Journal of Optimization Theory and Applications (2022) 194:1014–1041
https://doi.org/10.1007/s10957-022-02063-6

A Stochastic Subgradient Method for Distributionally Robust Non-convex and Non-smooth Learning

Mert Gürbüzbalaban[1] · Andrzej Ruszczyński[1] · Landi Zhu[1]

Received: 14 August 2020 / Accepted: 5 June 2022 / Published online: 8 July 2022

Abstract

We consider a distributionally robust formulation of stochastic optimization problems arising in statistical learning, where robustness is with respect to ambiguity in the underlying data distribution. Our formulation builds on risk-averse optimization techniques and the theory of coherent risk measures. It uses mean–semideviation risk for quantifying uncertainty, allowing us to compute solutions that are robust against perturbations in the population data distribution. We consider a broad class of generalized differentiable loss functions that can be non-convex and non-smooth, involving upward and downward cusps, and we develop an efficient stochastic subgradient method for distributionally robust problems with such functions. We prove that it converges to a point satisfying the optimality conditions. To our knowledge, this is the first method with rigorous convergence guarantees in the context of generalized differentiable non-convex and non-smooth distributionally robust stochastic optimization. Our method allows for the control of the desired level of robustness with little extra computational cost compared to population risk minimization with stochastic gradient methods. We also illustrate the performance of our algorithm on real datasets arising in convex and non-convex supervised learning problems.

Communicated by Zaid Harchaoui.

This work was partially supported by the National Science Foundation Awards DMS-1907522, CCF-1814888, and DMS-2053485, and by the Office of Naval Research Awards N00014-21-1-2161 and N00014-21-1-2244.

✉ Mert Gürbüzbalaban
mg1366@rutgers.edu

Andrzej Ruszczyński
rusz@rutgers.edu

Landi Zhu
lz401@scarletmail.rutgers.edu

[1] Rutgers University, Piscataway, NJ 08550, USA

Keywords Robust learning · Risk measures · Stochastic subgradient method · Non-smooth optimization · Composition optimization

Mathematics Subject Classification 90C15 · 90C48

1 Introduction

Statistical learning theory deals with the problem of making predictions and constructing models from a set of data. A typical statistical learning problem can be formulated as a stochastic optimization problem:

$$\min_{x\in X} \mathbb{E}_{D\sim\mathbb{P}}\left[\ell(x, D)\right], \tag{1}$$

where $\ell : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is the loss function of the predictor x on the random data D with an unknown distribution with probability law $\mathbb{P}$, and $X \subset \mathbb{R}^n$ is the feasible set (see, e.g., [60]). We consider loss functions that can be non-convex or non-differentiable (non-smooth). This framework includes a large class of problems in supervised learning including deep learning, linear and nonlinear regression and classification tasks [53].

A central problem in statistics is to make decisions that generalize well (i.e., work well on unseen data) as well as decisions that are robust to perturbations in the underlying data distribution [8]. Indeed, the statistical properties of the input data may be subject to some variations and distributional distortions, and a major goal is to build models that are not too sensitive to small changes in the input data distribution. This motivates the following distributionally robust version of problem (1):

$$\min_{x\in X} \max_{\mathbb{Q}\in\mathscr{M}(\mathbb{P})} \mathbb{E}_{D\sim\mathbb{Q}}\left[\ell(x, D)\right], \tag{2}$$

where $\mathscr{M}(\mathbb{P})$ is a weakly closed convex set of probability measures that model perturbations to the law $\mathbb{P}$, and the predictor x is chosen to accommodate worst-case perturbations. References [3, 35, 51] provide thorough discussion of the relevance of robustness in statistical learning. Problem (2) is related to quantifying risk of the random data distribution [20, 58]; its computational tractability depends on the underlying loss function and the uncertainty set $\mathscr{M}(\mathbb{P})$ [14, 17]. Existing approaches to the modeling of $\mathscr{M}(\mathbb{P})$ include conditional value at risk [58], f-divergence-based sets [14], Wasserstein balls around $\mathbb{P}$ [21, 55], and other statistical distance-based approaches (see, e.g., [21]). When $\ell(\cdot, D)$ is non-convex and non-differentiable, these formulations lead to non-convex min–max problems. To our knowledge, none of the existing algorithms admit provable convergence guarantees to a stationary point of (2) in this general case. Sinha et al. [55] consider the case when $\mathscr{M}(\mathbb{P})$ is defined as the ρ-neighborhood of the probability law $\mathbb{P}$ under the Wasserstein metric, where ρ is the desired level of robustness. The authors formulate a Lagrangian relaxation of this problem for a fixed penalty parameter $\gamma \geq 0$ and show that when the loss is smooth and the penalty parameter is large enough (or by duality if the desired level of robustness

ρ is small enough), the stochastic gradient descent (SGD) method achieves the same rate of convergence as the standard smooth non-convex optimization. The authors also provide a data-dependent upper bound for the worst-case population objective (2) for any robustness level ρ. Soma and Yoshida [56] proposed a conditional value-at risk (CVaR) formulation for robustness and show that for convex and smooth losses their algorithm based on SGD has $\mathscr{O}(1/\sqrt{n})$-convergence to the optimal CVaR, where n is the number of samples. For non-convex and smooth loss functions, they also show a generalization bound on the CVaR. However, none of these guarantees apply if the loss is non-smooth.

For some structured regression and classification problems of practical interest, distributionally robust formulations that result in finite-dimensional convex programs are known [17, 30, 37, 52] to be solvable in polynomial time; see also the reference [45] which contains a detailed list of tractable reformulations of distributionally robust constraints for several risk measures. For convex losses, conic interior point solvers or gradient descent with backtracking Armijo line-searches can also be used for solving a sample-based approximation of (2), when $\mathscr{M}(\mathbb{P})$ is defined via the f-divergences [14]. However, these approaches can be prohibitively expensive when the dimension or the number of samples are large. For smooth and convex losses, Namkoong and Duchi [41] showed that a sample-based approximation of (2) with f-divergences results in a min–max problem, which can then be solved with a bandit mirror descent algorithm with number of iterations comparable to that of the SGD for solving the sample-based approximation of problem (1). However, similar convergence guarantees for non-convex or non-smooth losses were not given. We also note that there are data-driven distributionally robust stochastic optimization formulations (see, e.g., [17, 20, 21]), which replace the population measure $\mathbb{P}$ with an empirical measure $\mathbb{P}^N$ constructed from samples of input data. A disadvantage is that the resulting set $\mathscr{M}(\mathbb{P}^N)$ becomes random.

We propose a new formulation of (2) based on the mean–semideviation measure of risk [43, 44]. We propose a specialized stochastic subgradient method for solving the resulting problem, which we call the *single-time-scale* (STS) method. Our method has local convergence guarantees for a large class of possibly non-convex and non-smooth loss functions.

Modeling the uncertainty set $\mathscr{M}(\mathbb{P})$ with mean semi-deviation risk Consider the random loss $Z(x) = \ell(x, D)$ defined on a sample space Ω equipped with a sigma algebra $\mathcal{F}$. We assume $\mathbb{E}[Z(x)]$ to be finite, *i.e.*, $Z(x) \in \mathcal{Z} = \mathcal{L}_1(\Omega, \mathcal{F}, \mathbb{P})$. Instead of problem (1), we propose to consider the *risk minimization problem*

$$\min_{x \in X} \rho\big[\ell(x, D)\big]. \tag{3}$$

with a coherent measure of risk $\rho[\cdot]$; see [2, 18, 54] and the references therein. Coherence means that $\rho : \mathcal{Z} \to \mathbb{R}$ satisfies the axioms of *convexity* ($\rho[\alpha Z + (1-\alpha)V] \le \alpha\rho[Z] + (1-\alpha)\rho[V]$, $\forall \alpha \in [0, 1]$), *monotonicity* ($\rho[Z] \le \rho[V]$ if $Z \le V$ a.s.), *translation equivariance* ($\rho[Z + c] = \rho[Z] + c$, $\forall c \in \mathbb{R}$), and *positive homogeneity* ($\rho[\gamma Z] = \gamma\rho[Z]$, $\forall \gamma \ge 0$).

Coherent measures of risk have the *dual representation* [48],

$$\rho[Z] = \max_{\mu \in \mathcal{A}} \int_{\Omega} Z(\omega)\mu(\omega)\ \mathbb{P}(d\omega) = \max_{\mathbb{Q}: \frac{d\mathbb{Q}}{d\mathbb{P}} \in \mathcal{A}} \int_{\Omega} Z(\omega)\ \mathbb{Q}(d\omega) = \max_{\mathbb{Q}: \frac{d\mathbb{Q}}{d\mathbb{P}} \in \mathcal{A}} \mathbb{E}_{\mathbb{Q}}[Z],$$

where $\mathcal{A}$ is a convex and closed set. Thus, problem (3) takes on the min–max form

$$\min_{x \in X} \max_{\mathbb{Q} \in \mathscr{M}(\mathbb{P})} \mathbb{E}_{\mathbb{Q}}[\ell(x, D)] \tag{4}$$

with the uncertainty set

$$\mathscr{M}(\mathbb{P}) = \left\{\mathbb{Q} : \frac{d\mathbb{Q}}{d\mathbb{P}} \in \mathcal{A}\right\}. \tag{5}$$

In this way, by using a coherent measure of risk, we achieve an implicit robust formulation. The set $\mathcal{A}$ depends on the measure of risk used, but it is essential that all probability measures in the uncertainty set (5) are absolutely continuous with respect to the original measure $\mathbb{P}$, thus excluding perturbations that are structurally impossible in the problem.

The usefulness of the formulation (3) is predicated on our ability to solve it in an efficient way and on the quality of the solutions obtained. We propose to use the first-order *mean–semideviation risk measure* [43, 44]:

$$\rho[Z] = \mathbb{E}[Z] + \varkappa \mathbb{E}\big[\max\big(0, Z - \mathbb{E}[Z]\big)\big], \qquad \varkappa \in [0, 1]. \tag{6}$$

The measure (6) has the set $\mathcal{A}$ defined as follows:

$$\mathcal{A} = \big\{\mu = \mathbb{1} + \xi - \mathbb{E}[\xi] :\ \xi \in \mathcal{L}_{\infty}(\Omega, \mathcal{F}, \mathbb{P}),\ \|\xi\|_{\infty} \le \varkappa,\ \xi \ge 0\big\},$$

(see, e.g., [48]). The level of robustness is controlled by the parameter $\varkappa$: For $\varkappa = 0$, the uncertainty set (5) contains only the original probability measure $\mathbb{P}$, while for $\varkappa > 0$ the measures $\mathbb{Q} \in \mathscr{M}(\mathbb{P})$ are distortions of $\mathbb{P}$. The range of relative distortions allowed, $\frac{d\mathbb{Q}}{d\mathbb{P}} - \mathbb{1}$, depends on $\varkappa$.

Problem (3) with the mean–semideviation risk measure (6) can be cast in the following form of a composition optimization problem:

$$\min_{x \in X} f(x, h(x)), \tag{7}$$

with the functions

$$f(x, u) = \mathbb{E}\Big[\ell(x, D) + \varkappa \max\big(0, \ell(x, D) - u\big)\Big], \tag{8}$$

$$h(x) = \mathbb{E}[\ell(x, D)]. \tag{9}$$

The main difficulty is that neither values nor (sub)gradients of $f(\cdot)$, $h(\cdot)$ and of their composition are available. Instead, we postulate access to their random estimates. Such estimates, however, may be biased, because estimating a (sub)gradient of the composition $F(x) = f(x, h(x))$ involves estimating $h(x)$. Although problem (7) can be further rewritten in the standard format of composition optimization,

$$\min_{x \in X} f(\bar{h}(x)), \tag{10}$$

with $\bar{h}(x) = (x, h(x))$, the more specific formulation (7) allows us to derive a more efficient specialized method, because x is observed.

