Stochastic Optimization for DC Functions and Non-smooth Non-convex Regularizers with Non-asymptotic Convergence

Yi Xu[1], Qi Qi[1], Qihang Lin[1], Rong Jin[2], Tianbao Yang[1]

1. The University of Iowa 2. Damo Academy at Alibaba

June 12, 2019
ICML, Long Beach, CA

Non-Convex and Non-smooth Optimization

- A family of **non-convex non-smooth** optimization problems:

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) := g(\mathbf{x}) - h(\mathbf{x}) + r(\mathbf{x}), \tag{1}$$

  - $g(\cdot)$, $h(\cdot)$: real-valued lower-semicontinuous convex
  - $r(\cdot)$: proper lower-semicontinuous

- $g(\mathbf{x}) = \mathrm{E}_{\xi}[g(\mathbf{x};\xi)]$, $h(\mathbf{x}) = \mathrm{E}_{\varsigma}[h(\mathbf{x};\varsigma)]$
  - Finite-sum (a special case): $g(\mathbf{x}) = \frac{1}{n_1}\sum_{i=1}^{n_1} g_i(\mathbf{x})$, $h(\mathbf{x}) = \frac{1}{n_2}\sum_{j=1}^{n_2} h_j(\mathbf{x})$.

- It covers many applications
  - Non-Convex Sparsity-Promoting Regularizers: LSP, MCP, SCAD, capped ℓ_1, transformed ℓ_1
  - Weakly convex
  - Least-squares Regression with ℓ_{1-2} Regularization
  - Positive-Unlabeled (PU) Learning

Main Goal

- **Critical Point**: a point $\bar{\mathbf{x}}$ s.t.

$$\partial h(\bar{\mathbf{x}}) \cap \hat{\partial}(g+r)(\bar{\mathbf{x}}) \neq \emptyset.$$

  - $\hat{\partial} f(\mathbf{x})$: Fréchet subgradient; $\partial f(\mathbf{x})$: limiting subgradient
- An ϵ-**Critical Point**: a point $\bar{\mathbf{x}}$ s.t.

$$\mathrm{dist}(\partial h(\bar{\mathbf{x}}), \hat{\partial}(g+r)(\bar{\mathbf{x}})) \leq \epsilon.$$

  - If $g+r$ is non-differentiable, finding an ϵ-critical point is challenging.
  - An example: $g = |x|, h = r = 0$, then $\mathrm{dist}(0, \partial|x|) = 1$ when $x \neq 0$.
- Goal: finding a **Nearly** ϵ-**Critical Point** $\mathbf{x}$: if there exists $\bar{\mathbf{x}}$ such that

$$\|\mathbf{x} - \bar{\mathbf{x}}\| \leq O(\epsilon), \quad \mathrm{dist}(\partial h(\bar{\mathbf{x}}), \hat{\partial}(g+r)(\bar{\mathbf{x}})) \leq \epsilon. \tag{2}$$

Stagewise Stochastic DC algorithm (SSDC-$\mathcal{A}$)

When $r(\mathbf{x})$ is convex, assume that the **proximal mapping** of $r(\mathbf{x})$ can be easily computed: $\text{prox}_{\eta r}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2\eta}\|\mathbf{x}-\mathbf{y}\|^2 + r(\mathbf{x})$.

Stagewise Stochastic DC (SSDC) Algorithm [1, 2, 3]

1: **for** $k = 1, \ldots, K$ **do**
2: $F^{\gamma}_{\mathbf{x}_k}(\mathbf{x}) = g(\mathbf{x}) + r(\mathbf{x}) - (h(\mathbf{x}_k) + \partial h(\mathbf{x}_k)^\top(\mathbf{x}-\mathbf{x}_k)) + \frac{\gamma}{2}\|\mathbf{x}-\mathbf{x}_k\|^2$.
3: $\mathbf{x}_{k+1} = \mathcal{A}(F^{\gamma}_{\mathbf{x}_k})$
4: **end for**

[1] Dinh, T.P., Souad, E.B. North-Holland Mathematics Studies, pp. 249-271, 1986.

[2] Thi, H. A. L., Le, H. M., Phan, D. N., and Tran, B. in ICML, pp. 3394-3403, 2017.

[3] Wen, B., Chen, X., and Pong, T. K. Computational Optimization and Applications, 69(2):297-324, 2018.

