-\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^n\right)\right|\right\}\right),
\end{aligned} \tag{285}$$

almost everywhere relative to $\mathcal{P}$. Let us consider each of the three terms on the RHS of (285) separately. Regarding the term related to z^n, exploiting condition **C1**, we may write

$$\begin{aligned}
&\mathbb{E}_{\mathscr{D}^n}\left\{\left(\left\|\underline{\nabla}F\left(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\right)\right\|_2+\left\|\underline{\nabla}F\left(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\right)\right\|_2\right)\right\} \\
&\equiv \left\|\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_1^{n+1}\right)\right\|_2+\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_2^{n+1}\right)\right\|_2\right\|_{\mathcal{L}_P}\Bigg|_{\boldsymbol{x}\equiv\boldsymbol{x}^n}\left|z^n-\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^n\right)\right| \\
&\le 2G\left|z^n-\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^n\right)\right|, \quad \mathcal{P}-a.e.
\end{aligned} \tag{286}$$

As far as the remaining term of (285) containing A_1^n and A_2^n is concerned, the situation is somewhat more complicated. Recall that, by assumption, we have $P \in [2,\infty]$, $P/(P-1) \le Q \le \infty$, and $P^{-1}+Q^{-1} \le 1$. Then, we may invoke the *generalized* Hölder's Inequality for finite measure spaces (on the *appropriate* Borel space), along with condition **C3**, obtaining

$$\begin{aligned}
&\mathbb{E}_{\mathscr{D}^n}\left\{\left(\left\|\underline{\nabla}F\left(\boldsymbol{x}^n,\boldsymbol{W}_1^{n+1}\right)\right\|_2+\left\|\underline{\nabla}F\left(\boldsymbol{x}^n,\boldsymbol{W}_2^{n+1}\right)\right\|_2\right)\left|\mathsf{A}_1^n\left(\boldsymbol{x}^n\right)-\mathsf{A}_2^n\left(\boldsymbol{x}^n,y^n\right)\right|\right\} \\
&\le \left\|\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_1^{n+1}\right)\right\|_2+\left\|\underline{\nabla}F\left(\boldsymbol{x},\boldsymbol{W}_2^{n+1}\right)\right\|_2\right\|_{\mathcal{L}_P}\Bigg|_{\boldsymbol{x}\equiv\boldsymbol{x}^n}\left\|\left|\mathsf{A}_1^n\left(\boldsymbol{x}\right)-\mathsf{A}_2^n\left(\boldsymbol{x},y\right)\right|\right\|_{\mathcal{L}_Q}\Bigg|_{\boldsymbol{x}\equiv\boldsymbol{x}^n,y\equiv y^n} \\
&\le 2G\left\|\left|\underline{\nabla}\left(\mathcal{R}\left(r\right)\right)^p\Big|_{r\equiv F\left(\boldsymbol{x},\boldsymbol{W}_2^{n+1}\right)-\mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}\right)}-\underline{\nabla}\left(\mathcal{R}\left(r\right)\right)^p\Big|_{r\equiv F\left(\boldsymbol{x},\boldsymbol{W}_2^{n+1}\right)-y}\right|\right\|_{\mathcal{L}_Q}\Bigg|_{\boldsymbol{x}\equiv\boldsymbol{x}^n,y\equiv y^n} \\
&\le 2GD\left|y-\mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}\right)\right|\Bigg|_{\boldsymbol{x}\equiv\boldsymbol{x}^n,y\equiv y^n} \equiv 2GD\left|y^n-\mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^n\right)\right|,
\end{aligned} \tag{287}$$

almost everywhere relative to $\mathcal{P}$. Combining (285) with (286) and (287), we readily obtain the upper bound

$$\mathbb{E}_{\mathscr{D}^n}\left\{\boldsymbol{U}^{n+1}\right\} \le 4\widetilde{\mathsf{B}}_p G c\alpha_n\left\|\boldsymbol{x}^n-\boldsymbol{x}^*\right\|_2\left(\left|y^n-\mathcal{S}^{\widetilde{F}}\left(\boldsymbol{x}^n\right)\right|+\left|z^n-\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^n\right)\right|\right), \quad \mathcal{P}-a.e., \tag{288}$$

where the constant $\widetilde{\mathsf{B}}_p < \infty$ is defined as

$$\widetilde{\mathsf{B}}_p \triangleq \max\left\{\left(\frac{1}{\varepsilon}\right)^{p-1}D, \left(p-1\right)\frac{\mathcal{E}^{p-1}}{\varepsilon^{2p-1}}\right\}. \tag{289}$$

As a next step, recalling Lemma 12, we observe that

$$\left|z^n-\mathcal{D}^{\widetilde{F}}\left(\boldsymbol{x}^n\right)\right|$$

$$
\begin{aligned}
&\equiv \left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n) + \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n) - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n)\right| \\
&\le \left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right| + \left|\mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n) - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n)\right| \\
&\equiv \left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right| \\
&\quad + \left|\mathbb{E}_{\mathscr{D}^n}\left\{\left(\mathcal{R}\left(F(\boldsymbol{x}^n,\boldsymbol{W}') - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right)\right)^p\right\} - \mathbb{E}_{\mathscr{D}^n}\left\{\left(\mathcal{R}\left(F(\boldsymbol{x}^n,\boldsymbol{W}') - y^n\right)\right)^p\right\}\right| \\
&\equiv \left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right| + \left|\mathbb{E}_{\mathscr{D}^n}\left\{\left(\mathcal{R}\left(F(\boldsymbol{x}^n,\boldsymbol{W}') - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right)\right)^p - \left(\mathcal{R}\left(F(\boldsymbol{x}^n,\boldsymbol{W}') - y^n\right)\right)^p\right\}\right| \\
&\le \left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right| + \mathbb{E}_{\mathscr{D}^n}\left\{\left|\left(\mathcal{R}\left(F(\boldsymbol{x}^n,\boldsymbol{W}') - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right)\right)^p - \left(\mathcal{R}\left(F(\boldsymbol{x}^n,\boldsymbol{W}') - y^n\right)\right)^p\right|\right\} \\
&\le \left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right| + \mathcal{E}^{p-1}p\left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|, \quad \mathcal{P}-a.e.. \qquad (290)
\end{aligned}
$$

This yields, in turn,

$$
\begin{aligned}
\mathbb{E}_{\mathscr{D}^n}\left\{\boldsymbol{U}^{n+1}\right\} &\le 4\mathsf{B}_p G c\alpha_n \left\|\boldsymbol{x}^n - \boldsymbol{x}^*\right\|_2 \left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right| \\
&\quad + 4\mathsf{B}_p G c\alpha_n \left\|\boldsymbol{x}^n - \boldsymbol{x}^*\right\|_2 \left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|, \qquad (291)
\end{aligned}
$$

where $\mathsf{B}_p \triangleq \left(1 + \mathcal{E}^{p-1}p\right)\widetilde{\mathsf{B}}_p$. Finally, invoking Lemmata 5 and 6, $\mathbb{E}_{\mathscr{D}^n}\left\{\boldsymbol{U}^{n+1}\right\}$ may be further bounded as

$$
\begin{aligned}
\mathbb{E}_{\mathscr{D}^n}\left\{\boldsymbol{U}^{n+1}\right\} &\le 4\mathsf{B}_p^2 G^2 c^2 \frac{\alpha_n^2}{\beta_n}\left\|\boldsymbol{x}^n - \boldsymbol{x}^*\right\|_2^2 + \beta_n\left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|^2 \\
&\quad + 4\mathsf{B}_p^2 G^2 c^2 \frac{\alpha_n^2}{\gamma_n}\left\|\boldsymbol{x}^n - \boldsymbol{x}^*\right\|_2^2 + \gamma_n\left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|^2, \qquad (292)
\end{aligned}
$$

almost everywhere relative to $\mathcal{P}$.

