ZEROTH-ORDER STOCHASTIC COMPOSITIONAL ALGORITHMS FOR RISK-AWARE LEARNING[*]

DIONYSIOS S. KALOGERIAS[†] AND WARREN B. POWELL[‡]

Abstract. We present *Free-MESSAGE*^{p}, the first zeroth-order algorithm for (weakly-)convex mean-semideviation-based risk-aware learning, which is also the first three-level zeroth-order compositional stochastic optimization algorithm whatsoever. Using a non-trivial extension of Nesterov's classical results on Gaussian smoothing, we develop the *Free-MESSAGE*^{p} algorithm from first principles, and show that it essentially solves a smoothed surrogate to the original problem, the former being a uniform approximation of the latter, in a useful, convenient sense. We then present a complete analysis of the *Free-MESSAGE*^{p} algorithm, which establishes convergence in a user-tunable neighborhood of the optimal solutions of the original problem for convex costs, as well as explicit convergence rates for convex, weakly convex, and strongly convex costs, and in a unified way. Orderwise, and for fixed problem parameters, our results demonstrate no sacrifice in convergence speed as compared to existing first-order methods, while striking a certain balance among the condition of the problem, its dimensionality, as well as the accuracy of the obtained results, naturally extending previous results in zeroth-order risk-neutral learning.

Key words. Risk-Averse Optimization, Risk-Aware Learning, Zeroth-order Methods, Risk Measures, Mean-Upper-Semideviation, Stochastic Gradient Methods, Compositional Optimization.

AMS subject classifications. 90-08, 90C25, 90C15, 90C59, 90C99

1. Introduction.

Statistical machine learning traditionally deals with the determination and characterization of optimal decision rules minimizing an expected cost criterion, quantifying, for instance, regression or misclassification error in relevant applications, on the basis of available training data [17,20,43]. Still, the expected cost paradigm is not appropriate, say, in applications involving *highly dispersive disturbances*, such as heavy tailed, skewed or multimodal noise, or in applications whose purpose is to *imitate uncertain human behavior*. In the first case, merely optimizing the expected cost is often statistically meaningless, since the resulting optimal prediction errors might exhibit unstable or erratic behavior, even with a small expected value. In the second case, as aptly put in [7], the fact is that human decision makers are inherently risk-averse, because they prefer consistent sequences of predictions instead of highly variable ones, even if the latter contain slightly better predictions.

Such situations motivate developments in the area of *risk-aware statistical learning*, in which expectation in the learning objective is replaced by more general functionals, called *risk measures* [39], whose purpose is to effectively quantify the statistical variability of the cost function considered, in addition to mean performance. Indeed, risk-awareness in learning and optimization has already been explored under various problem settings [1, 5, 7, 18, 21–23, 26, 31, 37, 41, 44, 47, 49], and has proved useful in many important applications, as well [5, 6, 24, 27, 34, 38].

In this paper, we study risk-aware learning problems in which expectation is generalized to the class of *mean-semideviation risk measures* developed in [23]. Specifically, given any complete probability space $(\Omega, \mathscr{F}, \mathcal{P})$, and a random element $\boldsymbol{W} : \Omega \rightarrow \mathbb{R}^M$ on $(\Omega, \mathscr{F})$ modeling abstractly all the uncertainty involved in the learning task, we

---

[*]Submitted to the editors DATE.

Funding: This material is based upon work supported by the U.S. Navy / SPAWAR Systems Center Pacific under Contract No. N66001-18-C-4031.

[†]Department of EE, Yale University, CT, USA (dionysis.kalogerias@yale.edu).

[‡]Department of ORFE, Princeton University, NJ, USA (powell@princeton.edu).

consider stochastic programs of the form

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \left\{\phi\left(\boldsymbol{x}\right) \triangleq \mathbb{E}\left\{F\left(\boldsymbol{x},\boldsymbol{W}\right)\right\} + c\left\|\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right) - \mathbb{E}\left\{F\left(\boldsymbol{x},\boldsymbol{W}\right)\right\}\right)\right\|_{\mathcal{L}_p}\right\}, \tag{1.1}$$

for $c \in [0,1]$ and *order* $p \in [1,2]$, and where $F : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$ is Borel in its second argument and either weakly convex, convex, or strongly convex in its first, $F\left(\cdot,\boldsymbol{W}\right) \in \mathcal{L}_p\left(\Omega,\mathscr{F},\mathcal{P};\mathbb{R}\right) \triangleq \mathcal{Z}_p$, $\|\cdot\|_{\mathcal{L}_p} : \mathcal{Z}_p \to \mathbb{R}_+$ is the corresponding standard norm on $\mathcal{Z}_p$, the set $\mathcal{X} \subseteq \mathbb{R}^N$ is nonempty, closed and convex, and $\mathcal{R} : \mathbb{R} \to \mathbb{R}$ is a *risk regularizer*, or *risk profile* [23], that is, any *convex*, *nonnegative*, *nondecreasing* and *nonexpansive* function. Hereafter, (1.1) will be called the *base problem.*

The objective ϕ evaluates the mean-semideviation risk measure $\rho\left(\cdot\right) \triangleq \mathbb{E}\left\{\cdot\right\} + c\|\mathcal{R}\left(\left(\cdot\right) - \mathbb{E}\left\{\cdot\right\}\right)\|_{\mathcal{L}_p}$ at $F\left(\cdot,\boldsymbol{W}\right)$, i.e., $\phi\left(\cdot\right) \equiv \rho\left(F\left(\cdot,\boldsymbol{W}\right)\right)$ [23]. The functional ρ generalizes the well-known *mean-upper-semideviation* [39], which is recovered by choosing $\mathcal{R}\left(\cdot\right) \equiv \left(\cdot\right)_+ \triangleq \max\left\{\cdot,0\right\}$, and is one of the most popular risk-measures in theory and practice [2,9,14,25,32,33,35,36]. For $c \in [0,1]$, ρ is a *convex risk measure* [23], ([39], Section 6) on $\mathcal{Z}_p$; thus, whenever F is convex, ϕ in (1.1) is convex on $\mathbb{R}^N$, as well.

In (1.1), the expected cost, called the *risk-neutral part* of the objective, is penalized by a *semideviation term*, called the *risk-averse part* of the objective. The latter explicitly quantifies, for each feasible decision, the deviation of the cost relative to its expectation, interpreted as a standardized statistical benchmark. The risk profile $\mathcal{R}$ acts on this central deviation as a weighting function, and its purpose is to reflect the particular risk preferences of the learner. As partially mentioned above, typical choices for $\mathcal{R}$ include the *hockey stick* $\left(\cdot\right)_+ + \eta$, also known as a *Rectified Linear Unit (ReLU)*, as well as its smooth approximations $(1/t)\log\left(1+\exp\left(t\left(\cdot\right)\right)\right)+\eta$, with $t > 0$, and $\eta \geq 0$. For a constructive characterization of mean-semideviation risk-measures, the reader is referred to [23].

Stochastic subgradient-based recursive optimization of mean-semideviation risk measures was recently considered in [23], where the so-called *MESSAGE^p algorithm* was proposed and analyzed for solving (1.1). The work of [23] is based on the fact that (1.1) can be expressed in nested form (see Section 2), and builds on previous results on general compositional stochastic optimization [45,46].

In this work, we are interested in solving (1.1) in a *zeroth-order setting*, using exclusively cost function evaluations, in absence of gradient information. Zeroth-order methods have a long history in both deterministic and risk-neutral stochastic optimization [4,12,15,16,19,29,40,48], and are of particular interest in applications where gradient information is very difficult, or even impossible to obtain, such as training of deep neural networks [8,42], nonsmooth optimization [30], clinical trials [7], and, more generally, machine learning *in the field*, simulation-based optimization [10,40], online auctions and search engines [12], and distributed learning [48]. Still, to the best of our knowledge, the development of zeroth-order methods for possibly nonsmooth risk-aware problems such as (1.1) and, *more generally*, compositional stochastic optimization problems, is completely unexplored. Our contributions are as follows:

- We present *Free-MESSAGE^p*, the first zeroth-order algorithm for solving (1.1) within a user-specified accuracy, which is also the first three-level zeroth-order compositional stochastic optimization algorithm, whatsoever. The *Free-MESSAGE^p* algorithm requires exactly *four* cost function evaluations per iteration, and is based on finite difference-based inexact quasigradients, in the spirit of [15,16,30]. By using a non-trivial extension of Nesterov's classical results on Gaussian smoothing [30], which we present and discuss (Section 3), we develop the *Free-MESSAGE^p* algorithm from first principles (Section 4), and we show that it essentially solves a *smoothed surrogate* to the original problem, the former provably being a uniform approximation to the latter (Lemma 5.2).

- We present a complete analysis of the *Free-MESSAGE*^{p} algorithm, establishing path convergence in a user-specified neighborhood of the optimal solutions of (1.1) for convex costs (Theorem 6.8), as well as explicit convergence rates for convex, weakly convex and strongly convex costs (Theorems 6.9, 6.10 and 6.11/6.12, respectively). *Orderwise*, and for *fixed* problem parameters, our results demonstrate no sacrifice in convergence speed as compared to the fully gradient-based *MESSAGE*^{p} algorithm [23], and explicitly quantify the effects of strong convexity on problem conditioning, reflected on the derived rates. Also, our results exhibit certain trade-offs between the size of the limiting neighborhood and the decision dimension N, and naturally extend core prior work on zeroth-order risk-neutral optimization [30].

Lastly, our results are supported by indicative numerical simulations (Section 7). As compared with prior works that assume access to stochastic gradients [23, 45, 46], passing to the zeroth-order setting is challenging for several reasons, on top of the corresponding convergence analysis (Section 6). First, the *key fact* that *Free-MESSAGE*^{p} can be designed in a way that it itself constitutes a stochastic gradient method tackling *directly* a well-defined and clearly identifiable smoothed surrogate to the original risk-aware problem is non-trivial (Section 5); this is because the objective ϕ in (1.1) does *not* admit an expectation representation, as otherwise standard in stochastic optimization. Of course, such a surrogate does not emerge in a gradient-based setting [23], at least as an essential entity.

At the same time, the *connection* between the smoothed surrogate and the original risk-aware problem is also not trivial: In fact, the analysis leading to our relevant uniform approximation bounds is substantially different from and more complex than that under the risk-neutral (expectation-based) setting [30], in regard to both the structure of our proofs (Lemma 5.2, Proposition 5.3), and the novel technical conditions imposed on the problem (Section 3, and Assumption 5.1). Those approximation bounds then make it possible to analyze convergence of *Free-MESSAGE*^{p} as a method for solving the smoothed surrogate, and subsequently relate the obtained results to the base problem (Section 6), in a transparent way. The corresponding analysis takes place under additional technical conditions (Assumption 6.1, which may be thought of as an evolution of Assumption 5.1, in turn following the discussion in Section 3), which are also new and different from those in [23, 45, 46].

Potentially Nonstandard Notation: We use **bold** letters to denote multidimensional quantities, such as vectors and matrices. Additionally, the symbol "$\triangleq$" denotes *equality by definition*, the symbol "$\equiv$" denotes *immediate equality/equivalence*, whereas the standard symbol "$=$" denotes possibly *not immediate equality/equivalence*. For a general vector/matrix-valued function $\boldsymbol{f} \in \mathcal{F}$, the *graph of* $\boldsymbol{f}$ *on a set* $\mathcal{G}$ is defined as the set $\mathrm{Graph}_{\mathcal{G}}(\boldsymbol{f}) \triangleq \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{G} \times \mathcal{F} \,|\, \boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})\}$. Lastly, within a given Cartesian product space, *tuples* are referred to as $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \ldots)$ or, in vector format, $[\boldsymbol{x}|\boldsymbol{y}|\boldsymbol{z}|\ldots]$.

2. Basic Properties of the Base Problem.

First, it will be convenient to express ϕ in *compositional (or nested) form*, as in [23]. By defining *expectation functions* $\varrho : \mathbb{R}_+ \to \mathbb{R}$, $g : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}_+$, $\boldsymbol{h} : \mathbb{R}^N \to \mathbb{R}^N \times \mathbb{R}$ and $s : \mathbb{R}^N \to \mathbb{R}$ as

$$\varrho(x) \triangleq x^{1/p},\ g(\boldsymbol{x}, y) \triangleq \mathbb{E}\{(\mathcal{R}(F(\boldsymbol{x}, \boldsymbol{W}) - y))^p\},\ \boldsymbol{h}(\boldsymbol{x}) \triangleq [\boldsymbol{x} \,|\, s(\boldsymbol{x}) \triangleq \mathbb{E}\{F(\boldsymbol{x}, \boldsymbol{W})\}],$$

respectively, and provided that the involved quantities are well-defined, ϕ may be reexpressed as

$$\phi(\boldsymbol{x}) \equiv s(\boldsymbol{x}) + c\varrho(g(\boldsymbol{h}(\boldsymbol{x}))), \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

Further, under appropriate conditions, differentiability of ϕ may be ensured as follows.

LEMMA 2.1 (**Differentiability of ϕ [23]**). *Let s and g be differentiable on $\mathcal{X}$ and $\mathrm{Graph}_{\mathcal{X}}(s)$, respectively, and let $\mathcal{R}:\mathbb{R}\to\mathbb{R}$ be such that $\{x\in\mathbb{R}\,|\mathcal{R}(x)\equiv 0\}\neq\mathbb{R}$. Also, if $p\in(1,2]$, and with $\kappa_{\mathcal{R}}\triangleq\sup\{x\in\mathbb{R}\,|\mathcal{R}(x)\equiv 0\}\in[-\infty,\infty)$, suppose that $\mathcal{P}(F(\boldsymbol{x},\boldsymbol{W})-s(\boldsymbol{x})\leq\kappa_{\mathcal{R}})<1$, for all $\boldsymbol{x}\in\mathcal{X}$. Then ϕ is differentiable on $\mathcal{X}$, and its gradient $\nabla\phi:\mathbb{R}^N\to\mathbb{R}^N$ may be expressed as*

$$\nabla\phi(\boldsymbol{x})\equiv\nabla s(\boldsymbol{x})+c\nabla\boldsymbol{h}(\boldsymbol{x})\nabla g(\boldsymbol{h}(\boldsymbol{x}))\nabla\varrho(g(\boldsymbol{h}(\boldsymbol{x}))),\quad\forall\boldsymbol{x}\in\mathcal{X}. \tag{2.1}$$

Lemma 2.1 states *carefully* the obvious: It verifies the composition rule for deriving the gradient of ϕ, properly handling the root ϱ. Also note that Lemma 2.1 is *not* concerned with actually determining $\nabla\boldsymbol{h}$ and ∇g; it just establishes the existence and intrinsically compositional structure of $\nabla\phi$.

3. Gaussian Smoothing and Its Properties. Let $f:\mathbb{R}^N\to\mathbb{R}$ be Borel. Also, for any $\mathbb{R}^N$-valued random element $\boldsymbol{U}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_N)$, and for $\mu\geq 0$, consider another Borel function $f_\mu:\mathbb{R}^N\to\mathbb{R}$, defined as $f_\mu(\cdot)\triangleq\mathbb{E}\{f((\cdot)+\mu\boldsymbol{U})\}$, provided that the involved integral is well-defined and finite for all $\boldsymbol{x}\in\mathbb{R}^N$. In many cases, the smoothed function f_μ may be shown to be differentiable on $\mathbb{R}^N$, even if f is not. A wide class of functions satisfying such a property is that of *Shift-Lipschitz functions*, or *SLipschitz functions*, for short, which are associated with two additional types of functions, which we call *divergences* and *normal remainders*, as introduced below.

DEFINITION 3.1 (**Divergences**). *A function $\mathsf{D}:\mathbb{R}^N\to\mathbb{R}$ is called a stationary divergence, or simply a divergence, if and only if $\mathsf{D}(\boldsymbol{u})\geq 0$, for all $\boldsymbol{u}\in\mathbb{R}^N$, and $\mathsf{D}(\boldsymbol{u})\equiv 0\iff\boldsymbol{u}\equiv\boldsymbol{0}$.*

DEFINITION 3.2 (**Normal Remainders**). *A function $\mathsf{T}:\mathbb{R}^{N_o}\times\mathbb{R}^N\to\mathbb{R}$ is called a normal remainder on $\mathcal{F}\subseteq\mathbb{R}^{N_o}$ if and only if, for $\boldsymbol{U}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_N)$, $\mathbb{E}\{\mathsf{T}(\boldsymbol{x},\mu\boldsymbol{U})\}\equiv 0$, for all $\boldsymbol{x}\in\mathcal{F}$ and $\mu\geq 0$.*

DEFINITION 3.3 (**Shift-Lipschitz Class**). *A function $f:\mathbb{R}^N\to\mathbb{R}$ is called Shift-Lipschitz with parameter $L<\infty$, relative to a divergence $\mathsf{D}:\mathbb{R}^N\to\mathbb{R}$ and a normal remainder $\mathsf{T}:\mathbb{R}^N\times\mathbb{R}^N\to\mathbb{R}$, or $(L,\mathsf{D},\mathsf{T})$-SLipschitz for short, on a subset $\mathcal{F}\subseteq\mathbb{R}^N$, if and only if, for every $\boldsymbol{u}\in\mathbb{R}^N$,*

$$\sup_{\boldsymbol{x}\in\mathcal{F}}|f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-\mathsf{T}(\boldsymbol{x},\boldsymbol{u})|\leq L\mathsf{D}(\boldsymbol{u}).$$

Apparently, every (real-valued) L-Lipschitz function on $\mathbb{R}^N$, with respect to some norm $\|\cdot\|_*:\mathbb{R}^N\to\mathbb{R}_+$, is $(L,\|\cdot\|_*,0)$-SLipschitz on $\mathbb{R}^N$. Similarly, every L-smooth function f on $\mathbb{R}^N$ is $\big(L/2,\|\cdot\|_2^2,\langle\nabla f(\bullet),\cdot\rangle\big)$-SLipschitz on $\mathbb{R}^N$; just recall that if f has L-Lipschitz gradient then

$$\big|f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2)-\langle\nabla f(\boldsymbol{x}_2),\boldsymbol{x}_1-\boldsymbol{x}_2\rangle\big|\leq\frac{L}{2}\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2,\quad\forall(\boldsymbol{x}_1,\boldsymbol{x}_2)\in\mathbb{R}^N\times\mathbb{R}^N.$$

But there are many non-Lipschitz or non-smooth functions, which can be shown to be SLipschitz, at least on some proper subset $\mathcal{F}\subset\mathbb{R}^N$, but where still $\boldsymbol{u}\in\mathbb{R}^N$ (see Definition 3.3). This is the main reason for working with the SLipschitz class and its extensions, as it provides substantially increased degrees of freedom regarding the choice of the cost function in (1.1).

We now formulate the next central result, providing several useful properties of f_μ. Simpler versions of this result have been presented earlier in the seminal paper [30], however under more restrictive conditions on f.

LEMMA 3.4 (**Properties of f_μ**). *Let $\boldsymbol{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$ and suppose that f satisfies the elementary growth condition*

$$\big[\mathbb{E}\{|f(\mu_\star \boldsymbol{U})|\} < \infty \iff f(\mu_\star \boldsymbol{U}) \in \mathcal{Z}_1\big], \quad \text{for some } \mu_\star \in (0, \infty). \tag{3.1}$$

Then, for any subset $\mathcal{F} \subseteq \mathbb{R}^N$, the following statements are true:

- *For every $0 \le \mu < \mu_\star$, f_μ is well-defined and finite on $\mathcal{F}$. Further, if f is $(L, \mathsf{D}, \mathsf{T})$-SLipschitz on $\mathcal{F}$,*

$$\sup_{\boldsymbol{x} \in \mathcal{F}} |f_\mu(\boldsymbol{x}) - f(\boldsymbol{x})| \le L\mathbb{E}\{\mathsf{D}(\mu \boldsymbol{U})\}. \tag{3.2}$$

- *If f is convex on $\mathbb{R}^N$, so is f_μ, and f_μ overestimates f everywhere on $\mathcal{F}$.*
- *For every $0 < \mu < \mu_\star$, f_μ is differentiable on $\mathcal{F}$, and its gradient $\nabla f_\mu : \mathbb{R}^N \to \mathbb{R}^N$ may be written as*

$$\nabla f_\mu(\boldsymbol{x}) \equiv \mathbb{E}\bigg\{\frac{f(\boldsymbol{x} + \mu \boldsymbol{U}) - f(\boldsymbol{x})}{\mu} \boldsymbol{U}\bigg\}, \quad \forall \boldsymbol{x} \in \mathcal{F}, \tag{3.3}$$

where integration is in the sense of Lebesgue. Further, if f is $(L, \mathsf{D}, \mathsf{T})$-SLipschitz on $\mathcal{F}$, then, for every $\boldsymbol{x} \in \mathcal{F}$,

$$\mathbb{E}\left\{\left\|\frac{f(\boldsymbol{x} + \mu \boldsymbol{U}) - f(\boldsymbol{x})}{\mu} \boldsymbol{U}\right\|_2^2\right\} \le \frac{1}{\mu^2} \mathbb{E}\big\{\big(L\mathsf{D}(\mu \boldsymbol{U}) + |\mathsf{T}(\boldsymbol{x}, \mu \boldsymbol{U})|\big)^2 \|\boldsymbol{U}\|_2^2\big\}. \tag{3.4}$$

Proof of Lemma 3.4. See Appendix A. ∎

Driven by Lemma 3.4, we also introduce a notion of *effectiveness* of a divergence-remainder pair, or (D, T)-*pair*, for short, which quantifies the accuracy of Gaussian smoothing, in general terms.

