A Spider-Type Stochastic Subgradient Method for Expectation-Constrained Nonconvex Nonsmooth Optimization

Wei Liu, *Yangyang Xu*

Rensselaer Polytechnic Institute

MOPTA 2026

partly supported by NSF grant DMS-2208394

Expectation constrained problem

$$\min_{\mathbf{x}\in\mathcal{X}} \; f(\mathbf{x}), \text{ s.t. } \underbrace{\mathbb{E}_{\xi_{\mathbf{g}}\sim\mathbb{P}_{\mathbf{g}}}[G(\mathbf{x},\xi_{\mathbf{g}})]}_{=:\mathbf{g}(\mathbf{x})} \le 0, \qquad \text{(P)}$$
$$\text{where } \mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})]$$

- $\mathcal{X}$ convex compact and admits easy projection
- f is ρ_f-weakly convex on $\mathcal{X}$ with $\rho_f > 0$, i.e., $f(\cdot) + \frac{\rho_f}{2}\|\cdot\|^2$ convex
- $\{g_i\}$ are ρ_g-weakly convex on $\mathcal{X}$ with $\rho_g \ge 0$ (convex when $\rho_g = 0$)
- f and $\{g_i\}$ have bounded subgrad on $\mathcal{X}$ but not 1st-order smooth
- ONLY stochastic subgradient of f is accessible
- stochastic subgradient of g_i is accessible or when $\mathbb{P}_{\mathbf{g}}$ is a uniform distribution on N data points, deterministic subgradient can be accessed if needed.

Applications and Motivations

- Many learning tasks involve **expectation constraints**, e.g.
  - Fairness-constrained machine learning
  - Neyman–Pearson classification
- Constraints depend on random or big amount of data; projection onto feasible set is **prohibitively expensive**.
- Need stochastic first-order methods that handle:
  - Nonconvex, nonsmooth objectives
  - Expectation constraints

Existing methods

proximal point method [Boob-Deng-Lan'23; Ma-Lin-Yang'20]:

$$\mathbf{x}^{k+1} \approx \underset{\mathbf{x}\in\mathcal{X}}{\arg\min} \left\{ f(\mathbf{x}) + \rho_f \|\mathbf{x} - \mathbf{x}^k\|^2, \text{ s.t. } g_i(\mathbf{x}) + \rho_g \|\mathbf{x} - \mathbf{x}^k\|^2 \le 0, \forall\, i \right\}$$

- k-th subproblem strongly convex and can be solved to near optimality
- **Issue:** requiring an inner solver without a numerically checkable stopping condition

Existing methods

proximal point method [Boob-Deng-Lan'23; Ma-Lin-Yang'20]:

$$\mathbf{x}^{k+1} \approx \operatorname*{arg\,min}_{\mathbf{x}\in\mathcal{X}} \left\{ f(\mathbf{x}) + \rho_f \|\mathbf{x} - \mathbf{x}^k\|^2, \text{ s.t. } g_i(\mathbf{x}) + \rho_g \|\mathbf{x} - \mathbf{x}^k\|^2 \leq 0, \forall\, i \right\}$$

- k-th subproblem strongly convex and can be solved to near optimality
- **Issue:** requiring an inner solver without a numerically checkable stopping condition

switching subgradient [Huang-Lin'23]:

$$\mathbf{x}^{k+1} = \begin{cases} \mathbf{proj}_{\mathcal{X}} \left(\mathbf{x}^k - \alpha_k \boldsymbol{\zeta}_f^{(k)} \right), & \text{if } \mathbf{x}^k \text{ is near feasible,} \\ \mathbf{proj}_{\mathcal{X}} \left(\mathbf{x}^k - \alpha_k \boldsymbol{\zeta}_g^{(k)} \right), & \text{otherwise} \end{cases}$$

- **Issue:** high complexity to check near feasibility, not for stochastic weakly convex cases

Proposed method: exact penalty + smoothing + acceleration

Proposed method: exact penalty + smoothing + acceleration

Exact penalty

$$(1.6) \qquad \min_{\mathbf{x}\in\mathcal{X}} \; F(\mathbf{x}) := f(\mathbf{x}) + \beta g_+(\mathbf{x}), \text{ with } g_+(\mathbf{x}) = h(\mathbf{g}(\mathbf{x})) = \sum_{i=1}^m [g_i(\mathbf{x})]_+,$$