As a result, invoking Lemma 11, (281) may be further bounded from above as

$$
\begin{aligned}
&\mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\right\|_2^2\right\} \\
&\le \left(1 + 4\mathsf{B}_p^2 G^2 c^2\left(\frac{\alpha_n^2}{\beta_n} + \frac{\alpha_n^2}{\gamma_n}\right)\right)\left\|\boldsymbol{x}^n - \boldsymbol{x}^*\right\|_2^2 + \alpha_n^2\left(2c\mathsf{R}_p + 1\right)^2 G^2 - 2\alpha_n\left(\phi^{\widetilde{F}}(\boldsymbol{x}^n) - \phi_*^{\widetilde{F}}\right) \\
&\quad + \beta_n\left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|^2 + \gamma_n\left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|^2 \qquad (293)
\end{aligned}
$$

almost everywhere relative to $\mathcal{P}$, yielding (118) in the statement of Lemma 6. The fact that $\mathbb{N}^+$ is countable completes the proof when $p > 1$. For the case where $p \equiv 1$, exactly the same procedure yields the constant $\mathsf{B}_p \equiv D$, whereas the ratio α_n^2/γ_n and last term on the RHS of (293) (left to right) *disappears.* ∎

7.9 A Generalization of Chung's Lemma

The following result is a generalization of Chung's Lemma [Chung, 1954], which is an old and well-known result for analyzing convergence rates of stochastic approximation algorithms.

Lemma 13. (Generalized Chung's Lemma) *Consider any nonnegative sequence* $\{S^n\}_{n\in\mathbb{N}}$, *such that*

$$S^{n+1} \le (1-\alpha_n) S^n + C\beta_n, \quad \forall n \in \mathbb{N}, \tag{294}$$

where $\{\alpha_n\}_{n\in\mathbb{N}}$, $\{\beta_n\}_{n\in\mathbb{N}}$ *are also nonnegative sequences, and* $C \ge 0$. *Suppose that*

$$n^+ \triangleq \min\left\{n \in \mathbb{N} \middle| \alpha_{n'} \le 1, \ \forall n' \in \mathbb{N}^n\right\} \in [0,\infty), \tag{295}$$

and choose $n_o \in \mathbb{N}^{n^+}$. *Then, for every* $n \in \mathbb{N}^{n_o}$, *it is true that*

$$S^{n+1} \le S^{n_o} \prod_{i\in\mathbb{N}_n^{n_o}} (1-\alpha_i) + C \sum_{i\in\mathbb{N}_n^{n_o}} \beta_i \prod_{j\in\mathbb{N}_n^{i+1}} (1-\alpha_j), \tag{296}$$

where, by convention, $\prod_{j\in\mathbb{N}_n^{n+1}}(\cdot) \equiv \prod_{j=n+1}^{n}(\cdot) \equiv 1$.

Proof of Lemma 13. Use simple induction; enough said. ∎

7.10 Proof of Lemma 9

In the following, we consider the case where $p > 1$. If $p \equiv 1$, the proof is similar, albeit simpler. From the proof of Lemma 7, we have already shown that

$$\mathbb{E}_{\mathscr{D}^n}\left\{\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\right\|_2^2\right\} \le \|\boldsymbol{x}^n - \boldsymbol{x}^*\|_2^2 + \alpha_n^2 (2c\mathsf{R}_p + 1)^2 G^2 - 2\alpha_n \left(\phi^{\widetilde{F}}(\boldsymbol{x}^n) - \phi_*^{\widetilde{F}}\right) + \mathbb{E}_{\mathscr{D}^n}\left\{\boldsymbol{U}^{n+1}\right\}, \tag{297}$$

and that

$$\mathbb{E}_{\mathscr{D}^n}\left\{\boldsymbol{U}^{n+1}\right\} \le 4\mathsf{B}_p G c\alpha_n \|\boldsymbol{x}^n - \boldsymbol{x}^*\|_2 \left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right| + 4\mathsf{B}_p G c\alpha_n \|\boldsymbol{x}^n - \boldsymbol{x}^*\|_2 \left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|, \tag{298}$$

almost everywhere relative to $\mathcal{P}$, for each $n \in \mathbb{N}$.

First, we exploit our assumption in regard to strong convexity of the objective $\phi^{\widetilde{F}}$, and taking expectations on both sides of (297), we get

$$\mathbb{E}\left\{\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\right\|_2^2\right\} \le (1 - 2\sigma\alpha_n)\mathbb{E}\left\{\|\boldsymbol{x}^n - \boldsymbol{x}^*\|_2^2\right\} + \alpha_n^2 (2c\mathsf{R}_p + 1)^2 G^2 + \mathbb{E}\left\{\boldsymbol{U}^{n+1}\right\}, \tag{299}$$

being true for all $n \in \mathbb{N}$. Let us focus on appropriately bounding the term $\mathbb{E}\left\{\boldsymbol{U}^{n+1}\right\}$, for $n \in \mathbb{N}$. It is true that

$$\begin{aligned}
\mathbb{E}_{\mathscr{D}^n}\left\{\boldsymbol{U}^{n+1}\right\} &\le \frac{\sigma}{2}\alpha_n \|\boldsymbol{x}^n - \boldsymbol{x}^*\|_2^2 + \alpha_n \frac{8\mathsf{B}_p^2 G^2 c^2}{\sigma}\left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|^2 \\
&\quad + \frac{\sigma}{2}\alpha_n \|\boldsymbol{x}^n - \boldsymbol{x}^*\|_2^2 + \alpha_n \frac{8\mathsf{B}_p^2 G^2 c^2}{\sigma}\left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|^2 \\
&\equiv \sigma\alpha_n \|\boldsymbol{x}^n - \boldsymbol{x}^*\|_2^2 + \alpha_n \frac{8\mathsf{B}_p^2 G^2 c^2}{\sigma}\left(\left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|^2 + \left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|^2\right),
\end{aligned} \tag{300}$$

almost everywhere relative to $\mathcal{P}$. Again, taking expectations on both sides, we have

$$\mathbb{E}\left\{\boldsymbol{U}^{n+1}\right\} \le \sigma\alpha_n \mathbb{E}\left\{\|\boldsymbol{x}^n - \boldsymbol{x}^*\|_2^2\right\} + \alpha_n \frac{8\mathsf{B}_p^2 G^2 c^2}{\sigma}\left(\mathbb{E}\left\{\left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|^2\right\} + \mathbb{E}\left\{\left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|^2\right\}\right), \tag{301}$$

for all $n \in \mathbb{N}$. For brevity, we hereafter make the identifications

$$A^n \triangleq \mathbb{E}\left\{\left\|\boldsymbol{x}^n - \boldsymbol{x}^*\right\|_2^2\right\}, \tag{302}$$

$$B^n \triangleq \mathbb{E}\left\{\left|y^n - \mathcal{S}^{\widetilde{F}}(\boldsymbol{x}^n)\right|^2\right\} \quad \text{and} \tag{303}$$

$$C^n \triangleq \mathbb{E}\left\{\left|z^n - \mathcal{D}^{\widetilde{F}}(\boldsymbol{x}^n, y^n)\right|^2\right\}, \quad \forall n \in \mathbb{N}. \tag{304}$$

As a result, we arrive at the inequalities

$$A^{n+1} \leq (1 - 2\sigma\alpha_n) A^n + \alpha_n^2 \left(2c\mathsf{R}_p + 1\right)^2 G^2 + \mathbb{E}\left\{\boldsymbol{U}^{n+1}\right\} \quad \text{and} \tag{305}$$

$$\mathbb{E}\left\{\boldsymbol{U}^{n+1}\right\} \leq \sigma\alpha_n A^n + \alpha_n \frac{8\mathsf{B}_p^2 G^2 c^2}{\sigma}\left(B^n + C^n\right), \quad \forall n \in \mathbb{N}, \tag{306}$$

which imply that

$$A^{n+1} \leq (1 - \sigma\alpha_n) A^n + \alpha_n^2 \left(2c\mathsf{R}_p + 1\right)^2 G^2 + \alpha_n \frac{8\mathsf{B}_p^2 G^2 c^2}{\sigma}\left(B^n + C^n\right), \quad \forall n \in \mathbb{N}. \tag{307}$$