The research on composition optimization problems of form (10) started from penalty functions for stochastic constraints and composite regression models in [15,Ch. V.4]. An established approach was to use two-level stochastic recursive algorithms with two stepsize sequences in different time scales: a slower one for updating the main decision variable x, and a faster one for filtering the value of the inner function h. References [28, 61–63] provide a detailed account of these techniques and existing results.

A central limit theorem for stochastic versions of problem (10) has been established in [10]. Large deviation bounds for the empirical optimal value were derived in [16]. A new single time-scale method for problem (10) with continuously differentiable functions has been recently proposed in [23]. It has the complexity of $\mathcal{O}(1/\epsilon^2)$ to obtain an ε-solution of the problem, the same as methods for one-level unconstrained stochastic optimization. However, the construction of the method and its analysis depend on the Lipschitz constants of the gradients of the functions involved. Our problem (7), unfortunately, involves a non-smooth function $\max(\cdot, \cdot)$ and may also involve a non-smooth (non-differentiable) loss function $\ell(\cdot, \cdot)$. Indeed, many key problems in machine learning involve non-convex and non-smooth loss functions. A prominent example is deep learning with ReLU activation functions (see e.g. [24]). There are many other statistical learning problems where the objective can be non-smooth and non-differentiable such as non-convex generalized linear models and non-convex regression and risk minimization (see, e.g., [1, 19, 26, 59]). The organic non-differentiability and non-convexity are additional challenges for the solution method.

Contributions We propose to model the perturbation to input data distribution by mean–semi-deviation risk, according to (5). Our formulation leads to the distributionally robust learning problem (4), which has the advantage that it results in a convex optimization problem when the loss ℓ is convex, in contrast to some alternative formulations, which result in min–max optimization problems (see, e.g., [41, 58]). When the loss is non-convex and non-smooth, we can still find a stationary point to (4), by our novel single time-scale parameter-free stochastic subgradient method, for a general class of loss functions that can be non-convex and non-differentiable. To our knowledge, our method is the first method with probability one convergence guarantees for solving a distributionally robust formulation of a population minimization problem, where the loss can be non-convex or non-differentiable.

We also note that the computational cost of stochastic first-order optimization algorithms is typically measured in terms of the number of stochastic gradient or subgradient evaluations they require (see, e.g., [6,Section 6], [22, 27]). Standard SGD methods (which go back to Robbins and Monro's pioneering work [46]) applied to the non-robust optimization problem (1) can operate with one stochastic subgradient evaluation under similar assumptions to ours; however, they are not applicable to the robust formulation (2) directly. In contrast, our method can converge to a stationary point of the robust formulation (2) with probability one requiring on average no more than $1+\varkappa$ stochastic subgradient evaluations at every iteration, where $\varkappa \in [0, 1]$ is the desired level of robustness. Therefore, comparing the numbers of stochastic gradient evaluations, the average computational cost of each iteration of our method is at most $1+\varkappa$ times larger than that of the standard SGD method, requiring little extra cost to compute robust solutions.

2 The Single Time-Scale (STS) Method with Subgradient Averaging

We present the method for problems of form (4), in which the loss function $\ell(x, D)$ is differentiable in a generalized sense [42] with respect to x and integrable with respect to D. This broad class of functions is contained in the set of locally Lipschitz functions and contains all semismooth locally Lipschitz loss functions that can be non-convex and non-differentiable [39]. We note that this class includes many of the losses arising in statistical learning problems, including population and empirical risk minimization with possibly non-convex and non-smooth regularizers [1, 19, 26, 59, 60], weakly convex and continuous losses [9, 34] as well as deep learning with ReLU activations [24]. In Appendix A, we provide the precise definition and recall the most important properties of the class of functions that are differentiable in a generalized sense.

Recall that the Clarke subdifferential $\partial_x \ell(x, D)$ is an inclusion-minimal generalized derivative of $\ell(\cdot, D)$ [42]. In what follows, however, we use the symbol $\hat{\partial} f$ to denote the generalized subdifferential of a function $f(\cdot)$ in the sense of Definition A.1 that we provide in Appendix A. We make the following assumptions.

(A1) The set $X \subset \mathbb{R}^n$ is convex and compact;

(A2) For almost every (a.e.) $\omega \in \Omega$, the function $\ell(\cdot, D(\omega))$ is differentiable in a generalized sense with the generalized differential $\hat{\partial}\ell(x, D(\omega))$, $x \in \mathbb{R}^n$. Moreover, for every compact set $K \in \mathbb{R}^n$ an integrable function $L_K : \Omega \to \mathbb{R}$ exists, satisfying $\sup_{x\in K} \sup_{g\in\hat{\partial}\ell(x,D(\omega))} \|g\| \leq L_K(\omega)$.

Under (A2), function (9) is also differentiable in a generalized sense. Although its generalized derivative is not readily available, we can draw $\widetilde{D}$ from the distribution of D and use an element of $\hat{\partial}\ell(x, \widetilde{D})$ as a *stochastic subgradient* (a random vector whose expected value is a subgradient). Furthermore, function (8) is also differentiable in a generalized sense with respect to (x, u). Its stochastic subgradient can be obtained as follows. First, we observe $\ell(x, \widetilde{D})$ and choose

$$r \in \begin{cases} \{0\} & \text{if } \ell(x, \widetilde{D}) < u, \\ [0, 1] & \text{if } \ell(x, \widetilde{D}) = u, \\ \{1\} & \text{if } \ell(x, \widetilde{D}) > u. \end{cases}$$

Then, the vector $\begin{bmatrix}\tilde{g}_x\\ \tilde{g}_u\end{bmatrix}$, where $\tilde{g}_x \in (1 + r\varkappa)\hat{\partial}\ell(x, \widetilde{D})$, $\tilde{g}_u = -r\varkappa$, is a stochastic subgradient of the function $f(x, u)$, which is defined by (8). These formulas follow from calculus rules for generalized subdifferentials of compositions and expected values (Theorems A.1 and A.2 in Appendix A). We can also use different samples for calculating stochastic subgradients of (8) and (9).

The STS method generates three random sequences: approximate solutions $\{x^k\}$, path-averaged stochastic subgradients $\{z^k\}$, and inner function estimates $\{u^k\}$, all defined on a certain probability space $(\Omega, \mathcal{F}, P)$. We let $\mathcal{F}_k$ to be the σ-algebra generated by $\{x^0, \ldots, x^k, z^0, \ldots, z^k, u^0, \ldots, u^k\}$. Starting from the initialization $x^0 \in X$, $z^0 \in \mathbb{R}^n$, $u^0 \in \mathbb{R}$, the method uses parameters $a > 0$, $b > 0$ and $c > 0$ to generate x^k, z^k, u^k for $k > 0$. At each iteration $k = 0, 1, 2, \ldots$, we compute

$$y^k = \operatorname*{argmin}_{y \in X} \left\{ \langle z^k, y - x^k \rangle + \frac{c}{2}\|y - x^k\|^2 \right\},$$

[1] and, with an $\mathcal{F}_k$-measurable stepsize $\tau_k \in \big(0, \min(1, 1/a)\big]$, we set

$$x^{k+1} = x^k + \tau_k(y^k - x^k). \tag{11}$$

Then, we obtain statistical estimates:

- $\tilde{g}^{k+1} = \begin{bmatrix}\tilde{g}_x^{k+1}\\ \tilde{g}_u^{k+1}\end{bmatrix}$ of an element $g^{k+1} = \begin{bmatrix}g_x^{k+1}\\ g_u^{k+1}\end{bmatrix} \in \hat{\partial} f(x^{k+1}, u^k)$,
- $\tilde{h}^{k+1}$ of $h(x^{k+1})$, and
- $\tilde{J}^{k+1}$ of an element $J^{k+1} \in \hat{\partial} h(x^{k+1})$ with the convention that J^{k+1} is a row vector,

and we update the running averages as

$$z^{k+1} = z^k + a\tau_k\Big(\tilde{g}_x^{k+1} + \big[\tilde{J}^{k+1}\big]^\top \tilde{g}_u^{k+1} - z^k\Big), \tag{12}$$

$$u^{k+1} = u^k + \tau_k \tilde{J}^{k+1}(y^k - x^k) + b\tau_k\big(\tilde{h}^{k+1} - u^k\big). \tag{13}$$

We assume the following conditions on the stepsizes and the stochastic estimates:

(A3) $\tau_k \in \big(0, \min(1, 1/a)\big]$ for all k, $\lim_{k\to\infty} \tau_k = 0$, $\sum_{k=0}^{\infty} \tau_k = \infty$, $\sum_{k=0}^{\infty} \mathbb{E}[\tau_k^2] < \infty$;

(A4) For all k,

(i) $\tilde{g}^{k+1} = g^{k+1} + e_g^{k+1} + \delta_g^{k+1}$, with
$g^{k+1} \in \hat{\partial} f(x^{k+1}, u^k)$, $\mathbb{E}\big\{e_g^{k+1}\big|\mathcal{F}_k\big\} = 0$, $\mathbb{E}\big\{\|e_g^{k+1}\|^2|\mathcal{F}_k\big\} \le \sigma_g^2$,
$\lim_{k\to\infty} \delta_g^{k+1} = 0$,

(ii) $\tilde{h}^{k+1} = h(x^{k+1}) + e_h^{k+1} + \delta_h^{k+1}$, with
$\mathbb{E}\big\{e_h^{k+1}\big|\mathcal{F}_k\big\} = 0$, $\mathbb{E}\big\{[e_h^{k+1}]^2|\mathcal{F}_k\big\} \le \sigma_h^2$, $\lim_{k\to\infty} \delta_h^{k+1} = 0$,

[1] From the update rule of y^k, it follows that the variable y^k is the projection of $x^k - z^k/c$ onto the constraint set X, where $1/c$ can be interpreted as the stepsize. This projection step ensures that the iterates y^k lie in the constraint set X.

(iii) $\tilde{J}^{k+1} = J^{k+1} + E^{k+1} + \Delta^{k+1}$, with $J^{k+1} \in \hat{\partial} h(x^{k+1})$, $\mathbb{E}\{E^{k+1} \big| \mathcal{F}_k\} = 0$, $\mathbb{E}\{\|E^{k+1}\|^2 | \mathcal{F}_k\} \leq \sigma_E^2$, $\lim_{k\to\infty} \Delta^{k+1} = 0$, and $\mathbb{E}\big[(E^{k+1})^\top e_{gu}^{k+1} \,\big|\, \mathcal{F}_k\big] = 0$ where $(e_{gx}^{k+1}, e_{gu}^{k+1})$ are the components of e_g^{k+1} that correspond to x and u.

