FAST AND STABLE CONVERGENCE OF ONLINE SGD FOR CV@R-BASED RISK-AWARE LEARNING

Dionysios S. Kalogerias

Department of Electrical Engineering
Yale University

ABSTRACT

Conditional Value-at-Risk (CV@R) is one of the most popular measures of risk, which has been recently considered as a performance criterion in supervised statistical learning, as it is related to desirable operational features in modern applications, such as safety, fairness, distributional robustness, and prediction error stability. However, due to its variational definition, CV@R is commonly believed to result in difficult optimization problems, even for smooth and strongly convex loss functions. In this work, we disprove this statement by establishing noisy (i.e., fixed-accuracy) linear convergence of stochastic gradient descent for sequential CV@R learning, for a large class of not necessarily strongly-convex (or even convex) loss functions satisfying a set-restricted Polyak-Łojasiewicz inequality. This class contains all smooth and strongly convex losses, confirming that classical problems, such as linear least squares regression, can be solved efficiently under the CV@R criterion, just as their risk-neutral versions. Our results are also illustrated empirically on an indicative risk-aware ridge regression task, verifying their validity.

***Index Terms*—** Statistical Learning, Risk-Aware Learning, Conditional Value-at-Risk, Stochastic Gradient Descent, Polyak-Łojasiewicz Inequality.

1. INTRODUCTION

Risk-awareness is steadily becoming an important aspect in modern statistical learning theory and practice, naturally driven by the need to meet strict reliability requirements in high-stakes, critical applications [1, 2, 3, 4, 5, 6, 7]. In such settings, risk-aware learning formulations are particularly appealing, since they can *explicitly balance* optimal predictor performance between average-case and "difficult" to learn, infrequent, or worst-case examples, providing a form of *statistical robustness* in the learning outcome [8, 9, 10, 4, 11, 12, 13]. The basic idea of risk-aware learning is to replace the standard expected loss objective by more general loss functionals, called *risk measures* [14], whose purpose is to effectively quantify the statistical variability of the particular random loss of choice, in addition to average performance. Popular examples of risk measures include mean-variance functionals [14, 15], mean-semideviations [16], and Conditional Value-at-Risk (CV@R) [17].

CV@R, in particular, plays a significant role in supervised statistical learning, as it is naturally connected not only to prediction error stability (see Section 7), but also to distributional robustness [14, 18, 19], fairness [20], as well as the formulation of classical learning problems, such as the celebrated (ν-)SVM [21, 22, 23]. Relevant generalization bounds were recently reported in [24] and [25], also establishing asymptotic consistency for CV@R learning.

But except for operational effectiveness and generalization performance, *computational methods* for actually obtaining optimal solutions to CV@R learning problems are of paramount importance, especially for practical considerations [12, 18, 19]. The design of such methods is partially facilitated by the variational definition of CV@R ([17], also see Section 2), allowing the reduction of any CV@R learning problem to a standard stochastic problem with a special loss function. This approach was followed recently in [12], where various averaged Stochastic Gradient Descent (SGD)-type algorithms were analyzed. Almost concurrently and under a batch setting (i.e., given a dataset available *a priori*), [18] proposed an adaptive sampling algorithm for CV@R learning, by exploiting the dual representation of CV@R [14]. In both works, convergence rates reported are *at best* of the order of $1/\sqrt{T}$, where T denotes the total runtime of the respective algorithm (iterations).

Such rates might seem to be nearly all we can get: Due to its construction, CV@R is commonly conjectured to result in potentially difficult and challenging stochastic problems, mainly because standard properties which enable fast convergence of gradient methods, such as strong convexity, are *not* preserved when transitioning from (data-driven) risk-neutral to CV@R learning, *even for* smooth and strongly convex losses. In this work, we disprove this argument by showing that SGD attains *noisy (i.e., fixed-tunable-accuracy) linear global convergence* for sequential CV@R learning (i.e., provided a datastream), for a large class of not necessarily strongly-convex (or even convex) loss functions satisfying a new *set-restricted Polyak-Łojasiewicz inequality* [26, 27]. As a byproduct, we also obtain noisy linear convergence of SGD for all smooth and strongly convex losses, since those belong to the aforementioned class.

Within the setting considered, our results confirm that CV@R learning is almost as easy as risk-neutral learning under certain natural conditions which we identify. This implies that CV@R learning can find widespread use in applications, since risk-aware versions of classical problems, such as linear least squares estimation, can be solved as efficiently as their risk-neutral counterparts, and with provable *and* equivalent rate guarantees. Numerical simulations on a basic ridge regression task illustrate the validity of the results.

Note: Proofs to all results presented are omitted and will be presented in a subsequent journal submission currently under preparation; the interested reader is also referred to the preprint [28].