By Lemmata 3, 5 and 6, we know that (by taking expectations on both sides), for every $n \in \mathbb{N}^+$, B^n and C^n satisfy the recursive inequalities

$$B^n \leq (1 - \beta_{n-1}) B^{n-1} + \frac{\alpha_{n-1}^2}{\beta_{n-1}} 2\left(2c\mathsf{R}_p + 1\right)^2 G^4 + \beta_{n-1}^2 2V \quad \text{and} \tag{308}$$

$$C^n \leq (1 - \gamma_{n-1}) C^{n-1} + \frac{\alpha_{n-1}^2}{\gamma_{n-1}} 16 G^4 \mathcal{E}^{2p-2} p^2 \left(2c\mathsf{R}_p + 1\right)^2 + \frac{\beta_{n-1}^2}{\gamma_{n-1}} 4\mathcal{E}^{2p-2} p^2 (m_h - m_l) + \gamma_{n-1}^2 2\mathcal{E}^{2p}. \tag{309}$$

Now, for simplicity and clarity, let us define a strictly *problem dependent* constant

$$\begin{aligned} \Sigma \triangleq \max\Big\{ & \left(2c\mathsf{R}_p + 1\right)^2 G^2, 8\mathsf{B}_p^2 G^2 c^2, 2\left(2c\mathsf{R}_p + 1\right)^2 G^4, 2V, \\ & 16 G^4 \mathcal{E}^{2p-2} p^2 \left(2c\mathsf{R}_p + 1\right)^2, 4\mathcal{E}^{2p-2} p^2 (m_h - m_l), 2\mathcal{E}^{2p}\Big\}. \end{aligned} \tag{310}$$

Then, it is true that

$$A^{n+1} \leq (1 - \sigma\alpha_n) A^n + \alpha_n^2 \Sigma + \alpha_n \frac{\Sigma}{\sigma} B^n + \alpha_n \frac{\Sigma}{\sigma} C^n, \quad \text{with} \tag{311}$$

$$B^n \leq (1 - \beta_{n-1}) B^{n-1} + \frac{\alpha_{n-1}^2}{\beta_{n-1}} \Sigma + \beta_{n-1}^2 \Sigma \quad \text{and} \tag{312}$$

$$C^n \leq (1 - \gamma_{n-1}) C^{n-1} + \frac{\alpha_{n-1}^2}{\gamma_{n-1}} \Sigma + \frac{\beta_{n-1}^2}{\gamma_{n-1}} \Sigma + \gamma_{n-1}^2 \Sigma, \quad \forall n \in \mathbb{N}^+. \tag{313}$$

Next, let $\{\boldsymbol{\Delta}_B^n \geq 0\}_{n \in \mathbb{N}^+}$ and $\{\boldsymbol{\Delta}_C^n \geq 0\}_{n \in \mathbb{N}^+}$ be two *auxiliary correction sequences*, to be determined shortly. It then follows that

$$\boldsymbol{\Delta}_B^n B^n \leq (1 - \beta_{n-1}) \boldsymbol{\Delta}_B^n B^n + \frac{\alpha_{n-1}^2}{\beta_{n-1}} \boldsymbol{\Delta}_B^n \Sigma + \beta_{n-1}^2 \boldsymbol{\Delta}_B^n \Sigma, \tag{314}$$

implying that, for every $n \in \mathbb{N}^+$,

$$\left(\boldsymbol{\Delta}_B^n + \alpha_n \frac{\Sigma}{\sigma}\right) B^n \le (1-\beta_{n-1}) \left(\boldsymbol{\Delta}_B^n + \alpha_n \frac{\Sigma}{\sigma}\right) B^{n-1} + \frac{\alpha_{n-1}^2}{\beta_{n-1}} \boldsymbol{\Delta}_B^n \Sigma + \beta_{n-1}^2 \boldsymbol{\Delta}_B^n \Sigma + \frac{\alpha_n \alpha_{n-1}^2}{\beta_{n-1}} \frac{\Sigma^2}{\sigma} + \alpha_n \beta_{n-1}^2 \frac{\Sigma^2}{\sigma}. \tag{315}$$

Exactly the same procedure for (313) yields

$$\left(\boldsymbol{\Delta}_C^n + \alpha_n \frac{\Sigma}{\sigma}\right) C^n \le (1-\gamma_{n-1}) \left(\boldsymbol{\Delta}_C^n + \alpha_n \frac{\Sigma}{\sigma}\right) C^{n-1} + \frac{\alpha_{n-1}^2}{\gamma_{n-1}} \boldsymbol{\Delta}_C^n \Sigma + \frac{\beta_{n-1}^2}{\gamma_{n-1}} \boldsymbol{\Delta}_C^n \Sigma + \gamma_{n-1}^2 \boldsymbol{\Delta}_C^n \Sigma + \frac{\alpha_n \alpha_{n-1}^2}{\gamma_{n-1}} \frac{\Sigma^2}{\sigma} + \frac{\alpha_n \beta_{n-1}^2}{\gamma_{n-1}} \frac{\Sigma^2}{\sigma} + \alpha_n \gamma_{n-1}^2 \frac{\Sigma^2}{\sigma}, \tag{316}$$

for all $n \in \mathbb{N}^+$. Simply combining (311), (315) and (316), we obtain

$$\begin{aligned}
&A^{n+1} + \left(\boldsymbol{\Delta}_B^n + \alpha_n \frac{\Sigma}{\sigma}\right) B^n + \left(\boldsymbol{\Delta}_C^n + \alpha_n \frac{\Sigma}{\sigma}\right) C^n \\
&\le (1-\sigma\alpha_n) A^n + (1-\beta_{n-1}) \left(\boldsymbol{\Delta}_B^n + \alpha_n \frac{\Sigma}{\sigma}\right) B^{n-1} + (1-\gamma_{n-1}) \left(\boldsymbol{\Delta}_C^n + \alpha_n \frac{\Sigma}{\sigma}\right) C^{n-1} \\
&+ \alpha_n^2 \Sigma + \alpha_n \frac{\Sigma}{\sigma} B^n + \alpha_n \frac{\Sigma}{\sigma} C^n \\
&+ \frac{\alpha_{n-1}^2}{\beta_{n-1}} \boldsymbol{\Delta}_B^n \Sigma + \beta_{n-1}^2 \boldsymbol{\Delta}_B^n \Sigma + \frac{\alpha_n \alpha_{n-1}^2}{\beta_{n-1}} \frac{\Sigma^2}{\sigma} + \alpha_n \beta_{n-1}^2 \frac{\Sigma^2}{\sigma} \\
&+ \frac{\alpha_{n-1}^2}{\gamma_{n-1}} \boldsymbol{\Delta}_C^n \Sigma + \frac{\beta_{n-1}^2}{\gamma_{n-1}} \boldsymbol{\Delta}_C^n \Sigma + \gamma_{n-1}^2 \boldsymbol{\Delta}_C^n \Sigma + \frac{\alpha_n \alpha_{n-1}^2}{\gamma_{n-1}} \frac{\Sigma^2}{\sigma} + \frac{\alpha_n \beta_{n-1}^2}{\gamma_{n-1}} \frac{\Sigma^2}{\sigma} + \alpha_n \gamma_{n-1}^2 \frac{\Sigma^2}{\sigma}, \ \forall n \in \mathbb{N}^+,
\end{aligned} \tag{317}$$