These assumptions are pretty standard in the study of stochastic gradient and stochastic approximation methods [32]. As discussed before, the stochastic estimates satisfying these conditions can be obtained by drawing at each iteration two independent samples: D_1^{k+1} and D_2^{k+1}, from the data. Then, we can take

$$\begin{aligned}
\tilde{g}_x^{k+1} &\in \begin{cases} \hat{\partial}\ell(x^{k+1}, D_1^{k+1}) & \text{if } \ell(x^{k+1}, D_1^{k+1}) < u^k, \\ (1+\varkappa)\hat{\partial}\ell(x^{k+1}, D_1^{k+1}) & \text{if } \ell(x^{k+1}, D_1^{k+1}) \geq u^k, \end{cases} \\
\tilde{g}_u^{k+1} &= \begin{cases} 0 & \text{if } \ell(x^{k+1}, D_1^{k+1}) < u^k, \\ -\varkappa & \text{if } \ell(x^{k+1}, D_1^{k+1}) \geq u^k, \end{cases} \\
\tilde{h}^{k+1} &= \ell(x^{k+1}, D_1^{k+1}), \\
\tilde{J}^{k+1} &\in \hat{\partial}\ell(x^{k+1}, D_2^{k+1}).
\end{aligned} \tag{14}$$

We also note that we can reduce the number of samples per iteration by randomization. Let β be an independent Bernoulli random variable with $\mathbb{P}[\beta = 1] = \varkappa$ and $\mathbb{P}[\beta = 0] = 1 - \varkappa$. After rewriting formula (8) as

$$f(x, u) = (1-\varkappa)\mathbb{E}\big[\ell(x, D)\big] + \varkappa\mathbb{E}\big[\ell(x, D) + \max\big(0, \ell(x, D) - u\big)\big],$$

we can interpret it as an expected value with respect to β. Therefore, its stochastic subgradient can be generated as follows. At each iteration, we sample β, independently of other samples in the method. Then, we set

$$\begin{aligned}
\tilde{g}_x^{k+1} &\in \begin{cases} \hat{\partial}\ell(x^{k+1}, D_1^{k+1}) & \text{if } \beta = 0 \text{ or } \ell(x^{k+1}, D_1^{k+1}) < u^k, \\ 2\hat{\partial}\ell(x^{k+1}, D_1^{k+1}) & \text{if } \beta = 1 \text{ and } \ell(x^{k+1}, D_1^{k+1}) \geq u^k, \end{cases} \\
\tilde{g}_u^{k+1} &= \begin{cases} 0 & \text{if } \beta = 0 \text{ or } \ell(x^{k+1}, D_1^{k+1}) < u^k, \\ -1 & \text{if } \beta = 1 \text{ and } \ell(x^{k+1}, D_1^{k+1}) \geq u^k, \end{cases} \\
\tilde{h}^{k+1} &= \ell(x^{k+1}, D_1^{k+1}), \\
\tilde{J}^{k+1} &\in \begin{cases} \hat{\partial}\ell(x^{k+1}, D_1^{k+1}) & \text{if } \beta = 0 \text{ or } \ell(x^{k+1}, D_1^{k+1}) < u^k, \\ \hat{\partial}\ell(x^{k+1}, D_2^{k+1}) & \text{if } \beta = 1 \text{ and } \ell(x^{k+1}, D_1^{k+1}) \geq u^k. \end{cases}
\end{aligned} \tag{15}$$

Since $\tilde{g}_u^{k+1} = 0$ when $\beta = 0$, assumption (A4)(iii) can be satisfied with $\tilde{J}^{k+1} \in \hat{\partial}\ell(x^{k+1}, D_1^{k+1})$. The need for the second sample from the data, D_2^{k+1}, may occur only if $\beta = 1$, that is, with probability $\varkappa$. Therefore, on average at most $1 + \varkappa$ samples are needed per iteration.

Our method refines and specializes the approach to multi-level stochastic optimization recently developed in [50]. We extend this approach to a new case in which the upper level function, $f(x, u)$, is not continuously differentiable with respect to u, and thus, the conditions of [50] are not satisfied. We establish the convergence in the new case as well, as detailed in the following section.

3 Convergence Analysis

To recall optimality conditions for problem (4) and analyze our method, we need to introduce relevant multifunctions. Consider the composition function

$$F(x) = f(x, h(x)), \quad x \in \mathbb{R}^n.$$

For a point $x \in \mathbb{R}^n$, we define the set:

$$\hat{\partial} F(x) = \operatorname{conv}\big\{s \in \mathbb{R}^n : s = g_x + J^\top g_u,\ g \in \hat{\partial} f(x, h(x)),\ J \in \hat{\partial} h(x)\big\}. \quad (16)$$

By [40,Thm. 1.6] (Theorem A.1 in Appendix A), the set $\hat{\partial} F(x)$ is a generalized subdifferential of $F(\cdot)$ at x. We call a point $x^* \in X$ *stationary* for problem (4), if

$$0 \in \hat{\partial} F(x^*) + N_X(x^*), \quad (17)$$

where $N_X(x)$ is the normal cone to X at x. The set of stationary points is denoted by X^*.

We start by considering the gap function $\eta : X \times \mathbb{R}^n \to (-\infty, 0]$,

$$\eta(x, z) = \min_{y \in X} \left\{ \langle z, y - x \rangle + \frac{c}{2} \|y - x\|^2 \right\}, \quad (18)$$

which admits the unique minimizer

$$\bar{y}(x, z) = \arg\min_{y \in X} \left\{ \langle z, y - x \rangle + \frac{c}{2} \|y - x\|^2 \right\}. \quad (19)$$

Since $\bar{y}(x, z)$ is a projection of $x - z/c$ on X,

$$\langle z, \bar{y}(x, z) - x \rangle + c \|\bar{y}(x, z) - x\|^2 \le 0, \quad (20)$$

for every $x \in X$ and $z \in \mathbb{R}^n$. Moreover, a point $x^* \in X^*$ if and only if $z^* \in \hat{\partial} F(x^*)$ exists such that $\eta(x^*, z^*) = 0$; see [47,Prop. 1].[2] Consider the multifunction $\Gamma : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \rightrightarrows \mathbb{R}^n \times \mathbb{R}$:

[2] This statement follows from the following argument: If $x^* \in X^*$, then by definition (17), there exists $z^* \in \hat{\partial} F(x^*)$ such that $-z^* \in N_X(x^*)$, which is equivalent to $\langle z^*, y - x^* \rangle \ge 0$ for every $y \in X$. This, together with the definition (18) of the gap function, implies that $\eta(x^*, z^*) \ge 0$, which yields $\eta(x^*, z^*) = 0$, due to (20). The other direction can be proved in a similar way. If $z^* \in \hat{\partial} F(x^*)$ exists such that $\eta(x^*, z^*) = 0$, then by definition (18), $\langle z^*, y - x^* \rangle \ge 0$ for every $y \in X$; otherwise, one gets a contradiction. The latter statement is equivalent to $-z^* \in N_X(x^*)$, and consequently, we obtain $x^* \in X$.

$$\begin{aligned}\Gamma(x,z,u) = &\big\{(R,v) : \exists g \in \hat{\partial} f(x,u), \exists J_1, J_2 \in \hat{\partial} h(x),\\ &v = J_1\big(\bar{y}(x,z)-x\big) + b(h(x)-u),\ R = a\big(g_x + J_2^\top g_u - z\big)\big\}.\end{aligned} \tag{21}$$

where $\bar{y}(x,z)$ is defined by (19). Here, (g_x, g_u) are the components of g that correspond to x and u. With this notation, we can write the updates (11)–(13) as follows:

$$\begin{aligned} x^{k+1} &= x^k + \tau_k(\bar{y}(x^{k+1},z^k) - x^{k+1}) + \tau_k \alpha_x^{k+1},\\ \begin{bmatrix} z^{k+1}\\ u^{k+1}\end{bmatrix} &\in \begin{bmatrix} z^{k}\\ u^{k}\end{bmatrix} + \tau_k \Gamma(x^{k+1}, z^k, u^k) + \tau_k \begin{bmatrix} \theta_R^{k+1}\\ \theta_v^{k+1}\end{bmatrix} + \tau_k \begin{bmatrix} \alpha_R^{k+1}\\ \alpha_v^{k+1}\end{bmatrix}, \end{aligned} \tag{22}$$

where

$$\begin{aligned} \theta_R^{k+1} &= a(e_{gx}^{k+1} + [E^{k+1}]^T g_u^{k+1} + [J^{k+1} + E^{k+1}]^T e_{gu}^{k+1}),\\ \theta_v^{k+1} &= E^{k+1}(\bar{y}(x^k,z^k) - x^k) + b e_h^{k+1},\\ \alpha_x^{k+1} &= \bar{y}(x^k,z^k) - \bar{y}(x^{k+1},z^k) + x^{k+1} - x^k,\\ \alpha_R^{k+1} &= a(\delta_{gx}^{k+1} + [\Delta^{k+1}]^T g_u^{k+1} + [\Delta^{k+1}]^T e_{gu}^{k+1} + [J^{k+1} + E^{k+1} + \Delta^{k+1}]^T \delta_{gu}^{k+1}),\\ \alpha_v^{k+1} &= \Delta^{k+1}(\bar{y}(x^k,z^k) - x^k) + b\delta_h^{k+1} + J^{k+1}\alpha_x^{k+1}. \end{aligned}$$

In this way, we formulate our algorithm as a perturbed differential inclusion model analyzed in [36], where the deterministic part on the right-hand side depends on the perturbed point (x^{k+1}, z^k, u^k); observe that $\|x^{k+1} - x^k\| \to 0$, due to (A1) and (A3). Furthermore, for $\theta^{k+1} = \begin{bmatrix} \theta_R^{k+1}\\ \theta_v^{k+1}\end{bmatrix}$ and $\alpha^{k+1} = \begin{bmatrix} \alpha_x^{k+1}\\ \alpha_R^{k+1}\\ \alpha_v^{k+1}\end{bmatrix}$ we have

$$\mathbb{E}\big[\theta^{k+1} \,\big|\, \mathcal{F}_k\big] = 0, \quad \mathbb{E}\big[\|\theta^{k+1}\|^2 \,\big|\, \mathcal{F}_k\big] \le C^\theta, \quad k = 0, 1, \ldots, \tag{23}$$

with some constant C^θ, and

$$\lim_{k\to\infty} \alpha^{k+1} = 0. \quad \text{a.s..} \tag{24}$$

The verification of relations (22)–(24) is straightforward from the description of the algorithm and assumptions (A3)–(A4).

Two technical results are needed for further analysis.

Lemma 3.1 *The multifunction Γ is upper-semicontinuous and compact and convex-valued.*

Proof By assumption (A2), for a.e. $\omega \in \Omega$, the loss function $\ell(x, D(\omega))$ is generalized differentiable, and therefore, the function $f(x,u)$ is also generalized differentiable where $\hat{\partial}_x f(x,u)$, $\hat{\partial}_u f(x,u)$ and $\hat{\partial} h(x)$ are all convex and compact-valued and upper-semicontinuous [42]. Therefore, Γ is upper-semicontinuous and compact-valued. It remains to verify the convexity.