DEFINITION 3.5 (**Effectiveness of Gaussian Smoothing**). *Let $\boldsymbol{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$ and fix $q \ge 2$. Then:*

- *A (D, T)-pair is called q-effective on $\mathcal{F} \subseteq \mathbb{R}^N$ if and only if there are Borel functions $\mathsf{d} : \mathbb{R}^N \to \mathbb{R}$ and $\mathsf{t}_q : \mathcal{F} \times \mathbb{R}^N \to \mathbb{R}$, such that, for some $\varepsilon \ge 0$, $\mu_o \in (0, \infty]$, and for every $\mu \le \mu_o$,*

$$\mathsf{D}(\mu \boldsymbol{u}) \le \mu^{1+\varepsilon} \mathsf{d}(\boldsymbol{u}) \text{ and } \|\mathsf{T}([\boldsymbol{x}, \boldsymbol{Q}], \mu \boldsymbol{u})\|_{\mathcal{L}_q} \le \mu \mathsf{t}_q(\boldsymbol{x}, \boldsymbol{u}), \ \forall (\boldsymbol{x}, \boldsymbol{u}) \in \mathcal{F} \times \mathbb{R}^N,$$

where $\boldsymbol{Q}$ is $\mathscr{F}$-measurable, $\mathsf{d}(\boldsymbol{U}) \in \mathcal{Z}_q$ and $\mathsf{t}_q(\cdot, \boldsymbol{U}) \in \mathcal{Z}_q$.
- *A (D, T)-pair is called q-stable on $\mathcal{F}$ if and only if it is q-effective on $\mathcal{F}$, with $\mathsf{d}(\boldsymbol{U}) \|\boldsymbol{U}\|_2^{2/\overline{q}} \in \mathcal{Z}_{\overline{q}}$ and $\mathsf{t}_q(\cdot, \boldsymbol{U}) \|\boldsymbol{U}\|_2^{2/\overline{q}} \in \mathcal{Z}_{\overline{q}}$, for all $\overline{q} \in [2, q]$.*
- *A (D, T)-pair is called uniformly q_o-effective (stable) on $\mathcal{F}$ if and only if it is q-effective (stable) on $\mathcal{F}$ and, additionally, it holds that $\sup_{\boldsymbol{x} \in \mathcal{F}} \|\mathsf{t}_q(\boldsymbol{x}, \boldsymbol{U})\|_{\mathcal{L}_{\overline{q}}} < \infty$ (plus $\sup_{\boldsymbol{x} \in \mathcal{F}} \|\mathsf{t}_q(\boldsymbol{x}, \boldsymbol{U}) \|\boldsymbol{U}\|_2^{2/\overline{q}}\|_{\mathcal{L}_{\overline{q}}} < \infty$), for $\overline{q} \in [2, q]$.*

In any case of the above, if $\varepsilon > 0$, then D is called an efficient divergence.

In the context of Lemma 3.4, effectiveness of a (D, T)-pair implies that $\mathbb{E}\{\mathsf{D}(\mu \boldsymbol{U})\}$ in (3.2) decreases *at least linearly in μ as $\mu \to 0$*, whereas stability implies that the right of (3.4) stays *bounded in μ as $\mu \to 0$*. If the (D, T)-pair is uniformly 2-stable, then the right-hand side of (3.3) is also bounded in $\boldsymbol{x}$. Further, if D is an efficient divergence, then $\mathbb{E}\{\mathsf{D}(\mu \boldsymbol{U})\}$ decreases *superlinearly in μ as $\mu \to 0$*. The additional conditions imposed by Definition 3.5 will be relevant shortly.

Typical examples of effective/stable (D, T)-pairs are the one where $\mathsf{D}(\cdot) \equiv \|\cdot\|_2$ and $\mathsf{T} \equiv 0$, associated with the Lipschitz class on $\mathbb{R}^N$, and that where $\mathsf{D}(\cdot) \equiv \|\cdot\|_2^2$ and $\mathsf{T}([\bullet, \star], \cdot) \equiv \mathsf{T}(\bullet, \cdot) \equiv \langle \nabla f(\bullet), \cdot \rangle$, associated with the smooth class on $\mathbb{R}^N$.

Algorithm 4.1 *Free-MESSAGE*^{p}

Input: Initial points $\boldsymbol{x}^0 \in \mathcal{X}$, $y^0 \in \mathcal{Y}$, $z^0 \in \mathcal{Z}$, stepsizes $\{\alpha_n\}_n$, $\{\beta_n\}_n$, $\{\gamma_n\}_n$, IID sequences $\{\boldsymbol{W}_1^n\}_n$, $\{\boldsymbol{W}_2^n\}_n$, penalty coefficient $c \in [0,1]$, smoothing parameter μ.
Output: Sequence $\{\boldsymbol{x}^n\}_{n\in\mathbb{N}}$.
1: **for** $n = 0, 1, 2, \ldots$ **do**
2: Sample $\boldsymbol{U}_1^{n+1} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$ and $F\big(\boldsymbol{x}^n + \mu \boldsymbol{U}_1^{n+1}, \boldsymbol{W}_1^{n+1}\big)$.
3: Update (First SA Level):

$$y^{n+1} = \Pi_{\mathcal{Y}}\big\{(1-\beta_n)\, y^n + \beta_n F\big(\boldsymbol{x}^n + \mu \boldsymbol{U}_1^{n+1}, \boldsymbol{W}_1^{n+1}\big)\big\}$$

4: Sample $\big[\big(\boldsymbol{U}_2^{n+1}\big)^{\boldsymbol{T}}\, U^{n+1}\big]^{\boldsymbol{T}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N+1})$ and $F\big(\boldsymbol{x}^n + \mu \boldsymbol{U}_2^{n+1}, \boldsymbol{W}_2^{n+1}\big)$.
5: Update (Second SA Level): If $p > 1$, set

$$z^{n+1} = \Pi_{\mathcal{Z}}\big\{(1-\gamma_n)\, z^n + \gamma_n \big(\mathcal{R}\big(F\big(\boldsymbol{x}^n + \mu \boldsymbol{U}_2^{n+1}, \boldsymbol{W}_2^{n+1}\big) - \mu U^{n+1} - y^n\big)\big)^p\big\}.$$

Otherwise, set $z^{n+1} = 1$.
6: Evaluate $F\big(\boldsymbol{x}^n, \boldsymbol{W}_1^{n+1}\big)$ and $F\big(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\big)$.
7: Define auxiliary variables:

$$\Delta_1 = \frac{F\big(\boldsymbol{x}^n + \mu \boldsymbol{U}_1^{n+1}, \boldsymbol{W}_1^{n+1}\big) - F\big(\boldsymbol{x}^n, \boldsymbol{W}_1^{n+1}\big)}{\mu}$$

$$\Delta_2 = \frac{\big(\mathcal{R}\big(F\big(\boldsymbol{x}^n + \mu \boldsymbol{U}_2^{n+1}, \boldsymbol{W}_2^{n+1}\big) - \mu U^{n+1} - y^n\big)\big)^p - \big(\mathcal{R}\big(F\big(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}\big) - y^n\big)\big)^p}{\mu}$$

$$\boldsymbol{\Delta} = p^{-1}\,(z^n)^{(1-p)/p}\,\big(\boldsymbol{U}_2^{n+1} + \Delta_1 \boldsymbol{U}_1^{n+1} U^{n+1}\big)\Delta_2$$

8: Update (Third SA Level):

$$\boldsymbol{x}^{n+1} = \Pi_{\mathcal{X}}\big\{\boldsymbol{x}^n - \alpha_n\big(\Delta_1 \boldsymbol{U}_1^{n+1} + c\boldsymbol{\Delta}\big)\big\}$$

9: **end for**

4. The *Free-MESSAGE*^{p} Algorithm. The basic idea is to carefully exploit Lemma 3.4, and replace the gradients involved in expression (2.1) of Lemma 2.1 by appropriate smoothed versions, which may be evaluated by exploiting only zeroth-order information. To this end, for $\mu \geq 0$, define functions $g_\mu : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}_+$ and $\boldsymbol{h}_\mu : \mathbb{R}^N \to \mathbb{R}^N \times \mathbb{R}$ and $s_\mu : \mathbb{R}^N \to \mathbb{R}$ as

$$g_\mu(\boldsymbol{x}, y) \triangleq \mathbb{E}\{(\mathcal{R}(F(\boldsymbol{x} + \mu\boldsymbol{U}, \boldsymbol{W}) - (y + \mu U)))^p\}, \quad \text{and}$$
$$\boldsymbol{h}_\mu(\boldsymbol{x}) \triangleq \big[\boldsymbol{x} \,|\, s_\mu(\boldsymbol{x}) \triangleq \mathbb{E}\{F(\boldsymbol{x} + \mu\boldsymbol{U}, \boldsymbol{W})\}\big],$$

where $\big[\boldsymbol{U}^{\boldsymbol{T}}\, U\big]^{\boldsymbol{T}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N+1})$, $\big[\boldsymbol{U}^{\boldsymbol{T}}\, U\big]^{\boldsymbol{T}}$ and $\boldsymbol{W}$ are mutually independent, and where, *temporarily*, we implicitly and arbitrarily assume that the involved expectations are well-defined and finite. Then, for $\mu > 0$, we may consider the *μ-smoothed quasigradient of ϕ*

$$\widehat{\nabla}_\mu \phi(\boldsymbol{x}) \equiv \nabla s_\mu(\boldsymbol{x}) + c\nabla \boldsymbol{h}_\mu(\boldsymbol{x})\, \nabla g_\mu(\boldsymbol{h}_\mu(\boldsymbol{x})) \nabla \varrho(g_\mu(\boldsymbol{h}_\mu(\boldsymbol{x}))), \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{4.1}$$

again provided that everything is well-defined and finite. If, further, the conditions of Lemma 3.4 are fulfilled, *and* with Fubini's permission, it must be true that, for every

$\boldsymbol{x} \in \mathcal{X}$,

$$\nabla \boldsymbol{h}_{\mu}(\boldsymbol{x}) \equiv \left[\boldsymbol{I}_N \middle| \nabla s_{\mu}(\boldsymbol{x})\right] = \left[\boldsymbol{I}_N \middle| \mathbb{E}\left\{\frac{F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W}) - F(\boldsymbol{x},\boldsymbol{W})}{\mu}\boldsymbol{U}\right\}\right], \tag{4.2}$$

and, for every $(\boldsymbol{x}, y) \in \mathrm{Graph}_{\mathcal{X}}(s_{\mu})$,

$$\nabla g_{\mu}(\boldsymbol{x}, y) = \mathbb{E}\left\{\frac{(\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W}) - (y+\mu U)))^p - (\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W}) - y))^p}{\mu}\begin{bmatrix}\boldsymbol{U}\\ U\end{bmatrix}\right\}. \tag{4.3}$$

The quasigradient $\widehat{\nabla}_{\mu}\phi$ suggests a compositional (nested) Stochastic Approximation (SA) scheme for *approximating* a stochastic gradient for ϕ. Similarly to [23,45,46], this scheme consists of *three* SA levels and presumes the existence of *two* mutually independent, Independent and Identically Distributed (IID) information streams, $\{\boldsymbol{W}_1^n\}_n$, $\{\boldsymbol{W}_2^n\}_n$, accessible by a $\mathcal{Z}$*eroth*-$\mathcal{O}$*rder* $\mathcal{S}$ampling $\mathcal{O}$racle ($\mathcal{ZOSO}$) for F. We also assume the existence of a *Gaussian sampler*, generating independent standard Gaussian elements on $\mathbb{R}^{N+1}$, mutually independently of $\{\boldsymbol{W}_1^n\}_n$ and $\{\boldsymbol{W}_2^n\}_n$.

The *Free-MESSAGE*^{p} algorithm is presented in Algorithm 4.1, where the updates of the first and second SA levels are clearly specified, and where $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{Z} \subseteq \mathbb{R}$ are closed intervals (to be properly selected later on; see Section 6). For the third SA level, given $F(\boldsymbol{x}^n, \boldsymbol{W}_1^{n+1})$ and $F(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1})$, and upon defining finite differences $\Delta_1^{n+1} : \mathbb{R}^N \times \Omega \to \mathbb{R}$ and $\Delta_2^{n+1} : \mathbb{R}^N \times \mathbb{R} \times \Omega \to \mathbb{R}$ as

$$\Delta_{1,\mu}^{n+1}(\boldsymbol{x}^n) \triangleq \frac{F(\boldsymbol{x}^n + \mu\boldsymbol{U}_1^{n+1}, \boldsymbol{W}_1^{n+1}) - F(\boldsymbol{x}^n, \boldsymbol{W}_1^{n+1})}{\mu} \quad \text{and}$$

$$\Delta_{2,\mu,p}^{n+1}(\boldsymbol{x}^n, y^n) \triangleq \frac{\left(\mathcal{R}\left(F(\boldsymbol{x}^n + \mu\boldsymbol{U}_2^{n+1}, \boldsymbol{W}_2^{n+1}) - \mu U^{n+1} - y^n\right)\right)^p}{\mu} - \frac{\left(\mathcal{R}\left(F(\boldsymbol{x}^n, \boldsymbol{W}_2^{n+1}) - y^n\right)\right)^p}{\mu},$$

a *stochastic* quasigradient $\widehat{\nabla}_{\mu}^{n+1}\phi : \mathbb{R}^N \times \mathbb{R} \times \mathbb{R} \times \Omega \to \mathbb{R}$ is formed as (cf. (4.1))

$$\begin{aligned}
&\widehat{\nabla}_{\mu}^{n+1}\phi(\boldsymbol{x}^n, y^n, z^n)\\
&\triangleq \Delta_{1,\mu}^{n+1}(\boldsymbol{x}^n)\boldsymbol{U}_1^{n+1} + c\frac{1}{p}(z^n)^{\frac{1-p}{p}}\left[\boldsymbol{I}_N \middle| \Delta_{1,\mu}^{n+1}(\boldsymbol{x}^n)\boldsymbol{U}_1^{n+1}\right]\Delta_{2,\mu,p}^{n+1}(\boldsymbol{x}^n, y^n)\begin{bmatrix}\boldsymbol{U}_2^{n+1}\\ U^{n+1}\end{bmatrix}\\
&\equiv \Delta_{1,\mu}^{n+1}(\boldsymbol{x}^n)\boldsymbol{U}_1^{n+1} + c\frac{1}{p}(z^n)^{\frac{1-p}{p}}\left(\boldsymbol{U}_2^{n+1} + \Delta_{1,\mu}^{n+1}(\boldsymbol{x}^n)\boldsymbol{U}_1^{n+1}U^{n+1}\right)\Delta_{2,\mu,p}^{n+1}(\boldsymbol{x}^n, y^n)\\
&\triangleq \Delta_{1,\mu}^{n+1}(\boldsymbol{x}^n)\boldsymbol{U}_1^{n+1} + c\boldsymbol{\Delta}_{\mu,p}^{n+1}(\boldsymbol{x}^n, y^n, z^n).
\end{aligned}$$

Finally, the current estimate $\boldsymbol{x}^n$ is updated via a projected quasigradient step as

$$\boldsymbol{x}^{n+1} \equiv \Pi_{\mathcal{X}}\left\{\boldsymbol{x}^n - \alpha_n \widehat{\nabla}_{\mu}^{n+1}\phi(\boldsymbol{x}^n, y^n, z^n)\right\}.$$

5. Smoothed Risk-Averse Surrogates. So far, most mathematical arguments presented in Section 4 have been imprecise, since we discussed neither well-definiteness of g_{μ}, $\boldsymbol{h}_{\mu}$ and $\widehat{\nabla}_{\mu}\phi$, nor fulfillment of the conditions of Lemma 3.4. Here, we resolve all technicalities, and reveal the actual usefulness of $\widehat{\nabla}_{\mu}\phi$ in solving problem (1.1). Our discussion will revolve around the *perturbed cost* $F((\cdot) + \mu\boldsymbol{U}, \boldsymbol{W}) - \mu U \in \mathcal{Z}_p$, ranked via the risk measure ρ. Accordingly, we consider the well-defined, finite-valued function $\phi_\mu : \mathbb{R}^N \to \mathbb{R}$ defined as

$$\phi_\mu(\boldsymbol{x}) \triangleq \rho\left(\left[F\left(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W}\right)-\mu U\right]\right).$$

We also impose regularity conditions on the cost F and risk profile $\mathcal{R}$, as follows.

ASSUMPTION 5.1. *F and $\mathcal{R}$ satisfy the following conditions:*
C0 *The functions s and g obey (3.1).*
C1 *There is $G<\infty$, and a (D,T)-pair, such that*

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\left\|F\left(\boldsymbol{x}+\boldsymbol{u},\boldsymbol{W}\right)-F\left(\boldsymbol{x},\boldsymbol{W}\right)-\mathsf{T}\left(\left[\boldsymbol{x},\boldsymbol{W}\right],\boldsymbol{u}\right)\right\|_{\mathcal{L}_2}\le G\mathsf{D}\left(\boldsymbol{u}\right),\ \forall\boldsymbol{u}\in\mathbb{R}^N.$$

C2 *There is $V<\infty$, such that $\sup_{\boldsymbol{x}\in\mathcal{X}}\left\|F\left(\boldsymbol{x},\boldsymbol{W}\right)\right\|_{\mathcal{L}_2}\le V$.*
C3 *The associated (D,T)-pair is uniformly 2-effective on $\mathcal{X}$, and we define $\mathcal{D}_i\triangleq\left\|\mathsf{d}\left(\boldsymbol{U}\right)\right\|_{\mathcal{L}_i}$, for $i\in\{1,2\}$, and $\mathcal{T}_2\triangleq\sup_{\boldsymbol{x}\in\mathcal{X}}\left\|\mathsf{t}_q\left(\boldsymbol{x},\boldsymbol{U}\right)\right\|_{\mathcal{L}_2}<\infty$.*
C4 *If $p\in(1,2]$, there is $\eta>0$, such that $\inf_{x\in\mathbb{R}}\mathcal{R}\left(x\right)\ge\eta$. Otherwise, $\eta\equiv0$.*

Under Assumption 5.1 and using Lemma 3.4, the next result establishes that ϕ_μ qualifies as a *surrogate* to the base problem (1.1). In the following, recall that, for $\sigma>0$, a function $f:\mathbb{R}^N\to\mathbb{R}$ is σ-strongly (σ-weakly) convex if and only if $f(\cdot)-(+)\sigma\|\cdot\|^2$ is convex [11].

LEMMA 5.2 (**Smoothed Surrogates).** *Suppose that Assumption 5.1 is in effect. Then, for $0\le\mu\le\mu_o$, if $F(\cdot,\boldsymbol{W})$ is (resp. σ-weakly, σ-strongly) convex, so is ϕ_μ, and ϕ_μ is differentiable on $\mathcal{X}$ with $\nabla\phi_\mu\equiv\widehat{\nabla}_\mu\phi$, where $\boldsymbol{h}_\mu$, g_μ are well-defined and the gradients $\nabla\boldsymbol{h}_\mu$, ∇g_μ are given by (4.2), (4.3), respectively. Further, it is true that*

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\phi_\mu\left(\boldsymbol{x}\right)-\phi\left(\boldsymbol{x}\right)\right|\le\mu^{1+\varepsilon}G\mathcal{D}_1+c\mathcal{C}\left(\mu\right)\left(\mu^{1+\varepsilon}G(\mathcal{D}_1+\mathcal{D}_2)+\mu(\mathcal{T}_2+1)\right),$$

with

$$\mathcal{C}\left(\mu\right)\triangleq\mathbb{1}_{\{p\equiv1\}}+\eta^{-p/2}\big(\mathcal{R}(0)+2V+\mu^{1+\varepsilon}G(2\mathcal{D}_1+\mathcal{D}_2)+\mu(\mathcal{T}_2+1)\big)^{p/2}\mathbb{1}_{\{p\in(1,2]\}}.$$

Additionally, for every $\boldsymbol{x}\in\mathcal{X}$, it holds that

$$\begin{aligned}\phi\left(\boldsymbol{x}\right)-\inf_{\boldsymbol{x}\in\mathcal{X}}\phi\left(\boldsymbol{x}\right)&\le\phi_\mu\left(\boldsymbol{x}\right)-\inf_{\boldsymbol{x}\in\mathcal{X}}\phi_\mu\left(\boldsymbol{x}\right)+2\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\phi_\mu\left(\boldsymbol{x}\right)-\phi\left(\boldsymbol{x}\right)\right|\\&\le\phi_\mu\left(\boldsymbol{x}\right)-\inf_{\boldsymbol{x}\in\mathcal{X}}\phi_\mu\left(\boldsymbol{x}\right)+\Sigma^o\mu(\mu^\varepsilon+c),\end{aligned}\tag{5.1}$$

where $\Sigma^o\triangleq2\max\{G\mathcal{D}_1,\mathcal{C}\left(\mu\right)\left(\mu^\varepsilon G(\mathcal{D}_1+\mathcal{D}_2)+(\mathcal{T}_2+1)\right)\}$.