or, equivalently,

$$\begin{aligned}
&A^{n+1} + \boldsymbol{\Delta}_B^n B^n + \boldsymbol{\Delta}_C^n C^n \\
&\le (1-\sigma\alpha_n) A^n + (1-\beta_{n-1}) \left(\boldsymbol{\Delta}_B^n + \alpha_n \frac{\Sigma}{\sigma}\right) B^{n-1} + (1-\gamma_{n-1}) \left(\boldsymbol{\Delta}_C^n + \alpha_n \frac{\Sigma}{\sigma}\right) C^{n-1} \\
&+ \alpha_n^2 \Sigma \\
&+ \frac{\alpha_{n-1}^2}{\beta_{n-1}} \boldsymbol{\Delta}_B^n \Sigma + \beta_{n-1}^2 \boldsymbol{\Delta}_B^n \Sigma + \frac{\alpha_n \alpha_{n-1}^2}{\beta_{n-1}} \frac{\Sigma^2}{\sigma} + \alpha_n \beta_{n-1}^2 \frac{\Sigma^2}{\sigma} \\
&+ \frac{\alpha_{n-1}^2}{\gamma_{n-1}} \boldsymbol{\Delta}_C^n \Sigma + \frac{\beta_{n-1}^2}{\gamma_{n-1}} \boldsymbol{\Delta}_C^n \Sigma + \gamma_{n-1}^2 \boldsymbol{\Delta}_C^n \Sigma + \frac{\alpha_n \alpha_{n-1}^2}{\gamma_{n-1}} \frac{\Sigma^2}{\sigma} + \frac{\alpha_n \beta_{n-1}^2}{\gamma_{n-1}} \frac{\Sigma^2}{\sigma} + \alpha_n \gamma_{n-1}^2 \frac{\Sigma^2}{\sigma}, \ \forall n \in \mathbb{N}^+.
\end{aligned} \tag{318}$$

Driven by the form of (318), we would first like to choose $\boldsymbol{\Delta}_B^n$ and $\boldsymbol{\Delta}_C^n$ such that, *at least for n sufficiently large*,

$$(1-\beta_{n-1}) \left(\boldsymbol{\Delta}_B^n + \alpha_n \frac{\Sigma}{\sigma}\right) \le (1-\sigma\alpha_n) \boldsymbol{\Delta}_B^n, \quad \text{and} \tag{319}$$

$$(1-\gamma_{n-1}) \left(\boldsymbol{\Delta}_C^n + \alpha_n \frac{\Sigma}{\sigma}\right) \le (1-\sigma\alpha_n) \boldsymbol{\Delta}_C^n. \tag{320}$$

The following procedure is exactly the same for both terms, so let us take, say, $\mathbf{\Delta}_B^n$. We may write (319) equivalently as

$$(1-\beta_{n-1})\,\alpha_n\frac{\Sigma}{\sigma}\le(\beta_{n-1}-\sigma\alpha_n)\,\mathbf{\Delta}_B^n. \tag{321}$$

Since, by condition **G1**,

$$\sigma\alpha_n\le\frac{K-1}{K}\min\{\beta_{n-1},\gamma_{n-1}\}<\beta_{n-1}\le 1, \tag{322}$$

for every $n\in\mathbb{N}^{n_o}$ and some globally fixed $n_o\in\mathbb{N}^+$, (319) is also equivalent to

$$\frac{(1-\beta_{n-1})\,\alpha_n}{\beta_{n-1}-\sigma\alpha_n}\frac{\Sigma}{\sigma}\le\mathbf{\Delta}_B^n,\quad\forall n\in\mathbb{N}^{n_o}. \tag{323}$$

Therefore, it suffices to choose

$$\mathbf{\Delta}_B^n\triangleq\frac{\alpha_n}{\beta_{n-1}-\sigma\alpha_n}\frac{\Sigma}{\sigma},\quad\forall n\in\mathbb{N}^{n_o}. \tag{324}$$

Additionally, by defining $\mathbf{\Delta}_B^{n_o-1}\triangleq\mathbf{\Delta}_B^{n_o}$, it is easy to see that conditions **G1** and **G2** imply that

$$\frac{\alpha_n}{\beta_{n-1}}\frac{\Sigma}{\sigma}\le\mathbf{\Delta}_B^n\le K\frac{\alpha_n}{\beta_{n-1}}\frac{\Sigma}{\sigma}\quad\text{and} \tag{325}$$

$$\mathbf{\Delta}_B^n\le\mathbf{\Delta}_B^{n-1},\quad\forall n\in\mathbb{N}^{n_o}, \tag{326}$$

respectively. Of course, similar results hold in a completely analogous fashion for $\mathbf{\Delta}_C^n$, for every $n\in\mathbb{N}^{n_o}$. Consequently, letting

$$J^n\triangleq A^n+\mathbf{\Delta}_B^{n-1}B^{n-1}+\mathbf{\Delta}_C^{n-1}C^{n-1},\quad\forall n\in\mathbb{N}^+, \tag{327}$$

and defining the constant

$$\widetilde{\Sigma}\triangleq\frac{\Sigma}{\sigma^2}\max\left\{(K+1)\,\frac{\Sigma}{\sigma^2},(K+1)\,\Sigma,1\right\}, \tag{328}$$

standard manipulations show that the RHS of expression (318) may be further bounded as

$$J^{n+1}\le(1-\sigma\alpha_n)\,J^n+\widetilde{\Sigma}\left(\sigma^2\alpha_n^2+\frac{\sigma^3\alpha_n\alpha_{n-1}^2}{\beta_{n-1}^2}+\sigma\alpha_n\beta_{n-1}+\frac{\sigma^3\alpha_n\alpha_{n-1}^2}{\gamma_{n-1}^2}+\frac{\sigma\alpha_n\beta_{n-1}^2}{\gamma_{n-1}^2}+\sigma\alpha_n\gamma_{n-1}\right), \tag{329}$$

for all $n\in\mathbb{N}^{n_o}$.

Lastly, let us now show the last part of Lemma 9. If, additionally, the assumptions of Lemma 8 are in effect, it follows that

$$\sup_{n\in\mathbb{N}}B^n<\infty\quad\text{and}\quad\sup_{n\in\mathbb{N}}C^n<\infty. \tag{330}$$

Thus, we may write

$$A^{n+1}\le(1-\sigma\alpha_n)\,A^n+\alpha_n^2\sigma^2\Lambda+\alpha_n\sigma\Lambda,\quad\forall n\in\mathbb{N}. \tag{331}$$

where

$$\Lambda \triangleq \frac{G^2}{\sigma^2} \max\left\{\left(2c\mathsf{R}_p + 1\right)^2, 8\mathsf{B}_p^2 c^2 \left(\sup_{n\in\mathbb{N}} B^n + \sup_{n\in\mathbb{N}} C^n\right)\right\} \tag{332}$$

We now make use of conditions **G1** and **G2**. It is true that

$$\begin{aligned} A^1 &\le (1-\sigma\alpha_0) A^0 + \alpha_0^2\sigma^2\Lambda + \alpha_0\sigma\Lambda \\ &\le A^0 + \max\left\{\sup_{n\in\mathbb{N}_{n_o-1}} \alpha_n^2\sigma^2, 1\right\}\Lambda + \sup_{n\in\mathbb{N}} \alpha_n\sigma\Lambda \triangleq D < \infty, \end{aligned} \tag{333}$$

and, of course, $A^0 \le D$. We use simple induction. Suppose that, for some $n \in \mathbb{N}$, $A^n \le D$. Then, there are two possibilities for $\sigma\alpha_n \ge 0$. Either $\sigma\alpha_n > 1$, which is of course only possible if $n \in \mathbb{N}_{n_o-1}$, in which case we have

$$\begin{aligned} A^{n+1} &\le (1-\sigma\alpha_n) A^n + \alpha_n^2\sigma^2\Lambda + \alpha_n\sigma\Lambda \\ &\le \alpha_n^2\sigma^2\Lambda + \alpha_n\sigma\Lambda \\ &\le \max\left\{\sup_{n\in\mathbb{N}_{n_o-1}} \alpha_n^2\sigma^2, 1\right\}\Lambda + \sup_{n\in\mathbb{N}} \alpha_n\sigma\Lambda \\ &\equiv D, \end{aligned} \tag{334}$$

or $\sigma\alpha_n \le 1$, which might happen for any $n \in \mathbb{N}$, yielding

$$\begin{aligned} A^{n+1} &\le (1-\sigma\alpha_n) A^n + \alpha_n^2\sigma^2\Lambda + \alpha_n\sigma\Lambda \\ &\le (1-\sigma\alpha_n) D + \alpha_n\sigma(\alpha_n\sigma\Lambda + \Lambda) \\ &\le (1-\sigma\alpha_n) D + \alpha_n\sigma\left(\sup_{n\in\mathbb{N}} \alpha_n\sigma\Lambda + \max\left\{\sup_{n\in\mathbb{N}_{n_o-1}} \alpha_n^2\sigma^2, 1\right\}\Lambda\right) \\ &\le D - \sigma\alpha_n D + \alpha_n\sigma D \\ &\equiv D. \end{aligned} \tag{335}$$