2. CV@R STATISTICAL LEARNING

Let $\mathcal{P}_{\mathcal{D}}$ be an *unknown* distribution over an *example space* $\mathcal{D} \triangleq \mathbb{R}^d \times \mathbb{R}$, and consider a *hypotheses class* $\mathcal{F} \triangleq \{\phi : \mathbb{R}^m \to \mathbb{R} | \phi(\cdot) \equiv f(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^m\}$. We are interested in discovering a hypothesis or predictor $f(\cdot, \boldsymbol{\theta}^*) \in \mathcal{F}$ that best approximates y when presented with $\boldsymbol{x}$, where $(\boldsymbol{x}, y) \sim \mathcal{P}_{\mathcal{D}}$. The quality of every predictor $f(\cdot, \boldsymbol{\theta})$ is captured by a loss $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ taking, for each example $(\boldsymbol{x}, y)$, the quantities $f(\boldsymbol{x}, \boldsymbol{\theta})$ and y and mapping them to an integrable random variable, $\ell(f(\boldsymbol{x}, \boldsymbol{\theta}), y)$. Posing the fitting problem

$$\inf_{\boldsymbol{\theta} \in \mathbb{R}^m} \left[\mathbb{E}_{\mathcal{P}_{\mathcal{D}}} \{\ell(f(\boldsymbol{x}, \boldsymbol{\theta}), y)\} \equiv \int_{\mathcal{D}} \ell(f(\boldsymbol{x}, \boldsymbol{\theta}), y) \mathrm{d}\mathcal{P}_{\mathcal{D}}(\boldsymbol{x}, y) \right], \quad (1)$$

is then standard and at the heart of machine learning and beyond, such as signal processing, statistics, and control.

Despite its wide popularity, though, a fundamental issue with the gold standard expected loss is its very nature: It is *risk-neutral*, i.e., it minimizes losses *only* on average. Because of this, it lacks robustness and essentially ignores *relatively infrequent but statistically significant* example instances, treating them as inconsequential. This is important from a practical point of view, since such more "difficult" or "extreme" examples will incur high and/or undesirable instantaneous losses, *even if* the optimal prediction error has minimal expected value [8, 14, 16, 5, 18, 12, 13].

As briefly explained in Section 1, the need for a systematic treatment of the shortcomings of the risk-neutral approach motivates and sets the premise of *risk-aware statistical learning*, in which expectation is replaced by more general loss functionals, called risk measures [14]. Their purpose is to induce risk-averse characteristics into the learning outcome by explicitly controlling the statistical variability of the random loss $\ell(f(\boldsymbol{x},\cdot),y)$, or, equivalently, its tail behavior. By far one of the most popular risk measures in theory and practice is CV@R, which for an integrable random loss Z is defined as [17]

$$\mathrm{CV@R}^{\alpha}(Z) \triangleq \inf_{t\in\mathbb{R}} \Big\{ t + \frac{1}{\alpha}\mathbb{E}\{(Z-t)_+\}\Big\}, \tag{2}$$

at confidence level $\alpha \in (0,1]$. Intuitively, $\mathrm{CV@R}^{\alpha}(Z)$ is the *mean of the worst* $(\alpha\cdot 100)\%$ of the values of Z, and is a strict generalization of expectation; in particular, it is true that

$$\mathrm{CV@R}^{1}(Z) \equiv \mathbb{E}\{Z\} \leq \mathrm{CV@R}^{\alpha}(Z), \forall \alpha \in (0,1], \text{ and} \tag{3}$$

$$\mathrm{CV@R}^{0}(Z) \triangleq \lim_{\alpha\downarrow 0} \mathrm{CV@R}^{\alpha}(Z) \equiv \operatorname{ess\,sup} Z. \tag{4}$$

One of the most important properties of CV@R is that it constitutes a *coherent* risk measure, meaning that it is a *convex*, *monotone*, *translation equivariant* and *positively homogeneous* functional of its argument; see ([14], Section 6.3).

By setting $Z \equiv \ell(f(\boldsymbol{x},\boldsymbol{\theta}),y), \boldsymbol{\theta}\in\mathbb{R}^m$, we may now formulate the CV@R statistical learning problem as

$$\boxed{\inf_{\boldsymbol{\theta}\in\mathbb{R}^m} \mathrm{CV@R}^{\alpha}_{\mathcal{P}_{\mathcal{D}}}[\ell(f(\boldsymbol{x},\boldsymbol{\theta}),y)].} \tag{5}$$

Observe that due to its defining properties, the CV@R problem is most intuitive, and allows for a *tunable* tradeoff between risk neutrality (for $\alpha\equiv 1$), and minimax robustness (as $\alpha\downarrow 0$). Additionally, because CV@R is a coherent risk measure, it follows that problem (5) is convex whenever $\ell(f(\boldsymbol{x},\cdot),y)$ is convex for each $(\boldsymbol{x},y)$, and strongly convex whenever $\ell(f(\boldsymbol{x},\cdot),y)$ is strongly convex for each $(\boldsymbol{x},y)$ [16]. Thus, problem (5) is favorably structured.

However, because CV@R is itself defined as the optimal value of a stochastic program, it is difficult to evaluate analytically, especially in a data-driven setting. Still, we may leverage the definition of CV@R and reformulate (5) as a risk-neutral stochastic program over *both* variables $(\boldsymbol{\theta},t)$ as

$$\boxed{\inf_{(\boldsymbol{\theta},t)\in\mathbb{R}^m\times\mathbb{R}} \mathbb{E}_{\mathcal{P}_{\mathcal{D}}}\Big\{ t + \frac{1}{\alpha}(\ell(f(\boldsymbol{x},\boldsymbol{\theta}),y)-t)_+\Big\}.} \tag{6}$$

Although problem (6) can now be tackled using standard methods of stochastic optimization, the structural benefits of the CV@R functional are largely gone: For instance, although it is true that (6) is convex whenever the composition $\ell(f(\boldsymbol{x},\cdot),y)$ is convex, it *might not* be strongly convex, even if $\ell(f(\boldsymbol{x},\cdot),y)$ is. This is important, because it would imply that classical setups, such as linear least squares, might result in badly behaving CV@R problems, for $\alpha\in(0,1)$. Of course, those issues can only get worse in a nonconvex setting, e.g., when the function f is a Deep Neural Network (DNN).