Lemma 5.2 suggests that ϕ_μ is useful as a *proxy* for studying *Free-MESSAGE^p* as a method to solve (1.1). Specifically, although a simple fact, inequality (5.1) is of key importance to the convergence analysis of the *Free-MESSAGE^p* algorithm, discussed later in Section 6. Lemma 5.2 will be proved in several stages, as follows.

5.1. Proof of Lemma 5.2. First, an immediate but very useful consequence of Assumption 5.1 is the following proposition. The proof is elementary and is omitted.

PROPOSITION 5.3 (**Implied Properties of $F(\cdot,\boldsymbol{W})$ I).** *Suppose that condition* **C1** *of Assumption 5.1 is in effect. Then the function $\mathsf{T}\left(\bullet,\cdot\right)\triangleq\mathbb{E}\{\mathsf{T}\left(\left[\bullet,\boldsymbol{W}\right],\cdot\right)\}$ is a normal remainder on $\mathcal{X}$. Further, it is true that, for every $\boldsymbol{u}\in\mathbb{R}^N$,*

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\left|s\left(\boldsymbol{x}+\boldsymbol{u}\right)-s\left(\boldsymbol{x}\right)-\mathsf{T}\left(\boldsymbol{x},\boldsymbol{u}\right)\right|$$

$$\leq \sup_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}\left\{\left|F\left(\boldsymbol{x}+\boldsymbol{u},\boldsymbol{W}\right)-F\left(\boldsymbol{x},\boldsymbol{W}\right)-\mathsf{T}\left([\boldsymbol{x},\boldsymbol{W}],\boldsymbol{u}\right)\right|\right\}\leq G\mathsf{D}\left(\boldsymbol{u}\right),$$

In other words, $\mathbb{E}\left\{F\left(\cdot,\boldsymbol{W}\right)\right\}$ *is* $(G,\mathsf{D},\mathsf{T})$*-SLipschitz on* $\mathcal{X}$*, and more. If, additionally, condition* **C2** *is in effect, it is true that, for every* $(\boldsymbol{x},\boldsymbol{u})\in\mathcal{X}\times\mathbb{R}^N$*,*

$$\begin{aligned}|s\left(\boldsymbol{x}+\boldsymbol{u}\right)|&\leq\mathbb{E}\left\{\left|F\left(\boldsymbol{x}+\boldsymbol{u},\boldsymbol{W}\right)\right|\right\}\\&\leq\left\|F\left(\boldsymbol{x}+\boldsymbol{u},\boldsymbol{W}\right)\right\|_{\mathcal{L}_2}\leq G\mathsf{D}\left(\boldsymbol{u}\right)+\left\|\mathsf{T}\left([\boldsymbol{x},\boldsymbol{W}],\boldsymbol{u}\right)\right\|_{\mathcal{L}_2}+V,\end{aligned}$$

For the rest of this section, define the set $\mathcal{Y}'\triangleq\left[-V-\mu^{1+\varepsilon}G\mathcal{D}_1,\mu^{1+\varepsilon}G\mathcal{D}_1+V\right]$. Leveraging Proposition 5.3, Assumption 5.1, and Lemma 3.4, we have the next result.

Lemma 5.4 (**Existence & Properties of s_μ and g_μ**). *Suppose that Assumption 5.1 is in effect. Then, for some* $\varepsilon\geq 0$ *and* $\mu_o\in(0,\infty]$ *according to Definition 3.5, the following statements are true:*

- *For every* $0\leq\mu\leq\mu_o$*,* s_μ *is well-defined and finite on* $\mathcal{X}$*, and*

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\left|s_\mu\left(\boldsymbol{x}\right)-s\left(\boldsymbol{x}\right)\right|\leq\mu^{1+\varepsilon}G\mathcal{D}_1.$$

  Further, if $F(\cdot,\boldsymbol{W})$ *is convex, then so is* s_μ*, and* $s_\mu\geq s$ *on* $\mathcal{X}$*.*
- *For every* $0<\mu\leq\mu_o$*,* s_μ *is differentiable on* $\mathcal{X}$*, and* ∇s_μ *is given by (4.2). Also,*

$$\mathbb{E}\left\{\left\|\frac{F\left(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W}\right)-F\left(\boldsymbol{x},\boldsymbol{W}\right)}{\mu}\boldsymbol{U}\right\|_2^2\right\}\leq\mathbb{E}\left\{\left(\mu^{\varepsilon}G\mathsf{d}\left(\boldsymbol{U}\right)+\mathsf{t}_2\left(\boldsymbol{x},\boldsymbol{U}\right)\right)^2\left\|\boldsymbol{U}\right\|_2^2\right\}.$$

- *For every* $0\leq\mu$*,* g_μ *is well-defined and finite on* $\mathcal{X}\times\mathcal{Y}'$*, and if* $F(\cdot,\boldsymbol{W})$ *is convex, then so is* g_μ*, and* $g_\mu\geq g$ *on* $\mathcal{X}\times\mathcal{Y}'$*. Further, if* $\mu\leq\mu_o$*, then for every* $(\boldsymbol{x},y_1,y_2)\in\mathcal{X}\times\mathcal{Y}'\times\mathcal{Y}'$*, and every* $\left[\boldsymbol{u}^{\boldsymbol{T}}\;u\right]^{\boldsymbol{T}}\in\mathbb{R}^{N+1}$*,* g *satisfies the Lipschitz-like property*

$$\begin{aligned}&\left|g\left(\boldsymbol{x}+\mu\boldsymbol{u},y_1+\mu u\right)-g\left(\boldsymbol{x},y_2\right)\right|\\&\quad\leq\mathcal{C}\left(\mu,\boldsymbol{x},\boldsymbol{u}\right)\left(\mu^{1+\varepsilon}G\mathsf{d}\left(\boldsymbol{u}\right)+\mu\mathsf{t}_2\left(\boldsymbol{x},\boldsymbol{u}\right)+\mu|u|+|y_1-y_2|\right),\end{aligned}$$

  where

$$\mathcal{C}\left(\mu,\boldsymbol{x},\boldsymbol{u}\right)\triangleq\begin{cases}1, & \text{if } p\equiv 1\\ p\eta^{(p-2)/2}[\mathcal{R}\left(0\right)+2V+2\mu^{1+\varepsilon}G\mathcal{D}_1 & \\ \quad+\mu^{1+\varepsilon}G\mathsf{d}\left(\boldsymbol{u}\right)+\mu\mathsf{t}_2\left(\boldsymbol{x},\boldsymbol{u}\right)+\mu\left|u\right|]^{p/2}, & \text{if } p\in(1,2]\end{cases}.$$

- *For every* $0<\mu\leq\mu_o$*,* g_μ *is differentiable on* $\mathcal{X}\times\mathcal{Y}'$*, where* ∇g_μ *is given by (4.3).*

Proof of Lemma 5.4. For the first part of the result, we know from Proposition 5.3 that the function $s\left(\cdot\right)\equiv\mathbb{E}\left\{F\left(\cdot,\boldsymbol{W}\right)\right\}$ is $(G,\mathsf{D},\mathsf{T})$-SLipschitz on $\mathcal{X}$. Then, for $0\leq\mu\leq\mu_o$, we may call the first part of Lemma 3.4, which implies that the function $\mathbb{E}\left\{s\left((\cdot)+\mu\boldsymbol{U}\right)\right\}\triangleq s'_\mu\left(\cdot\right)$ is well-defined and finite on $\mathcal{X}$, and

$$\sup_{x\in\mathcal{X}}\left|s'_\mu\left(\boldsymbol{x}\right)-s\left(\boldsymbol{x}\right)\right|\leq G\mathbb{E}\left\{\mathsf{D}\left(\mu\boldsymbol{U}\right)\right\}\leq\mu^{1+\varepsilon}G\mathbb{E}\left\{\mathsf{d}\left(\boldsymbol{U}\right)\right\}.$$

Additionally, if s is convex on $\mathcal{X}$, so is $\mathbb{E}\left\{s\left((\cdot)+\mu\boldsymbol{U}\right)\right\}$, and the latter overestimates the former. Observe, though, that s'_μ is by definition constructed as an *iterated expectation*, first relative to the distribution of $\boldsymbol{W}$, and then relative to that of $\boldsymbol{U}$, and *not* as an expectation relative to their product measure. Nevertheless, from Proposition 5.3 and condition **C3** we know that, for every $(\boldsymbol{x},\boldsymbol{u}) \in \mathcal{X} \times \mathbb{R}^N$,

$$\int |F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{w})| \mathcal{P}_{\boldsymbol{W}}(\mathrm{d}\boldsymbol{w}) \le \mu^{1+\varepsilon} G \mathsf{d}(\boldsymbol{u}) + \mu \mathsf{t}_2(\boldsymbol{x},\boldsymbol{u}) + V,$$

which in turn implies that, for every $\boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned}&\int \left[\int |F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{w})| \mathcal{P}_{\boldsymbol{W}}(\mathrm{d}\boldsymbol{w})\right] \mathcal{P}_{\boldsymbol{U}}(\mathrm{d}\boldsymbol{u}) \\ &\quad \le \mu^{1+\varepsilon} G \mathbb{E}\{\mathsf{d}(\boldsymbol{U})\} + \mu \mathbb{E}\{\mathsf{t}_2(\boldsymbol{x},\boldsymbol{U})\} + V < \infty.\end{aligned}$$

Then, by Fubini's Theorem (Corollary 2.6.5 and Theorem 2.6.6 in [3]), it follows that the function $\mathbb{E}\{F((\cdot)+\mu\boldsymbol{U},\boldsymbol{W})\} \equiv s_\mu(\cdot)$ is finite on $\mathcal{X}$, and that

$$\begin{aligned} s'_\mu(\boldsymbol{x}) &\equiv \int \left[\int F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{w}) \mathcal{P}_{\boldsymbol{W}}(\mathrm{d}\boldsymbol{w})\right] \mathcal{P}_{\boldsymbol{U}}(\mathrm{d}\boldsymbol{u}) \\ &\equiv \int F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{w}) [\mathcal{P}_{\boldsymbol{W}} \times \mathcal{P}_{\boldsymbol{U}}](\mathrm{d}[\boldsymbol{u},\boldsymbol{w}]) \equiv s_\mu(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathcal{X},\end{aligned}$$

since $\boldsymbol{W}$ and $\boldsymbol{U}$ are statistically independent. A similar procedure may be followed for the second part of the lemma, concerning the gradient of s_μ. Further, we have

$$\begin{aligned}&\mathbb{E}\left\{\left\|\frac{F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W}) - F(\boldsymbol{x},\boldsymbol{W})}{\mu}\boldsymbol{U}\right\|_2^2\right\} \\ &\equiv \frac{1}{\mu^2}\mathbb{E}\big\{\mathbb{E}\big\{|F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W}) - F(\boldsymbol{x},\boldsymbol{W}) \\ &\qquad - \mathsf{T}([\boldsymbol{x},\boldsymbol{W}],\mu\boldsymbol{U}) + \mathsf{T}([\boldsymbol{x},\boldsymbol{W}],\mu\boldsymbol{U})|^2 |\boldsymbol{U}\big\} \|\boldsymbol{U}\|_2^2\big\} \\ &\le \frac{1}{\mu^2}\mathbb{E}\big\{\big(\mu^{1+\varepsilon} G \mathsf{d}(\boldsymbol{U}) + \mu \mathsf{t}_2(\boldsymbol{x},\boldsymbol{U})\big)^2 \|\boldsymbol{U}\|_2^2\big\} \\ &\equiv \mathbb{E}\big\{\big(\mu^{\varepsilon} G \mathsf{d}(\boldsymbol{U}) + \mathsf{t}_2(\boldsymbol{x},\boldsymbol{U})\big)^2 \|\boldsymbol{U}\|_2^2\big\},\end{aligned}$$

which is what we wanted to show.

For the third part, because g is nonnegative, Fubini's Theorem implies that

$$\mathbb{E}\{g(\boldsymbol{x}+\mu\boldsymbol{U}, y+\mu U)\} \equiv \mathbb{E}\{(\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W}) - (y+\mu U)))^p\} \equiv g_\mu(\boldsymbol{x},y),$$

for all $(\boldsymbol{x},y) \in \mathcal{X} \times \mathcal{Y}'$, and for every $\mu \ge 0$, where the involved integrals always exist. Then, since g satisfies condition (3.1) of Lemma 3.4 by assumption (condition **C0**), it follows that g_μ inherits the respective properties. Next, we show that g is Lipschitz-like, as claimed. If $p \equiv 1$, we have, for every $(\boldsymbol{x},y_1,y_2) \in \mathcal{X} \times \mathcal{Y}' \times \mathcal{Y}'$ and $\left[\boldsymbol{u}^{\boldsymbol{T}}\, u\right]^{\boldsymbol{T}} \in \mathbb{R}^{N+1}$,

$$\begin{aligned}&|g(\boldsymbol{x}+\mu\boldsymbol{u}, y_1+\mu u) - g(\boldsymbol{x},y_2)| \\ &\quad \le \mathbb{E}\{|\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}) - (y_1+\mu u)) - \mathcal{R}(F(\boldsymbol{x},\boldsymbol{W}) - y_2)|\} \\ &\quad \le \mathbb{E}\{|F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}) - F(\boldsymbol{x},\boldsymbol{W})|\} + \mu|u| + |y_1 - y_2| \\ &\quad \le \mu^{1+\varepsilon} G \mathsf{d}(\boldsymbol{u}) + \mu \mathsf{t}_2(\boldsymbol{x},\boldsymbol{u}) + \mu|u| + |y_1 - y_2|,\end{aligned}$$

and we are done. When $p \in (1,2]$, we will exploit another uniform estimate

$$\begin{aligned}
&\|\mathcal{R}\left(F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-\mu u-y\right)\|_{\mathcal{L}_p}\\
&\le \|\mathcal{R}\left(F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-\mu u-y\right)\|_{\mathcal{L}_2}\\
&\le \|\mathcal{R}\left(0\right)+\left|F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-\mu u-y\right|\|_{\mathcal{L}_2}\\
&\le \mathcal{R}\left(0\right)+|y|+\mu\,|u|+\|F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)\|_{\mathcal{L}_2}\\
&\le \mathcal{R}\left(0\right)+2V+2\mu^{1+\varepsilon}G\mathcal{D}_1+\mu^{1+\varepsilon}G\mathsf{d}\left(\boldsymbol{u}\right)+\mu\mathsf{t}_2\left(\boldsymbol{x},\boldsymbol{u}\right)+\mu\,|u|\,,
\end{aligned}$$

which holds everywhere on $\mathcal{X}\times\mathcal{Y}'\times\mathbb{R}^N\times\mathbb{R}$. Similarly,

$$\|\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y\right)\|_{\mathcal{L}_p}\le\mathcal{R}\left(0\right)+2\mu^{1+\varepsilon}G\mathcal{D}_1+2V,$$

everywhere on $\mathcal{X}\times\mathcal{Y}'$. Then, for every $(\boldsymbol{x},y_1,y_2)\in\mathcal{X}\times\mathcal{Y}'\times\mathcal{Y}'$, and for every $\left[\boldsymbol{u}^{\boldsymbol{T}}\,u\right]^{\boldsymbol{T}}\in\mathbb{R}^{N+1}$, we may write (recall Assumption 5.1)

$$\begin{aligned}
&\left|g\left(\boldsymbol{x}+\mu\boldsymbol{u},y_1+\mu u\right)-g\left(\boldsymbol{x},y_2\right)\right|\\
&\le\mathbb{E}\big\{\left|\left(\mathcal{R}\left(F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-\mu u-y_1\right)\right)^p-\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y_2\right)\right)^p\right|\big\}\\
&\equiv\mathbb{E}\big\{\big|\left(\mathcal{R}\left(F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-\mu u-y_1\right)\right)^{p/2}-\left(\mathcal{R}\left(F\left(\boldsymbol{x}_2,\boldsymbol{W}\right)-y_2\right)\right)^{p/2}\big|\\
&\quad\times\big(\left(\mathcal{R}\left(F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-\mu u-y_1\right)\right)^{p/2}+\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y_2\right)\right)^{p/2}\big)\big\}\\
&\le\frac{p\eta^{(p-2)/2}}{2}\mathbb{E}\big\{\left|\mathcal{R}\left(F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-\mu u-y_1\right)-\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y_2\right)\right|\\
&\quad\times\big(\left(\mathcal{R}\left(F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-\mu u-y_1\right)\right)^{p/2}+\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y_2\right)\right)^{p/2}\big)\big\}\\
&\le\frac{p\eta^{(p-2)/2}}{2}\mathbb{E}\big\{\left(\left|F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-F\left(\boldsymbol{x},\boldsymbol{W}\right)\right|+\mu\,|u|+|y_1-y_2|\right)\\
&\quad\times\big(\left(\mathcal{R}\left(F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-\mu u-y_1\right)\right)^{p/2}+\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y_2\right)\right)^{p/2}\big)\big\}\\
&\le\frac{p\eta^{(p-2)/2}}{2}\big(\|F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-F\left(\boldsymbol{x},\boldsymbol{W}\right)\|_{\mathcal{L}_2}+\mu\,|u|+|y_1-y_2|\,\big)\\
&\quad\times\big(\|\left(\mathcal{R}\left(F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-\mu u-y_1\right)\right)^{p/2}\|_{\mathcal{L}_2}+\|\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y_2\right)\right)^{p/2}\|_{\mathcal{L}_2}\big)\\
&\equiv\frac{p\eta^{(p-2)/2}}{2}\big(\|F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-F\left(\boldsymbol{x},\boldsymbol{W}\right)\|_{\mathcal{L}_2}+\mu\,|u|+|y_1-y_2|\,\big)\\
&\quad\times\big(\|\left(\mathcal{R}\left(F\left(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W}\right)-\mu u-y_1\right)\right)\|_{\mathcal{L}_p}^{p/2}+\|\left(\mathcal{R}\left(F\left(\boldsymbol{x},\boldsymbol{W}\right)-y_2\right)\right)\|_{\mathcal{L}_p}^{p/2}\big)\\
&\le p\eta^{(p-2)/2}\big(\mu^{1+\varepsilon}G\mathsf{d}\left(\boldsymbol{u}\right)+\mu\mathsf{t}_2\left(\boldsymbol{x},\boldsymbol{u}\right)+\mu\,|u|+|y_1-y_2|\,\big)\\
&\quad\times\left[\mathcal{R}\left(0\right)+2V+2\mu^{1+\varepsilon}G\mathcal{D}_1+\mu^{1+\varepsilon}G\mathsf{d}\left(\boldsymbol{u}\right)+\mu\mathsf{t}_2\left(\boldsymbol{x},\boldsymbol{u}\right)+\mu\,|u|\right]^{p/2}.
\end{aligned}$$

Finally, the last part of Lemma 5.4 may be verified by another application of Fubini's Theorem, as in the first and second part discussed above, or the tower property, and another application of Lemma 3.4. ∎

We now prove Lemma 5.2 for $p\in(1,2]$; the case where $p\equiv 1$ is similar, albeit simpler. To start, for $0\le\mu\le\mu_o$, convexity of ϕ_μ on $\mathcal{X}$ follows from convexity of $F\left((\cdot)+\mu\boldsymbol{U},\boldsymbol{W}\right)-\mu U$ on $\mathcal{X}$, which may be shown trivially, based on the convexity of $F\left(\cdot,\boldsymbol{W}\right)$, whenever that is the case. If $F\left(\cdot,\boldsymbol{W}\right)$ is also σ-strongly convex of on $\mathbb{R}^N$, then this is equivalent to the approximate secant inequality

$$F\left(\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y},\boldsymbol{w}\right)\le\alpha F\left(\boldsymbol{x},\boldsymbol{w}\right)+(1-\alpha)F\left(\boldsymbol{y},\boldsymbol{w}\right)-\alpha(1-\alpha)\sigma\|\boldsymbol{x}-\boldsymbol{y}\|_2^2,$$

being true for all $(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w})\in\mathbb{R}^N\times\mathbb{R}^N\times\mathbb{R}^M$ and for all $\alpha\in[0,1]$. Then, for the randomly perturbed cost $F((\cdot)+\mu\boldsymbol{U},\boldsymbol{W})$ we have

$$\begin{aligned}F(\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}+\boldsymbol{u},\boldsymbol{w}) &\equiv F(\alpha(\boldsymbol{x}+\boldsymbol{u})+(1-\alpha)(\boldsymbol{y}+\boldsymbol{u}),\boldsymbol{w})\\ &\le \alpha F(\boldsymbol{x}+\boldsymbol{u},\boldsymbol{w})+(1-\alpha)F((\boldsymbol{y}+\boldsymbol{u}),\boldsymbol{w})\\ &\quad -\alpha(1-\alpha)\sigma\|(\boldsymbol{x}+\boldsymbol{u})-(\boldsymbol{y}+\boldsymbol{u})\|_2^2\\ &\equiv \alpha F(\boldsymbol{x}+\boldsymbol{u},\boldsymbol{w})+(1-\alpha)F(\boldsymbol{y}+\boldsymbol{u},\boldsymbol{w})\\ &\quad -\alpha(1-\alpha)\sigma\|\boldsymbol{x}-\boldsymbol{y}\|_2^2,\end{aligned}$$

for all $(\boldsymbol{x},\boldsymbol{y},\boldsymbol{u},\boldsymbol{w})\in\mathbb{R}^N\times\mathbb{R}^N\times\mathbb{R}^N\times\mathbb{R}^M$ and for all $\alpha\in[0,1]$. This demonstrates that, for every $\mu\ge0$, $F((\cdot)+\mu\boldsymbol{U},\boldsymbol{W})$ and thus $F((\cdot)+\mu\boldsymbol{U},\boldsymbol{W})-\mu U$ are both strongly convex on $\mathbb{R}^N$ with the same parameter σ, independent of μ. Consequently, ([23], Proposition 5) implies that ϕ_μ is σ-strongly convex on $\mathbb{R}^N$, as well. By exactly the same procedure we may show that ϕ_μ is σ-weakly convex whenever $F(\cdot,\boldsymbol{W})$ is σ-weakly convex; in the above inequalities, $\alpha(1-\alpha)\sigma\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$ is simply negated [11].