As a result, we have shown that $A^{n+1} \le D$, as well, implying that

$$\sup_{n\in\mathbb{N}} A^n \le D < \infty. \tag{336}$$

By definition of J^n, for $n \in \mathbb{N}^+$, we may write (utilizing condition **G2**)

$$\begin{aligned} \sup_{n\in\mathbb{N}^+} J^n &\equiv \sup_{n\in\mathbb{N}^+} A^n + \boldsymbol{\Delta}_B^{n-1} B^{n-1} + \boldsymbol{\Delta}_C^{n-1} C^{n-1} \\ &\le \sup_{n\in\mathbb{N}^+} A^n + \sup_{n\in\mathbb{N}^+} \boldsymbol{\Delta}_B^{n-1} \sup_{n\in\mathbb{N}^+} B^{n-1} + \sup_{n\in\mathbb{N}^+} \boldsymbol{\Delta}_C^{n-1} \sup_{n\in\mathbb{N}^+} C^{n-1} \\ &\equiv \sup_{n\in\mathbb{N}^+} A^n + \max\left\{\sup_{n\in\mathbb{N}^{n_o-2}} \boldsymbol{\Delta}_B^n, \boldsymbol{\Delta}_B^{n_o-1}\right\} \sup_{n\in\mathbb{N}^+} B^{n-1} + \max\left\{\sup_{n\in\mathbb{N}^{n_o-2}} \boldsymbol{\Delta}_C^n, \boldsymbol{\Delta}_C^{n_o-1}\right\} \sup_{n\in\mathbb{N}^+} C^{n-1} \\ &< \infty, \end{aligned} \tag{337}$$

and the proof is now complete. ■

7.11 Proof of Theorem 5

First, let us verify conditions **G1** and **G2** of Lemma 9. For **G1**, we perform, for every $n \in \mathbb{N}^2$, the equivalence test

$$\sigma\alpha_n \equiv \frac{1}{n} \le \frac{K-1}{K}\frac{1}{(n-1)^{\tau_2}} \equiv \frac{K-1}{K}\min\{\beta_{n-1}, \gamma_{n-1}\} \iff K \ge \frac{1}{1-\dfrac{(n-1)^{\tau_2}}{n}}, \tag{338}$$

which implies that any $K \ge 2$ works. Therefore, **G1** is satisfied for all $n \in \mathbb{N}^2$ by choosing, say, $K \equiv 2$. To verify **G2** for the sequence $\{\beta_n\}_{n\in\mathbb{N}}$, we would also like to show that

$$\alpha_{n+1}\beta_{n-1} \equiv \frac{1}{\sigma(n+1)}\frac{1}{(n-1)^{\tau_2}} \le \frac{1}{\sigma n}\frac{1}{n^{\tau_2}} \equiv \alpha_n\beta_n, \tag{339}$$

for all *sufficiently large* $n \in \mathbb{N}^2$. Indeed, it is a standard calculus exercise to show that

$$\frac{n^{\tau_2}}{(n-1)^{\tau_2}} \le \frac{n+1}{n}, \quad \forall n \in \left[\frac{1}{1-\tau_2^{1/(\tau_2+1)}}, \infty\right) \bigcap \mathbb{N} \subseteq \mathbb{N}^3. \tag{340}$$

To verify **G2** for the sequence $\{\gamma_n\}_{n\in\mathbb{N}}$, it suffices to observe that

$$\frac{n^{\tau_2}}{(n-1)^{\tau_2}} > \frac{n^{\tau_3}}{(n-1)^{\tau_3}}, \quad \forall n \in \mathbb{N}^2, \tag{341}$$

implying that

$$\frac{n^{\tau_3}}{(n-1)^{\tau_3}} \le \frac{n+1}{n} \iff \alpha_{n+1}\gamma_{n-1} \le \alpha_n\gamma_n, \; \forall n \in \left[\frac{1}{1-\tau_2^{1/(\tau_2+1)}}, \infty\right) \bigcap \mathbb{N}. \tag{342}$$

Therefore, we may apply Lemma 9 by choosing

$$n_o \equiv n_o(\tau_2) \equiv \left\lceil \frac{1}{1-\tau_2^{1/(\tau_2+1)}} \right\rceil, \tag{343}$$

in which case it must be true that, for every $n \in \mathbb{N}^{n_o}$,

$$\begin{aligned}
J^{n+1} &\le (1-\sigma\alpha_n)J^n + \widetilde{\Sigma}\left(\sigma^2\alpha_n^2 + \frac{\sigma^3\alpha_n\alpha_{n-1}^2}{\beta_{n-1}^2} + \sigma\alpha_n\beta_{n-1} + \frac{\sigma^3\alpha_n\alpha_{n-1}^2}{\gamma_{n-1}^2} + \frac{\sigma\alpha_n\beta_{n-1}^2}{\gamma_{n-1}^2} + \sigma\alpha_n\gamma_{n-1}\right) \\
&\le (1-\sigma\alpha_n)J^n + \widetilde{\Sigma}\left(\sigma^2\alpha_n^2 + \frac{\sigma^3\alpha_{n-1}^3}{\beta_{n-1}^2} + \sigma\alpha_{n-1}\beta_{n-1} + \frac{\sigma^3\alpha_{n-1}^3}{\gamma_{n-1}^2} + \frac{\sigma\alpha_{n-1}\beta_{n-1}^2}{\gamma_{n-1}^2} + \sigma\alpha_{n-1}\gamma_{n-1}\right) \\
&\le (1-\sigma\alpha_n)J^n + \widetilde{\Sigma}\left(\sigma^2\alpha_n^2 + 2\frac{\sigma^3\alpha_{n-1}^3}{\beta_{n-1}^2} + \frac{\sigma\alpha_{n-1}\beta_{n-1}^2}{\gamma_{n-1}^2} + 2\sigma\alpha_{n-1}\gamma_{n-1}\right) \\
&\equiv \left(1-\frac{1}{n}\right)J^n + \widetilde{\Sigma}\left(\frac{1}{n^2} + 2\frac{1}{(n-1)^{3-2\tau_2}} + \frac{1}{(n-1)^{1+2\tau_2-2\tau_3}} + 2\frac{1}{(n-1)^{1+\tau_3}}\right) \\
&\le \left(1-\frac{1}{n}\right)J^n + \widetilde{\Sigma}\left(\frac{1}{n^2} + 2\Lambda_2(\tau_2)\frac{1}{n^{3-2\tau_2}} + \Lambda_3(\tau_2)\frac{1}{n^{1+2\tau_2-2\tau_3}} + 2\Lambda_{23}(\tau_2)\frac{1}{n^{1+\tau_3}}\right),
\end{aligned} \tag{344}$$

where we have used the fact that, for our choice of the sequence $\{\alpha_n\}_{n\in\mathbb{N}}$, it is true that $\alpha_n \leq \alpha_{n-1}$, for all $n \in \mathbb{N}^+$, and the constants Λ_2,Λ_3 and Λ_{23} are defined as

$$\Lambda_2(\tau_2) \triangleq \left(\frac{n_o(\tau_2)}{n_o(\tau_2)-1}\right)^{3-2\tau_2} < 8, \tag{345}$$

$$\Lambda_3(\tau_2) \triangleq \left(\frac{n_o(\tau_2)}{n_o(\tau_2)-1}\right)^{1+2\tau_2-2\tau_3} < 8 \quad \text{and} \tag{346}$$

$$\Lambda_{23}(\tau_2) \triangleq \left(\frac{n_o(\tau_2)}{n_o(\tau_2)-1}\right)^{1+\tau_3} < 4 < 8. \tag{347}$$

Consequently, we obtain the bound

$$J^{n+1} \leq \left(1-\frac{1}{n}\right) J^n + 16\widetilde{\Sigma}\left(\frac{1}{n^2}+\frac{1}{n^{3-2\tau_2}}+\frac{1}{n^{1+2\tau_2-2\tau_3}}+\frac{1}{n^{1+\tau_3}}\right), \tag{348}$$

being true for all $n \in \mathbb{N}^{n_o}$, where n_o depends on τ_2 according to (343).