Consider the function $I(x, u; D) = \ell(x, D) + \varkappa \cdot \max(0, \ell(x, D) - u)$. We can calculate its generalized subdifferential with respect to (x, u):

$$\hat{\partial} I(x, u; D) = \begin{cases} \begin{bmatrix} (1+\varkappa)\hat{\partial}\ell(x, D) \\ -\varkappa \end{bmatrix}, \text{ if } u < \ell(x, D), \\ \left\{ \begin{bmatrix} (1+\varkappa r)\hat{\partial}\ell(x, D) \\ -\varkappa r \end{bmatrix} : r \in [0, 1] \right\}, \text{ if } u = \ell(x, D), \\ \begin{bmatrix} \hat{\partial}\ell(x, D) \\ 0 \end{bmatrix}, \text{ if } u > \ell(x, D). \end{cases}$$

Due to Assumption (A2), for every compact set $K \in \mathbb{R}^n$ an integrable function $M_K : \Omega \to \mathbb{R}$ exists, satisfying $\sup_{x \in K} \sup_{g \in \hat{\partial} I(x,u;D(\omega))} \|g\| \le M_K(\omega)$. By the interchangeability of the generalized subdifferential and integral operators (Theorem A.2 in Appendix A), we obtain:

$$\hat{\partial} f(x, u) = \mathbb{E}\big[\hat{\partial} I(x, u; D)\big]$$

Therefore,

$$\begin{aligned} \hat{\partial} f(x, u) = \Bigg\{ & \begin{bmatrix} p_1(1+\varkappa)l_1 + p_2(1+\varkappa r)l_2 + p_3 l_3 \\ -\varkappa(p_1 + p_2 r) \end{bmatrix} : \\ & r \in [0, 1], l_1 \in \mathbb{E}\big[\hat{\partial}\ell(x, D) \,\big|\, u < \ell(x, D)\big], \\ & l_2 \in \mathbb{E}\big[\hat{\partial}\ell(x, D) \,\big|\, u = \ell(x, D)\big], l_3 \in \mathbb{E}\big[\hat{\partial}\ell(x, D) \,\big|\, u > \ell(x, D)\big] \Bigg\}, \end{aligned} \tag{25}$$

where $p_1 = \mathbb{P}\{u < \ell(x, D)\}$, $p_2 = \mathbb{P}\{u = \ell(x, D)\}$, $p_3 = \mathbb{P}\{u > \ell(x, D)\}$. For a certain fixed input (x, z, u) to Γ, the quantities p_1, p_2, p_3 can be treated as scalar constants, the conditional subdifferentials $\mathbb{E}\big[\hat{\partial}\ell(x, D) \,\big|\, u < \ell(x, D)\big]$, $\mathbb{E}\big[\hat{\partial}\ell(x, D) \,\big|\, u = \ell(x, D)\big]$ and $\mathbb{E}\big[\hat{\partial}\ell(x, D) \,\big|\, u > \ell(x, D)\big]$ can be treated as fixed sets.

Now, in order to prove $\Gamma(x, z, u)$ is convex-valued, we notice by (21) that any point in $\Gamma(x, z, u)$ is a pair (R, v) generated by a triple (g, J_1, J_2) from $\hat{\partial} f(x, u) \times \hat{\partial} h(x) \times \hat{\partial} h(x)$. We can (arbitrarily) choose two points in $\Gamma(x, z, u)$: $A = (R_a, v_a)$ and $B = (R_b, v_b)$ and denote the triple generating the point A by (g_a, J_{1a}, J_{2a}), the triple generating the point B by (g_b, J_{1b}, J_{2b}).

By (21), for an arbitrary $s \in [0, 1]$, the convex combination $(R^s, v^s) = sA + (1-s)B$ of A and B can be expressed as:

$$\begin{aligned} R^s &= a(s g_{ax} + (1-s) g_{bx} + s J_{2a}^T g_{au} + (1-s) J_{2b}^T g_{bu} - z), \\ v^s &= (s J_{1a} + (1-s) J_{1b})(\bar{y}(x, z) - x) + b(h(x) - u). \end{aligned} \tag{26}$$

where (g_{ax}, g_{au}) are the components of g_a, and (g_{bx}, g_{bu}) are the components of g_b. If we can always find a triple $(g_c, J_{1c}, J_{2c}) \in \hat{\partial} f(x, u) \times \hat{\partial} h(x) \times \hat{\partial} h(x)$ that generates this convex combination, then $\Gamma(x, z, u)$ is convex-valued. That is what we show next.

From (25), we first deduce a simple relationship between g_x and g_u:

$$g_x = p_1 l_1 + p_2 l_2 + p_3 l_3 + \varkappa p_1 (l_1 - l_2) - l_2 g_u.$$

Therefore, choosing an element g from $\hat{\partial} f(x,u)$ is equivalent to choosing g_u from $\hat{\partial}_u f(x,u)$ and then choosing three conditional subgradients l_1, l_2, l_3 from $\mathbb{E}\big[\hat{\partial}\ell(x,D)\,\big|\,u < \ell(x,D)\big]$, $\mathbb{E}\big[\hat{\partial}\ell(x,D)\,\big|\,u = \ell(x,D)\big]$ and $\mathbb{E}\big[\hat{\partial}\ell(x,D)\,\big|\,u > \ell(x,D)\big]$.

Substitution into (26) yields:

$$\begin{aligned} R^s &= a(s(p_1 l_{1a} + p_2 l_{2a} + p_3 l_{3a} + \varkappa p_1 (l_{1a} - l_{2a}) - l_{2a} g_{au}) \\ &\quad + (1-s)(p_1 l_{1b} + p_2 l_{2b} + p_3 l_{3b} + \varkappa p_1 (l_{1b} - l_{2b}) - l_{2b} g_{bu}) \\ &\quad + s J_{2a}^T g_{au} + (1-s) J_{2b}^T g_{bu} - z), \\ v^s &= (s J_{1a} + (1-s) J_{1b})(\bar{y}(x,z) - x) + b(h(x) - u), \end{aligned}$$

where l_{1a}, l_{2a}, l_{3a} are the conditional subgradients corresponding to the point A, and l_{1b}, l_{2b}, l_{3b} are the conditional subgradients corresponding to the point B.

Notice that in the special case when $p_2 = 0$, the set $\hat{\partial} f(x,u)$ becomes a singleton, which is a convex set. Since the subdifferential $\hat{\partial} h(x)$ of the generalized differentiable function $h(x) = \mathbb{E}[\ell(x,D)]$ is a convex set and $\Gamma(x,z,u)$ is generated by the elements of the set $\hat{\partial} f(x,u) \times \hat{\partial} h(x) \times \hat{\partial} h(x)$, we can directly conclude that $\Gamma(x,z,u)$ is a convex set as well. In this case, there is nothing left to prove. Therefore, in the rest of the proof, we assume $p_2 \neq 0$. Noticing that $\hat{\partial}_u f(x,u)$ and $\hat{\partial} h(x)$ are convex sets, we consider the convex combinations

$$\begin{aligned} l_{1c} &:= s l_{1a} + (1-s) l_{1b}, \\ l_{2c} &:= \frac{s(p_2 - \varkappa p_1 - g_{au})}{p_2 - \varkappa p_1 - s g_{au} - (1-s) g_{bu}} l_{2a} + \frac{(1-s)(p_2 - \varkappa p_1 - g_{bu})}{p_2 - \varkappa p_1 - s g_{au} - (1-s) g_{bu}} l_{2b}, \\ l_{3c} &:= s l_{3a} + (1-s) l_{3b}, \\ g_{cu} &:= s g_{au} + (1-s) g_{bu}, \\ J_{1c} &:= s J_{1a} + (1-s) J_{1b}, \\ J_{2c} &:= \frac{s g_{au}}{s g_{au} + (1-s) g_{bu}} J_{2a} + \frac{(1-s) g_{bu}}{s g_{au} + (1-s) g_{bu}} J_{2b}; \end{aligned}$$

[3] with the convention that when $g_{au} = g_{bu} = 0$, we have $g_{cu} := 0$ and $J_{2c} := J_{2a}$. The corresponding point $C = (R_c, v_c)$ (generated by the triple (g_{cu}, J_{1c}, J_{2c})) will be:

$$\begin{aligned} R_c = a((1+\varkappa) & p_1 (s l_{1a} + (1-s) l_{1b}) + s(p_2 - \varkappa p_1 - g_{au}) l_{2a} \\ & + (1-s)(p_2 - \varkappa p_1 - g_{bu}) l_{2b} \\ & + p_3 (s l_{3a} + (1-s) l_{3b}) + s J_{2a}^T g_{au} + (1-s) J_{2b}^T g_{bu} - z), \end{aligned}$$

[3] Notice that in the definition of ℓ_{2c}, we have necessarily $p_2 - \varkappa p_1 - g_{au} > 0$ as $p_2 > 0$ and $-\varkappa p_1 - g_{au} \geq 0$ by (25). Similarly, $p_2 - \varkappa p_1 - g_{bu} > 0$. Therefore, the denominator $p_2 - \varkappa p_1 - s g_{au} - (1-s) g_{bu} > 0$.

$$v_c = (sJ_{1a} + (1-s)J_{1b})(\bar{y}(x,z) - x) + b(h(x) - u),$$

which is identical to the convex combination (R^s, v^s). Therefore, any convex combination of two arbitrary points in $\Gamma(x,z,u)$ remains in the set, and thus, $\Gamma(x,z,u)$ is convex-valued. □

Lemma 3.2 *Function (8) admits the chain rule (42) on every absolutely continuous path* $(x(t), u(t))$, $t \geq 0$, *such that* $x(\cdot)$ *is continuously differentiable.*

Proof Suppose $x(\cdot)$ is continuously differentiable and $u(\cdot)$ is absolutely continuous. Due to Assumption (A2) and Theorem A.3, for every D the function $L_D(t) = \ell(x(t), D)$ admits the chain rule. As it is absolutely continuous and the function $\max(0, \cdot)$ is convex, the function $t \mapsto \max(0, L_D(t) - u(t))$ admits the chain rule as well. By virtue of (A2), the expected value (8) admits the chain rule as claimed. □

Lemma 3.3 *The sequences* $\{z^k\}$ *and* $\{u^k\}$ *are bounded with probability 1.*

The proof is routine. For convenience of the reader, we provide it in Appendix B.

We analyze the method by the differential inclusion technique, by refining and specializing the approach adopted in [50]. Although our model does not fit the assumptions of [50], our result on the convexity of the multifunction $\Gamma(\cdot)$ allows for proving convergence in this case as well.

We need an additional technical assumption.

(A5) The set $F(X^*)$ does not contain an interval of nonzero length.

Theorem 3.1 *If the assumptions* (A1)–(A5) *are satisfied, then with probability 1 every accumulation point* $\hat{x}$ *of the sequence* $\{x^k\}$ *is stationary,* $\lim_{k\to\infty}(u^k - h(x^k)) = 0$, *and the sequence* $\{F(x^k)\}$ *is convergent.*

Proof We consider a specific trajectory of the method and divide the proof into three standard steps.

Step 1: The Limiting Dynamical System. We denote, by $p^k = (x^k, z^k, u^k)$, $k = 0, 1, 2, \ldots$, a realization of the sequence generated by the algorithm. We introduce the accumulated stepsizes $t_k = \sum_{j=0}^{k-1} \tau_j$, $k = 0, 1, 2 \ldots$, and we construct the interpolated trajectory

$$P_0(t) = p^k + \frac{t - t_k}{\tau_k}(p^{k+1} - p^k), \quad t_k \leq t \leq t_{k+1}, \quad k = 0, 1, 2, \ldots.$$

For an increasing sequence of positive numbers $\{s_k\}$ diverging to infinity, we define shifted trajectories $P_k(t) = P_0(t + s_k)$.

Relations (22), (23), and (24) fit the model of an algorithm analyzed in [36]. Our assumption (A3) is identical to the condition assumed there. Assumption (A1) and Lemma 3.3 guarantee the boundedness of the sequence $\{p^k\}$ and the functions $P_k(\cdot)$. Lemma 3.1 verifies the upper-semicontinuity of the multifunction $\Gamma(\cdot)$, and the convexity and the compactness of its values. All conditions of [36,Thm. 3.2] are thus satisfied. Therefore, by the statement (i) of that theorem, for any infinite set $\mathcal{K}$ of positive integers, there exist an infinite subset $\mathcal{K}_1 \subset \mathcal{K}$ and an absolutely continuous function $P : [0, +\infty) \to X \times \mathbb{R}^n \times \mathbb{R}$ such that for any $T > 0$

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}_1}} \sup_{t\in[0,T]} \left\| P_k(t) - P(t) \right\| = 0,$$

and $P(\cdot) = \big(X(\cdot), Z(\cdot), U(\cdot)\big)$ is a solution of the system of differential equations and inclusions corresponding to (11) and (22):

$$\overset{\bullet}{x}(t) = \bar{y}\big(x(t), z(t)\big) - x(t), \tag{27}$$

$$\big(\overset{\bullet}{z}(t), \overset{\bullet}{u}(t)\big) \in \varGamma(x(t), z(t), u(t)). \tag{28}$$

where $\bar{y}$ is as in (19). Moreover, [36,Thm. 3.2 (ii)] guarantees that for any $t \geq 0$, the triple $\big(X(t), Z(t), U(t)\big)$ is an accumulation point of the sequence $\{(x^k, z^k, u^k)\}$.