Next, to verify differentiability of ϕ_μ, it suffices to check the sufficient conditions of Lemma 2.1. Indeed, since, by condition **C4**, $\inf_{x\in\mathbb{R}}\mathcal{R}(x)\ge\eta>0$, it is true that $\kappa_{\mathcal{R}}\equiv-\infty$ and, thus, for every $\boldsymbol{x}\in\mathcal{X}$,

$$\mathcal{P}\left(F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})-\mu U-\mathbb{E}\{F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})-\mu U\}\le\kappa_{\mathcal{R}}\right)\equiv0<1.$$

Then, Lemma 2.1 implies that ϕ_μ is differentiable everywhere on $\mathcal{X}$, and also that $\nabla\phi_\mu(\boldsymbol{x})\equiv\widehat{\nabla}_\mu\phi(\boldsymbol{x})$, for all $\boldsymbol{x}\in\mathcal{X}$, which may easily shown by application of the composition rule to ϕ_μ, for which it is true that

$$\begin{aligned}\phi_\mu(\boldsymbol{x})&\equiv\mathbb{E}\{F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})-\mu U\}\\ &\qquad+c\,\|\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})-\mu U-\mathbb{E}\{F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})-\mu U\})\|_{\mathcal{L}_p}\\ &\equiv\mathbb{E}\{F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})\}\\ &\qquad+c\,\|\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})-(\mathbb{E}\{F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})\}+\mu U))\|_{\mathcal{L}_p}\\ &\equiv s_\mu(\boldsymbol{x})+c\varrho(g_\mu(\boldsymbol{h}_\mu(\boldsymbol{x}))),\quad\forall\boldsymbol{x}\in\mathcal{X}.\end{aligned}$$

Now, because of the fact that (see, for instance, Lemma 5.4)

$$-V-\mu^{1+\varepsilon}G\mathcal{D}_1<\inf_{\boldsymbol{x}\in\mathcal{X}}s_\mu(\boldsymbol{x})\le\sup_{\boldsymbol{x}\in\mathcal{X}}s_\mu(\boldsymbol{x})\le\mu^{1+\varepsilon}G\mathcal{D}_1+V\iff s_\mu\in\mathcal{Y}'\text{ on }\mathcal{X},$$

we may invoke Lemma 5.4, yielding, for every $\boldsymbol{x}\in\mathcal{X}$,

$$\begin{aligned}&\left|\phi_\mu(\boldsymbol{x})-\phi(\boldsymbol{x})\right|\\ &\le\left|s_\mu(\boldsymbol{x})-s(\boldsymbol{x})\right|+c|\varrho(g_\mu(\boldsymbol{h}_\mu(\boldsymbol{x})))-\varrho(g(\boldsymbol{h}(\boldsymbol{x})))|\\ &\le\mu^{1+\varepsilon}G\mathcal{D}_1+c|\varrho(g_\mu(\boldsymbol{h}_\mu(\boldsymbol{x})))-\varrho(g(\boldsymbol{h}(\boldsymbol{x})))|\\ &\le\mu^{1+\varepsilon}G\mathcal{D}_1+cp^{-1}\eta^{1-p}|\mathbb{E}\{g(\boldsymbol{x}+\mu\boldsymbol{U},s_\mu(\boldsymbol{x})+\mu U)\}-\mathbb{E}\{g(\boldsymbol{x},s(\boldsymbol{x}))\}|\\ &\le\mu^{1+\varepsilon}G\mathcal{D}_1+cp^{-1}\eta^{1-p}\mathbb{E}\{|g\left(\boldsymbol{x}+\mu\boldsymbol{U},s_\mu(\boldsymbol{x})+\mu U\right)-g(\boldsymbol{x},s(\boldsymbol{x}))|\}\\ &\le\mu^{1+\varepsilon}G\mathcal{D}_1+cp^{-1}\eta^{1-p}\mathbb{E}\big\{\mathcal{C}(\mu,\boldsymbol{x},\boldsymbol{U})\,(|s_\mu(\boldsymbol{x})-s(\boldsymbol{x})|\\ &\qquad\qquad+\mu^{1+\varepsilon}G\mathsf{d}(\boldsymbol{U})+\mu\mathsf{t}_2(\boldsymbol{x},\boldsymbol{U})+\mu|U|)\big\}\\ &\le\mu^{1+\varepsilon}G\mathcal{D}_1+cp^{-1}\eta^{1-p}\mathbb{E}\big\{\mathcal{C}(\mu,\boldsymbol{x},\boldsymbol{U})\,(\mu^{1+\varepsilon}G\mathcal{D}_1\end{aligned}\tag{5.2}$$

$$
\begin{aligned}
&\qquad\qquad + \mu^{1+\varepsilon} G\mathsf{d}(\boldsymbol{U}) + \mu \mathsf{t}_2(\boldsymbol{x},\boldsymbol{U}) + \mu |U|)\big\} \\
&\le \mu^{1+\varepsilon} G\mathcal{D}_1 + cp^{-1}\eta^{1-p} \|\mathcal{C}(\mu,\boldsymbol{x},\boldsymbol{U})\|_{\mathcal{L}_2} \big\| \mu^{1+\varepsilon} G\mathcal{D}_1 \\
&\qquad\qquad + \mu^{1+\varepsilon} G\mathsf{d}(\boldsymbol{U}) + \mu \mathsf{t}_2(\boldsymbol{x},\boldsymbol{U}) + \mu |U| \big\|_{\mathcal{L}_2} \\
&\le \mu^{1+\varepsilon} G\mathcal{D}_1 + cp^{-1}\eta^{1-p} \|\mathcal{C}(\mu,\boldsymbol{x},\boldsymbol{U})\|_{\mathcal{L}_2} (\mu^{1+\varepsilon} G\mathcal{D}_1 + \mu^{1+\varepsilon} G\mathcal{D}_2 + \mu \mathcal{T}_2 + \mu).
\end{aligned}
$$

Additionally, it is also true that

$$
\begin{aligned}
&\|\mathcal{C}(\mu,\boldsymbol{x},\boldsymbol{U})\|_{\mathcal{L}_2} \\
&\quad \equiv p\eta^{\frac{p-2}{2}} \big\| [\mathcal{R}(0) + 2V + 2\mu^{1+\varepsilon} G\mathcal{D}_1 + \mu^{1+\varepsilon} G\mathsf{d}(\boldsymbol{U}) + \mu \mathsf{t}_2(\boldsymbol{x},\boldsymbol{U}) + \mu |u|]^{p/2} \big\|_{\mathcal{L}_2} \\
&\quad \equiv p\eta^{\frac{p-2}{2}} \big\| \mathcal{R}(0) + 2V + 2\mu^{1+\varepsilon} G\mathcal{D}_1 + \mu^{1+\varepsilon} G\mathsf{d}(\boldsymbol{U}) + \mu \mathsf{t}_2(\boldsymbol{x},\boldsymbol{U}) + \mu |U| \big\|_{\mathcal{L}_p}^{p/2} \\
&\quad \le p\eta^{\frac{p-2}{2}} \big(\mathcal{R}(0) + 2V + 2\mu^{1+\varepsilon} G\mathcal{D}_1 + \mu^{1+\varepsilon} G\mathcal{D}_2 + \mu \mathcal{T}_2 + \mu \big)^{p/2}
\end{aligned}
$$

Therefore, for every $\boldsymbol{x} \in \mathcal{X}$, (5.2) may be further bounded from above as

$$
\begin{aligned}
\big|\phi_\mu(\boldsymbol{x}) - \phi(\boldsymbol{x})\big| \le \mu^{1+\varepsilon} G\mathcal{D}_1 + c\eta^{-p/2} \big(\mathcal{R}(0) + 2V + 2\mu^{1+\varepsilon} G\mathcal{D}_1 + \mu^{1+\varepsilon} G\mathcal{D}_2 + \mu \mathcal{T}_2 + \mu \big)^{p/2} \\
(\mu^{1+\varepsilon} G\mathcal{D}_1 + \mu^{1+\varepsilon} G\mathcal{D}_2 + \mu \mathcal{T}_2 + \mu),
\end{aligned}
$$

and we are done. Lastly, for every $(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{X} \times \mathcal{X}$, we may write

$$
\begin{aligned}
\phi(\boldsymbol{x}) - \phi(\boldsymbol{x}') &\le \phi_\mu(\boldsymbol{x}) + \sup_{\boldsymbol{x}\in\mathcal{X}} |\phi_\mu(\boldsymbol{x}) - \phi(\boldsymbol{x})| - \phi_\mu(\boldsymbol{x}') + \sup_{\boldsymbol{x}\in\mathcal{X}} |\phi_\mu(\boldsymbol{x}) - \phi(\boldsymbol{x})| \\
&\equiv \phi_\mu(\boldsymbol{x}) - \phi_\mu(\boldsymbol{x}') + 2 \sup_{\boldsymbol{x}\in\mathcal{X}} |\phi_\mu(\boldsymbol{x}) - \phi(\boldsymbol{x})| \\
&\le \phi_\mu(\boldsymbol{x}) - \inf_{\boldsymbol{x}\in\mathcal{X}} \phi_\mu(\boldsymbol{x}) + 2 \sup_{\boldsymbol{x}\in\mathcal{X}} |\phi_\mu(\boldsymbol{x}) - \phi(\boldsymbol{x})|.
\end{aligned}
$$

Enough said. □

Remark 5.5. We would like to note that although (weak, strong) convexity of ϕ_μ is guaranteed by (weak, strong) convexity of $F(\cdot, \boldsymbol{W})$ *and* provided that $c \in [0,1]$, the latter condition on c is by no means necessary for weak convexity of ϕ_μ, in particular. In fact, it is possible that ϕ_μ is weakly convex even if $c > 1$, despite that, in such a case, ρ is no longer a convex risk measure. This happens, for instance, when ϕ_μ is smooth, i.e., when its gradient $\nabla\phi_\mu$ is Lipschitz. □

6. Convergence Analysis. By Lemma 5.2, it follows that the compositional quasigradient $\widehat{\nabla}_\mu \phi$ (see (4.1)) is actually the gradient of the function ϕ_μ. Therefore, the *Free-MESSAGE*^{p} algorithm may be legitimately seen as a zeroth-order method to solve (*exactly* when convex) the mean-semideviation problem

$$
\inf_{\boldsymbol{x}\in\mathcal{X}} \big\{ \phi_\mu(\boldsymbol{x}) \equiv \rho([F(\boldsymbol{x} + \mu\boldsymbol{U}, \boldsymbol{W}) - \mu U]) \big\}, \tag{6.1}
$$

where $\mu > 0$ (if $\mu \equiv 0$, then $\phi_0 \equiv \phi$, and the situation is trivial). Lemma 5.2 explicitly quantifies the quality of the approximation of ϕ by ϕ_μ, as well. Consequently, it makes sense to *first* study the *Free-MESSAGE*^{p} algorithm as a method for solving the surrogate (6.1), and *then* attempt to relate the obtained results to the original problem, using Lemma 5.2. Our results follow this path. The behavior of the *Free-MESSAGE*^{p} algorithm will be characterized under the following conditions, extending Assumption 5.1 of the previous section.

ASSUMPTION 6.1. *Assumption 5.1 is in effect and conditions* **C1**-**C3** *are strengthened as follows:*

$\overline{\mathbf{C1}}$ *There is $G<\infty$, and a (D,T)-pair, as in condition* **C1**, *such that*

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\left\|F\left(\boldsymbol{x}+\boldsymbol{u},\boldsymbol{W}\right)-F\left(\boldsymbol{x},\boldsymbol{W}\right)-\mathsf{T}\left([\boldsymbol{x},\boldsymbol{W}],\boldsymbol{u}\right)\right\|_{\mathcal{L}_{4-2\mathbb{1}_{\{p\equiv 1\}}}}\le G\mathsf{D}\left(\boldsymbol{u}\right),\forall\boldsymbol{u}\in\mathbb{R}^{N}.$$

$\overline{\mathbf{C2}}$ *There is $V_p<\infty$, such that $\sup_{\boldsymbol{x}\in\mathcal{X}}\|F(\boldsymbol{x},\boldsymbol{W})\|_{\mathcal{L}_{2p}}\le V_p$. Thus, $V_1\equiv V$.*

$\overline{\mathbf{C3}}$ *The associated (D,T)-pair is uniformly $(4-2\mathbb{1}_{\{p\equiv 1\}})$-stable on $\mathcal{X}$.*

Additionally:

C5 *The sets $\mathcal{Y}$ and $\mathcal{Z}$ are chosen as*

$$\mathcal{Y}\triangleq\left[-V-\mu^{1+\varepsilon}G\mathcal{D}_1,\mu^{1+\varepsilon}G\mathcal{D}_1+V\right]\quad\text{and}$$
$$\mathcal{Z}\triangleq\left[\eta^{p},\infty\right).$$

C6 *There is $L<\infty$, such that g_μ satisfies the* <u>marginal</u> *smoothness condition*

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\|\nabla g_\mu(\boldsymbol{x},y_1)-\nabla g_\mu(\boldsymbol{x},y_2)\|_2\le L|y_1-y_2|,\quad\forall(y_1,y_2)\in\mathcal{Y}\times\mathcal{Y}.$$

Note that condition **C6** of Assumption 6.1 can be satisfied under various common circumstances, in particular when g is L-smooth globally on $\mathbb{R}^N\times\mathbb{R}$. Note, though, that condition **C6** is significantly weaker than demanding L-smoothness of g.

6.1. Main Implications of Assumption 6.1. As in the case of Assumption 5.1, an immediate consequence of Assumption 6.1 is the following proposition. The proof is omitted.

PROPOSITION 6.2 (**Implied Properties of $F(\cdot,\boldsymbol{W})$ II).** *Suppose that conditions $\overline{\mathbf{C1}}$ and $\overline{\mathbf{C2}}$ of Assumption 6.1 are in effect. Then, it is true that*

$$\|F(\boldsymbol{x}+\boldsymbol{u},\boldsymbol{W})\|_{\mathcal{L}_{2p}}\le G\mathsf{D}(\boldsymbol{u})+\|\mathsf{T}([\boldsymbol{x},\boldsymbol{W}],\boldsymbol{u})\|_{\mathcal{L}_{2p}}+V_p,$$

for every $(\boldsymbol{x},\boldsymbol{u})\in\mathcal{X}\times\mathbb{R}^N$. If condition $\overline{\mathbf{C3}}$ is also in effect, then, for every $\mu\in(0,\mu_o]$,

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\|F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})\|_{\mathcal{L}_{2p}}\le V_p'\triangleq\mu^{1+\varepsilon}G\|\mathsf{d}(\boldsymbol{U})\|_{\mathcal{L}_{2p}}+\mu\sup_{\boldsymbol{x}\in\mathcal{X}}\|\mathsf{t}_{2p}(\boldsymbol{x},\boldsymbol{U})\|_{\mathcal{L}_{2p}}+V_p.$$

The main purpose of Assumption 6.1 is to guarantee uniform boundedness of the gradients appearing in the *Free-MESSAGE^p* algorithm in a certain sense, *uniformly* on the respective feasible sets. In this respect, we have the next result.

LEMMA 6.3 (**Gradient Boundedness).** *Suppose that Assumption 6.1 is in effect. Then, for every $0<\mu\le\mu_o$, there exist problem dependent constants $B_1\equiv B_1^\mu<\infty$ and $B_2\equiv B_2^\mu<\infty$, both increasing and bounded in μ, such that*

$$B_1\ge\sup_{\boldsymbol{x}\in\mathcal{X}}\mathbb{E}\left\{\left\|\frac{F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})-F(\boldsymbol{x},\boldsymbol{W})}{\mu}\boldsymbol{U}\right\|_2^2\right\}\quad\text{and}\tag{6.2}$$

$$B_2\ge\sup_{\substack{\boldsymbol{x}\in\mathcal{X}\\ \boldsymbol{y}\in\mathcal{Y}}}\mathbb{E}\left\{\left\|\frac{(\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})-(y+\mu U)))^p-(\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y))^p}{\mu}\begin{bmatrix}\boldsymbol{U}\\ U\end{bmatrix}\right\|_2^2\right\}.\tag{6.3}$$

It thus follows that $\sup_{\boldsymbol{x}\in\mathcal{X}}\|\nabla s_\mu(\boldsymbol{x})\|_2^2\le B_1$ and $\sup_{(\boldsymbol{x},y)\in\mathcal{X}\times\mathcal{Y}}\|\nabla g_\mu(\boldsymbol{x},y)\|_2^2\le B_2$, implying that s_μ and g_μ are Lipschitz in the usual sense on $\mathcal{X}$ and $\mathcal{X}\times\mathcal{Y}$, respectively.

Proof of Lemma 6.3. We work assuming that $p \in (1,2]$. If $p \equiv 1$, the proof follows accordingly. Since (6.2) follows trivially from Lemma 5.4, we focus exclusively on showing (6.3). First, for every pair $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$, we may carefully write

$$\begin{aligned}
&\mathbb{E}\{|(\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-(y+\mu u)))^p-(\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y))^p|^2\} \\
&\equiv \mathbb{E}\{|(\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-(y+\mu u)))^{p/2}-(\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y))^{p/2}|^2 \\
&\quad \times |(\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-(y+\mu u)))^{p/2}+(\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y))^{p/2}|^2\} \\
&\le \frac{p^2\eta^{p-2}}{4}\mathbb{E}\{|\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-(y+\mu u))-\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y)|^2 \\
&\quad \times |(\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-(y+\mu u)))^{p/2}+(\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y))^{p/2}|^2\} \\
&\le \frac{p^2\eta^{p-2}}{4}\mathbb{E}\{(|F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-F(\boldsymbol{x},\boldsymbol{W})|+\mu|u|)^2 \\
&\quad \times |(\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-(y+\mu u)))^{p/2}+(\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y))^{p/2}|^2\} \\
&\le \frac{p^2\eta^{p-2}}{2}\|(|F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-F(\boldsymbol{x},\boldsymbol{W})|+\mu|u|)^2\|_{\mathcal{L}_2} \\
&\quad \times \left(\|(\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-(y+\mu u)))^p\|_{\mathcal{L}_2}+\|(\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y))^p\|_{\mathcal{L}_2}\right) \\
&\equiv \frac{p^2\eta^{p-2}}{2}\||F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-F(\boldsymbol{x},\boldsymbol{W})|+\mu|u|\|_{\mathcal{L}_4}^2 \\
&\quad \times \left(\|\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-(y+\mu u))\|_{\mathcal{L}_{2p}}^p+\|\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y)\|_{\mathcal{L}_{2p}}^p\right) \\
&\le \frac{p^2\eta^{p-2}}{2}\left(\||F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-F(\boldsymbol{x},\boldsymbol{W})|\|_{\mathcal{L}_4}+\mu|u|\right)^2 \\
&\quad \times \left(\|\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{W})-(y+\mu u))\|_{\mathcal{L}_{2p}}^p+\|\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y)\|_{\mathcal{L}_{2p}}^p\right) \\
&\le p^2\eta^{p-2}\left(\mu^{1+\varepsilon}G\mathsf{d}(\boldsymbol{u})+\mu\mathsf{t}_4(\boldsymbol{x},\boldsymbol{u})+\mu|u|\right)^2 \\
&\quad \times \left(\mathcal{R}(0)+2V_p+2\mu^{1+\varepsilon}G\mathcal{D}_1+\mu^{1+\varepsilon}G\mathsf{d}(\boldsymbol{u})+\mu\mathsf{t}_4(\boldsymbol{x},\boldsymbol{u})+\mu|u|\right)^p \\
&\le p^2\eta^{p-2}2^{p-1}\left(\mu^{1+\varepsilon}G\mathsf{d}(\boldsymbol{u})+\mu\mathsf{t}_4(\boldsymbol{x},\boldsymbol{u})+\mu|u|\right)^2 \\
&\quad \times \left(\left(\mathcal{R}(0)+2V_p+2\mu^{1+\varepsilon}G\mathcal{D}_1\right)^p+\left(\mu^{1+\varepsilon}G\mathsf{d}(\boldsymbol{u})+\mu\mathsf{t}_4(\boldsymbol{x},\boldsymbol{u})+\mu|u|\right)^p\right) \\
&\equiv \mu^2p^2\eta^{p-2}2^{p-1}\left(\mu^p\left(\mu^{\varepsilon}G\mathsf{d}(\boldsymbol{u})+\mathsf{t}_4(\boldsymbol{x},\boldsymbol{u})+|u|\right)^{p+2}\right. \\
&\quad \left.+\left(\mu^{\varepsilon}G\mathsf{d}(\boldsymbol{u})+\mathsf{t}_4(\boldsymbol{x},\boldsymbol{u})+|u|\right)^2\left(\mathcal{R}(0)+2V_p+2\mu^{1+\varepsilon}G\mathcal{D}_1\right)^p\right).
\end{aligned}$$