- stochastic subgradient method produces near ϵ-stationary solution in $O(\epsilon^{-4})$ iterations
- However, subgradient of $[\cdot]_+$ not Lipschitz around 0 $\Rightarrow$ expensive to have a good stochastic subgradient of $[g_i(\cdot)]_+$ at $\mathbf{x}^k$ if $g_i(\mathbf{x}^k)$ is close to 0

Proposed method: exact penalty + smoothing + acceleration

Exact penalty

$$(1.6) \qquad \min_{\mathbf{x}\in\mathcal{X}} \; F(\mathbf{x}) := f(\mathbf{x}) + \beta g_+(\mathbf{x}), \text{ with } g_+(\mathbf{x}) = h(\mathbf{g}(\mathbf{x})) = \textstyle\sum_{i=1}^m [g_i(\mathbf{x})]_+,$$

- stochastic subgradient method produces near ϵ-stationary solution in $O(\epsilon^{-4})$ iterations
- However, subgradient of $[\cdot]_+$ not Lipschitz around 0 $\Rightarrow$ expensive to have a good stochastic subgradient of $[g_i(\cdot)]_+$ at $\mathbf{x}^k$ if $g_i(\mathbf{x}^k)$ is close to 0

Smoothing penalty function

$$(\mathrm{PP}) \qquad \min_{\mathbf{x}\in\mathcal{X}} \; F^\nu(\mathbf{x}) = f(\mathbf{x}) + \beta h^\nu(\mathbf{g}(\mathbf{x})).$$

$$h^\nu(\mathbf{z}) := \sum_{i=1}^m H^\nu(z_i), \text{ with } H^\nu(z_i) = \max_{0\le y\le 1}\left[\langle y, z_i\rangle - \frac{\nu}{2}y^2\right] = \begin{cases} \dfrac{z_i^2}{2\nu}, & \text{if } 0 \le z_i \le \nu, \\ z_i - \dfrac{\nu}{2}, & \text{if } z_i > \nu, \\ 0, & \text{otherwise,} \end{cases}$$

- $\partial H^\nu(g_i(\mathbf{x})) = H^\nu_{z_i}(g_i(\mathbf{x}))\partial g_i(\mathbf{x})$ and 1-Lipschitzness of $H^\nu_{z_i}(\cdot) \Rightarrow$ enables more efficient way by variance reduction to estimate $H^\nu_{z_i}(g_i(\mathbf{x}^k))$

Near stationarity implying near KKT

Assume **Slater-type CQ** [Huang-Lin'23]: There exist $B_g > 0$, $B > 0$, and $\overline{\rho} > \rho_g \geq 0$, such that for any $\mathbf{x} \in \mathcal{X}$ satisfying $0 < g_+(\mathbf{x}) \leq B_g$, it holds

$$\max_{i\in[m]} g_i(\mathbf{y}) + \frac{\overline{\rho}}{2}\|\mathbf{y}-\mathbf{x}\|^2 \leq -B, \text{ for some } \mathbf{y} \in \mathrm{relint}(\mathcal{X}).$$

Then with penalty parameter $\beta = O(1)$ and $\nu \sim \epsilon$, we have

Lemma 2.3: ϵ-stationary point of (1.6) $\Downarrow$ ϵ-KKT point of (P)	**Theorem 2.5:** near-ϵ stationary point of (PP) $\Downarrow$ $(O(\epsilon), O(\epsilon))$-KKT point of problem (P)

- ϵ-stationary point of (1.6): $\mathrm{dist}\big(\mathbf{0}, \partial F(\mathbf{x}) + \mathcal{N}_{\mathcal{X}}(\mathbf{x})\big) \leq \epsilon$

Relation to Slater's condition

- For convex constraints, Slater's condition is often assumed, i.e.,

$$\exists\, \mathbf{y}_{\mathrm{feas}} \in \mathrm{relint}(\mathcal{X}) \text{ such that } \max_{i\in[m]} g_i(\mathbf{y}_{\mathrm{feas}}) < 0$$

Claim: when each g_i is convex, Slater-type CQ is equivalent to Slater's condition (as $\mathcal{X}$ is compact with diameter D)

$$\max_{i\in[m]} g_i(\mathbf{y}_{\mathrm{feas}}) + \frac{\overline{\rho}}{2}\|\mathbf{y}_{\mathrm{feas}} - \mathbf{x}\|^2 \le -B,\ \forall\, \mathbf{x} \in \mathcal{X}$$

with

$$\overline{\rho} = \frac{-\max_{i\in[m]} g_i(\mathbf{y}_{\mathrm{feas}})}{D^2} > \rho_g = 0,\ B = \frac{-\max_{i\in[m]} g_i(\mathbf{y}_{\mathrm{feas}})}{2}.$$