Nevertheless, it is intuitive that, due to the close relationship between problems (5) and (6), the good behavior of the former should carry through to the latter, and classical solution strategies, such as SGD, should exhibit good performance. Here we show that this is indeed the case.

3. CV@R STOCHASTIC GRADIENT DESCENT

Since $\mathcal{P}_{\mathcal{D}}$ is unknown, problem (1) (cf. (6)) is impossible to solve *a priori*. Instead, one should rely on *observable* example pairs (data). Here we are assuming a sequential setting, where a *stream of data* $\{(\boldsymbol{x}^n,y^n)\}_{n=0}^{\infty}$ is available, and the focus is on solving (1) via stochastic approximation, i.e., by applying the standard stochastic gradient descent algorithm to the equivalent CV@R problem (6). The sequential setting conforms with numerous real-time applications, and is standard in stochastic optimization. Throughout, we impose usual assumptions on the composition $\ell(f(\boldsymbol{x},\cdot),y)$.

Assumption 1. *Unless the function* $\ell(f(\boldsymbol{x},\cdot),y)$ *is convex on* $\mathbb{R}^m$ *for* $\mathcal{P}_{\mathcal{D}}$*-almost all* $(\boldsymbol{x},y)$*, then for each* $\boldsymbol{\theta}\in\mathbb{R}^m$*:*

1. $\ell(f(\boldsymbol{x},\cdot),y)$ *is* $C_{\boldsymbol{\theta}}(\boldsymbol{x},y)$*-Lipschitz on a neighborhood* $\boldsymbol{\theta}$ *for* $\mathcal{P}_{\mathcal{D}}$*-almost all* $(\boldsymbol{x},y)$*, and* $\mathbb{E}_{\mathcal{P}_{\mathcal{D}}}\{C_{\boldsymbol{\theta}}(\boldsymbol{x},y)\}<\infty$.
2. $\ell(f(\boldsymbol{x},\cdot),y)$ *is differentiable at* $\boldsymbol{\theta}$ *for* $\mathcal{P}_{\mathcal{D}}$*-almost all* $(\boldsymbol{x},y)$*, and* $\mathcal{P}_{\mathcal{D}}(\ell(f(\boldsymbol{x},\boldsymbol{\theta}),y)=t)\equiv 0$ *for all* $(\boldsymbol{\theta},t)\in\mathbb{R}^m\times\mathbb{R}$.

For convenience, let us define, for $(\boldsymbol{\theta},t)\in\mathbb{R}^m\times\mathbb{R}$,

$$G_{\alpha}(\boldsymbol{\theta},t) \triangleq \mathbb{E}_{\mathcal{P}_{\mathcal{D}}}\Big\{ t + \frac{1}{\alpha}(\ell(f(\boldsymbol{x},\boldsymbol{\theta}),y)-t)_+\Big\}. \tag{7}$$

Then it may be shown that, under Assumption 1, differentiation may be interchanged with expectation for G_{α} ([14], Section 7.2.4), yielding, for every $(\boldsymbol{\theta},t)$, the (sub)gradient representation

$$\nabla G_{\alpha}(\boldsymbol{\theta},t) = \begin{bmatrix} \frac{1}{\alpha}\mathbb{E}_{\mathcal{P}_{\mathcal{D}}}\{\mathbf{1}_{\mathcal{A}(\boldsymbol{\theta},t)}(\boldsymbol{x},y)\nabla_{\boldsymbol{\theta}}\ell(f(\boldsymbol{x},\boldsymbol{\theta}),y)\} \\ -\frac{1}{\alpha}\mathbb{E}_{\mathcal{P}_{\mathcal{D}}}\{\mathbf{1}_{\mathcal{A}(\boldsymbol{\theta},t)}(\boldsymbol{x},y)\}+1 \end{bmatrix}, \tag{8}$$

where for brevity and for later use we have defined the *event-valued* multifunction $\mathcal{A}:\mathbb{R}^m\times\mathbb{R}\rightrightarrows\mathcal{D}$ as

$$\mathcal{A}(\boldsymbol{\theta},t) \triangleq \{(\boldsymbol{x},y)\in\mathcal{D} | \ell(f(\boldsymbol{x},\boldsymbol{\theta}),y)-t>0\}, \tag{9}$$

for $(\boldsymbol{\theta},t)\in\mathbb{R}^m\times\mathbb{R}$. We note that, for each $(t,\boldsymbol{\theta})$, the set $\mathcal{A}(t,\boldsymbol{\theta})$ contains all examples corresponding to the *positive section* of the function $\ell(f(\bullet,\boldsymbol{\theta}),\cdot)-t$.

