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

Dionysis Kalogerias

Department of Electrical Engineering – Yale University

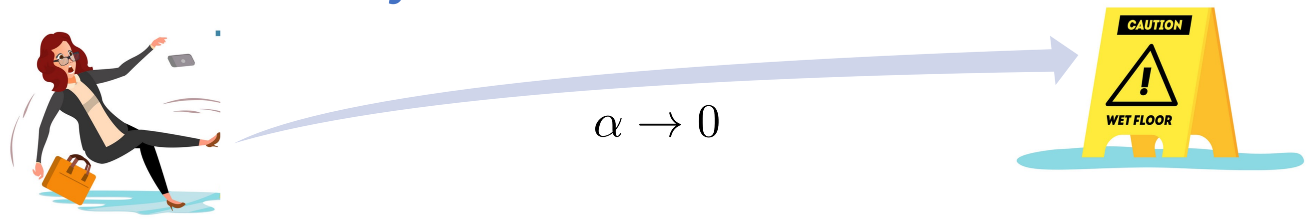
CV@R Risk-Aware Learning

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

- Definition for $\alpha \in (0,1]$:

$$\text{CV@R}^{\alpha}(Z) \triangleq \inf_{t\in\mathbb{R}} \left\{ t + \frac{1}{\alpha}\mathbb{E}\{(Z-t)_+\} \right\}$$

- Assess **tail loss events, not only mean losses**
- **Intuitive tradeoff between** risk neutrality and minimax robustness



Problem Formulation

- Reformulation as a risk-neutral program

$$\inf_{(\boldsymbol{\theta},t)\in\mathbb{R}^m\times\mathbb{R}} \left[G_{\alpha}(\boldsymbol{\theta},t) \triangleq \mathbb{E}_{\mathcal{P}_{\mathcal{D}}}\left\{ t + \frac{1}{\alpha}(\ell(f(\boldsymbol{x},\boldsymbol{\theta}),y)-t)_+ \right\} \right]$$

- **Structural benefits of the original CV@R problem are gone!**
- E.g., strong convexity of the loss does not imply strong convexity of the reformulated problem.
- Standard $\mathcal{O}(1/\sqrt{T})$ rates seem to be all we can get (prior work)
- Still, it is expected standard SGD schemes should work well.

- **Is this the case? Under which conditions?**

CV@R-SGD Algorithm

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

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

$$t^{n+1} = t^n - \gamma\left[1 - \frac{1}{\alpha}\mathbf{1}_{\mathcal{A}(\boldsymbol{\theta}^n,t^n)}(\boldsymbol{x}^{n+1},y^{n+1})\right]$$

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

Technical Framework

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 of $\boldsymbol{\theta}$ 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 $(\boldsymbol{\theta},t)\in\mathbb{R}^m\times\mathbb{R}$.*

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 $\boldsymbol{z}\in\Sigma$,*

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

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

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*

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

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

Main Result

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*

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

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

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

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

Numerical Example

- We consider the quadratic loss

$$\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}),y) = (y - \boldsymbol{\theta}^T\boldsymbol{x})^2 + \lambda\|\boldsymbol{\theta}\|^2$$

where $y = \boldsymbol{\theta}_o^T\boldsymbol{x}$.

- **Risk-aware ridge regression problem**

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

$$\boldsymbol{\theta}_o \in \mathbb{R}^7, \quad \boldsymbol{x}\in\mathbb{R}^7 \text{ indep. unif. in } [0,2], \quad \lambda \equiv 0.1$$