In order to analyze the equilibrium points of the system (27)–(28), we first study the dynamics of the functions $H(t) = h(X(t))$ and $F(t) = f(X(t), U(t))$. It follows from (27) that the path $X(\cdot)$ is continuously differentiable. By virtue of assumption (A2) and [49,Thm. 1] (Theorem A.3), for any $J(t) \in \hat{\partial} h(X(t))$,

$$\overset{\bullet}{H}(t) = J(t)\overset{\bullet}{X}(t). \tag{29}$$

Again, Assumption (A2) and Theorem A.3 imply that for any $G(t) \in \hat{\partial} f(X(t), U(t))$,

$$\overset{\bullet}{F}(t) = G_x(t)^\top \overset{\bullet}{X}(t) + G_u(t)\overset{\bullet}{U}(t). \tag{30}$$

To understand the dynamics of $U(\cdot)$, from (28) and (22) we deduce that

$$\overset{\bullet}{U}(t) = \hat{J}_1(t)\overset{\bullet}{X}(t) + b[H(t) - U(t)], \tag{31}$$

with some $\hat{J}_1(t) \in \hat{\partial} h(X(t))$. Therefore, using $J(\cdot) = \hat{J}_1(\cdot)$ in (29), we obtain

$$\overset{\bullet}{U}(t) = \overset{\bullet}{H}(t) + b[H(t) - U(t)]. \tag{32}$$

Consequently, the solution of (30)–(31) has the form:

$$\overset{\bullet}{F}(t) = \hat{G}_1(t)^\top \overset{\bullet}{X}(t) + bG_u(t)[H(t) - U(t)]. \tag{33}$$

with $\hat{G}_1(t) = G_x(t) + \hat{J}_1(t)^\top G_u(t)$. These observations will help us study the stability of the system.

Step 2: Descent Along a Path. We use the Lyapunov function

$$W(x, z, u) = af(x, u) - \eta(x, z) + \gamma \left\| h(x) - u \right\|, \tag{34}$$

with the coefficient $\gamma > 0$ to be specified later.

Directly from (33), we obtain

$$f(X(T),U(T)) - f(X(0),U(0)) = \int_0^T \hat{G}_1(t)^\top \dot{X}(t)\,\mathrm{d}t + b\int_0^T G_u(t)\big[H(t)-U(t)\big]\,\mathrm{d}t. \tag{35}$$

We now estimate the change of $\eta(X(\cdot),Z(\cdot))$ from 0 to T. Since $\bar{y}(x,z)$ is unique, the function $\eta(\cdot,\cdot)$ is continuously differentiable. Therefore, the chain formula holds for it as well:

$$\begin{aligned}&\eta(X(T),Z(T)) - \eta(X(0),Z(0))\\ &\quad= \int_0^T \big\langle \nabla_x \eta(X(t),Z(t)), \dot{X}(t)\big\rangle\,\mathrm{d}t + \int_0^T \big\langle \nabla_z \eta(X(t),Z(t)), \dot{Z}(t)\big\rangle\,\mathrm{d}t.\end{aligned}$$

From (28), we obtain

$$\dot{Z}(t) = a\big(\hat{G}_2(t) - Z(t)\big),$$

with $\hat{G}_2(t) = G_x(t) + \hat{J}_2(t)^\top G_u(t)$ and $\hat{J}_2(t) \in \hat{\partial} h(X(t))$. The function $\eta(x,z)$ defined in (18), as the optimal value of an optimization problem, can be differentiated with respect to the parameters (x,z) at the unique optimal solution $\bar{y}(x,z)$ directly [4,Thm. 4.13]. Substituting $\nabla_x\eta(x,z) = -z + c(x - \bar{y}(x,z))$, $\nabla_z\eta(x,z) = \bar{y}(x,z) - x$, and using the inequality (20) twice, we obtain

$$\begin{aligned}&\eta(X(T),Z(T)) - \eta(X(0),Z(0))\\ &= \int_0^T \big\langle -Z(t) + c(X(t) - \bar{y}(X(t),Z(t)))\,,\ \bar{y}(X(t),Z(t)) - X(t)\big\rangle\,\mathrm{d}t\\ &+ a\int_0^T \big\langle \bar{y}(X(t),Z(t)) - X(t)\,,\ \hat{G}_2(t) - Z(t)\big\rangle\,\mathrm{d}t\\ &\geq\ a\int_0^T \big\langle \bar{y}(X(t),Z(t)) - X(t)\,,\ \hat{G}_2(t) - Z(t)\big\rangle\,\mathrm{d}t\\ &\geq\ a\int_0^T \hat{G}_2(t)^\top\big(\bar{y}(X(t),Z(t)) - X(t)\big)\,\mathrm{d}t + ac\int_0^T \big\|\bar{y}(X(t),Z(t)) - X(t)\big\|^2\,\mathrm{d}t.\end{aligned}$$

With a view at (27), we conclude that

$$\eta(X(T),Z(T)) - \eta(X(0),Z(0)) \geq a\int_0^T \hat{G}_2^\top(t)\dot{X}(t)\,\mathrm{d}t + ac\int_0^T \big\|\dot{X}(t)\big\|^2\,\mathrm{d}t. \tag{36}$$

We now estimate the increment of $\big|H(\cdot) - U(\cdot)\big|$ from 0 to T. As $|\cdot|$ is convex and $H(\cdot)$ and $U(\cdot)$ are absolutely continuous, the chain rule applies as well: for any $\lambda(t) \in \partial|H(t) - U(t)|$, we have

$$\big|H(T)-U(T)\big|-\big|H(0)-U(0)\big|=\int_0^T \lambda(t)\big(\overset{\bullet}{H}(t)-\overset{\bullet}{U}(t)\big)\,\mathrm{d}t.$$

By (32), $\overset{\bullet}{H}(t)-\overset{\bullet}{U}(t)=b\big[U(t)-H(t)\big]$ for almost all t. As $\lambda(t)=\operatorname{sign}\big(H(t)-U(t)\big)$, we obtain

$$\big|H(T)-U(T)\big|-\big|H(0)-U(0)\big|=-b\int_0^T \big|H(t)-U(t)\big|\,\mathrm{d}t. \tag{37}$$

We can now combine (35), (36), and (37) to estimate the change of the function (34):

$$\begin{aligned}&W\big(X(T),Z(T),U(T)\big)-W\big(X(0),Z(0),U(0)\big)\\&\quad\le a\int_0^T G_u(t)(\hat{J}_1(t)-\hat{J}_2(t))\overset{\bullet}{X}(t)\,\mathrm{d}t+ab\int_0^T G_u(t)\big[H(t)-U(t)\big]\,\mathrm{d}t\\&\qquad-ac\int_0^T \big\|\overset{\bullet}{X}(t)\big\|^2\,\mathrm{d}t-b\gamma\int_0^T \big|H(t)-U(t)\big|\,\mathrm{d}t.\end{aligned}$$

It follows from (25) that $\big|G_u(t)\big|\le 1$. Furthermore, $\hat{J}_1(t)\overset{\bullet}{X}(t)=\hat{J}_2(t)\overset{\bullet}{X}(t)=\overset{\bullet}{H}(t)$, by virtue of (29). The last estimate entails:

$$\begin{aligned}&W\big(X(T),Z(T),U(T)\big)-W\big(X(0),Z(0),U(0)\big)\\&\quad\le -ac\int_0^T \big\|\overset{\bullet}{X}(t)\big\|^2\,\mathrm{d}t-b(\gamma-a)\int_0^T \big|H(t)-U(t)\big|\,\mathrm{d}t.\end{aligned} \tag{38}$$

By choosing $\gamma>a$, we ensure that $W(\cdot)$ has the descent property to be used in our stability analysis at Step 3.

Step 3: Analysis of the Limit Points. Define the set

$$\mathcal{S}=\big\{(x,z,u)\in X^*\times\mathbb{R}^n\times\mathbb{R}:\eta(x,z)=0,\ u=h(x)\big\}.$$

Our analysis uses similar ideas to [13, 36], with modifications due to the complex form of our Lyapunov function $W(\cdot)$. As the sequence $\{p^k\}$ is bounded, the quantity

$$L=\liminf_{k\to\infty}\ W(p^k) \tag{39}$$

is finite. Suppose $\bar{p}=(\bar{x},\bar{z},\bar{u})$ is the limit of a convergent subsequence $\{p^{n_k}\}$ such that $L=\lim_{k\to\infty}W(p^{n_k})$. If $\eta(\bar{x},\bar{z})<0$, then $\bar{y}(\bar{x},\bar{z})\neq\bar{x}$ and thus every solution $(X(t),Z(t),U(t))$ of the system (27)–(28), starting from $p(0)=(\bar{x},\bar{z},\bar{u})$ has $\|\overset{\bullet}{X}(0)\|>0$. If $\bar{u}\neq h(\bar{x})$, then $|H(0)-U(0)|>0$. In any case, it follows from (38) that $\delta>0$ exists, such that $\overset{\bullet}{W}(P(0))\le-2\delta$. Therefore, $\tau>0$ exists, such that $W(P(\tau))\le L-\delta\tau$. As $W(P(\tau))$ is also an accumulation point of the sequence $\{p^k\}$ (by [36,Thm. 3.2], already used in Step 1), we obtain a contradiction with (39).

Therefore, we must have $\eta(\bar{x}, \bar{z}) = 0$ and $\bar{u} = h(\bar{x})$. Suppose $\bar{x} \notin X^*$. Then

$$\operatorname{dist}\big(0, \hat{\partial} F(\bar{x}) + N_X(\bar{x})\big) > 0. \tag{40}$$

Suppose the system (27)–(28) starts from $(\bar{x}, \bar{z}, \bar{u})$ and $X(t) = \bar{x}$ for all $t \geq 0$. From (28) and (21), in view of the equations $\bar{y}(\bar{x}, \bar{z}) = \bar{x}$ and $\bar{u} = h(\bar{x})$, we obtain $U(t) = h(\bar{x})$ for all $t \geq 0$. The inclusion (28), in view of (16), simplifies

$$\overset{\bullet}{z}(t) \in a\big(\hat{\partial} F(\bar{x}) - z(t)\big).$$

For the convex Lyapunov function $V(z) = \operatorname{dist}\big(z, \hat{\partial} F(\bar{x})\big)$, we apply the classical chain formula [5] on the path $Z(\cdot)$:

$$V((Z(T)) - V(Z(0)) = \int_0^T \big\langle \partial V(Z(t)), \overset{\bullet}{Z}(t) \big\rangle \, \mathrm{d}t.$$

For $Z(t) \notin \hat{\partial} F(\bar{x})$, we have

$$\partial V(Z(t)) = \frac{Z(t) - \operatorname{Proj}_{\hat{\partial} F(\bar{x})}(Z(t))}{\|Z(t) - \operatorname{Proj}_{\hat{\partial} F(\bar{x})}(Z(t))\|}$$

and $\overset{\bullet}{Z}(t) = a(d(t) - Z(t))$ with some $d(t) \in G(\bar{x})$. Therefore,

$$\big\langle \partial V(Z(t)), \overset{\bullet}{Z}(t) \big\rangle \leq -a\|Z(t) - \operatorname{Proj}_{G(\bar{x})}(Z(t))\| = -aV(Z(t)).$$

It follows that

$$V((Z(T)) - V(Z(0)) \leq -a \int_0^T V(Z(t)) \, \mathrm{d}t,$$

and thus

$$\lim_{t \to \infty} \operatorname{dist}\big(Z(t), \hat{\partial} F(\bar{x})\big) = 0. \tag{41}$$

It follows from (40)–(41) that $T > 0$ exists, such that $-Z(T) \notin N_X(\bar{x})$, which yields $\overset{\bullet}{X}(T) \neq 0$. Consequently, the path $X(t)$ starting from $\bar{x}$ cannot be constant (our supposition made right after (40) cannot be true). But if is not constant, then again $T > 0$ exists, such that $\overset{\bullet}{X}(T) \neq 0$. By [36,Thm. 3.2], already used in Step 1, the triple $(X(T), Z(T), U(T))$ would have to be an accumulation point of the sequence $\{(x^k, z^k, u^k)\}$. We have already excluded the case of $\overset{\bullet}{X}(T) \neq 0$ at the beginning of Step 3, because then $\overset{\bullet}{W}(T) < 0$ and (39) are violated. We conclude that the accumulation point $(\bar{x}, \bar{z}, \bar{u})$, corresponding to L, is in $\mathcal{S}$.