Leveraging (8), and provided an independent and identically distributed datastream $\{(\boldsymbol{x}^n,y^n)\}_{n=0}^{\infty}$, we can now outline the simplest and most obvious scheme for possibly tackling the CV@R problem (6), i.e., the standard SGD rule, described via the recursive updates

$$t^{n+1} = t^n - \gamma\Big[1-\frac{1}{\alpha}\mathbf{1}_{\mathcal{A}(\boldsymbol{\theta}^n,t^n)}(\boldsymbol{x}^{n+1},y^{n+1})\Big] \quad \text{and} \tag{10}$$

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n - \beta\frac{1}{\alpha}\mathbf{1}_{\mathcal{A}(\boldsymbol{\theta}^n,t^n)}(\boldsymbol{x}^{n+1},y^{n+1}) \times \nabla_{\boldsymbol{\theta}}\ell(f(\boldsymbol{x}^{n+1},\boldsymbol{\theta}^n),y^{n+1}), \tag{11}$$

where $n\in\mathbb{N}$ is an iteration index, $\beta>0$ and $\gamma>0$ are constant stepsizes, and where $(\boldsymbol{\theta}^0,t^0)$ are appropriately chosen initial values.

We observe that the SGD updates (10) and (11) can be regarded as a modification of the standard risk-neutral SGD (solving (1)), but where learning happens *if and only if* $\ell(f(\boldsymbol{x}^{n+1},\boldsymbol{\theta}^n),y^{n+1})-t^n\geq 0$, for each n. The update in t controls the frequency of learning, as well as the proportion of examples that participate in learning. Also note that if $\alpha\equiv 1$, then t^n is nonincreasing, and therefore $\boldsymbol{\theta}^n$ should approach a risk-neutral solution. In the following, we suggestively refer to the algorithm comprised by (10) and (11) as CV@R-SGD.

4. POLYAK-ŁOJASIEWICZ CONDITIONS

We next present the standard Polyak-Łojasiewicz (PŁ) inequality, first appeared in [26].

Definition 1. (PŁ [26]) We say that a function $\varphi:\mathbb{R}^L\to\mathbb{R}$ satisfies the *Polyak-Łojasiewicz (PŁ) inequality with parameter* $\mu>0$ on $\Sigma\subseteq\mathbb{R}^L$, if and only if φ is differentiable on Σ and, for every $\boldsymbol{x}\in\Sigma$,

$$\frac{1}{2}\|\nabla\varphi(\boldsymbol{x})\|_2^2\geq\mu(\varphi(\boldsymbol{x})-\varphi^\star), \tag{12}$$

where $\varphi^\star\triangleq\inf_{\boldsymbol{x}\in\Sigma}\varphi(\boldsymbol{x})$.

In a recent seminal article [27], the PŁ inequality was exploited to show linear convergence of gradient methods under multiple interesting and useful setups. Further, [27] shows that strong convexity implies the PŁ inequality, but also that there are lots of *nonconvex* functions obeying the PŁ inequality. This indeed implies that S(GD) converges *globally and linearly* for such functions.

For our purposes, unfortunately, the standard PŁ inequality (Definition 1) will not suffice. Instead, we introduce and rely on a generalization, which we call the *set-restricted PŁ inequality*, as follows.

Definition 2. (Set-Restricted PŁ) Consider a measurable function $\varphi:\mathbb{R}^L\times\mathbb{R}^M\to\mathbb{R}$, a Borel-valued multifunction $\mathcal{B}:\mathbb{R}^L\rightrightarrows\mathbb{R}^M$, and a probability measure $\mathcal{M}$ on $\mathscr{B}(\mathbb{R}^M)$. We say that φ satisfies the (diagonal) *$\mathcal{B}$-restricted Polyak-Łojasiewicz (PŁ) inequality with parameter* $\mu>0$, relative to $\mathcal{M}$ and on a subset $\Sigma\subseteq\mathbb{R}^L$, if and only if $\varphi(\cdot,\boldsymbol{w})$ is subdifferentiable on Σ for $\mathcal{M}$-almost every $\boldsymbol{w}\in\mathbb{R}^M$, and it is true that, for every $\boldsymbol{z}\in\Sigma$,

$$\frac{1}{2}\|\mathbb{E}_{\mathcal{M}}\{\nabla_{\boldsymbol{z}}\varphi(\boldsymbol{z},\boldsymbol{w})|\mathcal{B}(\boldsymbol{z})\}\|_2^2 \geq\mu\mathbb{E}_{\mathcal{M}}\{\varphi(\boldsymbol{z},\boldsymbol{w})-\varphi^\star(\boldsymbol{z})|\mathcal{B}(\boldsymbol{z})\}, \tag{13}$$

where $\varphi^\star(\cdot)\triangleq\inf_{\widetilde{\boldsymbol{z}}\in\Sigma}\mathbb{E}_{\mathcal{M}}\{\varphi(\widetilde{\boldsymbol{z}},\boldsymbol{w})|\mathcal{B}(\cdot)\}$.

Although admittedly somewhat mysterious at first sight, the set-restricted PŁ inequality is basically the same as the classical PŁ inequality [27], with the important difference that expectation is replaced by conditional expectation relative to an event *varying* in the argument of the function involved (i.e., an event-valued multifunction). In fact, the set-restricted PŁ inequality quantifies the curvature of the loss surface by restricting attention on sets of learning examples that matter (in Definition 2, $\mathcal{B}$ plays this role).

