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

**Yuhan Ye**
Peking University; incoming PhD student, MIT CCSE

Joint work with **Professor Ying Cui** (UC Berkeley) and
**Jingyi Wang** (Lawrence Livermore National Laboratory)

International Conference on Machine Learning (ICML)
July 2025

# 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 $\omega^1,\ldots,\omega^n \overset{\text{i.i.d.}}{\sim} \omega$, $\frac{1}{n}\sum_{k=1}^n \tau\left(x,\omega^k\right)$ 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 $\varphi$ 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:

$$\mathbb{E}_{\bar{\omega}^n} \left| \frac{1}{n} \sum_{k=1}^{n} \nabla_x \varphi(x, \omega^k) - \nabla \mathbb{E}_\omega[\varphi(x, \omega)] \right|^2 \leq \frac{\sigma^2}{n}.$$

- When $\varphi$ 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 $\partial_x$ are interchangeable when $\varphi(\cdot, \omega)$ is Clarke regular.
- **The problem is:**
  - $\mathbb{E}_\omega\left[\tau\left(x, \omega\right)\right] \in \partial \mathbb{E}_\omega\left[\varphi(x, \omega)\right]$ only if $\tau(x, \cdot)$ is measurable[1];
  - Such measurable selectors may be difficult to compute.

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

# Convergence Rate for the SAA of Subdifferential Mappings

- Define the SAA error for $\partial\varphi(x, \cdot) : \Omega \to 2^{\mathbb{R}^d}$ by the Hausdorff distance:

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

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

- $\tau(x, \cdot)$ no longer needs to be measurable if $\Delta_n$ is bounded well.
- **Existing work:**
  - $O(\sqrt[4]{d/n})$ uniform rate for the gradients of the Moreau envelopes.[2]
  - $O(\sqrt{d/n})$ uniform rate under convex-smooth composite structure and further subgaussian assumptions on distributions.[3]

[2] 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.

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

# Our Result

- 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* $L_\varphi$ *uniformly in* $\omega$, *for any* $\alpha \in (0, 1/2)$, $\alpha' \in (\alpha, 1/2)$, *we have*

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

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

- This is 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**

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

*where* $\sigma(u, S) \triangleq \sup_{s \in S} u^T s$.

2. There is an $O(\sqrt{d/n})$ convergence rate (modulo logarithmic factors) for the SAA error of $\sigma\left(u, \partial_x \varphi\left(x, \omega\right)\right)$, since $\sigma\left(\cdot, \partial_x \varphi\left(x, \omega\right)\right)$ are bounded and Lipschitz continuous in $u \in \mathbb{B}(0, 1)$ uniformly. [4]

[4] 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.

# Some Details of the Lemma

- **Proof technique: analyze through support functions.**

**Lemma**

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

*where* $\sigma(u, S) \triangleq \sup_{s \in S} u^T s$.

Some key points:

- $\sigma(u, S) = \sigma(u, \operatorname{conv} S)$.
- $\sigma(u, S + S') = \sigma(u, S) + \sigma(u, S')$.
- Hömander's formula[5]: $\mathbb{D}(A, B) = \max_{\|u\| \leqslant 1}(\sigma(u, A) - \sigma(u, B))$, where $A$ and $B$ are nonempty convex and compact subsets of $\mathbb{R}^p$.
- $\mathbb{E}_\omega$ and $\sigma$ are interchangeable: $\mathbb{E}_\omega\left[\sigma\left(u, \partial_x \varphi(x, \omega)\right)\right] = \sigma\left(u, \mathbb{E}_\omega\left[\partial_x \varphi(x, \omega)\right]\right)$ [6].

---

[5] C. Castaing and M. Valadier. "Measurable multifunctions." In: *Convex Analysis and Measurable Multifunctions*. Springer, Berlin, Heidelberg, 1977, pp. 59–90.

[6] N. S. Papageorgiou. "On the theory of Banach space valued multifunctions. I. Integration and conditional expectation." *Journal of Multivariate Analysis*, 17(2):185–206, 1985.

# Application: Stochastic DC Optimization

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

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

1. The feasible set $C$ is convex and closed, $f(x)$ is bounded below;
2. For all $\xi, \zeta$, $G(\cdot,\xi)$ 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 $P_\xi$ and $P_\zeta$ but converge to them over time in terms of the cumulative Wasserstein-1 distance:

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

## An Online Adaptive Stochastic Proximal DCA

- **Online**:

  The method is robust to distribution shifts since it never aggregates stale samples.