The convergence of the entire sequence $\big\{W(x^k, z^k, u^k)\big\}$ to L then follows in the same way as [36,Thm. 3.5, Steps 3-4]. The proof can be reproduced *verbatim* here.

Our Assumption (A5) corresponds to the condition (ii) of that result and is required at this step of the analysis only. It is the same as [13,Ass. C].

Since every accumulation point $(\bar{x}, \bar{z}, \bar{u})$ of the sequence $\{(x^k, z^k, u^k)\}$ is in $\mathcal{S}$, then $\eta(x^k, z^k) \to 0$ and $h(x^k) - u^k \to 0$. Then, the convergence of $\{f(x^k, u^k)\}$ follows from the convergence of $\left\{W(x^k, z^k, u^k)\right\}$. With $h(x^k) - u^k \to 0$, the sequence $\{F(x^k)\}$ is convergent as well. □

4 Numerical Experiments

In this section, we report results of numerical experiments that illustrate the performance of our single time-scale (STS) method for deep learning and logistic regression. For both applications, we consider perturbations in the training data set, which leads to a distributional shift in the population measure $\mathbb{P}$, whereas we do not perturb the test data. We run the STS algorithm on the contaminated training data and investigate the robustness of the solution found by STS by considering different samples from the test data and the corresponding distribution of the test loss.

Both versions of the method, with stochastic subgradients calculated by (14) or (15), were tested, and both converged to the same solutions of the risk-averse models. Both required similar numbers of iterations as the SGD method for the expected value model.

Our numerical results were obtained using Python (Version 3.7) on an Alienware Aurora R8 desktop with a 3.60 GHz CPU (i7-2677M) and 16GB memory.

4.1 Deep Learning

We consider a fully-connected network on two benchmark datasets: MNIST [33] and CIFAR10 [29], where the model has the depth (the number of layers) of 3 and the width (the number of neurons per hidden layer) of 100. The MNIST dataset consists of black and white images of handwritten digits, with a 28×28 format. It is split into a training dataset of 60,000 examples and a test dataset of 10,000 examples. The CIFAR10 dataset consists of color images of various objects, with a $3 \times 32 \times 32$ format. It is split into a training part of 50,000 examples and a test part of 10,000 examples. The model for the MNIST dataset has 99,710 parameters in total: 78,500 parameters for the first layer, 10,100 parameters for the second layer, 10,100 parameters for the third layer and 1010 parameters for the output layer. The model for the CIFAR10 dataset has 328,510 parameters in total: 307,300 parameters for the first layer, 10,100 parameters for the second layer, 10,100 parameters for the third layer and 1010 parameters for the output layer. For both MNIST and CIFAR10 datasets, the task is to classify the images with an integer label valued from 0 to 9, and we use the cross-entropy loss during training (see reference here). The resulting loss function $\ell(x, D)$ is a composition of the fully-connected network and the cross-entropy loss. We distort the distributions of MNIST and CIFAR10 training datasets by deleting all the data points with the y value equal to 0 (such points account for approximately 10% of the whole dataset). Based on the contaminated data, we train our model with different robustness levels $\varkappa$ for 4000 iterations. To test the robustness of the model found by STS, we sample 100 points from the test dataset and compute the corresponding loss and repeat this procedure 200 times for both datasets to generate a histogram of the test loss. We then report the corresponding cumulative distribution function (CDF) of the test loss in Figs. 1 and 2 for different values of $\varkappa$, compared with results from a model trained by SGD.[4]

If the training data are not contaminated at all, we have observed in our experiments that STS generates a similar or slightly worse solution than SGD. This is expected as STS optimizes a penalized (robust) loss (4), which is different than the empirical loss. The numerical details are omitted for the sake of brevity. On the other hand, when the data suffer from distributional distortion, we see a clear advantage of the STS method over the SGD method.

4.2 Logistic Regression

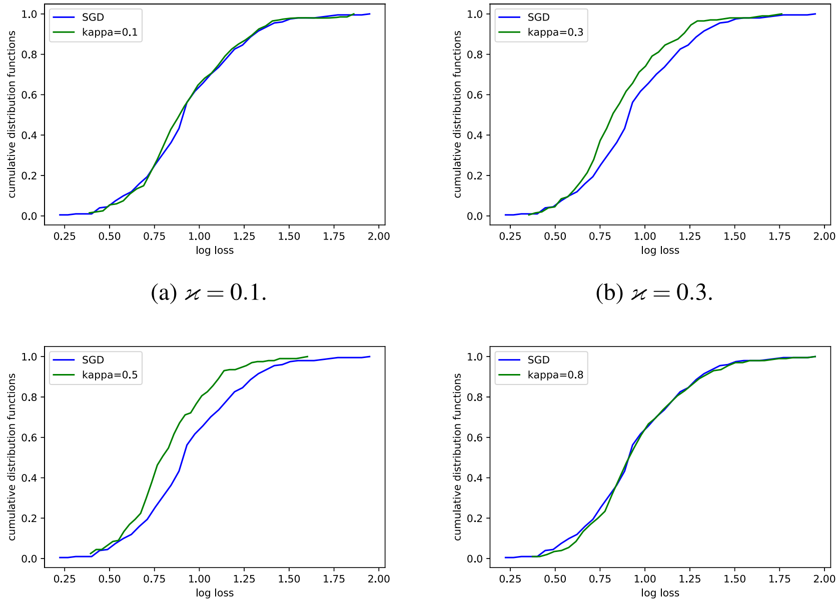
We consider binary logistic regression on the Adult dataset [12] where the loss function has the form $\ell(x, D) = \big[\log(1 + \exp(-b\, a^T x))\big]$ where $D = (a, b)$ is the input data. The problem is to predict whether the annual income of a person will be above \$50,000 or not, based on $n = 123$ predictor variables. The dataset has 32,561 training examples and 16,281 test examples. We follow a similar methodology as before, where we distort the training data by deleting 80% of the data points with the corresponding income below \$50,000. We trained our model with STS and another state-of-the-art method Bandit mirror descent (BMD) developed in [41], allowing both methods to execute the same numbers of iterations, which corresponds to 80,000 iterations of the STS method. We then compare the cdf of the loss of the trained models based on 3000 samples from the test data. Test data are sampled from the original (uncontaminated) data. The results are reported in Fig. 3 for different values of the robustness level $\varkappa$. We see that STS results in smaller errors and conclude that our method has desirable robustness properties with respect to perturbations in the input distribution.

4.3 Remarks on the Assumptions

In the numerical examples, we replace the population quantities with their empirical counterparts. The subgradient noise arises as the subgradients $\tilde{J}^{k+1}$, $\tilde{g}^{k+1} = \begin{bmatrix} \tilde{g}_x^{k+1} \\ \tilde{g}_u^{k+1} \end{bmatrix}$ and the estimate $\tilde{h}^{k+1}$ of the empirical risk are obtained from randomly sampled subset of data points in mini-batches (instead of considering all the data points to estimate the actual subgradients), by formulas (14) or (15).

For logistic regression, similar to [38], we take the feasible set X to be a (closed) Euclidean ball with a radius R chosen large enough to contain the minimum. Clearly, X is convex and compact in this case and Assumption (A1) holds. Since the loss $\ell(x, D)$ is a continuous function of both arguments, x and D, the input data are normalized and bounded, and the iterates stay in the compact set X (where the loss ℓ and its subgradients

[4] There are also adversarial learning methods [25, 31, 35, 64] where the aim is to be resistant to norm-bounded perturbations of the input before we have access to it; however, we do not compare with these methods as our formulation (4) focuses on a distributional distortion.



(a) $\varkappa = 0.1$.

(b) $\varkappa = 0.3$.

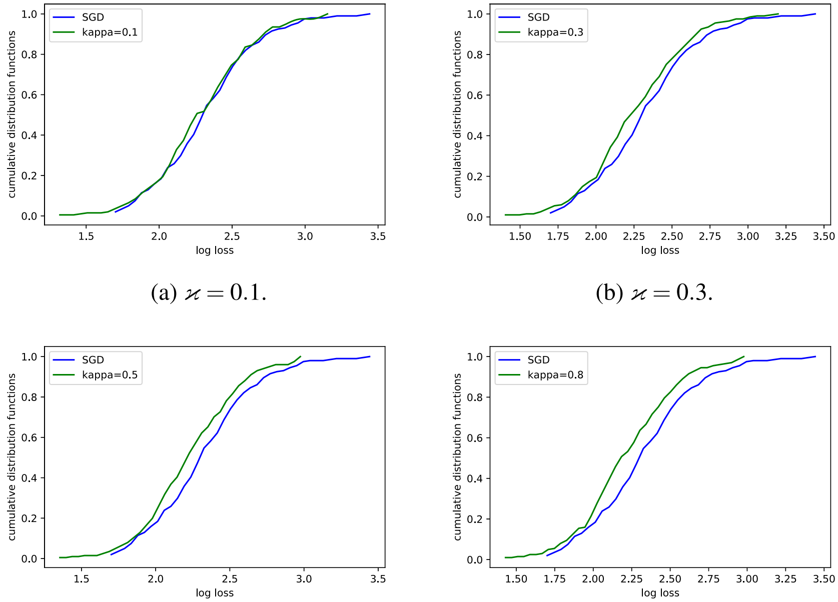
(c) $\varkappa = 0.5$.

(d) $\varkappa = 0.8$.

Fig. 1 The CDFs of the loss of the SGD solution and the STS solution on the test data. Test data is the original (uncontaminated) MNIST data, whereas the models are trained with the contaminated data

with respect to x are bounded), Assumption (A2) holds. Furthermore, we observe from (14) that the sequences $\tilde{J}^{k+1}$, $\tilde{g}^{k+1}$ and $\tilde{h}^{k+1}$ stay uniformly bounded over k; therefore, their variance (conditioned on the natural filtration $\mathcal{F}_k$) is bounded. Moreover, $\tilde{J}^{k+1}$, $\tilde{g}^{k+1}$ and $\tilde{h}^{k+1}$ are unbiased estimates. This is a direct consequence of the fact that D_1^{k+1} and D_2^{k+1} are i.i.d. samples from the empirical data distribution. If we take the expectation of these estimates, as the generalized subdifferentials are bounded, we can interchange the subdifferential and the expectation operators (Theorem A.2). From these observations, we conclude that Assumption (A4) is satisfied. In our experiments, we take $\tau_k = c_1/(1 + c_2 k)$ where c_1, c_2 are positive constants; we choose c_1 small enough so that Assumption (A3) holds. We conclude that all the assumptions (A1)–(A4) hold in our experiments for the logistic regression. The Sard-like assumption (A5) is purely technical; it is hard to imagine a practical problem that violates it.

