An Online Adaptive Stochastic DCA via Sharp SAA Convergence Rates for Subdifferential Mappings

Yuhan Ye[1] Ying Cui[2] Jingyi Wang[3]

[1]Peking University; incoming PhD student, MIT CCSE [2]University of California, Berkeley [3]Lawrence Livermore National Laboratory

Background: SAA for Subgradients

Nonsmooth optimization problem:

$$\min_{x\in\mathbb{R}^d} \mathbb{E}_\omega[\varphi(x,\omega)] \tag{1}$$

- $\varphi(\cdot,\omega)$ is regular: (weakly) convex, (locally) Lipschitz continuous.
- $\tau(x,\omega)\in\partial_x\varphi(x,\omega)$ is a subgradient selector.
- Given $\omega^1,\dots,\omega^n \overset{\text{i.i.d.}}{\sim} \omega$, $\frac{1}{n}\sum_{k=1}^n \tau\left(x,\omega^k\right)$ is the **Sample Average Approximation (SAA)** for the subgradient of $\mathbb{E}_\omega[\varphi(x,\omega)]$.

In general, stochastic subgradient methods rely on subgradient selectors whose expectations are valid: $\mathbb{E}_\omega\left[\tau\left(x,\omega^k\right)\right]\in\partial\mathbb{E}_\omega[\varphi(x,\omega)]$.

Challenge: Set-valued Subdifferentials

- When φ is smooth at x, $\partial_x\varphi\left(x,\omega^k\right)=\{\nabla_x\varphi(x,\omega^k)\}$.
  - Unbiased: $\mathbb{E}_\omega[\nabla_x\varphi(x,\omega)]=\nabla\mathbb{E}_\omega[\varphi(x,\omega)]$;
  - Classic variance reduction rate: $\mathbb{E}_{\bar\omega^n}\left|\frac{1}{n}\sum_{k=1}^n\nabla_x\varphi(x,\omega^k)-\nabla\mathbb{E}_\omega[\varphi(x,\omega)]\right|^2\le\frac{\sigma^2}{n}$.
- When φ is nonsmooth at x, $\partial_x\varphi(x,\omega)$ is set-valued.
  - $\mathbb{E}_\omega\partial_x\varphi(x,\omega)$ is the set of $\mathbb{E}_\omega\left[\tau\left(x,\omega^k\right)\right]$ over all integrable selection;
  - $\mathbb{E}_\omega$ and ∂_x are interchangeable when $\varphi(\cdot,\omega)$ is Clarke regular.
- **The problem is:**
  - $\mathbb{E}_\omega[\tau(x,\omega)]\in\partial\mathbb{E}_\omega[\varphi(x,\omega)]$ only if $\tau(x,\cdot)$ is measurable[a];
  - Such measurable selectors may be difficult to compute.

SAA Convergence Rate for Subdifferential Mappings

- Define the SAA error by the Hausdorff distance:

$$\triangle_n(\varphi,x,\bar\omega^n)\triangleq\mathbb{H}\left(\frac{1}{n}\sum_{i=1}^n\partial_x\varphi\left(x,\omega^i\right),\mathbb{E}_\omega\partial_x\varphi(x,\omega)\right),$$

where $\mathbb{H}(A,C)\triangleq\max\{\mathbb{D}(A,C),\mathbb{D}(C,A)\}$, $\mathbb{D}(A,C)\triangleq\sup_{x\in A}\operatorname{dist}(x,C)$.

- **Existing work:**
  - $O(\sqrt[4]{d/n})$ uniform rate for the gradients of the Moreau envelopes;[b]
  - $O(\sqrt{d/n})$ uniform rate under convex-smooth composite structure and further subgaussian assumptions on distributions.[c]

[a] F. H. Clarke. *Optimization andonsmooth Analysis*. SIAM, 1990.

[b] D. Davis and D. Drusvyatskiy, "Graphical Convergence of Subgradients in Nonconvex Optimization and Learning," *Mathematics of Operations Research*, vol. 47, no. 1, pp. 209–231, 2022.

[c] F. Ruan, "Subgradient Convergence Implies Subdifferential Convergence on Weakly Convex Functions: With Uniform Rate Guarantees," *arXiv preprint arXiv:2405.10289*, 2024.

Our Result: A novel $O(\sqrt{d/n})$ Pointwise Rate

- We derive a **clean $O(\sqrt{d/n})$ pointwise** convergence rate (modulo logarithmic factors), almost matching the smooth case.

Theorem

If $\varphi(\cdot,\omega)$ is (weakly) convex and Lipschitz continuous with Lipschitz constant L_φ uniformly in ω, for any $\alpha\in(0,1/2)$, $\alpha'\in(\alpha,1/2)$, we have

$$\sup_{x\in\mathcal{D}_\varphi}\mathbb{E}_{\bar\omega^n}\left[\triangle_n(\varphi,x,\bar\omega^n)\right]\le\frac{\hat c}{n^\alpha},\ \text{and}\ \sup_{x\in\mathcal{D}_\varphi}\mathbb{E}_{\bar\omega^n}\left[\triangle_n(\varphi,x,\bar\omega^n)^2\right]\le\frac{c}{n^{2\alpha}},$$

where $c\triangleq\hat c\left(\hat c+L_\varphi\frac{\sqrt{\alpha'}}{\sqrt{2(\alpha'-\alpha)e}}\right)+L_\varphi^2$, $\hat c\triangleq\sqrt{d}(2L_\varphi+L_\varphi/\sqrt{(1-2\alpha')e})$.