Let us now apply the Generalized Chung's Lemma (Lemma 13) to the recursion (348). For every $n \in \mathbb{N}^{n_o}$, we have

$$\begin{aligned}
J^{n+1} &\leq J^{n_o}\prod_{i\in\mathbb{N}^{n_o}_n}\left(1-\frac{1}{i}\right) + 16\widetilde{\Sigma}\sum_{i\in\mathbb{N}^{n_o}_n}\left(\frac{1}{i^2}+\frac{1}{i^{3-2\tau_2}}+\frac{1}{i^{1+2\tau_2-2\tau_3}}+\frac{1}{i^{1+\tau_3}}\right)\prod_{j\in\mathbb{N}^{i+1}_n}\left(1-\frac{1}{j}\right)\\
&\equiv J^{n_o}\prod_{i\in\mathbb{N}^{n_o}_n}\left(\frac{i-1}{i}\right) + 16\widetilde{\Sigma}\sum_{i\in\mathbb{N}^{n_o}_{n-1}}\left(\frac{1}{i^2}+\frac{1}{i^{3-2\tau_2}}+\frac{1}{i^{1+2\tau_2-2\tau_3}}+\frac{1}{i^{1+\tau_3}}\right)\prod_{j\in\mathbb{N}^{i+1}_n}\left(\frac{j-1}{j}\right)\\
&\qquad + 16\widetilde{\Sigma}\left(\frac{1}{n^2}+\frac{1}{n^{3-2\tau_2}}+\frac{1}{n^{1+2\tau_2-2\tau_3}}+\frac{1}{n^{1+\tau_3}}\right)\\
&\equiv J^{n_o}\frac{n_o-1}{n} + 16\widetilde{\Sigma}\sum_{i\in\mathbb{N}^{n_o}_{n-1}}\left(\frac{1}{i^2}+\frac{1}{i^{3-2\tau_2}}+\frac{1}{i^{1+2\tau_2-2\tau_3}}+\frac{1}{i^{1+\tau_3}}\right)\frac{i}{n}\\
&\qquad + 16\widetilde{\Sigma}\left(\frac{1}{n^2}+\frac{1}{n^{3-2\tau_2}}+\frac{1}{n^{1+2\tau_2-2\tau_3}}+\frac{1}{n^{1+\tau_3}}\right)\\
&\equiv J^{n_o}\frac{n_o-1}{n} + \frac{16\widetilde{\Sigma}}{n}\sum_{i\in\mathbb{N}^{n_o}_{n-1}}\left(\frac{1}{i}+\frac{1}{i^{2-2\tau_2}}+\frac{1}{i^{2\tau_2-2\tau_3}}+\frac{1}{i^{\tau_3}}\right)\\
&\qquad + 16\widetilde{\Sigma}\left(\frac{1}{n^2}+\frac{1}{n^{3-2\tau_2}}+\frac{1}{n^{1+2\tau_2-2\tau_3}}+\frac{1}{n^{1+\tau_3}}\right).
\end{aligned} \tag{349}$$

Due to our assumption that $1/2 \leq \tau_3 < \tau_2 < 1$, it holds that $2-2\tau_2 \neq 1$ and $2\tau_2-2\tau_3 \neq 1$. Consequently, it is true that

$$\sum_{i\in\mathbb{N}^{n_o}_{n-1}}\frac{1}{i^{2-2\tau_2}} < \frac{1}{n_o^{2-2\tau_2}} + \frac{n^{1-(2-2\tau_2)}}{1-(2-2\tau_2)}, \tag{350}$$

$$\sum_{i\in\mathbb{N}^{n_o}_{n-1}}\frac{1}{i^{2\tau_2-2\tau_3}} < \frac{1}{n_o^{2\tau_2-2\tau_3}} + \frac{n^{1-(2\tau_2-2\tau_3)}}{1-(2\tau_2-2\tau_3)} \quad \text{and} \tag{351}$$

$$\sum_{i\in\mathbb{N}_{n-1}^{n_o}} \frac{1}{i^{\tau_3}} < \frac{1}{n_o^{\tau_3}} + \frac{n^{1-\tau_3}}{1-\tau_3}, \tag{352}$$

whereas

$$\sum_{i\in\mathbb{N}_{n-1}^{n_o}} \frac{1}{i} < \frac{1}{n_o} + \log(n). \tag{353}$$

By defining the quantity

$$\mathsf{R} \equiv \mathsf{R}(\tau_2,\tau_3) \triangleq \frac{1}{1-\max\{2-2\tau_2, 2\tau_2-2\tau_3, \tau_3\}} > 1, \tag{354}$$

we may further bound (349) from above as

$$\begin{aligned}
J^{n+1} &\le J^{n_o}\frac{n_o-1}{n} + 16\widetilde{\Sigma}\mathsf{R}\left(\frac{\log(n)}{n} + \frac{1}{n^{2-2\tau_2}} + \frac{1}{n^{2\tau_2-2\tau_3}} + \frac{1}{n^{\tau_3}}\right) \\
&\quad + \frac{64\widetilde{\Sigma}}{n} + 16\widetilde{\Sigma}\left(\frac{1}{n^2} + \frac{1}{n^{3-2\tau_2}} + \frac{1}{n^{1+2\tau_2-2\tau_3}} + \frac{1}{n^{1+\tau_3}}\right) \\
&\le \frac{J^{n_o}n_o + 64\widetilde{\Sigma}}{n} + 32\widetilde{\Sigma}\mathsf{R}\left(\frac{\log(n)}{n} + \frac{1}{n^{2-2\tau_2}} + \frac{1}{n^{2\tau_2-2\tau_3}} + \frac{1}{n^{\tau_3}}\right) \\
&\le \frac{n_o\left(J^{n_o} + 64\widetilde{\Sigma}\right)}{n} + 32\widetilde{\Sigma}\mathsf{R}\left(\frac{1}{n^{1/2}} + \frac{1}{n^{2-2\tau_2}} + \frac{1}{n^{2\tau_2-2\tau_3}} + \frac{1}{n^{\tau_3}}\right) \\
&\le \frac{n_o\left(J^{n_o} + 64\widetilde{\Sigma}\right)}{n} + 128\widetilde{\Sigma}\mathsf{R}\frac{1}{n^{\min\{1/2, 2-2\tau_2, 2\tau_2-2\tau_3, \tau_3\}}} \\
&\equiv \frac{n_o\left(J^{n_o} + 64\widetilde{\Sigma}\right)}{n} + 128\widetilde{\Sigma}\mathsf{R}\frac{1}{n^{\min\{2-2\tau_2, 2\tau_2-2\tau_3\}}},
\end{aligned} \tag{355}$$

For our stepsize choices, it is trivial to see that the remaining condition **G3** of Lemma 9 is also satisfied. Therefore, it must be true that

$$J^{n+1} \le \frac{n_o\left(\sup_{n\in\mathbb{N}^+} J^n + 64\widetilde{\Sigma}\right)}{n} + \frac{128\widetilde{\Sigma}\mathsf{R}}{n^{\min\{2-2\tau_2, 2\tau_2-2\tau_3\}}}, \quad \forall n \in \mathbb{N}^{n_o}, \tag{356}$$

where $\sup_{n\in\mathbb{N}^+} J^n < \infty$, and defining another constant $\widehat{\Sigma} \triangleq \max\left\{\left(\sup_{n\in\mathbb{N}^+} J^n + 64\widetilde{\Sigma}\right), 128\widetilde{\Sigma}\right\}$, we end up with the inequality

$$\mathbb{E}\left\{\left\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\right\|_2^2\right\} \le J^{n+1} \le \frac{\widehat{\Sigma}n_o}{n} + \frac{\widehat{\Sigma}\mathsf{R}}{n^{\min\{2-2\tau_2, 2\tau_2-2\tau_3\}}}, \quad \forall n \in \mathbb{N}^{n_o}, \tag{357}$$

completing the proof for first part Theorem 5.