Therefore, the tower property implies that

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\frac{(\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})-(y+\mu U)))^p-(\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y))^p}{\mu}\begin{bmatrix}\boldsymbol{U}\\U\end{bmatrix}\right\|_2^2\right\} \\
&\equiv \frac{1}{\mu^2}\mathbb{E}\left\{\mathbb{E}\{|(\mathcal{R}(F(\boldsymbol{x}+\mu\boldsymbol{U},\boldsymbol{W})-(y+\mu U)))^p\right. \\
&\qquad\qquad \left.-(\mathcal{R}(F(\boldsymbol{x},\boldsymbol{W})-y))^p|^2|\boldsymbol{U},U\}\left\|\begin{bmatrix}\boldsymbol{U}\\U\end{bmatrix}\right\|_2^2\right\} \\
&\le p^2\eta^{(p-2)}2^{p-1}\left(\mu^p\mathbb{E}\{\left(\mu^{\varepsilon}G\mathsf{d}(\boldsymbol{U})+\mathsf{t}_4(\boldsymbol{x},\boldsymbol{U})+|U|\right)^{p+2}(\|\boldsymbol{U}\|_2^2+U^2)\}\right. \\
&\quad \left.+\left(\mathcal{R}(0)+2V_p+2\mu^{1+\varepsilon}G\mathcal{D}_1\right)^p\mathbb{E}\{\left(\mu^{\varepsilon}G\mathsf{d}(\boldsymbol{U})+\mathsf{t}_4(\boldsymbol{x},\boldsymbol{U})+|U|\right)^2(\|\boldsymbol{U}\|_2^2+U^2)\}\right),
\end{aligned} \tag{6.4}$$

for all $\boldsymbol{x} \in \mathcal{X}$. The proof is now complete, but let us consider the two expectations on the right-hand side of (6.4) separately. For the first one, we may write

$$\begin{aligned}
&\mathbb{E}\big\{\big(\mu^{\varepsilon}G\mathsf{d}\,(\boldsymbol{U})+\mathsf{t}_4\,(\boldsymbol{x},\boldsymbol{U})+|U|\,\big)^{p+2}(\|\boldsymbol{U}\|_2^2+U^2)\big\}\\
&\le 2^{p+1}\mathbb{E}\big\{\big((\mu^{\varepsilon}G\mathsf{d}\,(\boldsymbol{U})+\mathsf{t}_4\,(\boldsymbol{x},\boldsymbol{U}))^{p+2}+|U|^{p+2}\big)(\|\boldsymbol{U}\|_2^2+U^2)\big\}\\
&\equiv 2^{p+1}\big(\mathbb{E}\big\{\big(\mu^{\varepsilon}G\mathsf{d}\,(\boldsymbol{U})+\mathsf{t}_4\,(\boldsymbol{x},\boldsymbol{U})\big)^{p+2}\|\boldsymbol{U}\|_2^2\big\}\\
&\quad+\mathbb{E}\big\{\big(\mu^{\varepsilon}G\mathsf{d}\,(\boldsymbol{U})+\mathsf{t}_4\,(\boldsymbol{x},\boldsymbol{U})\big)^{p+2}|U|^{p+2}\,\big\}+\mathbb{E}\big\{\,|U|^{p+2}\,\|\boldsymbol{U}\|_2^2\big\}+\mathbb{E}\big\{\,|U|^{p+4}\,\big\}\big)\\
&\le 2^{p+1}\big(2^{p+1}\big(\mu^{\varepsilon(p+2)}\mathbb{E}\big\{\big(G\mathsf{d}\,(\boldsymbol{U})\big)^{p+2}\|\boldsymbol{U}\|_2^2\big\}+\sup_{\boldsymbol{x}\in\mathcal{X}}\mathbb{E}\big\{\big(\mathsf{t}_4\,(\boldsymbol{x},\boldsymbol{U})\big)^{p+2}\|\boldsymbol{U}\|_2^2\big\}\big)\\
&\quad+2^{p+1}\big(\mu^{\varepsilon(p+2)}\mathbb{E}\big\{\big(G\mathsf{d}\,(\boldsymbol{U})\big)^{p+2}\big\}\mathbb{E}\big\{|U|^{p+2}\,\big\}\\
&\quad+\sup_{\boldsymbol{x}\in\mathcal{X}}\mathbb{E}\big\{\big(\mathsf{t}_4\,(\boldsymbol{x},\boldsymbol{U})\big)^{p+2}\big\}\mathbb{E}\big\{|U|^{p+2}\,\big\}\big)+\mathbb{E}\big\{\,|U|^{p+2}\,\big\}\mathbb{E}\big\{\|\boldsymbol{U}\|_2^2\big\}+\mathbb{E}\big\{\,|U|^{p+4}\,\big\}\big).
\end{aligned}$$

For the second one, the situation is similar. Enough said. □

6.2. Recursions. We follow the approach taken previously in ([23], Section 4.4), but with appropriate technical modifications in the proofs of the corresponding results, reflecting the problem setting and assumptions considered herein. Because the techniques utilized are similar to ([23], Section 4.4), the proofs are omitted. Still, we would like to emphasize that the results presented below crucially exploit gradient boundedness ensured by Lemma 6.3, which follows as a result of Assumption 6.1.

Hereafter, let $\{\mathscr{D}^n\subseteq\mathscr{F}\}_{n\in\mathbb{N}}$ be the filtration generated from all data observed so far, *by both the user and the* $\mathcal{ZOSO}$, with $\mathscr{D}^n\triangleq\sigma\big\{\boldsymbol{x}^i,y^i,z^i,\boldsymbol{W}_1^i,\boldsymbol{W}_2^i,\boldsymbol{U}_1^i,\boldsymbol{U}_2^i,U^i,\forall i\in\mathbb{N}_n\big\}$, $n\in\mathbb{N}$. Also, if $\mathscr{C}$ is a sub σ-algebra of $\mathscr{F}$, we compactly write $\mathbb{E}\{\cdot\,|\mathscr{C}\}\equiv\mathbb{E}_{\mathscr{C}}\{\cdot\}$. Our first basic result follows.

Lemma 6.4 (**Iterate Increment Growth).** *Suppose that Assumption 6.1 is in effect. Then, for every $0<\mu\le\mu_o$, there exists a problem dependent constant $\Sigma_p^1<\infty$, increasing and bounded in μ, such that the process $\{\boldsymbol{x}^n\}_{n\in\mathbb{N}}$ generated by the Free-MESSAGE^p algorithm satisfies the inequality*

$$\mathbb{E}_{\mathscr{D}^n}\big\{\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^n\|_2^2\big\}\le\Sigma_p^1\alpha_n^2,$$

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

Using Lemma 6.4, we have the next result on the growth of $|y^n-s_\mu(\boldsymbol{x}^n)|^2$.

Lemma 6.5 (**1$^{\text{st}}$ Zeroth-order SA Level: Error Growth).** *Suppose that Assumption 6.1 is in effect. Also, let $\beta_n\in(0,1]$, for all $n\in\mathbb{N}$. Then, for every $0<\mu\le\mu_o$, there exists a problem dependent constant $\Sigma_p^2<\infty$, increasing and bounded in μ, such that the process $\{(\boldsymbol{x}^n,y^n)\}_{n\in\mathbb{N}}$ generated by the Free-MESSAGE^p algorithm satisfies the inequality*

$$\mathbb{E}_{\mathscr{D}^n}\big\{|y^{n+1}-s_\mu(\boldsymbol{x}^{n+1})|^2\big\}\le(1-\beta_n)\,|y^n-s_\mu(\boldsymbol{x}^n)|^2+\Sigma_p^2(\beta_n^2+\beta_n^{-1}\alpha_n^2),$$

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

Similarly, when $p>1$, the growth of $|z^n-g_\mu(\boldsymbol{x}^n,y^n)|^2$ may be characterized as follows.

Lemma 6.6 (**2$^{\text{nd}}$ Zeroth-order SA Level: Error Growth).** *Suppose that Assumption 6.1 is 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, for every $0<\mu\le\mu_o$, there exists a problem dependent constant*

$\Sigma_p^3 < \infty$, *increasing and bounded in* μ, *such that the process* $\{(\boldsymbol{x}^n, y^n, z^n)\}_{n\in\mathbb{N}}$ *generated by the Free-MESSAGE*^{p} *algorithm satisfies the inequality*

$$\mathbb{E}_{\mathscr{D}^n}\{|z^{n+1} - g_\mu(\boldsymbol{x}^{n+1}, y^{n+1})|^2\}$$
$$\le (1-\gamma_n)|z^n - g_\mu(\boldsymbol{x}^n, y^n)|^2 + \Sigma_p^3(\gamma_n^2 + \gamma_n^{-1}\alpha_n^2 + \gamma_n^{-1}\beta_n^2),$$

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

Now note that, as in the original *MESSAGE*^{p} algorithm [23], it is true that, for every $(n, \boldsymbol{x}) \in \mathbb{N}^+ \times \mathcal{X}$,

$$\mathbb{E}\{\widehat{\nabla}_\mu^{n+1}\phi(\boldsymbol{x}, s_\mu(\boldsymbol{x}), g_\mu(\boldsymbol{x}, s_\mu(\boldsymbol{x})))\} \equiv \widehat{\nabla}_\mu\phi(\boldsymbol{x}),$$

implying that $\widehat{\nabla}_\mu^{n+1}\phi$ constitutes an unbiased estimator of $\widehat{\nabla}_\mu\phi$, that is, a valid stochastic gradient associated with the latter. Using this fact, we now characterize the evolution of $\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^\star\|_2^2$ for arbitrary $\boldsymbol{x}^\star \in \mathcal{X}$, to be properly selected later.

LEMMA 6.7 (**3$^{\text{rd}}$ Zeroth-order SA Level: Error Growth).** *Suppose that Assumption 6.1 is in effect, and let* $\beta_n \in (0,1]$, $\gamma_n \in (0,1]$, *for all* $n \in \mathbb{N}$. *Then, for every* $0 < \mu \le \mu_o$ *and an arbitrary* $\boldsymbol{x}^\star \in \mathcal{X}$, *there exists another problem dependent constant* $0 < \Sigma_p^4 < \infty$, *also increasing and bounded in* μ, *such that the process* $\{(\boldsymbol{x}^n, y^n, z^n)\}_{n\in\mathbb{N}}$ *generated by the Free-MESSAGE*^{p} *algorithm satisfies*

$$\mathbb{E}_{\mathscr{D}^n}\{\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^\star\|_2^2\}$$
$$\le \|\boldsymbol{x}^n - \boldsymbol{x}^\star\|_2^2 - 2\alpha_n(\boldsymbol{x}^n - \boldsymbol{x}^\star)^{\boldsymbol{T}}\nabla\phi_\mu(\boldsymbol{x}^n)$$
$$+ \Sigma_p^1\alpha_n^2 + 2\sqrt{\Sigma_p^4}c\alpha_n\|\boldsymbol{x}^n - \boldsymbol{x}^\star\|_2\big(|y^n - s_\mu(\boldsymbol{x}^n)| + \mathbb{1}_{\{p>1\}}|z^n - g_\mu(\boldsymbol{x}^n, y^n)|\big)$$

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

At this point, it is important to observe that Lemmata 6.4, 6.5, 6.6 and 6.7 share essentially the same structure with the corresponding results used in the analysis of the gradient-based *MESSAGE*^{p} algorithm of [23]; see, in particular, ([23], Section 4.4). Therefore, the behavior of the *Free-MESSAGE*^{p} algorithm as a method to solve *the surrogate problem* (6.1) can be analyzed *almost automatically*, by calling the respective convergence results developed in [23], which are based exclusively on the counterparts of Lemmata 6.4, 6.5, 6.6 and 6.7, presented therein. Then, the obtained results can be related back to the base problem (1.1), via Lemma 5.2. This is the path taken for proving our main results, as discussed below.

Also note that the constants Σ_p^1, Σ_p^2, Σ_p^3 and Σ_p^4 involved in Lemmata 6.4, 6.5, 6.6 and 6.7, respectively, *are all increasing and bounded in the smoothing parameter* $\mu \in (0, \mu_o]$. Therefore, when deriving convergence rates of the expected value type, based exclusively on Lemmata 6.4, 6.5, 6.6 and 6.7, similarly to ([23], Section 4.4), and under appropriate stepsize initialization, all resulting constants will also be increasing and bounded functions of $\mu \in (0, \mu_o]$.

6.3. Path Convergence for Convex Surrogates. When the smoothed surrogate ϕ_μ is convex (ensured if the cost $F(\cdot, \boldsymbol{W})$ is convex; see Lemma 5.2), the path behavior of the *Free-MESSAGE*^{p} algorithm may be characterized via the following result. Hereafter, let $\phi^* \triangleq \inf_{\boldsymbol{x}\in\mathcal{X}}\phi(\boldsymbol{x}) \in \mathbb{R}$.

THEOREM 6.8 (**Path Convergence *Free-MESSAGE*^{p} | Convex Surrogate).** *Suppose that Assumption 6.1 is in effect, and let* $\beta_n \in (0,1]$, $\gamma_n \in (0,1]$, *for all* $n \in \mathbb{N}$.

Also, suppose that

$$\sum_{n\in\mathbb{N}}\alpha_n \equiv \infty,\ \sum_{n\in\mathbb{N}}\alpha_n^2+\beta_n^2+\frac{\alpha_n^2}{\beta_n}<\infty, \textit{and if } p>1,\ \sum_{n\in\mathbb{N}}\gamma_n^2+\frac{\alpha_n^2}{\gamma_n}+\frac{\beta_n^2}{\gamma_n}<\infty.$$

Then, for $0<\mu\le\mu_o$, and provided that ϕ_μ is convex and $\mathcal{X}_\mu^o \triangleq \arg\min_{\boldsymbol{x}\in\mathcal{X}}\phi_\mu(\boldsymbol{x}) \neq \emptyset$, there is an event $\Omega'\subseteq\Omega$ with $\mathcal{P}(\Omega')\equiv 1$, such that, for every $\omega\in\Omega'$, the process $\{\boldsymbol{x}^n(\omega)\}_{n\in\mathbb{N}}$ generated by the Free-MESSAGE^p algorithm converges as

$$\boldsymbol{x}^n(\omega) \underset{n\to\infty}{\longrightarrow} \boldsymbol{x}^o(\omega)\in\mathcal{X}_\mu^o, \tag{6.5}$$

also implying that

$$\lim_{n\to\infty}\phi(\boldsymbol{x}^n(\omega))-\phi^* \le 2\sup_{\boldsymbol{x}\in\mathcal{X}}|\phi_\mu(\boldsymbol{x})-\phi(\boldsymbol{x})| \equiv \Sigma^o\mu(\mu^\varepsilon+c). \tag{6.6}$$

In words, almost everywhere relative to $\mathcal{P}$, $\{\boldsymbol{x}^n\}_{n\in\mathbb{N}}$ converges in the set of optimal solutions of (6.1), and $\{\phi(\boldsymbol{x}^n)\}_{n\in\mathbb{N}}$ converges to a μ-neighborhood of ϕ^*.

Proof of Theorem 6.8. First, for every $\boldsymbol{x}^o\in\mathcal{X}_\mu^o\neq\emptyset$ (as assumed), by convexity, and after standard manipulations, Lemma 6.7 readily implies that

$$\begin{aligned}
&\mathbb{E}_{\mathscr{D}^n}\{\|\boldsymbol{x}^{n+1}-\boldsymbol{x}^o\|_2^2\}\\
&\le\left[1+\Sigma_p^4c^2\left(\frac{\alpha_n^2}{\beta_n}+\frac{\alpha_n^2}{\gamma_n}\mathbb{1}_{\{p>1\}}\right)\right]\|\boldsymbol{x}^n-\boldsymbol{x}^o\|_2^2+\Sigma_p^1\alpha_n^2\\
&\quad-2\alpha_n\big(\phi_\mu(\boldsymbol{x}^n)-\phi_\mu^o\big)+\beta_n|y^n-s_\mu(\boldsymbol{x}^n)|^2+\gamma_n|z^n-g_\mu(\boldsymbol{x}^n,y^n)|^2\mathbb{1}_{\{p>1\}},
\end{aligned} \tag{6.7}$$

for all $n\in\mathbb{N}$, almost everywhere relative to $\mathcal{P}$, where we define $\phi_\mu^o\triangleq\inf_{\boldsymbol{x}\in\mathcal{X}}\phi_\mu(\boldsymbol{x})$.

Then, the proof of (6.5) follows directly from ([23], Section 4.4, Theorem 3), based on an application of the *T-level almost-supermartingale convergence lemma* [46]. To prove (6.6), note that, for every $\omega\in\Omega'$, continuity of ϕ on $\mathcal{X}$ implies that

$$\lim_{n\to\infty}\phi(\boldsymbol{x}^n(\omega))-\phi^*\equiv\phi(\boldsymbol{x}^o(\omega))-\phi^*.$$

Then, since $\boldsymbol{x}^o(\omega)\in\mathcal{X}_\mu^o$, Lemma 5.2 implies that

$$\begin{aligned}
\phi(\boldsymbol{x}^o(\omega))-\phi^* &\equiv \phi(\boldsymbol{x}^o(\omega))-\inf_{\boldsymbol{x}\in\mathcal{X}}\phi(\boldsymbol{x})\\
&\le \phi_\mu(\boldsymbol{x}^o(\omega))-\inf_{\boldsymbol{x}\in\mathcal{X}}\phi_\mu(\boldsymbol{x})+\Sigma^o\mu(\mu^\varepsilon+c)\\
&\equiv \phi_\mu(\boldsymbol{x}^o(\omega))-\phi_\mu(\boldsymbol{x}^o(\omega))+\Sigma^o\mu(\mu^\varepsilon+c)\\
&\equiv \Sigma^o\mu(\mu^\varepsilon+c),
\end{aligned}$$

and we are done. ∎

6.4. Convergence Rates.

6.4.1. Convex Surrogate. For the case of a generic convex surrogate ϕ_μ (obtained, e.g., whenever the cost $F(\cdot,\boldsymbol{W})$ is convex; see Lemma 5.2), we have the following result on the rate of convergence of the *Free-MESSAGE^p* algorithm, concerning smoothened iterates of the form [45, 46]

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

THEOREM 6.9 (**Rate | Convex Surrogate | Subharmonic Stepsizes).** *Let Assumption 6.1 be in effect, set $\alpha_0 \equiv \beta_0 \equiv \gamma_0 \equiv 1$, and for $n \in \mathbb{N}^+$, choose $\alpha_n \equiv n^{-\tau_1}$, $\beta_n \equiv n^{-\tau_2}$ and $\gamma_n \equiv n^{-\tau_3}$, where, for fixed $\epsilon \in [0,1)$, $\delta \in (0,1)$ and $\zeta \in (0,1)$ such that $\delta \geq \zeta$,*

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

Additionally, for $0 < \mu \leq \mu_o$, suppose that ϕ_μ is convex, and that $\sup_{n\in\mathbb{N}} \mathbb{E}\{\|\boldsymbol{x}^n - \boldsymbol{x}^o\|_2^2\} \leq E_\mu < \infty$, where $\boldsymbol{x}^o \in \mathcal{X}_\mu^o$. Then, for every $n \in \mathbb{N}^+$, the Free-MESSAGE^p algorithm satisfies

$$\mathbb{E}\{\phi(\widehat{\boldsymbol{x}}^n) - \phi^*\} \leq \mathcal{K}_p^{E_\mu} n^{-(1-\epsilon)/(4\mathbb{1}_{\{p\in(1,2]\}}+4)} + \Sigma^o\mu(\mu^\varepsilon + c), \tag{6.8}$$

where $\mathcal{K}_p^{E_\mu} \in (0,\infty)$ is increasing and bounded in μ, whenever E_μ is in fact independent of μ.