The usefulness of the set-restricted PŁ inequality stems from the fact that, interestingly, it is satisfied by every strongly convex smooth function, as the next result suggests.

Proposition 1. (Strong Convexity $\implies$ Set-Restricted PŁ) *Suppose that the loss $\ell(f(\boldsymbol{x},\cdot),y)$ is L-smooth and μ-strongly convex for $\mathcal{P}_{\mathcal{D}}$-almost all $(\boldsymbol{x},y)$. Then, for every pair $(\boldsymbol{\theta},\mathcal{B})\in\mathbb{R}^m\times\mathscr{B}(\mathcal{D})$ such that $\mathcal{P}_{\mathcal{D}}(\mathcal{B})>0$, it is true that*

$$\frac{1}{2}\|\mathbb{E}\{\nabla_{\boldsymbol{\theta}}\ell(f(\boldsymbol{x},\boldsymbol{\theta}),y)|\mathcal{B}\}\|_2^2\geq\mu\mathbb{E}\{\ell(f(\boldsymbol{x},\boldsymbol{\theta}),y)-\ell^\star(\mathcal{B})|\mathcal{B}\}, \tag{14}$$

where $\ell^\star(\mathcal{B})\equiv\inf_{\widetilde{\boldsymbol{\theta}}}\mathbb{E}\{\ell(f(\boldsymbol{x},\widetilde{\boldsymbol{\theta}}),y)|\mathcal{B}\}$.

From Proposition 1, it follows that every smooth strongly convex loss satisfies the set-restricted PŁ inequality relative to *every* qualifying event-valued multifunction of choice. For instance, in the notation of Proposition 1, one may set $\mathcal{B}\equiv\mathcal{A}(\boldsymbol{\theta},t)$, for every fixed pair $(\boldsymbol{\theta},t)$. This choice is particularly important, as we will see in the next section.

5. LINEAR CONVERGENCE OF CV@R-SGD

Hereafter, let $\{\mathscr{D}_n\}_{n\in\mathbb{N}}$ be the history (i.e., filtration) generated by CV@R-SGD and the available datastream. Our main result follows, showing linear convergence of CV@R-SGD under the $\mathcal{A}$-restricted PŁ inequality, in particular.

Theorem 1. (Linear Convergence of CV@R-SGD) *Fix $\alpha\in(0,1)$, let Assumption 1 be in effect and suppose that, for a set $\Delta\equiv\Delta_m\times[-\infty,\bar{t}]$, with $\Delta_m\subseteq\mathbb{R}^m$, it holds that $(\boldsymbol{\theta}^*,t^*)\in\arg\min_\Delta G_\alpha(\boldsymbol{\theta},t)\neq\emptyset$, and that the loss $\ell(f(\boldsymbol{x},\cdot),y)$ obeys the $\mathcal{A}$-restricted PŁ inequality with parameter $\mu>0$ relative to $\mathcal{P}_{\mathcal{D}}$ on Δ. Further, for fixed $T\in\mathbb{N}$, let γ be small enough such that*

$$\mathbb{E}_n\{t^{n+1}|\mathscr{D}_n\}\geq t^n+2\gamma\mu(t^*-t^n)_+,\quad\forall n\in\mathbb{N}_T. \tag{15}$$

As long as $\Delta_T\triangleq\{\boldsymbol{\theta}^n,t^n\}_{n\in\mathbb{N}_T}\subseteq\Delta$, G_α is $L\equiv L_\alpha$-smooth on Δ_T, and $2\mu\min\{\beta,\gamma\}<1$, it is true that

$$\boxed{\begin{aligned}&\mathbb{E}\big\{G_\alpha(\boldsymbol{\theta}^{T+1},t^{T+1})-G_\alpha(\boldsymbol{\theta}^*,t^*)\big\}\\&\quad\leq(1-2\mu\min\{\beta,\gamma\})^T\big(G_\alpha(\boldsymbol{\theta}^0,t^0)-G_\alpha(\boldsymbol{\theta}^*,t^*)\big)\\&\qquad+\frac{(\max\{\beta,\gamma\})^2}{\min\{\beta,\gamma\}}\frac{L(1+C_T^2)}{4\alpha^2\mu},\end{aligned}} \tag{16}$$

where $\sup_{n\in\mathbb{N}_T}\mathbb{E}\{\|\nabla_{\boldsymbol{\theta}}\ell(f(\boldsymbol{x}^{n+1},\boldsymbol{\theta}^n),y^{n+1})\|_2^2\}\leq C_T^2$.

Some remarks regarding the assumptions and conclusions of Theorem 1 are essential at this point. First, we should discuss the existence of an appropriate γ satisfying condition (15), which is of central importance in the proof the theorem. Indeed, if we assume that there are choices of $\varepsilon>0$ and γ such that, for every $n\in\mathbb{N}_T$, the inequality

$$\alpha\Big(1+\frac{\varepsilon}{\gamma}\Big)\leq\mathcal{P}_{\mathcal{D}}(\mathcal{A}(\boldsymbol{\theta}^n\equiv\boldsymbol{\theta}_{\alpha}^n,t^n\equiv t_{\alpha,\gamma}^n)) \tag{17}$$

is satisfied, then it may be shown that (15) will be (conservatively) satisfied as long as [28]

$$\frac{\alpha\varepsilon}{1-\alpha}\leq\gamma<\frac{\varepsilon}{2\mu(t^*-l)+1}, \tag{18}$$

where $l\in\mathbb{R}$ denotes the lowest possible value of the loss function under consideration, provided such value exists. Note that conditions (17) and (18) can indeed be satisfied for particular choices of ε and γ when α is small enough.