Stagewise Stochastic DC algorithm (SSDC-$\mathcal{A}$)

When $r(\mathbf{x})$ is convex, assume that the **proximal mapping** of $r(\mathbf{x})$ can be easily computed: $\text{prox}_{\eta r}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2\eta}\|\mathbf{x}-\mathbf{y}\|^2 + r(\mathbf{x})$.

Basic idea: solving a convex majorant function in stage-wise

Stagewise Stochastic DC (SSDC) Algorithm [1, 2, 3]

1: **for** $k = 1, \ldots, K$ **do**
2: $F^{\gamma}_{\mathbf{x}_k}(\mathbf{x}) = g(\mathbf{x}) + r(\mathbf{x}) - (h(\mathbf{x}_k) + \partial h(\mathbf{x}_k)^\top(\mathbf{x}-\mathbf{x}_k)) + \frac{\gamma}{2}\|\mathbf{x}-\mathbf{x}_k\|^2$.
3: $\mathbf{x}_{k+1} = \mathcal{A}(F^{\gamma}_{\mathbf{x}_k})$
4: **end for**

[1] Dinh, T.P., Souad, E.B. North-Holland Mathematics Studies, pp. 249-271, 1986.
[2] Thi, H. A. L., Le, H. M., Phan, D. N., and Tran, B. in ICML, pp. 3394-3403, 2017.
[3] Wen, B., Chen, X., and Pong, T. K. Computational Optimization and Applications, 69(2):297-324, 2018.

Stagewise Stochastic DC algorithm (SSDC-$\mathcal{A}$)

When $r(\mathbf{x})$ is convex, assume that the **proximal mapping** of $r(\mathbf{x})$ can be easily computed: $\text{prox}_{\eta r}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2\eta}\|\mathbf{x}-\mathbf{y}\|^2 + r(\mathbf{x})$.

Basic idea: solving a convex majorant function in stage-wise

Stagewise Stochastic DC (SSDC) Algorithm [1, 2, 3]

1: **for** $k = 1, \ldots, K$ **do**
2: $F^{\gamma}_{\mathbf{x}_k}(\mathbf{x}) = g(\mathbf{x}) + r(\mathbf{x}) - (h(\mathbf{x}_k) + \partial h(\mathbf{x}_k)^\top(\mathbf{x}-\mathbf{x}_k)) + \frac{\gamma}{2}\|\mathbf{x}-\mathbf{x}_k\|^2$.
3: $\mathbf{x}_{k+1} = \mathcal{A}(F^{\gamma}_{\mathbf{x}_k})$
4: **end for**

- $\mathcal{A}$: stochastic algorithms (e.g., SPG, AdaGrad, SVRG) apply to $F^{\gamma}_{\mathbf{x}_k}(\mathbf{x})$

[1] Dinh, T.P., Souad, E.B. North-Holland Mathematics Studies, pp. 249-271, 1986.

[2] Thi, H. A. L., Le, H. M., Phan, D. N., and Tran, B. in ICML, pp. 3394-3403, 2017.

[3] Wen, B., Chen, X., and Pong, T. K. Computational Optimization and Applications, 69(2):297-324, 2018.

When $r(\mathbf{x})$ is convex, assume that the **proximal mapping** of $r(\mathbf{x})$ can be easily computed: $\text{prox}_{\eta r}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2\eta}\|\mathbf{x}-\mathbf{y}\|^2 + r(\mathbf{x})$.

Basic idea: solving a convex majorant function in stage-wise

Stagewise Stochastic DC (SSDC) Algorithm [1, 2, 3]

1: **for** $k = 1, \ldots, K$ **do**
2: $F^{\gamma}_{\mathbf{x}_k}(\mathbf{x}) = g(\mathbf{x}) + r(\mathbf{x}) - (h(\mathbf{x}_k) + \partial h(\mathbf{x}_k)^\top(\mathbf{x}-\mathbf{x}_k)) + \frac{\gamma}{2}\|\mathbf{x}-\mathbf{x}_k\|^2$.
3: $\mathbf{x}_{k+1} = \mathcal{A}(F^{\gamma}_{\mathbf{x}_k})$
4: **end for**

- $\mathcal{A}$: stochastic algorithms (e.g., SPG, AdaGrad, SVRG) apply to $F^{\gamma}_{\mathbf{x}_k}(\mathbf{x})$
- Finding $\mathbf{x}_{k+1}$ s.t. $\mathrm{E}\left[F^{\gamma}_{\mathbf{x}_k}(\mathbf{x}_{k+1}) - \min_{\mathbf{x}\in\mathbb{R}^d} F^{\gamma}_{\mathbf{x}_k}(\mathbf{x})\right] \leq \frac{c}{k}$.