Relation to a strong feasibility condition

- To establish its $O(\epsilon^{-6})$ complexity for the case of nonconvex constraints, [Boob-Deng-Lan'23] assumes a strong feasibility condition:

$$\exists\, \mathbf{y}_{\text{feas}} \in \mathcal{X} \text{ such that } \max_{i\in[m]} g_i(\mathbf{y}_{\text{feas}}) \le -2\rho_g D^2.$$

Claim: Slater-type CQ is strictly weaker than strong feasibility

- Consider $g(x) = -\frac{7}{2} + 8x + x^2 - x^3$ on $\mathcal{X} = [0, 1]$
  - ⋄ $g''(x) = 2 - 6x \ge -4, \forall\, x \in [0, 1]$, thus 4-weakly convex on $\mathcal{X}$.
  - ⋄ For any $x \in \mathcal{X}$, $\min_{y\in\mathcal{X}} g(y) + \frac{5}{2}(y - x)^2 \le g(0) + \frac{5x^2}{2} \le -1$.
  - ⋄ But $\min_{x\in\mathcal{X}} g(x) = -\frac{7}{2} > -4 = -\rho_g D^2$.

Direct applying stochastic subgradient method

Direct applying stochastic subgradient method

For the smoothed penalty problem (PP):

- Notice $\partial h^\nu(\mathbf{g}(\mathbf{x})) = \sum_{i=1}^m \mathrm{Proj}_{[0,1]}\left(\frac{g_i(\mathbf{x})}{\nu}\right)\partial g_i(\mathbf{x})$
- Hence, an unbiased stochastic subgradient method performs update

$$\mathbf{x}^{(k+1)} = \mathrm{Proj}_{\mathcal{X}}\left(\mathbf{x}^{(k)} - \alpha_k\left(\boldsymbol{\zeta}_f^{(k)} + \beta\sum_{i=1}^m \mathrm{Proj}_{[0,1]}\left(\frac{g_i(\mathbf{x}^k)}{\nu}\right)\boldsymbol{\zeta}_{g_i}^{(k)}\right)\right)$$

  where $\boldsymbol{\zeta}_f^{(k)}$ and $\boldsymbol{\zeta}_{g_i}^{(k)}$ are unbiased stochastic subgradients of f and g_i.
- Direct using sample average to estimate $g_i(\mathbf{x}^k)$ is inefficient

Spider-accelerated stochastic subgradient method

- Utilize the Lipschitz continuity of each g_i and estimate $\mathbf{g}(\mathbf{x}^k)$ by

$$\mathbf{u}^{(k)} = \begin{cases} \mathbf{g}(\mathbf{x}^{(k)}, \mathcal{B}_k), & \text{with } |\mathcal{B}_k| = S_1, \text{if } \operatorname{mod}(k, q) = 0, \\ \mathbf{u}^{(k-1)} + \mathbf{g}(\mathbf{x}^{(k)}, \mathcal{B}_k) - \mathbf{g}(\mathbf{x}^{(k-1)}, \mathcal{B}_k), & \text{with } |\mathcal{B}_k| = S_2, \text{ otherwise,} \end{cases}$$

- Proposed update:

$$\mathbf{x}^{(k+1)} = \operatorname{Proj}_{\mathcal{X}}\left(\mathbf{x}^{(k)} - \alpha_k \left(\boldsymbol{\zeta}_f^{(k)} + \beta \sum_{i=1}^{m} \operatorname{Proj}_{[0,1]}\left(\frac{u_i^{(k)}}{\nu}\right) \boldsymbol{\zeta}_{g_i}^{(k)}\right)\right)$$

Complexity results

Pick $q = S_2 \sim \sqrt{S_1}$. Then for the smooth penalty problem,

- can find a near-ϵ stationary solution of (PP) in expectation with $O(\epsilon^{-4})$ stochastic subgradient and $O(\epsilon^{-5}\nu^{-1})$ sample evaluations on $\mathbf{g}$ or $O(N + \sqrt{N}\epsilon^{-4})$ samples for the finite-sum case

For the original problem (P), with $\nu \sim \epsilon$,

- can find a near-ϵ KKT solution in expectation with $O(\epsilon^{-4})$ stochastic subgradient and $O(\epsilon^{-6})$ sample evaluations on $\mathbf{g}$ or $O(N + \sqrt{N}\epsilon^{-4})$ samples for the finite-sum case.