Although these dependencies might seem fairly restrictive, they are very reasonable, since in order for CV@R-SGD to converge fast, the condition $\ell(f(\boldsymbol{x}^{n+1},\boldsymbol{\theta}^n),y^{n+1})-t^n\geq 0$ needs to be satisfied sufficiently often. But all this is reasonable from a practical perspective as well: If α is closer to 1 (risk-neutral setting), risky events are effectively smoothened, whereas, if α approaches zero, only rare events matter and an essentially robust solution is sought, which does not really exhibit the dynamic character of a risk-aware solution. Therefore, depending on the problem, α should be chosen modestly, providing *both* non-trivial results *and* fast linear convergence; from

a conceptual point of view, there is a certain logical *balance to be respected between moderatism and conservatism*.

Second, since G_α depends on the choice of the CV@R level α, it is expected that the smoothness parameter L will be dependent on α as well. We discuss this issue further in Section 6.

Third, by combining Proposition 1 with Theorem 1, it follows that CV@R-SGD *converges linearly to fixed, user-tunable accuracy whenever* $\ell(f(\boldsymbol{x},\cdot),y)$ *is strongly convex and smooth for every* $(\boldsymbol{x},y)$, even though G_α might not be strongly convex (only smoothness of G_α is required). This is nice, because it shows that classical problems, such as linear least squares regression, can *provably* be solved most efficiently using SGD under risk-aware performance criteria, i.e., the CV@R, just as their risk-neutral counterparts (e.g., via the celebrated Least-Mean-Squares (LMS) algorithm for linear least squares problems) –also see our Section 7 later on.

6. ENFORCING SMOOTHNESS

There are two potential issues associated with the CV@R problem (6) and the assumptions ensuring linear convergence of CV@R-SGD, as suggested in Theorem 1. The *first* is that there are useful cases where the demand that $\mathcal{P}_{\mathcal{D}}(\ell(f(\boldsymbol{x},\bullet),y)=(\cdot))\equiv 0$ on $\mathbb{R}^m\times\mathbb{R}$ (see Assumption 1.2) might not be satisfied; this happens, e.g., in classification problems where the hypothesis class $\mathcal{F}$ contains hard classifiers, i.e., functions with binary or discrete range. The *second* issue is that the smoothness assumption on G_α, essential to obtain the rate promised by Theorem 1, might not be easy to verify or even hold by merely assuming that the loss $\ell(f(\boldsymbol{x},\cdot),y)$ is smooth; this is due to the presence of the indicator $\mathbf{1}_{\mathcal{A}(\bullet,\cdot)}(\boldsymbol{x},y)$ next to $\nabla_{\boldsymbol{\theta}}\ell(f(\boldsymbol{x},\bullet),y)$ in (8). It turns out that these two issues are related, and both may be mitigated by a rather simple strategy, which we now outline. For more details, the reader is referred to [28].

Consider an *augmented example* $(\boldsymbol{x},y,w)$, where $w\sim\mathcal{N}(0,\sigma^2)$, $\sigma^2>0$, is a *fictitious target*, independent of $(\boldsymbol{x},y)$, which we choose to use *adversarially* during the training process. In particular, we do that by defining the *surrogate loss* $\widetilde{\ell}:\mathbb{R}\times\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ as

$$\widetilde{\ell}(f(\boldsymbol{x},\boldsymbol{\theta}),y,w)\triangleq\ell(f(\boldsymbol{x},\boldsymbol{\theta}),y)-w, \tag{19}$$

Although such a surrogate loss is meaningless in the risk-neutral setting (since $\mathbb{E}\{w\}\equiv 0$), it *provides regularization* in risk-aware and, in particular, CV@R statistical learning. In fact, it can be easily shown that, by choosing $\widetilde{\ell}$ as the loss, Assumption 1.2 is always satisfied, and the resulting objective function G_α in problem (6) is L'-smooth whenever $\ell(f(\boldsymbol{x},\cdot),y)$ is C-Lipschitz and L-smooth, with

$$L'\equiv L'_\alpha\equiv(L\sigma\sqrt{2\pi}+C^2)/(\alpha\sigma\sqrt{2\pi}). \tag{20}$$