- Useful for convergence analysis in stochastic nonsmooth optimization.
- **Sketch of proof:**

1. Transform the Hausdorff distance of set-valued subdifferentials into the SAA error of support functions by the following lemma.

   Lemma.

   $$\triangle_n(\varphi,x,\bar\omega^n)=\max_{\|u\|\leqslant1}\left|\frac{1}{n}\sum_{i=1}^n\sigma\left(u,\partial_x\varphi\left(x,\omega^i\right)\right)-\mathbb{E}_\omega\left[\sigma\left(u,\partial_x\varphi(x,\omega)\right)\right]\right|,$$

   where $\sigma(u,S)\triangleq\sup_{s\in S}u^Ts$.
2. There is an $O(\sqrt{d/n})$ convergence rate (modulo logarithmic factors) for the SAA error of $\sigma\left(u,\partial_x\varphi(x,\omega)\right)$, since $\sigma\left(\cdot,\partial_x\varphi(x,\omega)\right)$ are bounded and Lipschitz continuous in $u\in\mathbb{B}(0,1)$ uniformly. [a]

Application: Stochastic DC Optimization

Online decision-making with stochastic difference-of-convex objective:

$$\min_{x\in C} f(x)\triangleq\underbrace{\mathbb{E}_{\xi\sim P_\xi}[G(x,\xi)]}_{\triangleq g(x)}-\underbrace{\mathbb{E}_{\zeta\sim P_\zeta}[H(x,\zeta)]}_{\triangleq h(x)}. \tag{2}$$

- Assumptions for functions and distributions:
  1. The feasible set C is convex and closed, $f(x)$ is bounded below;
  2. For all ξ,ζ, $G(\cdot,\xi)$ and $H(\cdot,\zeta)$ are convex and L_1-Lipschitz continuous;
  3. For all $x\in C$, $G(x,\cdot)$ and $H(x,\cdot)$ are L_2-Lipschitz continuous;
  4. **The underlying data-generating distribution is time-varying:** At time t, samples are drawn from $P_{\xi,t}$ and $P_{\zeta,t}$, which may differ from the true distributions P_ξ and P_ζ but converge to them over time in terms of the cumulative Wasserstein-1 distance:

$$\sum_{t=1}^\infty W_1(P_{\xi,t},P_{\xi,t-1})<\infty,\quad\sum_{t=1}^\infty W_1(P_{\zeta,t},P_{\zeta,t-1})<\infty.$$

[a] This result is derived from the Rademacher average of function families, see Y. M. Ermoliev and V. I. Norkin, "Sample Average Approximation Method for Compound Stochastic Optimization Problems," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2231–2263, 2013., and Ying Cui and Jong-Shi Pang, *Modern Nonconvex Nondifferentiable Optimization*, SIAM, 2021.

Our Work: An Online Adaptive Stochastic Proximal DCA

Online: Instead of aggregating past samples to compute current solutions, the algorithm is **robust to distributional shifts** during iterations.

Algorithm 1 The ospDCA framework

1: Initialize $x_0,\mu_0,N_{g,0},N_{h,0}$.
2: **for** $t=0,1,2,\dots$ **do**
3: Generate i.i.d. samples $S_{g,t}=\{\xi^{t,i}\}_{i=1}^{N_{g,t}}$ and $S_{h,t}=\{\zeta^{t,i}\}_{i=1}^{N_{h,t}}$ from $P_{\xi,t}$ and $P_{\zeta,t}$, which are **independent** of the past samples.
4: Set $\bar g_t(x)=\frac{1}{N_{g,t}}\sum_{i=1}^{N_{g,t}}G(x,\xi^{t,i})$, $\bar h_t(x)=\frac{1}{N_{h,t}}\sum_{i=1}^{N_{h,t}}H(x,\zeta^{t,i})$, and select $\bar y_t\in\partial\bar h_t(x_t)=\frac{1}{N_{h,t}}\sum_{i=1}^{N_{h,t}}\partial_xH(x_t,\zeta^{t,i})$.
5: Solve the convex subproblem to obtain $\bar d_t$:

$$\begin{aligned}&\underset{d}{\text{minimize}}\ \ \bar g_t(x_t+d)-\bar h_t(x_t)-\bar y_t^Td+\frac{1}{2}\mu_t\|d\|^2\\&\text{subject to}\ \ x_t+d\in C.\end{aligned}$$

6: Set $x_{t+1}=x_t+\bar d_t$.
7: Update $\mu_{t+1},N_{g,t+1},N_{h,t+1}$ **adaptively**.
8: **end for**