By a similar reasoning, assumptions (A1)–(A4) hold in the setting of our deep learning experiments as well. The only difference is that in deep learning we chose the feasible set X differently: $X = \{x : \|x\|_\infty \leq B\}$ for $B = 10$. This amounts to requiring that the weights of the hidden units are all bounded. Such constraints are employed frequently in practice for regularization purposes [57]. In practice, with zero initialization and over multiple sample paths of our algorithm, we have never observed the constraint set being violated.



(a) $\varkappa = 0.1$.

(b) $\varkappa = 0.3$.

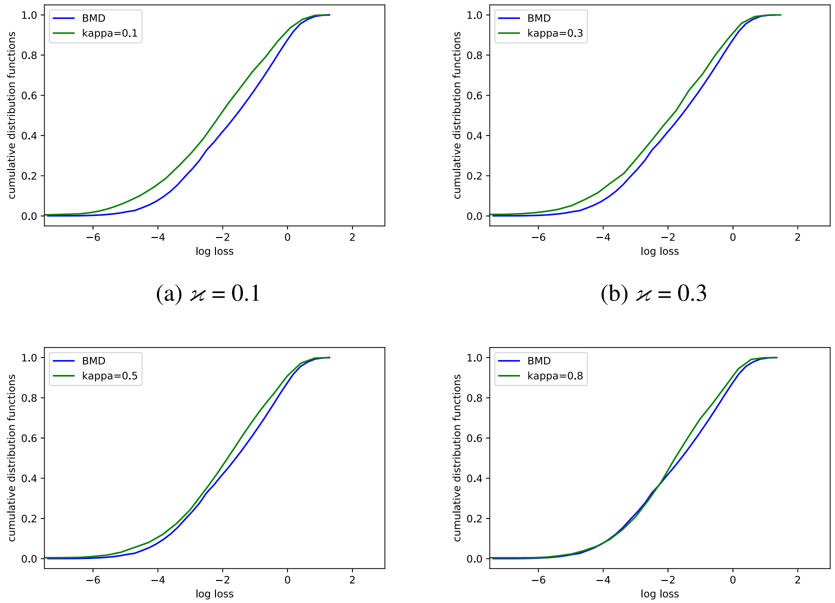
(c) $\varkappa = 0.5$.

(d) $\varkappa = 0.8$.

Fig. 2 The CDFs of the loss of the SGD solution and the STS solution on the test data. Test data is the original (uncontaminated) CIFAR10 data, whereas the models are trained with the contaminated data

5 Contributions

We considered a distributionally robust formulation of stochastic optimization problems arising in statistical learning, where robustness is with respect to ambiguity in the underlying data distribution. We focused on a broad class of generalized differentiable loss functions that can be non-convex and non-smooth, such as those arising in deep learning with ReLU activations. We developed an efficient single-time-scale stochastic subgradient method and showed rigorously that under some assumptions it converges to a point satisfying the optimality conditions with probability one. Our method allows learning predictive models from data while being robust with respect to uncertainty in the underlying data distribution and requires little extra computational effort compared to population risk minimization with stochastic gradient methods. We also provided numerical experiments that illustrates the efficiency of our method on logistic regression and deep learning problems.



(a) $\varkappa = 0.1$

(b) $\varkappa = 0.3$

(c) $\varkappa = 0.5$

(d) $\varkappa = 0.8$

Fig. 3 The CDFs of the loss of the BMD solution and the STS solution on the test data. The test data are the original (uncontaminated) Adult data, whereas the models are trained with the contaminated data

Appendix A: Generalized Differentiability of Functions

Norkin [42] introduced the following class of functions.

Definition A.1 A function $f : \mathbb{R}^n \to \mathbb{R}$ is *differentiable in a generalized sense at a point* $x \in \mathbb{R}^n$, if an open set $U \subset \mathbb{R}^n$ containing x, and a non-empty, convex, compact valued, and upper semicontinuous multifunction $\hat{\partial} f : U \rightrightarrows \mathbb{R}^n$ exist, such that for all $y \in U$ and all $g \in \hat{\partial} f(y)$ the following equation is true:

$$f(y) = f(x) + \langle g(y), y - x \rangle + o(x, y, g),$$

with

$$\lim_{y \to x} \sup_{g \in G(y)} \frac{o(x, y, g)}{\|y - x\|} = 0.$$

The set $\hat{\partial} f(y)$ is the *generalized subdifferential* of f at y. If a function is differentiable in a generalized sense at every $x \in \mathbb{R}^n$ with the same generalized subdifferential mapping $\hat{\partial} f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, we call it *differentiable in a generalized sense*.

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable in a generalized sense, if each of its component functions, $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, has this property.

The class of such functions is contained in the set of locally Lipschitz functions and contains all subdifferentially regular functions [7], Whitney stratifiable Lipschitz functions [11], semismooth functions [39], and their compositions. The Clarke subdifferential $\partial f(x)$ is an inclusion-minimal generalized subdifferential, but the generalized sub-differential mapping $\hat{\partial} f(\cdot)$ is not uniquely defined in Definition A.1. However, if $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable in a generalized sense, then for almost all $x \in \mathbb{R}^n$ we have $\hat{\partial} f(x) = \{\nabla f(x)\}$.

Compositions of generalized differentiable functions are crucial in our analysis.

Theorem A.1 [40,Thm. 1.6] *If $h : \mathbb{R}^m \to \mathbb{R}$ and $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \dots, m$, are differentiable in a generalized sense, then the composition $\psi(x) = h\big(f_1(x), \dots, f_m(x)\big)$ is differentiable in a generalized sense, and at any point $x \in \mathbb{R}^n$ we can define the generalized subdifferential of ψ as follows:*

$$\hat{\partial}\psi(x) = conv\big\{g \in \mathbb{R}^n : g = \big[g_1 \cdots g_m\big]\, g_0,$$
$$with\ g_0 \in \hat{\partial}h\big(f_1(x), \dots, f_m(x)\big)\ and\ g_j \in \hat{\partial} f_j(x),\ \ j = 1, \dots, m\big\}.$$

Even if we take $\hat{\partial}h(\cdot) = \partial h(\cdot)$ and $\hat{\partial} f_j(\cdot) = \partial f_j(\cdot)$, $j = 1, \dots, m$, we may obtain $\hat{\partial}\psi(\cdot) \neq \partial\psi(\cdot)$, but $\hat{\partial}\psi$ defined above satisfies Definition A.1.

For stochastic optimization, essential is the closure of the class functions differentiable in a generalized sense with respect to expectation.

Theorem A.2 [40,Thm. 23.1] *Suppose $(\Omega, \mathcal{F}, P)$ is a probability space and a function $f : \mathbb{R}^n \times \Omega \to \mathbb{R}$ is differentiable in a generalized sense with respect to x for all $\omega \in \Omega$ and integrable with respect to ω for all $x \in \mathbb{R}^n$. Let $\hat{\partial} f : \mathbb{R}^n \times \Omega \rightrightarrows \mathbb{R}^n$ be a multifunction, which is measurable with respect to ω for all $x \in \mathbb{R}^n$, and which is a generalized subdifferential mapping of $f(\cdot, \omega)$ for all $\omega \in \Omega$. If for every compact set $K \subset \mathbb{R}^n$ an integrable function $L_K : \Omega \to \mathbb{R}$ exists, such that $\sup_{x \in K} \sup_{g \in \hat{\partial} f(x,\omega)} \|g\| \le L_K(\omega)$, $\omega \in \Omega$, then the function*

$$F(x) = \int_{\Omega} f(x, \omega)\ P(d\omega), \quad x \in \mathbb{R}^n,$$

is differentiable in a generalized sense, and the multifunction

$$\hat{\partial} F(x) = \int_{\Omega} \hat{\partial} f(x, \omega)\ P(d\omega), \quad x \in \mathbb{R}^n,$$

is its generalized subdifferential mapping.

A key step in the analysis of stochastic recursive algorithms by the differential inclusion method is the *chain rule on a path* (see [9] and the references therein). For an absolutely continuous function $p : [0, \infty) \to \mathbb{R}^n$, we denote by $\overset{\bullet}{p}(\cdot)$ its weak derivative: a measurable function such that

$$p(t) = p(0) + \int_0^t \overset{\bullet}{p}(s)\ \mathrm{d}s, \quad \forall\, t \ge 0.$$

Theorem A.3 [49,Thm. 1] *If a function $f : \mathbb{R}^n \to \mathbb{R}^m$ and a path $p : [0, \infty) \to \mathbb{R}^n$ are differentiable in a generalized sense, then*

$$f(p(T)) - f(p(0)) = \int_0^T g(p(t))\, \overset{\bullet}{p}(t)\, dt, \tag{42}$$

for all selections $g(\cdot) \in \hat{\partial} f(\cdot)$, and all $T > 0$.

Appendix B: Proof of Lemma 3.3

Proof Formula (13) and assumptions (A4)(ii) and (iii) yield:

$$u^{k+1} = u^k + \tau_k\big[J^{k+1}\big(\bar{y}(x^k, z^k) - x^k\big) + b\big(h(x^{k+1}) - u^k\big)\big] + \tau_k\theta_u^{k+1} + \tau_k\epsilon_u^{k+1}, \tag{43}$$

with the errors

$$\theta_u^{k+1} = E^{k+1}\big(\bar{y}(x^k, z^k) - x^k\big) + be_h^{k+1},$$
$$\epsilon_u^{k+1} = \Delta^{k+1}\big(\bar{y}(x^k, z^k) - x^k\big) + b\delta_h^{k+1}.$$

Due to assumption (A4), for some constant C_u^θ,

$$\mathbb{E}\big[\theta_u^{k+1} \,\big|\, \mathcal{F}_k\big] = 0, \quad \mathbb{E}\big[\|\theta_u^{k+1}\|^2 \,\big|\, \mathcal{F}_k\big] \le C_u^\theta, \quad k = 0, 1, \ldots \tag{44}$$

and

$$\lim_{k\to\infty} \epsilon_u^{k+1} = 0 \quad \text{a.s..}$$

To verify the boundedness of $\{u^k\}$, we define the quantities

$$\tilde{u}^k = u^k + \sum_{j=k}^{\infty} \tau_j\theta_u^{j+1}.$$

Owing to (A3) and (44), by virtue of the martingale convergence theorem, the series in the formula above is convergent a.s., and thus, $\tilde{u}^k - u^k \to 0$ a.s., when $k \to \infty$. We can now use (43) to establish the following recursive relation:

$$\tilde{u}^{k+1} = (1 - b\tau_k)\tilde{u}^k + b\tau_k\Big[\frac{1}{b}J^{k+1}\big(\bar{y}(x^k, z^k) - x^k\big) + h(x^{k+1}) + \frac{1}{b}\epsilon_u^{k+1} + (\tilde{u}^k - u^k)\Big].$$

By (A1), the sequences $\{J^k\}$ and $\{h(x^k)\}$ are bounded. Since $\tilde{u}^k - u^k \to 0$ and $\epsilon_u^k \to 0$ a.s., the elements in the brackets in the formula above constitute an almost surely bounded sequence. Consequently, the sequence $\{\tilde{u}^k\}$ of their convex combinations is almost surely bounded as well. The same is true for the sequence $\{u^k\}$, because $\tilde{u}^k - u^k \to 0$ a.s.