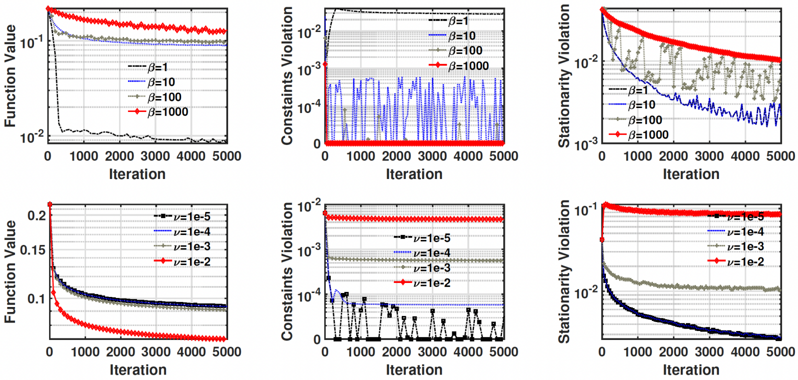
Comparison to existing work

Methods	RIS[¶]	structure of $\mathbf{g}$	CQ on $\mathbf{g}$	OGC	CGC	CFC
IPP	yes	stochastic convex	Slater's condition	$O(\epsilon^{-6})$	$O(\epsilon^{-6})$	
		finite-sum[†] convex		$O(\epsilon^{-4})$	$O(\min\{N\epsilon^{-4}, \epsilon^{-6}\})$	
		stochastic WC[¶]	strong feasibility	$O(\epsilon^{-6})$	$O(\epsilon^{-6})$	
		finite-sum[†] WC		$O(\epsilon^{-4})$	$O(\min\{N\epsilon^{-4}, \epsilon^{-6}\})$	
SSG	no	stochastic convex	Slater-type CQ*	$O(\epsilon^{-4})$	$O(\epsilon^{-4})$	$O(\epsilon^{-8})$
		finite-sum[†] convex		$O(\epsilon^{-4})$	$O(\epsilon^{-4})$	$O(N\epsilon^{-4})$
		WC			none[‡]	
Ours	no	stochastic convex	Slater-type CQ*	$O(\epsilon^{-4})$	$O(\epsilon^{-4})$	$O(\epsilon^{-6})$
		finite-sum[†] convex			$O(\epsilon^{-4})$	$O(\sqrt{N}\epsilon^{-4})$
		stochastic WC	Slater-type CQ, with $B_g > m\rho_g D^{2*}$	$O(\epsilon^{-4})$	$O(\epsilon^{-4})$	$O(\epsilon^{-6})$
		finite-sum[†] WC			$O(\epsilon^{-4})$	$O(\sqrt{N}\epsilon^{-4})$

- RIS indicates whether requiring inner solver; WC is for weakly convex
- **Observation:** our method achieves lower complexity in either subgradient or function evaluations

Numerical Results on ROC-based fairness classification

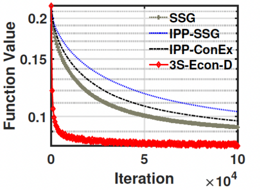
Effects of β and ν



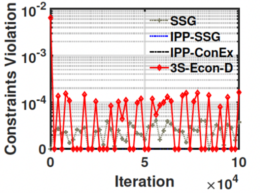
Observations:

- big β easily yields feasible solution but slow convergence
- final stationarity violation proportional to ν

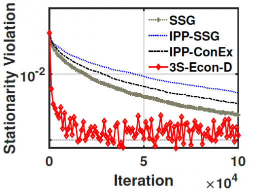
Comparison to existing methods



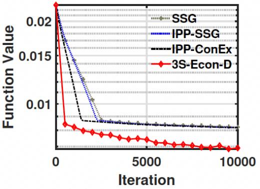
(a) Problem 1, a9a



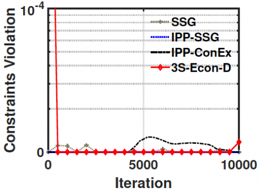
(b) Problem 1, a9a



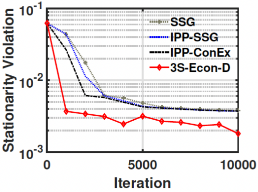
(c) Problem 1, a9a



(d) Problem 1, COMPAS



(e) Problem 1, COMPAS



(f) Problem 1, COMPAS

Observations: all methods can deliver feasible solution but proposed method converges faster

Reference

W. Liu, Y. Xu. A SPIDER-type stochastic subgradient method for expectation-constrained nonconvex nonsmooth optimization. SIOPT, 2026.

Thank you!!!