- **An adaptive sampling strategy:**

Given sample size upper bound sequence $\{\hat N_{g,t}\}$ and $\{\hat N_{h,t}\}$ which satisfy $\sum_{t\ge0}\left(\hat N_{h,t}^{-\alpha_h}+\hat N_{g,t}^{-\alpha_g}\right)<\infty$, predetermined proximal parameters $\{\mu_t\}$ with upper bound $\hat\mu$ and lower bound $\check\mu$, update $N_{g,t+1}$ and $N_{h,t+1}$ such that one of the followings stands:

1. $\left(\mu_t-\frac{\check\mu}{2}\right)\left\|\bar d_t\right\|^2\ge\frac{C_g}{\mu_{t+1}N_{g,t+1}^{\alpha_g}}+\frac{C_h}{(2\rho_g+2\rho_h+\check\mu)N_{h,t+1}^{\alpha_h}}$;
2. $N_{g,t+1}\ge\hat N_{g,t+1}$, and $N_{h,t}\ge\hat N_{h,t+1}$.

Why adaptive sampling?

- Far from a critical point: *cheap, low-accuracy* estimates suffice.
- Near a critical point: *higher accuracy* is essential for convergence.

Convergence Property and Sample Sizes Requirement

- It converges to DC critical points along a subsequence, almost surely.
- The sample size **matches the results achieved in the smooth case** under static distributions.

Method	Assumption		Sample size	
	Convex part	Concave part	Convex part	Concave part
Previous work[a]	Nonsmooth	Nonsmooth	$O(k^2)$	$O(k^2)$
	Nonsmooth	Smooth	$O(k^2)$	$O(k)$
Ours	**Nonsmooth**	**Nonsmooth**	$O(k^2)$	$O(k)$

[a] Le Thi, Hoai An, Luu, Hoang Phuc Hau, and Dinh, Tao Pham. "Online Stochastic DCA with Applications to Principal Component Analysis." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 5, 2024, pp. 7035–7047.

Application: Online Sparse Robust Regression

$$\min_{\beta\in\mathbb{R}^d}\mathbb{E}_{(x,y)\sim\mathcal{D}_t}\left[|y-\langle\beta,x\rangle|\right]+\lambda\sum_{j=1}^d\min(1,\alpha|\beta_j|).$$

- $\{(x_i,y_i)\}_{i=1}^\infty$ is drawn from **unknown and varying** distributions $\mathcal{D}_t$.
- The capped-ℓ_1 penalty $\sum_{j=1}^d\min(1,\alpha|\beta_j|)$ approximates the ℓ_0-norm.
- DC decomposition:

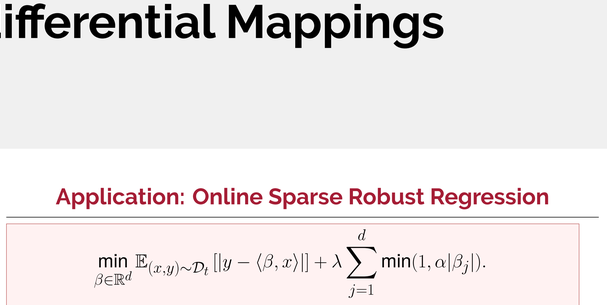
$$\min_{\beta\in\mathbb{R}^d}\mathbb{E}_{(x,y)\sim\mathcal{D}_t}\left[G(\beta,x,y)\right]-h(\beta),$$

  - $G(\beta,x,y)=|y-\langle\beta,x\rangle|+\lambda\sum_{j=1}^d\left(1+\alpha|\beta_j|\right)$,
  - $h(\beta)=\sum_{j=1}^d\max(1,\alpha|\beta_j|)$.

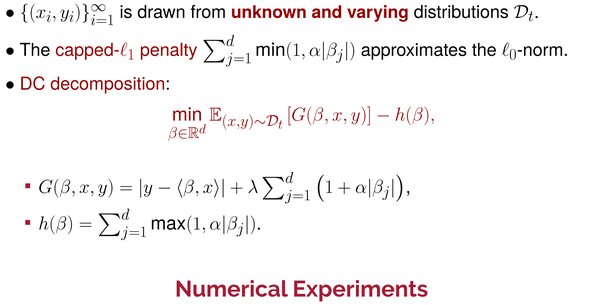
Numerical Experiments

Experiment setup:

- x_t is sampled uniformly from $[-1,1]^d$.
- $y_t=x_t^\top(\beta_{\text{opt}}+\delta_t)+\varepsilon$, where $\varepsilon\sim N(0,1)$, δ_t is the distribution shift.
- We set $\delta_t=(-1)^t100t^{-2}\mathbf{1}_d$, since $W_1(\mathcal{D}_t,\mathcal{D}_{t+1})\le\|\delta_t-\delta_{t+1}\|_1$.



(a) $d=50$, $\beta_{\text{opt}}=[10,-15,0,0,\cdots,0]$.



(b) $d=200$, $\beta_{\text{opt}}=[10,-15,0,0,\cdots,0]$.

yyh03@mit.edu