The boundedness of $\{z^k\}$ can be established in a similar way. We rewrite (12) as

$$z^{k+1} = z^k + a\tau_k\Big(g_x^{k+1} + \big[J^{k+1}\big]^\top g_u^{k+1} - z^k\Big) + a\tau_k\theta_z^{k+1} + a\tau_k\epsilon_z^{k+1}, \tag{45}$$

with the errors

$$\theta_z^{k+1} = e_{gx}^{k+1} + \big[J^{k+1}\big]^\top e_{gu}^{k+1} + \big[E^{k+1}\big]^\top g_u^{k+1} + \big[E^{k+1}\big]^\top e_{gu}^{k+1},$$
$$\epsilon_z^{k+1} = \delta_{gx}^{k+1} + \big[\tilde{J}^{k+1}\big]^\top \delta_{gu}^{k+1} + \big[\Delta^{k+1}\big]^\top \tilde{g}_u^{k+1}.$$

Due to assumption (A4) (note the statistical independence of E^{k+1} and e_{gu}^{k+1}), for some constant C_z^θ,

$$\mathbb{E}\big[\theta_z^{k+1} \,\big|\, \mathcal{F}_k\big] = 0, \quad \mathbb{E}\big[\|\theta_z^{k+1}\|^2 \,\big|\, \mathcal{F}_k\big] \le C_z^\theta, \quad k = 0, 1, \ldots$$

and

$$\lim_{k\to\infty} \epsilon_z^{k+1} = 0 \quad \text{a.s.}.$$

The remaining proof is the same as that for $\{u^k\}$, with relation (45) replacing (43). □

References

1. Allen-Zhu, Z., Elad, H.: Variance reduction for faster non-convex optimization. In: Maria Florina, B., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning, vol. 48 of Proceedings of Machine Learning Research, pp. 699–707. New York, New York, USA, 20–22 Jun 2016. PMLR
2. Artzner, P., Delbaen, F., Eber, J.-M., Heath, D.: Coherent measures of risk. Math. Finance **9**, 203–228 (1999)
3. Baker, J.W., Schubert, M., Faber, M.H.: On the assessment of robustness. Struct. Safety **30**(3), 253–267 (2008)
4. Bonnans, J.F., Alexander, S.: Perturbation Analysis of Optimization Problems. Springer (2013)
5. Brézis, H.: Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations. In: Contributions to Nonlinear Functional Analysis, pp. 101–156. Elsevier (1971)
6. Bubeck, S.: Convex optimization: Algorithms and complexity. Found. Trends$\mathring{R}$ Mach. Learn. **8**(3–4), 231–357 (2015)
7. Clarke, F.H.: Generalized gradients and applications. Trans. Am. Math. Soc. **205**, 247–262 (1975)
8. Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., Walczak, B.: Robust statistics in data analysis—a review: basic concepts. Chemometr. Intell. Lab. Syst. **85**(2), 203–219 (2007)
9. Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Optim. **29**(1), 207–239 (2019)
10. Dentcheva, D., Penev, S., Ruszczyński, A.: Statistical estimation of composite risk functionals and risk optimization problems. Ann. Inst. Stat. Math. **69**(4), 737–760 (2017)
11. Drusvyatskiy, D., Ioffe, A.D., Lewis, A.S.: Curves of descent. SIAM J. Control Optim. **53**(1), 114–138 (2015)
12. Dheeru, D., Casey, G.: UCI Machine Learning Repository (2017) https://archive.ics.uci.edu/ml/index.php

13. Duchi, J.C., Ruan, F.: Stochastic methods for composite and weakly convex optimization problems. SIAM J. Optim. **28**(4), 3229–3259 (2018)
14. Duchi, J.C., Namkoong, H.: Learning models with uniform performance via distributionally robust optimization. Ann. Stat. **49**(3), 1378–1406 (2021)
15. Ermoliev, Y.M.: Methods of Stochastic Programming. Nauka, Moscow (1976)
16. Ermoliev, Y.M., Norkin, V.I.: Sample average approximation method for compound stochastic optimization problems. SIAM J. Optim. **23**(4), 2231–2263 (2013)
17. Esfahani, P.M., Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Math. Program. **171**(1–2), 115–166 (2018)
18. Föllmer, H., Schied, A.: Stochastic Finance: An Introduction in Discrete Time. Walter de Gruyter (2011)
19. Foster, D.J., Sekhari, A., Sridharan, K.: Uniform convergence of gradients for non-convex learning and optimization. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31, pp. 8745–8756. Curran Associates, Inc. (2018)
20. Gao, R., Chen, X., Kleywegt, A.J.: Wasserstein distributional robustness and regularization in statistical learning (2017). arXiv preprint arXiv:1712.06050
21. Gao, R., Kleywegt, A.J.: Distributionally robust stochastic optimization with Wasserstein distance (2016). arXiv preprint arXiv:1604.02199. https://arxiv.org/pdf/1712.06050.pdf
22. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. SIAM J. Optim. **23**(4), 2061–2089 (2013)
23. Ghadimi, S., Ruszczynski, A., Wang, M.: A single timescale stochastic approximation method for nested stochastic optimization. SIAM J. Optim. **30**(1), 960–979 (2020)
24. Goodfellow, I., Yoshua, B., Aaron, C.: Deep Learning. MIT Press (2016)
25. Goodfellow, I.J, Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2014). arXiv preprint arXiv:1412.6572
26. Hastie, T., Tibshirani, R., Wainwright, M.: The Lasso and Generalizations. CRC Press, Statistical learning with sparsity (2015)
27. Jain, P., Kakade, S.M., Kidambi, R., Netrapalli, P., Sidford, A.: Accelerating stochastic gradient descent for least squares regression. In: Sébastien, B., Vianney, P., Philippe, R. (eds.) Proceedings of the 31st Conference On Learning Theory, vol. 75 of Proceedings of Machine Learning Research, pp. 545–604 (2018) (PMLR, 06–09 Jul 2018)
28. Kalogerias, D.S., Powell, W.B.: Recursive optimization of convex risk measures: mean-semideviation models (2018). arXiv preprint arXiv:1804.00636
29. Krizhevsky, A.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)
30. Kuhn, D., Peyman Mohajerin, E., Viet Anh, N., Soroosh, S.-A.: Wasserstein distributionally robust optimization: theory and applications in machine learning. In: Operations Research & Management Science in the Age of Analytics, pp. 130–166. INFORMS (2019)
31. Kurakin, A., Ian, G., Samy, B.: Adversarial machine learning at scale (2016). arXiv preprint arXiv:1611.01236
32. Kushner, H., Yin, G.G.: Stochastic Approximation Algorithms and Applications. Springer, New York (2003)
33. LeCun, Y.L., Corinna, C., Burges, C.J.: MNIST handwritten digit database. ATT Labs **2** (2010). http://yann.lecun.com/exdb/mnist
34. Li, X., Zhihui, Z., Anthony, M.-C.S., Lee, J.D.: Incremental Methods for Weakly Convex Optimization (2019). arXiv e-prints arXiv:1907.11687
35. Madry, A., Aleksandar, M., Ludwig, S., Dimitris, T., Adrian, V.: Towards deep learning models resistant to adversarial attacks (2017). arXiv preprint arXiv:1706.06083
36. Majewski, S., Miasojedow, B., Moulines, E.: Analysis of nonsmooth stochastic approximation: the differential inclusion approach (20118). arXiv preprint arXiv:1805.01916
37. Mehrotra, S., Zhang, H.: Models and algorithms for distributionally robust least squares problems. Math. Program. **146**(1), 123–141 (2014)
38. Mei, S., Yu, B., Andrea, M.: The landscape of empirical risk for nonconvex losses. Ann. Stat. **46**(6A), 2747–2774 (2018)

39. Mifflin, R.: Semismooth and semiconvex functions in constrained optimization. SIAM J. Control Optim. **15**(6), 959–972 (1977)
40. Mikhalevich, V.S., Gupal, A.M., Norkin, V.I.: Nonconvex Optimization Methods. Nauka, Moscow (1987)
41. Namkoong, H., Duchi, J.C: Stochastic gradient methods for distributionally robust optimization with f-divergences. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 29. Curran Associates, Inc. (2016)
42. Norkin, V.I.: Generalized-differentiable functions. Cybern. Syst. Anal. **16**(1), 10–12 (1980)
43. Ogryczak, W., Ruszczyński, A.: From stochastic dominance to mean-risk models: semideviations as risk measures. Eur. J. Oper. Res. **116**, 33–50 (1999)
44. Ogryczak, W., Ruszczyński, A.: On consistency of stochastic dominance and mean-semideviation models. Math. Program. **89**, 217–232 (2001)
45. Postek, K., den Hertog, D., Melenberg, B.: Computationally tractable counterparts of distributionally robust constraints on risk measures. SIAM Rev. **58**(4), 603–650 (2016)
46. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. 400–407 (1951)
47. Ruszczyński, A.: A linearization method for nonsmooth stochastic programming problems. Math. Oper. Res. **12**(1), 32–49 (1987)
48. Ruszczyński, A., Shapiro, A.: Optimization of convex risk functions. Math. Oper. Res. **31**, 433–452 (2006)
49. Ruszczyński, A.: Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization. Optim. Lett. **14**, 1615–1625 (2020)
50. Ruszczynski, A.: A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization. SIAM J. Control Optim. **59**(3), 2301–2320 (2021)
51. Seidman, J.H., Fazlyab, M., Preciado, V.M., Pappas, G.J.: Robust deep learning as optimal control: Insights and convergence guarantees. In: Bayen, A.M., Jadbabaie, A., Pappas, G., Parrilo, P.A., Benjamin, R., Claire, T., Melanie, Z. (eds.) Proceedings of the 2nd Conference on Learning for Dynamics and Control, vol. 120 of Proceedings of Machine Learning Research, pp. 884–893. PMLR, 10–11 (2020)
52. Soroosh, S.-A., Peyman, M., Esfahani, D.K.: Distributionally robust logistic regression. In: Proceedings of the 28th International Conference on Neural Information Processing Systems—vol. 1. NIPS'15, pp. 1576–1584. Cambridge, MA, USA, 2015. MIT Press (2015)
53. Shai, S.-S., Shai, B.-D.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press (2014)
54. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming: Modeling and Theory. SIAM, Philadelphia (2009)
55. Sinha, A., Hongseok, N., John, D.: Certifying some distributional robustness with principled adversarial training. In: International Conference on Learning Representations (2018). https://openreview.net/forum?id=Hk6kPgZA-
56. Soma, T., Yuichi, Y.: Statistical learning with conditional value at risk (2020). arXiv preprint arXiv:2002.05826
57. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(56), 1929–1958 (2014)
58. Takeda, A., Kanamori, T.: A robust approach based on conditional value-at-risk measure to statistical learning problems. Eur. J. Oper. Res. **198**(1), 287–296 (2009)
59. Teo, C.H., Vishwanthan, S.V.N., Smola, Alex J., Le, Quoc V.: Bundle methods for regularized risk minimization. J. Mach. Learn. Res. **11**(10), 311–365 (2010)
60. Vladimir, V.: The Nature of Statistical Learning Theory. Springer Science & Business Media (2013)
61. Wang, M., Fang, E.X., Liu, B.: Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. Math. Program. **161**(1–2), 419–449 (2017)
62. Wang, M., Liu, J., Fang, E.X.: Accelerating stochastic composition optimization. J. Mach. Learn. Res. **18**, 1–23 (2017)
63. Yang, S., Wang, M., Fang, E.X.: Multilevel stochastic gradient methods for nested composition optimization. SIAM J. Optim. **29**(1), 616–659 (2019)

64. Zhang, D., Tianyuan, Z., Yiping, L., Zhanxing, Z., Bin, D.: You only propagate once: Accelerating adversarial training via maximal principle. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.