Proof of Theorem 6.9. By exploiting (6.7) as in the proof of Theorem 6.8, the result follows in part from ([23], Section 4.4, Theorem 4 and its proof), which applied to our setting yields

$$\mathbb{E}\{\phi_\mu(\widehat{\boldsymbol{x}}^n) - \phi_\mu^o\} \leq \mathcal{K}_p^{E_\mu} n^{-(1-\epsilon)/(4\mathbb{1}_{\{p\in(1,2]\}}+4)}, \quad \forall n \in \mathbb{N}^+,$$

where $\mathcal{K}_p^{E_\mu} \in (0,\infty)$ is increasing and bounded in μ, whenever E_μ is not dependent on μ (e.g., if $\mathcal{X}$ is compact). Then, for any $\boldsymbol{x}^o \in \mathcal{X}_\mu^o$, Lemma 5.2 implies that

$$\begin{aligned} \phi(\widehat{\boldsymbol{x}}^n) - \phi^* &\equiv \phi(\widehat{\boldsymbol{x}}^n) - \inf_{\boldsymbol{x}\in\mathcal{X}} \phi(\boldsymbol{x}) \\ &\leq \phi_\mu(\widehat{\boldsymbol{x}}^n) - \inf_{\boldsymbol{x}\in\mathcal{X}} \phi_\mu(\boldsymbol{x}) + \Sigma^o\mu(\mu^\varepsilon + c) \\ &\equiv \phi_\mu(\widehat{\boldsymbol{x}}^n) - \phi_\mu(\boldsymbol{x}^o) + \Sigma^o\mu(\mu^\varepsilon + c) \\ &\equiv \phi_\mu(\widehat{\boldsymbol{x}}^n) - \phi_\mu^o + \Sigma^o\mu(\mu^\varepsilon + c), \quad \forall n \in \mathbb{N}^+, \end{aligned}$$

everywhere on Ω. Taking expectations completes the proof. ∎

6.4.2. Weakly Convex Objective and Surrogate. Next, we investigate the case of *both* a weakly convex objective ϕ *and* a weakly convex surrogate ϕ_μ (simultaneously obtained, e.g., whenever the cost $F(\cdot, \boldsymbol{W})$ is weakly convex; see Lemma 5.2). Here, following the approach taken in [11], our figure of merit will rely on the *Moreau envelope* associated with the risk function ϕ, $\phi^\lambda : \mathbb{R}^N \to \mathbb{R}$, defined for $\lambda > 0$ as

$$\phi^\lambda(\boldsymbol{x}) \triangleq \inf_{\boldsymbol{y}\in\mathcal{X}} \Big\{\phi(\boldsymbol{y}) + \frac{1}{2\lambda}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2\Big\},$$

as well as the closely related *proximal mapping* $\mathrm{prox}_{\lambda\phi} : \mathbb{R}^N \to \mathbb{R}^N$, defined as

$$\mathrm{prox}_{\lambda\phi}(\boldsymbol{x}) \triangleq \operatorname*{arg\,min}_{\boldsymbol{y}\in\mathcal{X}} \Big\{\phi(\boldsymbol{y}) + \frac{1}{2\lambda}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2\Big\}.$$

From [28], or ([11], Lemma 2.2), we know that if ϕ is σ-weakly convex (in the sense used in the proof of Lemma 5.2), ϕ^λ is continuously differentiable on $\mathbb{R}^N$ for every $\lambda \in (0, (2\sigma)^{-1})$, and its gradient may be expressed as

$$\nabla\phi^\lambda(\boldsymbol{x}) \equiv \frac{1}{\lambda}[\boldsymbol{x} - \mathrm{prox}_{\lambda\phi}(\boldsymbol{x})].$$

This is an important fact, because, as thoroughly explained in [11], the quantity $\|\nabla\phi^{\lambda}(\cdot)\|$ constitutes a reasonable measure of near-stationarity of ϕ: A small-valued $\|\nabla\phi^{\lambda}(\boldsymbol{x})\|$ implies that the particular $\boldsymbol{x}$ is *close to* another point $\hat{\boldsymbol{x}}_{\lambda} \triangleq \mathrm{prox}_{\lambda\phi}(\boldsymbol{x})$ which is *nearly stationary* for ϕ [11]. Therefore, we may adopt $\|\nabla\phi^{\lambda}(\cdot)\|$ as a stationarity measure (i.e., a figure of merit) for the base problem (1.1).

Essentially, what we do here is that we replace the original risk-aware objective ϕ by the *Moreau surrogate* ϕ^{λ}, and we study the rate of convergence of *Free-MESSAGE^p* for the resulting surrogate problem, instead of the original. This exactly matches the approach taken in [11]. However, the additional challenge here will be to understand the interplay between the Moreau surrogate ϕ^{λ} and the smoothed surrogate ϕ_{μ}, the latter naturally related to *Free-MESSAGE^p* by construction. In this respect, we have the next result.

THEOREM 6.10 (**Rate | Weakly Convex Objective/Surrogate | Subharmonic Stepsizes).** *Let Assumption 6.1 be in effect, and consider the stepsize selection of Theorem 6.9. Suppose that both ϕ and ϕ_{μ} are σ-weakly convex for some $0<\mu\le\mu_o$ and, additionally, suppose that $\sup_{n\in\mathbb{N}}\mathbb{E}\{\|\boldsymbol{x}^n\|_2^2\}\le E_{\mu}<\infty$. Then, for any fixed $\overline{\sigma}>\sigma$ and for every $n\in\mathbb{N}^+$, the Free-MESSAGE^p algorithm satisfies*

$$\frac{1}{n}\sum_{i\in\mathbb{N}_n^+}\mathbb{E}\big\{\|\nabla\phi^{1/2\overline{\sigma}}(\boldsymbol{x}^i)\|^2\big\}\le\frac{\overline{\sigma}}{\overline{\sigma}-\sigma}\big[\mathcal{K}_{p,\overline{\sigma}}^{E_{\mu}}n^{-(1-\epsilon)/(4\mathbb{1}_{\{p\in(1,2]\}}+4)}+2\overline{\sigma}\Sigma^o\mu(\mu^{\varepsilon}+c)\big],$$

where $\mathcal{K}_{p,\overline{\sigma}}^{E_{\mu}}\in(0,\infty)$ is increasing and bounded in μ, as long as E_{μ} is independent of μ.

Proof of Theorem 6.10. By weak convexity of ϕ_{μ} and by invoking Lemma 5.2, it is true that, for every $(\boldsymbol{x},\boldsymbol{x}')\in\mathcal{X}\times\mathcal{X}$,

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{x}')^{\boldsymbol{T}}\nabla\phi_{\mu}(\boldsymbol{x})&\ge\phi_{\mu}(\boldsymbol{x})-\phi_{\mu}(\boldsymbol{x}')-\sigma\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2\\&\ge\phi(\boldsymbol{x})-\phi(\boldsymbol{x}')-\sigma\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2-\Sigma^o\mu(\mu^{\varepsilon}+c).\end{aligned}$$

Setting $(\boldsymbol{x},\boldsymbol{x}')\equiv(\boldsymbol{x}^n,\hat{\boldsymbol{x}}^n_{1/2\overline{\sigma}})$, and for every choice of $\overline{\sigma}>\sigma$, it is a *key fact* that

$$\begin{aligned}(\boldsymbol{x}^n-\hat{\boldsymbol{x}}^n_{1/2\overline{\sigma}})^{\boldsymbol{T}}\nabla\phi_{\mu}(\boldsymbol{x}^n)&\ge\phi(\boldsymbol{x}^n)-\phi(\hat{\boldsymbol{x}}^n_{1/2\overline{\sigma}})-\sigma\|\boldsymbol{x}^n-\hat{\boldsymbol{x}}^n_{1/2\overline{\sigma}}\|_2^2-\Sigma^o\mu(\mu^{\varepsilon}+c)\\&\equiv\phi(\boldsymbol{x}^n)+\overline{\sigma}\|\boldsymbol{x}^n-\boldsymbol{x}^n\|_2^2-(\phi(\hat{\boldsymbol{x}}^n_{1/2\overline{\sigma}})+\overline{\sigma}\|\boldsymbol{x}^n-\hat{\boldsymbol{x}}^n_{1/2\overline{\sigma}}\|_2^2)\\&\quad+(\overline{\sigma}-\sigma)\|\boldsymbol{x}^n-\hat{\boldsymbol{x}}^n_{1/2\overline{\sigma}}\|_2^2-\Sigma^o\mu(\mu^{\varepsilon}+c)\\&\ge2(\overline{\sigma}-\sigma)\|\boldsymbol{x}^n-\hat{\boldsymbol{x}}^n_{1/2\overline{\sigma}}\|_2^2-\Sigma^o\mu(\mu^{\varepsilon}+c)\\&\equiv\frac{(\overline{\sigma}-\sigma)}{2\overline{\sigma}^2}\|\nabla\phi^{1/2\overline{\sigma}}(\boldsymbol{x}^n)\|_2^2-\Sigma^o\mu(\mu^{\varepsilon}+c),\end{aligned}$$

where in the second inequality we have used the fact that the function $\phi(\cdot)+\overline{\sigma}\|(\cdot)-\boldsymbol{x}^n\|_2^2$ is $(\overline{\sigma}-\sigma)$-strongly convex and minimized at the proximal point $\hat{\boldsymbol{x}}^n_{1/2\overline{\sigma}}$, and in the second equivalence we have used the representation of the gradient of the Moreau envelope $\phi^{1/2\overline{\sigma}}$ at $\boldsymbol{x}^n$, due to weak convexity of ϕ. Consequently, by definition of the Moreau envelope and Lemma 6.7 it follows that $\phi^{1/2\overline{\sigma}}(\boldsymbol{x}^n)-\phi^*\ge0$ is uniformly bounded in expectation relative to n and that it satisfies the recursion

$$\begin{aligned}&\mathbb{E}_{\mathscr{D}^n}\{\phi^{1/2\overline{\sigma}}(\boldsymbol{x}^{n+1})-\phi^*\}\\&\le\left[1+\Sigma_p^4c^2\left(\frac{\alpha_n^2}{\beta_n}+\frac{\alpha_n^2}{\gamma_n}\mathbb{1}_{\{p>1\}}\right)\right](\phi^{1/2\overline{\sigma}}(\boldsymbol{x}^n)-\phi^*)+\overline{\sigma}\Sigma_p^1\alpha_n^2\end{aligned}$$

$$- \alpha_n \frac{\overline{\sigma} - \sigma}{\overline{\sigma}} \|\nabla \phi^{1/2\overline{\sigma}}(\boldsymbol{x}^n)\|_2^2 + \overline{\sigma}\beta_n |y^n - s_\mu(\boldsymbol{x}^n)|^2 + \overline{\sigma}\gamma_n |z^n - g_\mu(\boldsymbol{x}^n, y^n)|^2 \mathbb{1}_{\{p>1\}}$$
$$+ 2\alpha_n \overline{\sigma} \Sigma^o \mu(\mu^\varepsilon + c),$$

and the proof may be completed by following essentially the same procedure as in the proof of Theorem 6.9 (excluding the very last step), with the additional need for "dragging" the bias term $2\overline{\sigma}\Sigma^o\mu(\mu^\varepsilon + c)$ through all subsequent arguments. □

It is interesting to see that the rate is of the running average type (i.e., does not require the knowledge of a finite iteration horizon *a priori*), and of exactly the same order as in the convex case (Theorem 6.9). This is expected and in perfect agreement with the results reported in [11].

6.4.3. Strongly Convex Objective and Surrogate. Lastly, we assume that *both* ϕ *and* ϕ_μ are strongly convex (this happens in particular whenever the cost $F(\cdot, \boldsymbol{W})$ is strongly convex; see Lemma 5.2). In this case, we can formulate the next result, significantly improving upon Theorem 6.9, whenever subharmonic stepsizes are used. Hereafter, let us also define $\boldsymbol{x}^* \triangleq \arg\min_{\boldsymbol{x}\in\mathcal{X}} \phi(\boldsymbol{x})$, which, by strong convexity, is of course unique.

THEOREM 6.11 (**Rate | Strongly Convex Objective/Surrogate | Subharmonic Stepsizes).** *Let Assumption 6.1 be in effect, set* $\alpha_0 \equiv \sigma^{-1}$ *and* $\beta_0 \equiv \gamma_0 \equiv 1$, *and for* $n \in \mathbb{N}^+$, *choose* $\alpha_n \equiv (\sigma n)^{-1}$, $\beta_n \equiv n^{-\tau_2}$ *and* $\gamma_n \equiv n^{-\tau_3}$, *where, if* $p \equiv 1$, $\tau_2 \equiv 2/3$, *whereas if* $p > 1$, *and for fixed* $\epsilon \in [0, 1)$, *and* $\delta \in (0, 1)$,

$$\tau_2 \equiv (3 + \epsilon)/4 \quad \textit{and} \quad \tau_3 \equiv (1 + \delta\epsilon)/2.$$

Also define the quantity $n_o(\tau_2) \triangleq \lceil (1 - \tau_2^{1/(\tau_2+1)})^{-1} \rceil \in \mathbb{N}^3$. *Additionally, suppose that both* ϕ *and* ϕ_μ *are* σ*-strongly convex, for some* $0 < \mu \le \mu_o$. *Then. for every* $n \in \mathbb{N}^{n_o(\tau_2)}$, *the Free-MESSAGE*^{p} *algorithm satisfies*

$$\begin{aligned} &\mathbb{E}\{\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\|_2^2\} \\ &\le \Sigma_p^\sigma \times \begin{Bmatrix} (n_o(\tau_2) + 3) n^{-2/3}, & \textit{if } p \equiv 1 \\ (n_o(\tau_2) + 2(1-\epsilon)^{-1}) n^{-(1-\epsilon)/2}, & \textit{if } p \in (1, 2] \end{Bmatrix} + \frac{2\Sigma^o \mu(\mu^\varepsilon + c)}{\sigma}, \end{aligned} \tag{6.9}$$

where $\Sigma_p^\sigma \in (0, \infty)$ *is increasing and bounded in* μ, *and if* $\sigma \ge 1$, $\Sigma_p^\sigma \le \Sigma_p / \sigma^2 < \infty$.

Proof of Theorem 6.11. We focus on the case where $p \in (1, 2]$; when $p \equiv 1$, the steps to the proof of the theorem are similar. By strong convexity of ϕ_μ, and specifically the fact that (recall that $\phi_\mu^o \equiv \inf_{\boldsymbol{x}\in\mathcal{X}} \phi_\mu(\boldsymbol{x})$)

$$\phi_\mu(\boldsymbol{x}) - \phi_\mu^o \ge \sigma \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2, \quad \forall \boldsymbol{x} \in \mathcal{X},$$

Lemma 6.7 once again implies that

$$\begin{aligned} \mathbb{E}\{\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^o\|_2^2\} &\le (1 - \sigma\alpha_n) \mathbb{E}\{\|\boldsymbol{x}^n - \boldsymbol{x}^o\|_2^2\} + \Sigma_p^1 \alpha_n^2 \\ &\quad + \frac{\Sigma_p^4 c^2}{\sigma} \alpha_n \left(\mathbb{E}\{|y^n - s_\mu(\boldsymbol{x}^n)|^2\} + \mathbb{E}\{|z^n - g_\mu(\boldsymbol{x}^n, y^n)|^2\}\right), \end{aligned}$$

where we recall that $\boldsymbol{x}^o \triangleq \arg\min_{\boldsymbol{x}\in\mathcal{X}} \phi_\mu(\boldsymbol{x})$.

Observe that, by our assumptions (in particular, Condition **C5**), in addition to the constants Σ_p^2, Σ_p^3 and Σ_p^4 involved in Lemmata 6.5, 6.6 and 6.7 being bounded and increasing in $\mu \in (0, \mu_o]$, the average errors $\mathbb{E}\{|y^n - s_\mu(\boldsymbol{x}^n)|^2\}$ and $\mathbb{E}\{|z^n - g_\mu(\boldsymbol{x}^n, y^n)|^2\}$ are *both* uniformly bounded relative to $n \in \mathbb{N}$ *and* $\sigma > 0$ *and* $\mu \in (0, \mu_o]$, and increasing relative to the latter, as well (uniform boundedness of the term $\mathbb{E}\{|z^n - g_\mu(\boldsymbol{x}^n, y^n)|^2\}$ may be shown along the lines of ([46], arXiv version, Proof of Lemma 2.3(c))). Additionally, we may show that $\mathbb{E}\{\|\boldsymbol{x}^n - \boldsymbol{x}^o\|_2^2\}$ is also uniformly bounded relative to $n \in \mathbb{N}^+$ and increasing and bounded in $\mu \in (0, \mu_o]$, given our choice of $\alpha_0 \equiv \sigma^{-1}$. Indeed, there is another constant $\Sigma_p^5 < \infty$, increasing and bounded in μ and independent of σ, such that

$$\mathbb{E}\{\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^o\|_2^2\} \leq (1 - \sigma\alpha_n)\,\mathbb{E}\{\|\boldsymbol{x}^n - \boldsymbol{x}^o\|_2^2\} + \Sigma_p^1\alpha_n^2 + \Sigma_p^5 c^2 \frac{\alpha_n}{\sigma},$$

for all $n \in \mathbb{N}$. By using the same inductive argument as in ([23], Section 4.4, last part of proof of Lemma 9), and by noting that

$$\begin{aligned}\mathbb{E}\{\|\boldsymbol{x}^1 - \boldsymbol{x}^o\|_2^2\} &\leq (1 - \sigma\alpha_0)\,\mathbb{E}\{\|\boldsymbol{x}^0 - \boldsymbol{x}^o\|_2^2\} + \Sigma_p^1\alpha_0^2 + \Sigma_p^5 c^2 \frac{\alpha_0}{\sigma} \\ &\equiv \Sigma_p^1\sigma^{-2} + \Sigma_p^5 c^2\sigma^{-2},\end{aligned}$$

where the right-hand side is increasing and bounded in μ, it easily follows that

$$\sup_{n\in\mathbb{N}^+} \mathbb{E}\{\|\boldsymbol{x}^n - \boldsymbol{x}^o\|_2^2\} \leq \Sigma_p^1\sigma^{-2} + \Sigma_p^5 c^2\sigma^{-2}. \tag{6.10}$$

Now, by another closer inspection of ([23], Section 4.4, Lemma 9, Theorem 5 and the respective proofs), it follows that for $\mu \in (0, \mu_o]$ and for every $n \in \mathbb{N}^{n_o(\tau_2)} \subseteq \mathbb{N}^3$,

$$\mathbb{E}\{\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^o\|_2^2\} \leq \overline{\Sigma}_p^\sigma(n_o(\tau_2) + 2\,(1-\epsilon)^{-1})n^{-(1-\epsilon)/2},$$

for a problem dependent constant $\overline{\Sigma}_p^\sigma < \infty$, which, in case $\sigma \geq 1$, may be bounded as $\overline{\Sigma}_p^\sigma \leq \overline{\Sigma}_p/\sigma^2$, for some other constant $\overline{\Sigma}_p$ (independent of σ). The constant $\overline{\Sigma}_p^\sigma$ is also increasing and bounded in μ, since it is dependent only on Σ_p^1, Σ_p^2, Σ_p^3 and Σ_p^4, as well as the uniform bounds of $\mathbb{E}\{|y^n - s_\mu(\boldsymbol{x}^n)|^2\}$, $\mathbb{E}\{|z^n - g_\mu(\boldsymbol{x}^n, y^n)|^2\}$, and $\mathbb{E}\{\|\boldsymbol{x}^n - \boldsymbol{x}^o\|_2^2\}$. Finally, we may exploit Lemma 5.2, and the fact that

$$\phi\,(\boldsymbol{x}) - \phi^* \geq \sigma\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2, \quad \forall \boldsymbol{x} \in \mathcal{X},$$

which of course follows by strong convexity of ϕ, to obtain

$$\begin{aligned}\mathbb{E}\{\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\|_2^2\} &\leq 2\mathbb{E}\{\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^o\|_2^2\} + 2\|\boldsymbol{x}^o - \boldsymbol{x}^*\|_2^2 \\ &\leq 2\mathbb{E}\{\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^o\|_2^2\} + 2\frac{1}{\sigma}(\phi\,(\boldsymbol{x}^o) - \phi^*) \\ &\leq \Sigma_p^\sigma(n_o(\tau_2) + 2\,(1-\epsilon)^{-1})n^{-(1-\epsilon)/2} + \frac{2\Sigma^o\mu(\mu^\varepsilon + c)}{\sigma},\end{aligned} \tag{6.11}$$

being true for all $n \in \mathbb{N}^{n_o(\tau_2)}$, where $\Sigma_p^\sigma \triangleq 2\overline{\Sigma}_p^\sigma$. ∎

We also provide a rate result for constant stepsize selection, very popular and reasonable in practical considerations. This is useful in particular when the distribution of $\boldsymbol{W}$ changes during the operation of the algorithm, and the goal is to make the *Free-MESSAGE*^{p} algorithm *adaptive* to such changes.