To prove the second part, let

$$\tau_2 \equiv \frac{3+\epsilon}{4} \quad \text{and} \quad \tau_3 \equiv \frac{1+\delta\epsilon}{2}, \tag{358}$$

for some $\epsilon \in [0,1)$ and $\delta \in (0,1)$. Then, for the exponents of the corresponding terms in (357), we have the identities

$$2 - 2\tau_2 \equiv \frac{1-\epsilon}{2} \quad \text{and} \tag{359}$$

$$2\tau_2 - 2\tau_3 \equiv \frac{1-\epsilon\,(2\delta-1)}{2}, \tag{360}$$

out of which the first is the smallest. Additionally, it also true that

$$\begin{aligned}
\mathsf{R}\,(\tau_2,\tau_3) \equiv \mathsf{R}\,(\delta,\epsilon) &\equiv \frac{1}{1-\max\left\{\dfrac{1-\epsilon}{2}, \dfrac{1-\epsilon\,(2\delta-1)}{2}, \dfrac{1+\delta\epsilon}{2}\right\}} \\
&\equiv \frac{1}{1-\dfrac{1}{2}\max\left\{1-\epsilon, 1-\epsilon\,(2\delta-1), 1+\delta\epsilon\right\}} \\
&\equiv \frac{2}{1-\epsilon\max\left\{1-2\delta,\delta\right\}} \\
&< \frac{2}{1-\epsilon}.
\end{aligned} \tag{361}$$

As a result, we may further bound (357) as

$$\begin{aligned}
\mathbb{E}\left\{\left\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^*\right\|_2^2\right\} \le J^{n+1} &\le \frac{\widehat{\Sigma} n_o\,(\epsilon)}{n} + \frac{\widehat{\Sigma}\mathsf{R}\,(\delta,\epsilon)}{n^{\min\{(1-\epsilon)/2,(1-\epsilon(2\delta-1))/2\}}} \\
&\equiv \frac{\widehat{\Sigma} n_o\,(\epsilon)}{n} + \frac{\widehat{\Sigma}\mathsf{R}\,(\delta,\epsilon)}{n^{(1-\epsilon)/2}} \\
&\le \frac{\widehat{\Sigma} n_o\,(\epsilon)}{n^{(1-4\epsilon)/2}} + \frac{\widehat{\Sigma}\dfrac{2}{1-\epsilon}}{n^{(1-\epsilon)/2}} \\
&\equiv \frac{\widehat{\Sigma}\left(n_o\,(\epsilon)+\dfrac{2}{1-\epsilon}\right)}{n^{(1-\epsilon)/2}},
\end{aligned} \tag{362}$$

for every $n \in \mathbb{N}^{n_o(\epsilon)}$. ■

☺

References

Shabbir Ahmed, Ulaş Çakmak, and Alexander Shapiro. Coherent risk measures in inventory problems. *European Journal of Operational Research*, 182(1):226–238, October 2007. ISSN 03772217. doi: 10.1016/j.ejor.2006.07.016.

Fredrik Andersson, Helmut Mausser, Dan Rosen, and Stanislav Uryasev. Credit risk optimization with conditional value-at-risk criterion. *Mathematical Programming*, 89(2):273–291, January 2001. ISSN 0025-5610. doi: 10.1007/PL00011399.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, July 1999. ISSN 0960-1627. doi: 10.1111/1467-9965.00068.

Tsvetan Asamov and Andrzej Ruszczyński. Time-consistent approximations of risk-averse multistage stochastic optimization problems. *Mathematical Programming*, 153(2):459–493, November 2015. ISSN 14364646. doi: 10.1007/s10107-014-0813-x.

Robert B Ash and Catherine Doléans-Dade. *Probability and Measure Theory*. Academic Press, 2000. ISBN 0120652021. doi: 10.2307/2291440.

V. I. Bogachev. *Measure theory, Vol. 1*. Springer, 2007. ISBN 9783540345145.

Sergio Bruno, Shabbir Ahmed, Alexander Shapiro, and Alexandre Street. Risk neutral and risk-averse approaches to multistage renewable investment planning under uncertainty. *European Journal of Operational Research*, 250(3):979–989, May 2016. ISSN 03772217. doi: 10.1016/j.ejor.2015.10.013.

Özlem Çavuş and Andrzej Ruszczyński. Risk-averse control of undiscounted transient markov models. *SIAM Journal on Control and Optimization*, 52(6):3935–3966, January 2014a. ISSN 0363-0129. doi: 10.1137/13093902X.

Özlem Çavuş and Andrzej Ruszczyński. Computational methods for risk-averse undiscounted transient markov models. *Operations Research*, 62(2):401–417, April 2014b. ISSN 0030-364X. doi: 10.1287/opre.2013.1251.

Xin Chen, Melvyn Sim, David Simchi-Levi, and Peng Sun. Risk aversion in inventory management. *Operations Research*, 55(5):828–842, 2007. ISSN 0030-364X. doi: 10.1287/opre.1070.0429.

Zhiping Chen and Yi Wang. Two-sided coherent risk measures and their application in realistic portfolio optimization. *Journal of Banking and Finance*, 32(12):2667–2673, 2008. ISSN 03784266. doi: 10.1016/j.jbankfin.2008.07.004.

Zhiping Chen and Li Yang. Nonlinearly weighted convex risk measure and its application. *Journal of Banking and Finance*, 35(7):1777–1793, July 2011. ISSN 03784266. doi: 10.1016/j.jbankfin.2010.12.004.

K. L. Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25 (3):463–483, September 1954. ISSN 0003-4851. doi: 10.1214/aoms/1177728716.

Ricardo A. Collado, Dávid Papp, and Andrzej Ruszczyński. Scenario decomposition of risk-averse multistage stochastic programming problems. *Annals of Operations Research*, 200(1):147–170, November 2012. ISSN 02545330. doi: 10.1007/s10479-011-0935-y.

Darinka Dentcheva and Andrzej Ruszczyński. Risk-averse control of continuous-time markov chains. In *2017 Proceedings of the Conference on Control and its Applications*, pages 78–85. Society for Industrial and Applied Mathematics, Philadelphia, PA, January 2017. doi: 10.1137/1.9781611975024.11.

Darinka Dentcheva, Spiridon Penev, and Andrzej Ruszczyński. Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, 69(4):737–760, August 2017. ISSN 15729052. doi: 10.1007/s10463-016-0559-8.

Rick Durrett. *Probability: Theory & Examples.* Cambridge university press, 2010.

Jingnan Fan and Andrzej Ruszczyński. Risk measurement and risk-averse control of partially observable discrete-time markov systems. *Mathematical Methods of Operations Research*, pages 1–24, February 2018. ISSN 1432-2994. doi: 10.1007/s00186-018-0633-5.

Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447, October 2002. ISSN 09492984. doi: 10.1007/s007800200072.

Peter I. Frazier, Warren B. Powell, and Savas Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, January 2008. ISSN 0363-0129. doi: 10.1137/070693424.

Tianwen Fu, Xinkai Zhuang, Yongchang Hui, and Jia Liu. Convex risk measures based on generalized lower deviation and their applications. *International Review of Financial Analysis*, 52:27–37, 2017. ISSN 10575219. doi: 10.1016/j.irfa.2017.04.008.