[1] Dinh, T.P., Souad, E.B. North-Holland Mathematics Studies, pp. 249-271, 1986.
[2] Thi, H. A. L., Le, H. M., Phan, D. N., and Tran, B. in ICML, pp. 3394-3403, 2017.
[3] Wen, B., Chen, X., and Pong, T. K. Computational Optimization and Applications, 69(2):297-324, 2018.

Summary of Results (r is convex)

Table: Summary of results for finding a (nearly) ϵ-critical point of the problem (1)

g	h	r	Algorithm $\mathcal{A}$	Complexity
-	SM	CX	SPG, AdaGrad	$O(1/\epsilon^4)$
SM	SM	CX	SVRG	$O(n/\epsilon^2)$
SM	-	CX, SM	SPG, AdaGrad	$O(1/\epsilon^4)$
SM	-	CX, SM	SVRG	$O(n/\epsilon^2)$

- SM: smooth; CX: convex.
- n: the total number of components in a finite-sum problem.

Non-Smooth Non-Convex Regularization

- When $r(\mathbf{x})$ is non-convex, the challenge is the presence of non-smooth non-convex function r.
- The Moreau envelope of r ($\mu > 0$) is a DC function [4]:

$$r_\mu(\mathbf{x}) = \min_{\mathbf{y}\in\mathbb{R}^d} \left\{ \frac{1}{2\mu}\|\mathbf{y}-\mathbf{x}\|^2 + r(\mathbf{y}) \right\}$$
$$= \frac{1}{2\mu}\|\mathbf{x}\|^2 - \underbrace{\max_{\mathbf{y}\in\mathbb{R}^d} \left\{ \frac{1}{\mu}\mathbf{y}^\top\mathbf{x} - \frac{1}{2\mu}\|\mathbf{y}\|^2 - r(\mathbf{y}) \right\}}_{R_\mu(\mathbf{x})},$$

- Key idea: solving the following DC problem,

$$\min_{\mathbf{x}\in\mathbb{R}^d} F_\mu(\mathbf{x}) := g(\mathbf{x}) - h(\mathbf{x}) + \frac{1}{2\mu}\|\mathbf{x}\|^2 - R_\mu(\mathbf{x}).$$

[4] Liu, T., Pong, T. K., and Takeda, A. Mathematical Programming, 2018.

Summary of Results (r is non-convex)

Table: Summary of results for finding a (nearly) ϵ-critical point of the problem (1)

g	h	r	Algorithm $\mathcal{A}$	Complexity
SM	SM	NC, NS, LP	SPG	$O(1/\epsilon^8)$
SM	SM	NC, NS, FV, LB	SPG	$O(1/\epsilon^{12})$
SM	SM	NC, NS, LP	SVRG	$O(n/\epsilon^8)$
SM	SM	NC, NS, FV, LB	SVRG	$O(n/\epsilon^6)$
SM	SM	NC, NS, FVC	SVRG	$O(n/\epsilon^6)$

- SM: smooth; CX: convex; NC: non-convex; NS: non-smooth; LP: Lipchitz continuous function; LB: lower bounded over $\mathbb{R}^d$; FV: finite-valued over $\mathbb{R}^d$; FVC: finite-valued over a compact set.

Thank You!

Poster #109, Pacific Ballroom, 06:30-09:00 PM

Summary of Results (r is non-convex)

Table: Summary of results for finding a (nearly) ϵ-critical point of the problem (1)

g	h	r	Algorithm $\mathcal{A}$	Complexity
SM	SM	NC, NS, LP	SPG	$O(1/\epsilon^8)$
SM	SM	NC, NS, FV, LB	SPG	$O(1/\epsilon^{12})$
SM	SM	NC, NS, LP	SVRG	$O(n/\epsilon^8)$
SM	SM	NC, NS, FV, LB	SVRG	$O(n/\epsilon^6)$
SM	SM	NC, NS, FVC	SVRG	$O(n/\epsilon^6)$

- SM: smooth; CX: convex; NC: non-convex; NS: non-smooth; LP: Lipchitz continuous function; LB: lower bounded over $\mathbb{R}^d$; FV: finite-valued over $\mathbb{R}^d$; FVC: finite-valued over a compact set.

Thank You!

Poster #109, Pacific Ballroom, 06:30-09:00 PM