Further, we have *uniform* estimates in $(\boldsymbol{\theta},t)$

$$\begin{aligned}G_\alpha(\boldsymbol{\theta},t)\leq\widetilde{G}_\alpha(\boldsymbol{\theta},t)&\triangleq\mathbb{E}_{\mathcal{P}_{\widetilde{\mathcal{D}}}}\Big\{t+\frac{1}{\alpha}(\widetilde{\ell}(f(\boldsymbol{x},\boldsymbol{\theta}),y,w)-t)_+\Big\}\\&\leq G_\alpha(\boldsymbol{\theta},t)+\sigma(\alpha\sqrt{2\pi})^{-1},\end{aligned} \tag{21}$$

where $\widetilde{\mathcal{D}}\triangleq\mathbb{R}^d\times\mathbb{R}\times\mathbb{R}$. Then, similarly to Theorem 1, we obtain linear convergence up to fixed accuracy

$$\frac{(\max\{\beta,\gamma\})^2}{\min\{\beta,\gamma\}}\frac{(1+C_T^2)}{4\alpha^2\mu}\frac{L\sigma\sqrt{2\pi}+C^2}{\alpha\sigma\sqrt{2\pi}}+\frac{\sigma}{\alpha\sqrt{2\pi}}, \tag{22}$$

which by proper choice of σ results in a quantity of the order of

$$\Big(\sqrt{(\max\{\beta,\gamma\})^2/\min\{\beta,\gamma\}}\Big)/\alpha^2. \tag{23}$$

Lastly, observe that the smoothness parameter of G_α in Theorem 1, here L', depends on the CV@R level α, as would be expected.

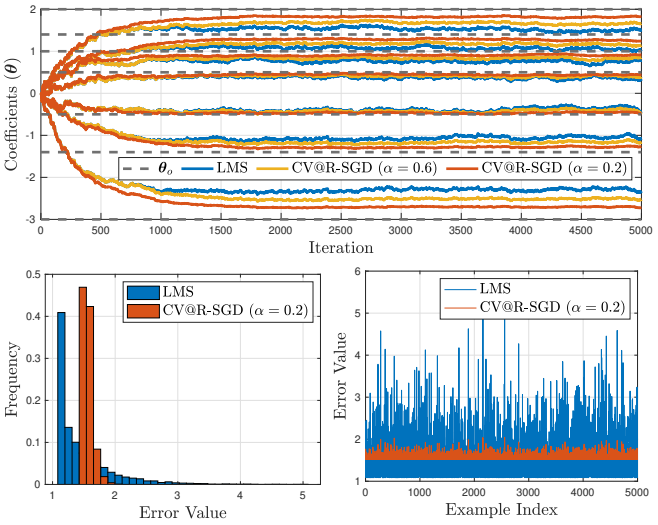

Fig. 1. Comparison between risk-neutral (LMS) and risk-aware (CV@R-SGD) ridge regression. Top: Evolution of iterates $\{\boldsymbol{\theta}^n\}_n$. Bottom: Histogram (left) and actual values (right) of the test error.

7. AN INDICATIVE NUMERICAL EXAMPLE

In this section, we numerically demonstrate the behavior of CV@R-SGD, confirming the validity of Theorem 1. To this end, we consider the λ-strongly convex, risk-aware ridge regression problem

$$\inf_{\boldsymbol{\theta}\in\mathbb{R}^m}\text{CV@R}^{\alpha}_{\mathcal{P}_{\mathcal{D}}}\left[(y-\langle\boldsymbol{\theta},\boldsymbol{x}\rangle)^2+\lambda\|\boldsymbol{\theta}\|_2^2\right], \tag{24}$$

where $y\equiv\langle\boldsymbol{\theta}_o,\boldsymbol{x}\rangle\in\mathbb{R}$ for constant $\boldsymbol{\theta}_o\in\mathbb{R}^7$ and with the elements of $\boldsymbol{x}\in\mathbb{R}^7$ being independent uniform in $[0,2]$, $\lambda\equiv 0.1$ and for two values of α, i.e., $\alpha\equiv 0.2$ and $\alpha\equiv 0.6$. Our goal is to find a $\boldsymbol{\theta}^*$ which minimizes the mean of the worst 20% (if $\alpha=0.2$) or 60% (if $\alpha=0.6$) of all possible values of the random error $(y-\langle\cdot,\boldsymbol{x}\rangle)^2+\lambda\|\cdot\|_2^2$. Note that, for $\alpha\equiv 1$, problem (24) reduces to ordinary ridge regression, and may be solved via the LMS algorithm, which in this special case coincides with the CV@R-SGD algorithm by setting $t^0=0$ (implying that $t^n=0$, for all $n\in\mathbb{N}$, since $\alpha\equiv 1$).

Fig. 1 shows the iterate evolution as well as the behavior of the optimal prediction (test) error for both CV@R-SGD (with stepsizes $\beta\equiv\alpha\times 0.01$ and $\gamma\equiv 0.001$) and the LMS scheme (with stepsize $\beta\equiv 0.01$). We observe that both algorithms converge at an essentially identical *noisy linear rate*, in line with Theorem 1. However, the solutions are radically different. In fact, the risk-aware solutions discovered by CV@R-SGD dramatically reduce the volatility of prediction error, and improve prediction stability. Although this apparently comes at the cost sacrificing mean performance, such sacrifice is fully user-customizable by varying the CV@R level α.

8. CONCLUSION

In this work, we established noisy linear convergence of SGD for sequential CV@R learning, for a large class of possibly nonconvex loss functions satisfying a set-restricted PŁ inequality, also including all smooth and strongly convex losses as special cases. This result disproves the belief that CV@R learning is fundamentally difficult, and shows that classical learning problems can be solved efficiently under CV@R criteria, just as their risk-neutral versions. Our theory was also illustrated via an indicative numerical example. Future work includes the consideration of special learning settings such as linear least squares, as well as other risk measures beyond CV@R.