Xianghua Gan, Suresh P Sethi, and Houmin Yan. Coordination of supply chains with risk-averse agents. *Production and Operations Management*, 13(2):135–149, 2004. ISSN 0144-3577. doi: 10.1108/01443579310023963.

Bogdan Grechuk and Michael Zabarankin. Sensitivity analysis in applications with deviation, risk, regret, and error measures. *SIAM Journal on Optimization*, 27(4):2481–2507, January 2017. ISSN 1052-6234. doi: 10.1137/16M1105165.

Vincent Guigues, Volker Krätschmer, and Alexander Shapiro. Statistical inference and hypotheses testing of risk averse stochastic programs. *Arxiv*, pages 1–27, 2016.

Wenjie Huang and William B. Haskell. Risk-aware q-learning for markov decision processes. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 4928–4933. IEEE, December 2017. ISBN 978-1-5090-2873-3. doi: 10.1109/CDC.2017.8264388.

Daniel R. Jiang and Warren B. Powell. Practicality of nested risk measures for dynamic electric vehicle charging. *Arxiv*, May 2016.

Daniel R. Jiang and Warren B. Powell. Risk-averse approximate dynamic programming with quantile-based risk measures. *Mathematics of Operations Research*, page moor.2017.0872, November 2017. ISSN 0364-765X. doi: 10.1287/moor.2017.0872.

Masaaki Kijima and Masamitsu Ohnishi. Mean-risk analysis of risk aversion and wealth effects on optimal portfolios with multiple investment opportunities. *Annals of Operations Research*, 45(1): 147–163, December 1993. ISSN 02545330. doi: 10.1007/BF02282046.

Pavlo Krokhmal, Jonas Palmquist, and Stanislav Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of Risk*, 4(2):1–36, March 2001. ISSN 1465-1211. doi: 10.1.1.4.1197.

Harold J. (Harold Joseph) Kushner and George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003. ISBN 9780387008943.

Wann-Jiun Ma, Chanwook Oh, Yang Liu, Darinka Dentcheva, and Michael M. Zavlanos. Risk-averse access point selection in wireless communication networks. *IEEE Transactions on Control of Network Systems*, 5870(c):1–1, 2018. ISSN 2325-5870. doi: 10.1109/TCNS.2018.2792309.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, March 1952. ISSN 15406261. doi: 10.1111/j.1540-6261.1952.tb01525.x.

Somayeh Moazeni, Warren B. Powell, and Amir H. Hajimiragha. Mean-conditional value-at-risk optimal energy storage operation in the presence of transaction costs. *IEEE Transactions on Power Systems*, 30(3):1222–1232, May 2015. ISSN 08858950. doi: 10.1109/TPWRS.2014.2341642.

Somayeh Moazeni, Warren B. Powell, Boris Defourny, and Belgacem Bouzaiene-Ayari. Parallel nonstationary direct policy search for risk-averse stochastic optimization. *INFORMS Journal on Computing*, 29(2):332–349, May 2017. ISSN 15265528. doi: 10.1287/ijoc.2016.0733.

Włodzimierz Ogryczak and Andrzej Ruszczyński. From stochastic dominance to mean-risk models: Semideviations as risk measures. *European Journal of Operational Research*, 116(1):33–50, July 1999. ISSN 03772217. doi: 10.1016/S0377-2217(98)00167-2.

Włodzimierz Ogryczak and Andrzej Ruszczyński. Dual stochastic dominance and related mean-risk models. *SIAM Journal on Optimization*, 13(1):60–78, 2002. ISSN 1052-6234. doi: 10.1137/S1052623400375075.

Warren B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley, 2011. ISBN 9780470604458.

H. Robbins and D. Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier, 1971. ISBN 9780126045505. doi: 10.1016/B978-0-12-604550-5.50015-8.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729586.

R. T. Rockafellar and R. J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2004.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–41, 1997. ISSN 10769986. doi: 10.2307/1165345.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26(7):1443–1471, July 2002. ISSN 03784266. doi: 10.1016/S0378-4266(02)00271-6.

R. Tyrrell Rockafellar, Stan Uryasev, and Michael Zabarankin. Generalized deviations in risk analysis. *Finance and Stochastics*, 10(1):51–74, January 2006. ISSN 09492984. doi: 10.1007/s00780-005-0165-8.

Tyrrell R. Rockafellar, Stanislav P. Uryasev, and Michael Zabarankin. Deviation measures in risk analysis and optimization. *SSRN Electronic Journal*, December 2003. ISSN 1556-5068. doi: 10.2139/ssrn.365640.

Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical Programming*, 125(2):235–261, October 2010. ISSN 00255610. doi: 10.1007/s10107-010-0393-3.

Andrzej Ruszczyński and Alexander Shapiro. Optimization of convex risk functions. *Mathematics of Operations Research*, 31(3):433–452, 2006a. ISSN 0364-765X. doi: 10.1287/moor.1050.0186.

Andrzej Ruszczyński and Alexander Shapiro. Conditional risk mappings. *Mathematics of Operations Research*, 31(3):544–561, 2006b. ISSN 0364-765X. doi: 10.1287/moor.1060.0204.

Ilya O Ryzhov. On the convergence rates of expected improvement methods. *Operations Research*, 64(6):1515–1528, 2016. ISSN 0030-364X. doi: 10.1287/opre.2016.1494.

Tadeusz Sawik. On the risk-averse optimization of service level in a supply chain under disruption risks. In *International Journal of Production Research*, volume 54, pages 98–113, 2016. ISBN 0020-7543. doi: 10.1080/00207543.2015.1016192.

Danjue Shang, Victor Kuzmenko, and Stan Uryasev. Cash flow matching with risks controlled by buffered probability of exceedance and conditional value-at-risk. *Annals of Operations Research*, 260(1-2):501–514, January 2018. ISSN 15729338. doi: 10.1007/s10479-016-2354-6.

Alexander Shapiro. Consistency of sample estimates of risk-averse stochastic programs. *Journal of Applied Probability*, 50(2):533–541, June 2013. ISSN 00219002. doi: 10.1239/jap/1371648959.

Alexander Shapiro. Interchangeability principle and dynamic equations in risk averse stochastic programming. *Operations Research Letters*, 45(4):377–381, July 2017. ISSN 01676377. doi: 10.1016/j.orl.2017.05.008.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics, 2nd edition, 2014. ISBN 089871687X. doi: http://dx.doi.org/10.1137/1.9780898718751.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Sequential decision making with coherent risk. *IEEE Transactions on Automatic Control*, 62(7):3323–3338, July 2017. ISSN 00189286. doi: 10.1109/TAC.2016.2644871.

Mengdi Wang and Dimitri P. Bertsekas. Stochastic first-order methods with random constraint projection. *SIAM Journal on Optimization*, 26(1):681–717, 2016. ISSN 1052-6234, 1095-7189. doi: 10.1137/130931278.

Mengdi Wang, Ethan X. Fang, and Han Liu. Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2): 419–449, January 2017. ISSN 14364646. doi: 10.1007/s10107-016-1017-3.

Xu Xinsheng, Meng Zhiqing, Shen Rui, Jiang Min, and Ji Ping. Optimal decisions for the loss-averse newsvendor problem under cvar. *International Journal of Production Economics*, 164:146–159, June 2015. ISSN 09255273. doi: 10.1016/j.ijpe.2015.03.019.

Shuoguang Yang, Mengdi Wang, and Ethan X. Fang. Multi-level stochastic gradient methods for nested composition optimization. *Arxiv*, January 2018.

Pengqian Yu, William B. Haskell, and Huan Xu. Approximate value iteration for risk-aware markov decision processes. *IEEE Transactions on Automatic Control*, pages 1–1, 2018. ISSN 00189286. doi: 10.1109/TAC.2018.2790261.

Dali Zhang, Huifu Xu, and Yue Wu. Single and multi-period optimal inventory control models with risk-averse constraints. *European Journal of Operational Research*, 199(2):420–434, December 2009. ISSN 03772217. doi: 10.1016/j.ejor.2008.11.047.

§