THEOREM 6.12 (**Rate | Strongly Convex Objective/Surrogate | Constant Stepsizes).** *Let Assumption 6.1 be in effect and, for $n \in \mathbb{N}^+$, choose the stepsizes as $\alpha_n \equiv \alpha\sigma^{-1}, \alpha \in (0,1)$, $\beta_n \equiv \beta \in (0,1]$ and $\gamma_n \equiv \gamma \in (0,1]$, such that $\alpha < \min\{\beta, \gamma\}$. Additionally, suppose that both ϕ and ϕ_μ are σ-strongly convex, for some $0 < \mu \le \mu_o$. Then, for every $n \in \mathbb{N}^+$, the Free-MESSAGE^p algorithm satisfies*

$$\begin{aligned}(6.12)\qquad &\mathbb{E}\{\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\|_2^2\} \\ &\le (1-\alpha)^n\left(2\|\boldsymbol{x}^0 - \boldsymbol{x}^o\|_2^2 + \frac{\widehat{\Sigma}_p^1}{\sigma^2}\right) + \widehat{\Sigma}_p^\sigma \mathcal{H}(\alpha,\beta,\gamma) + \frac{2\Sigma^o\mu(\mu^\varepsilon + c)}{\sigma},\end{aligned}$$

where $\widehat{\Sigma}_p^1 \in (0,\infty)$ is independent of σ, $\widehat{\Sigma}_p^\sigma \in (0,\infty)$ is such that if $\sigma \ge 1$, $\widehat{\Sigma}_p^\sigma \le \widehat{\Sigma}_p^0/\sigma^2 < \infty$, both are increasing and bounded in μ, and $\mathcal{H}(\alpha,\beta,\gamma) \triangleq \alpha + \beta + \alpha^2\beta^{-2} + (\gamma + \alpha^2\gamma^{-2} + \beta^2\gamma^{-2})\mathbb{1}_{\{p\in(1,2]\}}$.

Proof of Theorem 6.12. Once more, we explicitly present the proof whenever $p \in (1,2]$. Let $J_s^n \triangleq \mathbb{E}\{|y^n - s_\mu(\boldsymbol{x}^n)|^2\}$, $J_g^n \triangleq \mathbb{E}\{|z^n - g_\mu(\boldsymbol{x}^n, y^n)|^2\}$, and $J_o^n \triangleq \mathbb{E}\{\|\boldsymbol{x}^n - \boldsymbol{x}^o\|_2^2\}$, $n \in \mathbb{N}$, and for nonnegative sequences $\{H_s^n\}_{n\in\mathbb{N}}$ and $\{H_g^n\}_{n\in\mathbb{N}}$, define

$$J^n \triangleq J_o^n + H_s^{n-1}J_s^{n-1} + H_g^{n-1}J_g^{n-1}, \quad n \in \mathbb{N}^+.$$

Then, by our assumptions, and from ([23], Section 4.4, Lemma 9), it follows that $\{H_s^n\}_{n\in\mathbb{N}}$ and $\{H_g^n\}_{n\in\mathbb{N}}$ may be chosen in a way such that, for every $n \in \mathbb{N}^+$,

$$J^{n+1} \le (1-\alpha)J^n + \widetilde{\Sigma}_p^\sigma\left(\alpha^2 + \frac{\alpha^3}{\beta^2} + \alpha\beta + \frac{\alpha^3}{\gamma^2} + \frac{\alpha\beta^2}{\gamma^2} + \alpha\gamma\right).$$

where $0 < \widetilde{\Sigma}_p^\sigma < \infty$ is increasing and bounded in μ. Proceeding inductively, we have

$$\begin{aligned}J^{n+1} &\le (1-\alpha)J^n + \widetilde{\Sigma}_p^\sigma\left(\alpha^2 + \frac{\alpha^3}{\beta^2} + \alpha\beta + \frac{\alpha^3}{\gamma^2} + \frac{\alpha\beta^2}{\gamma^2} + \alpha\gamma\right) \\ &\le (1-\alpha)^n J^1 + \widetilde{\Sigma}_p^\sigma\left(\alpha^2 + \frac{\alpha^3}{\beta^2} + \alpha\beta + \frac{\alpha^3}{\gamma^2} + \frac{\alpha\beta^2}{\gamma^2} + \alpha\gamma\right)\sum_{i\in\mathbb{N}_{n-1}}(1-\alpha)^i \\ &\equiv (1-\alpha)^n J^1 + \widetilde{\Sigma}_p^\sigma\left(\alpha^2 + \frac{\alpha^3}{\beta^2} + \alpha\beta + \frac{\alpha^3}{\gamma^2} + \frac{\alpha\beta^2}{\gamma^2} + \alpha\gamma\right)\frac{1-(1-\alpha)^n}{\alpha} \\ &\le (1-\alpha)^n J^1 + \widetilde{\Sigma}_p^\sigma\left(\alpha + \frac{\alpha^2}{\beta^2} + \beta + \frac{\alpha^2}{\gamma^2} + \frac{\beta^2}{\gamma^2} + \gamma\right).\end{aligned}$$

Now, again from ([23], Section 4.4, Lemma 9 *and* its proof), and as in Theorem 6.11, it follows that, whenever $\sigma \ge 1$, $\widetilde{\Sigma}_p^\sigma \le \widetilde{\Sigma}_p^0/\sigma^2$, for some $\widetilde{\Sigma}_p^0 < \infty$, and the same type of argument holds for H_s^0 and H_g^0, as well, but for all $\sigma > 0$. Therefore, it is true that

$$\begin{aligned}J^1 &\equiv J_o^1 + H_s^0J_s^0 + H_g^0J_g^0 \\ &\le (1-\alpha)\,J_o^0 + \Sigma_p^1\frac{\alpha^2}{\sigma^2} + c^2\Sigma_p^5\frac{\alpha}{\sigma^2} + H_s^0J_s^0 + H_g^0J_g^0 \le J_o^0 + \frac{\widetilde{\Sigma}_p^1}{\sigma^2},\end{aligned}$$

where $0 < \widetilde{\Sigma}_p^1 < \infty$ is independent of σ, and increasing and bounded in μ. As a result, we get

$$J_o^{n+1} \le J^{n+1} \le (1-\alpha)^n\left(J_o^0 + \frac{\widetilde{\Sigma}_p^1}{\sigma^2}\right) + \widetilde{\Sigma}_p^\sigma\left(\alpha + \beta + \gamma + \frac{\alpha^2}{\beta^2} + \frac{\alpha^2}{\gamma^2} + \frac{\beta^2}{\gamma^2}\right),$$

being true for all $n \in \mathbb{N}^+$. Finally, using the same argument as in (6.11), we get

$$\mathbb{E}\{\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\|_2^2\} \le (1-\alpha)^n \left(2\|\boldsymbol{x}^0 - \boldsymbol{x}^o\|_2^2 + \frac{\widehat{\Sigma}_p^1}{\sigma^2}\right) + \widehat{\Sigma}_p^\sigma\left(\alpha + \beta + \gamma + \frac{\alpha^2}{\beta^2} + \frac{\alpha^2}{\gamma^2} + \frac{\beta^2}{\gamma^2}\right) + \frac{2\Sigma^o \mu(\mu^\varepsilon + c)}{\sigma},$$

for every $n \in \mathbb{N}^+$, where $\widehat{\Sigma}_p^1 \triangleq 2\widetilde{\Sigma}_p^1$, $\widehat{\Sigma}_p^\sigma \triangleq 2\widetilde{\Sigma}_p^\sigma$ and, whenever $\sigma \ge 1$, $\widehat{\Sigma}_p^\sigma \le \widehat{\Sigma}_p^0/\sigma^2 \triangleq 2\widetilde{\Sigma}_p^0/\sigma^2$. The proof is now complete. ∎

6.5. Discussion. First, we comment on the role of $\epsilon \in [0,1)$ on the rates of Theorems 6.9, 6.10 and 6.11. For $\epsilon \equiv 0$, the rates are of the orders of $\mathcal{O}(n^{-1/(4\mathbb{1}_{\{p\in(1,2]\}}+4)} + \mu)$ (roughly) and $\mathcal{O}(n^{-1/2} + \mu)$, as $\mu \to 0$, respectively, the latter when $p \in (1,2]$. However, if $\epsilon \equiv 0$, the resulting stepsizes do *not* satisfy the conditions of Theorem 6.8, and path convergence of the *Free-MESSAGE*^{p} algorithm is not guaranteed, at least for the case of a convex cost (see also [23]). Nevertheless, if $\epsilon \in (0,1)$, rates *arbitrarily close* to the ones above can be achieved, while path convergence is simultaneously guaranteed, ensuring better algorithmic stability.

We may also finalize all rate results developed in Theorems 6.9, 6.10, 6.11 and 6.12 by explicitly choosing μ appropriately in each case, as follows:

- *Convex and weakly convex case (subharmonic stepsizes, Theorems 6.9 and 6.10)*: Assuming a fixed iteration horizon $T \in \mathbb{N}^+$, a compact feasible set (for simplicity), and relative to the appropriate figure of merit, choosing $\mu \equiv \mathcal{O}(T^{-1/(4\mathbb{1}_{\{p\in(1,2]\}}+4)})$ results in a rate of the order of $\mathcal{O}(T^{-(1-\epsilon)/(4\mathbb{1}_{\{p\in(1,2]\}}+4)})$, for every $\epsilon \in [0,1)$. Regarding stepsize selection, we may simply set $\delta \equiv \zeta \equiv 1/2$ (where applicable).
- *Strongly convex case with subharmonic stepsizes (Theorem 6.11)*: Again, we assume a fixed iteration horizon $T \in \mathbb{N}^{n_o(\tau_2)}$ (i.e., sufficiently large). For $p \equiv 1$, choosing $\mu \equiv \mathcal{O}(T^{-2/3})$ results in a rate of the order of $\mathcal{O}(T^{-2/3})$. For $p \in (1,2]$, the choice $\mu \equiv \mathcal{O}(T^{-1/2})$ gives a rate of the order of $\mathcal{O}(T^{-(1-\epsilon)/2})$, for every $\epsilon \in [0,1)$. Again, the stepsize choice $\delta \equiv \zeta \equiv 1/2$ works fine, as above.
- *Strongly convex case with constant stepsizes (Theorem 6.12)*: For $p \equiv 1$, choosing $\beta \in (0,1)$, $\alpha \equiv \beta^{3/2}$ and $\mu \equiv \mathcal{O}(\beta)$ (as $\beta \downarrow 0$) results in the bound

$$\mathbb{E}\{\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\|_2^2\} \le \mathcal{O}\big((1-\beta^{3/2})^n + \beta\big), \quad \text{as } \gamma \to 0, \quad \forall n \in \mathbb{N}^+.$$

  Lastly, for $p \in (1,2]$, we may choose $\gamma \in (0,1)$, $\beta \equiv \gamma^{3/2}$, $\alpha \equiv \gamma^{9/4}$ and $\mu \equiv \mathcal{O}(\gamma)$ (as $\gamma \downarrow 0$); in this case, we obtain the bound

$$\mathbb{E}\{\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^*\|_2^2\} \le \mathcal{O}\big((1-\gamma^{9/4})^n + \gamma\big), \quad \text{as } \gamma \to 0, \quad \forall n \in \mathbb{N}^+.$$

  Observe that these bounds establish *noisy linear convergence of Free-MESSAGE*^{p} within a neighborhood around the solution of the base problem and of predictable diameter, and are very similar (though slower) to well-known bounds for the standard, risk-neutral stochastic gradient algorithm; also see Section 7 for a numerical demonstration of this result.

Further, we would like to emphasize the explicit dependence on σ on both terms appearing on the right of (6.9) and (6.12), implying that *strong convexity benefits both algorithmic and smoothing stability.* More generally, all rates in (6.8), (6.9) and (6.12) present certain tradeoffs among μ, σ and N. In particular, the dependence on N appears of both terms on the right of (6.8), (6.9) and (6.12), and varies relative to the associated (D, T)-pair.

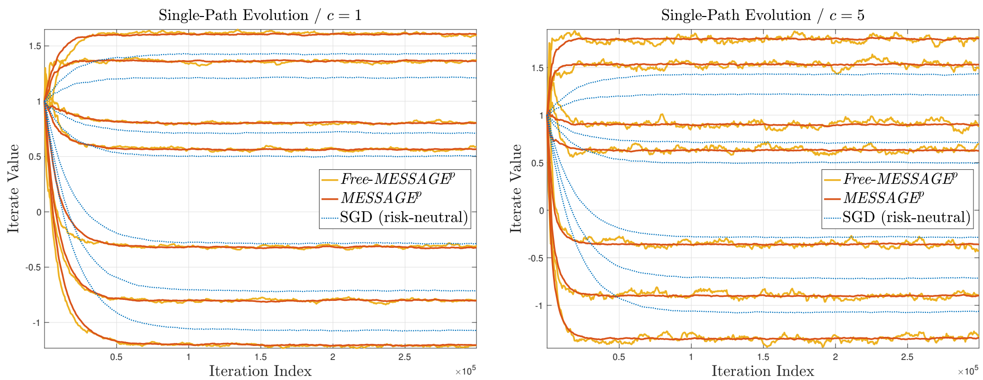

FIGURE 7.1. *Single-path evolution of Free-MESSAGE^p, MESSAGE^p and Stochastic Gradient Descent (SGD) for the problem setting considered in Section 7, and for two values of c (left, right).*

7. Numerical Simulations. Here, we evaluate the empirical performance of the *Free-MESSAGE^p* algorithm on a synthetic numerical setting, and compare its practical performance to that of the fully gradient-based *MESSAGE^p* algorithm of [23]. For our evaluation, we consider a regularized, strongly convex, linear-quadratic risk regression cost defined as

$$F_\sigma(\boldsymbol{x}, \boldsymbol{W}) \triangleq \frac{1}{2}(y - \langle \boldsymbol{h}, \boldsymbol{x}\rangle)^2 + \frac{\sigma}{2}\|\boldsymbol{x}\|_2^2, \quad \boldsymbol{W} \triangleq (\boldsymbol{h}, y),$$

where $y \equiv \langle \boldsymbol{h}, \boldsymbol{x}_o\rangle \in \mathbb{R}$ for a constant $\boldsymbol{x}_o \equiv [-0.4, -1, 1.7, 0.7, 2, -1.5, 1]^{\boldsymbol{T}} \in \mathbb{R}^7$ and with the elements of $\boldsymbol{h} \in \mathbb{R}^7$ being independent Gaussian with zero mean and variance 0.5^2, and $\sigma \equiv 0.1$. Then, by choosing $p \equiv 2$ and $\mathcal{R}(\cdot) \equiv (\cdot)_+ + 1/2$ (i.e., $\eta \equiv 1/2$), the resulting learning problem (cf. (1.1)) may be expressed as

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{2}\left[\mathbb{E}\{(y - \langle \boldsymbol{h}, \boldsymbol{x}\rangle)^2\} + c\sqrt{\mathbb{E}\{(((y - \langle \boldsymbol{h}, \boldsymbol{x}\rangle)^2 - \mathbb{E}\{(y - \langle \boldsymbol{h}, \boldsymbol{x}\rangle)^2\})_+ + 1)^2\}} + \sigma\|\boldsymbol{x}\|_2^2\right],$$

which clearly constitutes an instance of a *risk-aware ridge regression task*. Of course, if $c \equiv 0$, we recover standard (risk-neutral) σ-regularized ridge regression.

To test the effectiveness of *Free-MESSAGE^p*, we execute it concurrently with its gradient-based sibling *MESSAGE^p* [23] with identical constant stepsizes selected as

$$\gamma \equiv 0.02, \quad \beta \equiv \gamma^{3/2} \equiv 0.0028, \quad \alpha \equiv \gamma^{9/4} \approx 0.00015,$$

and with $\mu \equiv 0.001$, in line with our discussion in Section 6.5, and over a *single* datastream comprised of $2T \equiv 6 \times 10^5$ IID learning example realizations, $\{(\boldsymbol{h}^n, y^n)\}_{n\in\mathbb{N}_{2T}}$, taken *sequentially in pairs*. In other words, both algorithms are executed for T iterations, and on exactly the same dataset. Also, we choose $\mathcal{X} \equiv [-20, 20]^7$, but we arbitrarily set $\mathcal{Y} \equiv \mathbb{R}$ and $\mathcal{Z} \equiv \mathbb{R}$, reflecting the fact that the exact values of the parameters required for rigorously defining $\mathcal{Y}$ (see Assumption 6.1, condition **C5**) are typically unknown in applications (and which would be the case in our ridge regression example), and to better evaluate iterate stability of both algorithms in practice. Under this setting, observe that *Free-MESSAGE^p* is implemented in a completely gradient-free and parameter-free fashion.

Fig. 7.1 shows the evolution of all seven entries of the regressor process $\{\boldsymbol{x}^n\}_{n\in\mathbb{N}_T}$, for both algorithms considered and for two values of c, namely, $c \equiv 1$ (left) and $c \equiv 5$ (right). Recall that if $c \in [0, 1]$, then the risk-aware ridge regression problem at hand

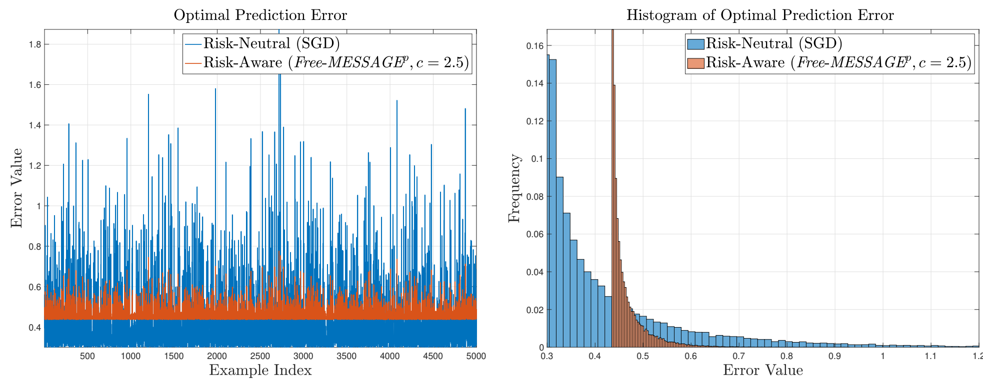

FIGURE 7.2. *Prediction (test) errors achieved by Free-MESSAGE^p and SGD, for the problem setting considered in Section 7 with $c \equiv 2.5$ (left: error values, right: corresponding histogram).*

is strongly convex. Also, although convexity is not guaranteed if $c > 1$, in such a case the problem is expected to be at least weakly convex (at least approximately); this is due to the smoothness of F_σ and the Gaussian smoothing involved in the construction of the surrogate ϕ_μ (in the case of *Free-MESSAGE^p*, pertaining to our analysis). In the figure, we also provide comparatively the paths generated by standard Stochastic Gradient Descent (SGD) with constant stepsize α, to highlight the substantial difference in the solutions of the learning problem between the risk-neutral (i.e., $c \equiv 0$) and risk-aware settings, respectively.

For both values of c, Fig. 7.1 clearly demonstrates that *Free-MESSAGE^p* closely mimics *MESSAGE^p*, and that both algorithms converge to a stable neighborhood of the optimal regressor (hopefully for $c \equiv 5$) at an *identical* linear rate. In particular, for $c \equiv 1$ (where strong convexity is ensured), this behavior is in agreement with our theoretical results. We further observe that the price paid for the lack of first-order information is noisier and more sensitive zeroth-order quasigradients, as indicated by the larger fluctuations of the iterates generated by *Free-MESSAGE^p*. Those fluctuations increase when $c \equiv 5$; this is expected, because the magnitude of the quasigradients of *Free-MESSAGE^p* is proportional to c. Still, we clearly observe that *Free-MESSAGE^p* exhibits very consistent behavior as compared with its gradient-based counterpart.

At this point, we would also like to note that in worse-conditioned problems than our indicative example where two-sample-based zeroth-order quasigradients might be too noisy, one can reformulate *Free-MESSAGE^p* in an almost straightforward way by incorporating *Gaussian minibatching* for quasigradient stabilization (see, e.g., [16], Section 5), and with little additional effort in the corresponding convergence analysis. However, minibatching comes at the expense of additional sampling requirements.

Additionally, *Free-MESSAGE^p* is favorably comparable to *MESSAGE^p* in terms of computational requirements. While the throughput of the two algorithms is exactly the same ($1:2$, since each iteration requires two learning examples), the complexity per iteration of *Free-MESSAGE^p* is expected to be smaller, since *Free-MESSAGE^p* relies only on four function evaluations and elementary vector (*not* matrix) operations. This holds under the reasonable assumption that that full gradient evaluations are generally more complex than evaluations of cost function values. The only additional computational requirement of *Free-MESSAGE^p* over *MESSAGE^p* is that of a Gaussian sampler, which is really rather trivial for most practical considerations.

Lastly, the effects of the solutions achieved by *Free-MESSAGE^p* (or *MESSAGE^p*) and SGD on the resulting optimal prediction errors are shown in Fig. 7.2 (for $c \equiv 2.5$).

The premise of risk-aware statistical learning is to effectively *control the statistical dispersion of the random cost* associated with a particular learning task. In statistical regression, this translates to a desire to ensure *optimal prediction error stability*, also reasonably trading with keeping as small mean prediction error as possible; this is exactly what Fig. 7.2 illustrates for our risk-aware regression example. We observe that the reduction in the volatility of the instantaneous prediction errors achieved by the risk-aware solution is rather drastic as compared with the risk-neutral solution (left), also translating to a much tighter corresponding empirical distribution (right). Of course, the price to be paid for an optimal risk-aware regressor is a higher average regression cost; this is natural and expected, since the ultimately minimum average cost is achieved by the risk-neutral solution, which is recovered by setting $c \equiv 0$.