9. REFERENCES

[1] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, oct 2018.

[2] 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*, vol. 5870, no. c, pp. 1–1, 2018.

[3] 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*, vol. 4, no. 3, pp. 2493–2500, jul 2019.

[4] A. R. Cardoso and H. Xu, "Risk-Averse Stochastic Convex Bandit," in *International Conference on Artificial Intelligence and Statistics*, apr 2019, vol. 89, pp. 39–47.

[5] A. Koppel, A. S. Bedi, and K. Rajawat, "Controlling the Bias-Variance Tradeoff via Coherent Risk for Robust Learning with Kernels," in *Proceedings of the American Control Conference*. jul 2019, vol. 2019-July, pp. 3519–3525, Institute of Electrical and Electronics Engineers Inc.

[6] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, and P. Popovski, "Risk-Based Optimization of Virtual Reality over Terahertz Reconfigurable Intelligent Surfaces," in *IEEE International Conference on Communications*. jun 2020, vol. 2020-June, Institute of Electrical and Electronics Engineers Inc.

[7] Y. Li, D. Guo, Y. Zhao, X. Cao, and H. Chen, "Efficient Risk-Averse Request Allocation for Multi-Access Edge Computing," *IEEE Communications Letters*, pp. 1–1, sep 2020.

[8] A. Takeda and T. Kanamori, "A Robust Approach Based on Conditional Value-at-Risk Measure to Statistical Learning Problems," *European Journal of Operational Research*, vol. 198, no. 1, pp. 287–296, oct 2009.

[9] 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*. dec 2018, vol. 2018-Janua, pp. 4928–4933, IEEE.

[10] C. A. Vitt, D. Dentcheva, and H. Xiong, "Risk-Averse Classification," *Annals of Operations Research*, aug 2019.

[11] L. Zhou and P. Tokekar, "An Approximation Algorithm for Risk-Averse Submodular Optimization," in *Springer Proceedings in Advanced Robotics, vol. 14*, pp. 144–159. Springer, Cham, dec 2020.

[12] T. Soma and Y. Yoshida, "Statistical Learning with Conditional Value at Risk," *arXiv preprint, arXiv:2002.05826*, feb 2020.

[13] M. Gürbüzbalaban, A. Ruszczyński, and L. Zhu, "A Stochastic Subgradient Method for Distributionally Robust Non-Convex Learning," *arXiv preprint, arXiv:2006.04873*, jun 2020.

[14] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, Society for Industrial and Applied Mathematics, 2nd edition, 2014.

[15] H. Markowitz, "Portfolio Selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, mar 1952.

[16] D. S. Kalogerias and W. B. Powell, "Recursive Optimization of Convex Risk Measures: Mean-Semideviation Models," *arXiv preprint, arXiv:1804.00636*, apr 2018.

[17] R. T. Rockafellar and S. Uryasev, "Optimization of Conditional Value-at-Risk," *Journal of Risk*, vol. 2, pp. 21–41, 2000.

[18] S. Curi, K. Y. Levy, S. Jegelka, and A. Krause, "Adaptive Sampling for Stochastic Risk-Averse Learning," *Advances in Neural Information Processing Systems*, vol. 2020-Decem, oct 2020.

[19] Y. Laguel, J. Malick, and Z. Harchaoui, "First-order Optimization for Superquantile-based Supervised Learning," in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, sep 2020.

[20] R. C. Williamson and A. K. Menon, "Fairness Risk Measures," in *36th International Conference on Machine Learning, ICML 2019*, 2019, vol. 2019-June, pp. 11763–11774.

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

[22] A. Takeda and M. Sugiyama, "ν-Support Vector Machine as Conditional Value-at-Risk Minimization," in *Proceedings of the 25th International Conference on Machine Learning*, New York, New York, USA, 2008, pp. 1056–1063, Association for Computing Machinery (ACM).

[23] J. Y. Gotoh and A. Takeda, "CVaR Minimizations in Support Vector Machines," in *Financial Signal Processing and Machine Learning*, pp. 233–265. John Wiley & Sons, Ltd, Chichester, UK, apr 2016.

[24] Z. Mhammedi, B. Guedj, and R. C. Williamson, "PAC-Bayesian Bound for the Conditional Value at Risk," *arXiv preprint, arXiv:2006.14763*, jun 2020.

[25] J. Lee, S. Park, and J. Shin, "Learning Bounds for Risk-sensitive Learning," *arXiv preprint, arXiv:2006.08138*, jun 2020.

[26] B. T. Polyak, "Gradient Methods for Minimizing Functionals," *USSR Computational Mathematics and Mathematical Physics*, vol. 3, no. 4, pp. 864–878, jan 1963.

[27] H. Karimi, J. Nutini, and M. Schmidt, "Linear Convergence of Gradient and Proximal-Gradient Methods under the Polyak-Łojasiewicz Condition," in *Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2016, Lecture Notes in Computer Science*. 2016, vol. 9851 LNAI, pp. 795–811, Springer Verlag.

[28] D. S. Kalogerias, "Noisy Linear Convergence of Stochastic Gradient Descent for CV@R Statistical Learning under Polyak-Łojasiewicz Conditions," *arXiv preprint, arXiv:2012.07785*, 2020.