- **Adaptive**:

  Both sample and step sizes are set from current estimates of the stochastic quantities.

**Why adaptive sampling?**

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

**Algorithm** 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 $\bar{y}_t \in \partial \bar{h}_t(x_t) = \frac{1}{N_{h,t}} \sum_{i=1}^{N_{h,t}} \partial_x H(x_t, \zeta^{t,i})$.
5: Solve the convex subproblem to obtain $\bar{d}_t$:

$$\begin{aligned} \underset{d}{\text{minimize}} \quad & \bar{g}_t(x_t + d) - \bar{h}_t(x_t) - \bar{y}_t^T d + \frac{1}{2}\mu_t \|d\|^2 \\ \text{subject to} \quad & x_t + d \in C. \end{aligned}$$

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 $\sum_{t\geq 0} \left( \hat{N}_{h,t}^{-\alpha_h} + \hat{N}_{g,t}^{-\alpha_g} \right) < \infty$, 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. $\left(\mu_t - \frac{\check{\mu}}{2}\right) \left\|\bar{d}_t\right\|^2 \geq \frac{C_g}{\mu_{t+1} N_{g,t+1}^{\alpha_g}} + \frac{C_h}{(2\rho_g + 2\rho_h + \hat{\mu}) N_{h,t+1}^{\alpha_h}}$;
2. $N_{g,t+1} \geq \hat{N}_{g,t+1}$, and $N_{h,t} \geq \hat{N}_{h,t+1}$.

# Convergence Property and Sample Sizes Requirement

- The algorithm converges subsequentially to DC critical points almost surely.
- The sample size of our algorithm **matches the results achieved in the smooth case** under static distributions.

Table: Online stochastic DCA: Sample size at iteration $k$ (modulo logarithmic factors)

| Method | Assumption | | Sample size | |
|---|---|---|---|---|
| | Convex part | Concave part | Convex part | Concave part |
| Previous work[7] | Nonsmooth | Nonsmooth | $O(k^2)$ | $O(k^2)$ |
| | Nonsmooth | Smooth | $O(k^2)$ | $O(k)$ |
| **Ours** | **Nonsmooth** | **Nonsmooth** | $O(k^2)$ | $O(k)$ |

[7] 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

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

- $\{(x_i, y_i)\}_{i=1}^{\infty}$ are drawn from **unknown and varying** distributions $\mathcal{D}_t$.
- The capped-$\ell_1$ penalty $\sum_{j=1}^{d} \min(1, \alpha|\beta_j|)$ approximates the $\ell_0$-norm.
- **DC decomposition**:

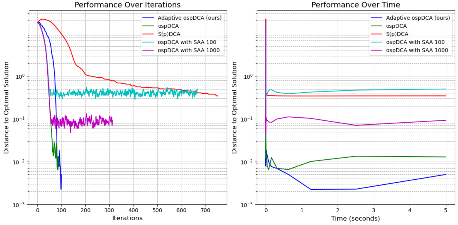
$$\min_{\beta \in \mathbb{R}^d} \mathbb{E}_{(x,y)\sim\mathcal{D}_t} [G(\beta, x, y)] - h(\beta),$$

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

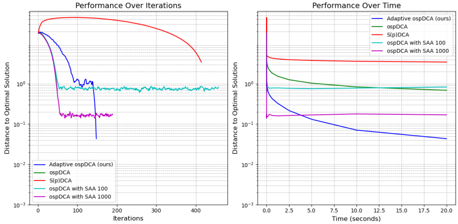
**Experiment Setup.**

- $x_t$ is sampled uniformly from $[-1, 1]^d$.
- The label $y_t = x_t^\top(\beta_{\text{opt}} + \delta_t) + \varepsilon$, where $\varepsilon \sim N(0, 1)$, $\delta_t$ is the distribution shift.
- Set $\delta_t = (-1)^t 100 t^{-2} \mathbf{1}_d$, since $W_1(\mathcal{D}_t, \mathcal{D}_{t+1}) \le \|\delta_t - \delta_{t+1}\|_1$.

# Numerical Experiments




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




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