Remark 7.1. The fact that convergence of *(Free-)MESSAGE*^{p} appears to be faster than that of stochastic gradient descent in Fig. 7.1 does not of course imply that risk-aware ridge regression is in general simpler and/or easier than risk-neutral ridge regression, as the two problems are structurally very different. In fact, the opposite is most probably true, especially for higher-dimensional problems. Also, the convergence rate achieved by stochastic gradient descent for our ridge regression example can be significantly and stably accelerated by using a more aggressive stepsize. □

8. Future Work. There are several interesting topics for future work, building on the results presented in this paper; indicatively, we discuss some. First, although our rate results quantify explicitly the dependence on μ and σ, we have not paid much attention to the decision dimension, N. Indeed, if $c \equiv 0$, then, orderwise relative to N, our bounds are equivalent to those in [30], known to be order-suboptimal (see, e.g., [12]). Therefore, it would be of interest to see if order improvement relative to N is possible, by potentially exploiting ideas from more ingenious methods for risk-neutral zeroth-order optimization, such as those with diminishing μ, multi-point finite differences, and/or minibatching. Second, also driven by [12], another challenging topic is the development of lower complexity bounds for risk-aware learning, which would be useful in the design of optimal algorithms and, of course, as complexity benchmarks. Lastly, further relaxing the convexity of the base problem is of particular interest, as the resulting setting fits more accurately many application settings in modern artificial intelligence and deep learning.

Appendix A. Proof of Lemma 3.4. If $\mu \equiv 0$, the situation is trivial. So, for the rest of the proof, we assume that $\mu > 0$. Let $\mathcal{N} : \mathbb{R}^N \to \mathbb{R}$ be the standard Gaussian density on $\mathbb{R}^N$. We first make the observation that, for every finite $B > 0$,

$$\begin{aligned}\mathcal{N}\left(\frac{\boldsymbol{u}}{\mu}\right)\exp\left(\frac{\|\boldsymbol{u}\|_2 B}{\mu^2}\right)\max\{1,\|\boldsymbol{u}\|_2\} &\propto \exp\left(-\frac{\|\boldsymbol{u}\|_2^2}{2\mu^2}\right)\exp\left(\frac{\|\boldsymbol{u}\|_2 B}{\mu^2}\right)\max\{1,\|\boldsymbol{u}\|_2\}\\ &\le \exp\left(-\frac{\|\boldsymbol{u}\|_2^2}{2\mu^2}+\frac{\|\boldsymbol{u}\|_2 B}{\mu^2}+\|\boldsymbol{u}\|_2\right)\\ &\le \exp\left(-\frac{\|\boldsymbol{u}\|_2^2}{2\mu_\star^2}\right),\end{aligned}$$

provided that $\mu < \mu_\star$ and $\|\boldsymbol{u}\|_2 \ge \big(2(B+\mu^2)\mu_\star^2\big)/\big(\mu_\star^2-\mu^2\big)$. Consequently, as long as condition (3.1) is in effect, it readily follows that

$$\int \mathcal{N}\left(\frac{\boldsymbol{u}}{\mu}\right)\exp\left(\frac{\|\boldsymbol{u}\|_2 B}{\mu^2}\right)\max\{1,\|\boldsymbol{u}\|_2\}|f(\boldsymbol{u})|\mathrm{d}\boldsymbol{u} < \infty.$$

To see why this is important, recall the definition of $f_\mu(\cdot) \equiv \mathbb{E}\{f((\cdot)+\mu\boldsymbol{U})\}$, for which is must be true that

$$\begin{aligned}\mathbb{E}\{|f(\boldsymbol{x}+\mu\boldsymbol{U})|\} &\equiv \mu^{-N}\int |f(\boldsymbol{u})|\mathcal{N}\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{\mu}\right)\mathrm{d}\boldsymbol{u}\\ &\leq \mu^{-N}\int |f(\boldsymbol{u})|\mathcal{N}\left(\frac{\boldsymbol{u}}{\mu}\right)\exp\left(\frac{\|\boldsymbol{u}\|_2\|\boldsymbol{x}\|_2}{\mu^2}\right)\mathrm{d}\boldsymbol{u}<\infty,\end{aligned}$$

from where it follows that the random function $f(\boldsymbol{x}+\mu\boldsymbol{U})$ in $\mathcal{Z}_1$, for all $\boldsymbol{x}\in\mathbb{R}^N$. Equivalently, we have shown that the function $f_\mu(\cdot)\equiv\mathbb{E}\{f((\cdot)+\mu\boldsymbol{U})\}$ is well-defined and finite, everywhere on $\mathbb{R}^N$. The rest of the first part, and the second part of Lemma 3.4 may be developed along the lines of [30], where we explicitly use the identity $\mathbb{E}\{\mathsf{T}(\boldsymbol{x},\boldsymbol{U})\}\equiv 0$, for all $\boldsymbol{x}\in\mathcal{F}$, since T is a normal remainder on $\mathcal{F}$.

For the third part, the result on the existence and representation of ∇f_μ will follow by a careful application of the Dominated Convergence Theorem, which provides an *extension* of the standard Leibniz rule of Riemann integration, and permits interchangeability of differentiation and integration. Specifically, we will exploit a multidimensional version of ([13], Theorem 2.27). To this end, for $\mu>0$, define

$$\varphi(\boldsymbol{x},\boldsymbol{u})\triangleq f(\boldsymbol{u})\,\mu^{-N}\mathcal{N}\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{\mu}\right),\quad(\boldsymbol{x},\boldsymbol{u})\in\mathcal{F}\times\mathbb{R}^N.$$

By our construction, $\varphi(\boldsymbol{x},\cdot)$ is Lebesgue integrable on $\mathbb{R}^N$ for every $\boldsymbol{x}\in\mathbb{R}^N$, and $\varphi(\cdot,\boldsymbol{u})$ is differentiable everywhere on $\mathbb{R}^N$ for every $\boldsymbol{u}\in\mathbb{R}^N$, with

$$\nabla_{\boldsymbol{x}}\varphi(\boldsymbol{x},\boldsymbol{u})\equiv\mu^{-N-2}f(\boldsymbol{u})\,\mathcal{N}\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{\mu}\right)(\boldsymbol{u}-\boldsymbol{x}).$$

Now, consider any compact box $\mathcal{B}\subseteq\mathbb{R}^N$. Choosing $B\triangleq\sup_{\boldsymbol{x}\in\mathcal{B}}\|\boldsymbol{x}\|_2$ and for every $\boldsymbol{u}\in\mathbb{R}^N$, we may write

$$\begin{aligned}\|\nabla_{\boldsymbol{x}}\varphi(\boldsymbol{x},\boldsymbol{u})\|_2 &\leq \mu^{-N-2}\mathcal{N}\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{\mu}\right)|f(\boldsymbol{u})|\left(\|\boldsymbol{u}\|_2+\|\boldsymbol{x}\|_2\right)\\ &\leq \mu^{-N-2}(1+B)\,\mathcal{N}\left(\frac{\boldsymbol{u}}{\mu}\right)\exp\left(\frac{\|\boldsymbol{u}\|_2 B}{\mu^2}\right)\max\{1,\|\boldsymbol{u}\|_2\}|f(\boldsymbol{u})|\triangleq\psi_{\mathcal{B}}(\boldsymbol{u}).\end{aligned}$$

Note that the use of the ℓ_2-norm is arbitrary; any (equivalent) vector norm works. The analysis in the beginning of the proof implies that $\psi_{\mathcal{B}}$ has a finite Lebesgue integral on $\mathbb{R}^N$. Therefore, it is true that $\sup_{\boldsymbol{x}\in\mathcal{B}}\|\nabla_{\boldsymbol{x}}\varphi(\boldsymbol{x},\cdot)\|_2\leq\psi_{\mathcal{B}}(\cdot)\in\mathcal{L}_1\big(\mathbb{R}^N,\mathscr{B}\big(\mathbb{R}^N\big),\lambda;\mathbb{R}\big)$, where $\lambda:\mathscr{B}\big(\mathbb{R}^M\big)\to\mathbb{R}_+$ denotes the corresponding Lebesgue measure. It then follows that the function $f_\mu(\cdot)\equiv\int\varphi(\cdot,\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}$ is differentiable on $\mathcal{B}$, and that

$$\begin{aligned}\nabla f_\mu(\boldsymbol{x}) &\equiv \int\nabla_{\boldsymbol{x}}\varphi(\boldsymbol{x},\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}\\ &= \int\mu^{-1}f(\boldsymbol{x}+\mu\boldsymbol{u})\,\mathcal{N}(\boldsymbol{u})\,\boldsymbol{u}\mathrm{d}\boldsymbol{u}-\int\mu^{-1}f(\boldsymbol{x})\,\mathcal{N}(\boldsymbol{u})\,\boldsymbol{u}\mathrm{d}\boldsymbol{u}\\ &\equiv \int\frac{f(\boldsymbol{x}+\mu\boldsymbol{u})-f(\boldsymbol{x})}{\mu}\boldsymbol{u}\mathcal{N}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u},\end{aligned}$$

for every $\boldsymbol{x} \in \mathcal{B}$ (Theorem 2.27 in [13]). But the box $\mathcal{B}$ is arbitrary, and any $\boldsymbol{x} \in \mathbb{R}^N$ is contained in a compact box. For the rest of the third part of Lemma 3.4, if f is $(L, \mathsf{D}, \mathsf{T})$-SLipschitz on $\mathcal{F}$, we may write

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\frac{f(\boldsymbol{x}+\mu\boldsymbol{U})-f(\boldsymbol{x})}{\mu}\boldsymbol{U}\right\|_2^2\right\} \\
&\quad\equiv \frac{1}{\mu^2}\mathbb{E}\left\{|f(\boldsymbol{x}+\mu\boldsymbol{U})-f(\boldsymbol{x})-\mathsf{T}(\boldsymbol{x},\mu\boldsymbol{U})+\mathsf{T}(\boldsymbol{x},\mu\boldsymbol{U})|^2\,\|\boldsymbol{U}\|_2^2\right\} \\
&\quad\le \frac{1}{\mu^2}\mathbb{E}\left\{\left(L\mathsf{D}(\mu\boldsymbol{U})+|\mathsf{T}(\boldsymbol{x},\mu\boldsymbol{U})|\right)^2\|\boldsymbol{U}\|_2^2\right\},
\end{aligned}$$

for all $\boldsymbol{x} \in \mathcal{F}$. Enough said. □

REFERENCES

[1] P. L. A. and M. Fu, *Risk-sensitive reinforcement learning: A constrained optimization viewpoint*, arXiv preprint, arXiv:1810.09126, (2018), https://arxiv.org/abs/1810.09126.

[2] S. Ahmed, U. Çakmak, and A. Shapiro, *Coherent risk measures in inventory problems*, European Journal of Operational Research, 182 (2007), pp. 226–238, https://doi.org/10.1016/j.ejor.2006.07.016.

[3] R. B. Ash and C. Doléans-Dade, *Probability and Measure Theory*, Academic Press, 2000, https://doi.org/10.2307/2291440, https://arxiv.org/abs/arXiv:1011.1669v3.

[4] K. Balasubramanian and S. Ghadimi, *Zeroth-rder (non)-convex stochastic optimization via conditional gradient and gradient updates*, in Advances in Neural Information Processing Systems, vol. 2018-Decem, 2018, pp. 3455–3464.

[5] A. S. Bedi, A. Koppel, and K. Rajawat, *Nonparametric compositional stochastic optimization*, arXiv preprint, arXiv:1902.06011, (2019), https://arxiv.org/abs/1902.06011.

[6] S. Bruno, S. Ahmed, A. Shapiro, and A. Street, *Risk neutral and risk-averse approaches to multistage renewable investment planning under uncertainty*, European Journal of Operational Research, 250 (2016), pp. 979–989, https://doi.org/10.1016/j.ejor.2015.10.013.

[7] A. R. Cardoso and H. Xu, *Risk-averse stochastic convex bandit*, in International Conference on Artificial Intelligence and Statistics, vol. 89, Apr. 2019, pp. 39–47.

[8] Y. Chen, H. Chang, J. Meng, and D. Zhang, *Ensemble neural networks (enn): A gradient-free stochastic method*, Neural Networks, 110 (2019), pp. 170–185, https://doi.org/10.1016/j.neunet.2018.11.009.

[9] Z. Chen and Y. Wang, *Two-sided coherent risk measures and their application in realistic portfolio optimization*, Journal of Banking and Finance, 32 (2008), pp. 2667–2673, https://doi.org/10.1016/j.jbankfin.2008.07.004.

[10] A. R. A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-Free Optimization*, Society for Industrial and Applied Mathematics/Mathematical Programming Society, 2009.

[11] D. Davis and D. Drusvyatskiy, *Stochastic model-based minimization of weakly convex functions*, SIAM Journal on Optimization, 29 (2019), pp. 207–239, https://doi.org/10.1137/18M1178244, https://arxiv.org/abs/1803.06523.

[12] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, *Optimal rates for zero-order convex optimization: The power of two function evaluations*, IEEE Transactions on Information Theory, 61 (2015), pp. 2788–2806, https://doi.org/10.1109/TIT.2015.2409256.

[13] G. B. Folland, *Real Analysis: Modern Techniques and their Applications*, John Wiley & Sons, 2nd ed., 1999.

[14] T. Fu, X. Zhuang, Y. Hui, and J. Liu, *Convex risk measures based on generalized lower deviation and their applications*, International Review of Financial Analysis, 52 (2017), pp. 27–37, https://doi.org/10.1016/j.irfa.2017.04.008.

[15] S. Ghadimi and G. Lan, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, SIAM Journal on Optimization, 23 (2013), pp. 2341–2368, https://doi.org/10.1137/120880811, https://arxiv.org/abs/1309.5549.

[16] S. Ghadimi, G. Lan, and H. Zhang, *Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization*, Mathematical Programming, 155 (2016), pp. 267–305, https://doi.org/10.1007/s10107-014-0846-1.

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT Press, 2016.
[18] J.-y. Gotoh and S. Uryasev, *Support vector machines based on convex risk functions and general norms*, Annals of Operations Research, 249 (2017), pp. 301–328, https://doi.org/10.1007/s10479-016-2326-x.
[19] D. Hajinezhad and M. M. Zavlanos, *Gradient-free multi-agent nonconvex nonsmooth optimization*, in Proceedings of the IEEE Conference on Decision and Control, vol. 2018-Decem, IEEE, Dec. 2019, pp. 4939–4944, https://doi.org/10.1109/CDC.2018.8619333.
[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York, New York, NY, 2009, https://doi.org/10.1007/978-0-387-84858-7.
[21] W. Huang and W. B. Haskell, *Risk-aware q-learning for markov decision processes*, in 2017 IEEE 56th Annual Conference on Decision and Control, CDC 2017, vol. 2018-Janua, IEEE, Dec. 2018, pp. 4928–4933, https://doi.org/10.1109/CDC.2017.8264388.
[22] D. R. Jiang and W. B. Powell, *Risk-averse approximate dynamic programming with quantile-based risk measures*, Mathematics of Operations Research, 43 (2018), pp. 554–579, https://doi.org/10.1287/moor.2017.0872, https://arxiv.org/abs/1509.01920.
[23] D. S. Kalogerias and W. B. Powell, *Recursive optimization of convex risk measures: Mean-semideviation models*, arXiv preprint, arXiv:1804.00636, (2018), https://arxiv.org/abs/1804.00636.
[24] S.-K. Kim, R. Thakker, and A.-a. Agha-mohammadi, *Bi-directional value learning for risk-aware planning under uncertainty*, IEEE Robotics and Automation Letters, 4 (2019), pp. 2493–2500, https://doi.org/10.1109/LRA.2019.2903259, https://arxiv.org/abs/1902.05698.
[25] W.-J. Ma, C. Oh, Y. Liu, D. Dentcheva, and M. M. Zavlanos, *Risk-averse access point selection in wireless communication networks*, IEEE Transactions on Control of Network Systems, 5870 (2018), pp. 1–1, https://doi.org/10.1109/TCNS.2018.2792309.
[26] S. Moazeni, W. B. Powell, B. Defourny, and B. Bouzaiene-Ayari, *Parallel nonstationary direct policy search for risk-averse stochastic optimization*, INFORMS Journal on Computing, 29 (2017), pp. 332–349, https://doi.org/10.1287/ijoc.2016.0733.
[27] S. Moazeni, W. B. Powell, and A. H. Hajimiragha, *Mean-conditional value-at-risk optimal energy storage operation in the presence of transaction costs*, IEEE Transactions on Power Systems, 30 (2015), pp. 1222–1232, https://doi.org/10.1109/TPWRS.2014.2341642.
[28] J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bulletin de la Société mathématique de France, 79 (1965), pp. 273–299, https://doi.org/10.24033/bsmf.1625.
[29] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley & Sons, New York, 1983.
[30] Y. Nesterov and V. Spokoiny, *Random gradient-free minimization of convex functions*, Foundations of Computational Mathematics, 17 (2017), pp. 527–566, https://doi.org/10.1007/s10208-015-9296-2.
[31] M. Norton, A. Mafusalov, and S. Uryasev, *Soft margin support vector classification as buffered probability minimization*, Journal of Machine Learning Research, 18 (2017), pp. 1–43.
[32] W. Ogryczak and A. Ruszczyński, *From stochastic dominance to mean-risk models: Semideviations as risk measures*, European Journal of Operational Research, 116 (1999), pp. 33–50, https://doi.org/10.1016/S0377-2217(98)00167-2.
[33] W. Ogryczak and A. Ruszczyński, *Dual stochastic dominance and related mean-risk models*, SIAM Journal on Optimization, 13 (2002), pp. 60–78, https://doi.org/10.1137/S1052623400375075.
[34] A. A. Pereira, J. Binney, G. A. Hollinger, and G. S. Sukhatme, *Risk-aware path planning for autonomous underwater vehicles using predictive ocean models*, Journal of Field Robotics, 30 (2013), pp. 741–762, https://doi.org/10.1002/rob.21472.
[35] R. T. Rockafellar, S. Uryasev, and M. Zabarankin, *Generalized deviations in risk analysis*, Finance and Stochastics, 10 (2006), pp. 51–74, https://doi.org/10.1007/s00780-005-0165-8.
[36] T. R. Rockafellar, S. P. Uryasev, and M. Zabarankin, *Deviation measures in risk analysis and optimization*, SSRN Electronic Journal, (2003), https://doi.org/10.2139/ssrn.365640.
[37] A. Sani, A. Lazaric, and R. Munos, *Risk-aversion in multi-armed bandits*, in Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012, pp. 3275–3283.
[38] D. Shang, V. Kuzmenko, and S. Uryasev, *Cash flow matching with risks controlled by buffered probability of exceedance and conditional value-at-risk*, Annals of Operations Research, 260 (2018), pp. 501–514, https://doi.org/10.1007/s10479-016-2354-6.

[39] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, Society for Industrial and Applied Mathematics, 2nd ed., 2014, https://doi.org/http://dx.doi.org/10.1137/1.9780898718751.

[40] J. C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, Wiley-Interscience, 2003.

[41] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, *Sequential decision making with coherent risk*, IEEE Transactions on Automatic Control, 62 (2017), pp. 3323–3338, https://doi.org/10.1109/TAC.2016.2644871.

[42] G. Taylor, R. Burmeister, Z. Xu, B. Singh, A. Patel, and T. Goldstein, *Training neural networks without gradients: A scalable admm approach*, in 33rd International Conference on Machine Learning (ICML 2016), June 2016, pp. 2722–2731, https://arxiv.org/abs/1605.02026.

[43] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 2000.

[44] C. A. Vitt, D. Dentcheva, and H. Xiong, *Risk-averse classification*, Annals of Operations Research, (2019), https://doi.org/10.1007/s10479-019-03344-6, https://arxiv.org/abs/1805.00119.

[45] M. Wang, E. X. Fang, and H. Liu, *Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions*, Mathematical Programming, 161 (2017), pp. 419–449, https://doi.org/10.1007/s10107-016-1017-3, https://arxiv.org/abs/1411.3803.

[46] S. Yang, M. Wang, and E. X. Fang, *Multilevel stochastic gradient methods for nested composition optimization*, SIAM Journal on Optimization, 29 (2019), pp. 616–659, https://doi.org/10.1137/18M1164846.

[47] P. Yu, W. B. Haskell, and H. Xu, *Approximate value iteration for risk-aware markov decision processes*, IEEE Transactions on Automatic Control, 63 (2018), pp. 3135–3142, https://doi.org/10.1109/TAC.2018.2790261, https://arxiv.org/abs/1701.01290.

[48] D. Yuan and D. W. Ho, *Randomized gradient-free method for multiagent optimization over time-varying networks*, IEEE Transactions on Neural Networks and Learning Systems, 26 (2015), pp. 1342–1347, https://doi.org/10.1109/TNNLS.2014.2336806.

[49] L. Zhou and P. Tokekar, *An approximation algorithm for risk-averse submodular optimization*, arXiv preprint, arXiv:1807.09358, (2018), https://doi.org/10.1007/978-3-030-44051-0_9, https://arxiv.org/abs/1807.09358.