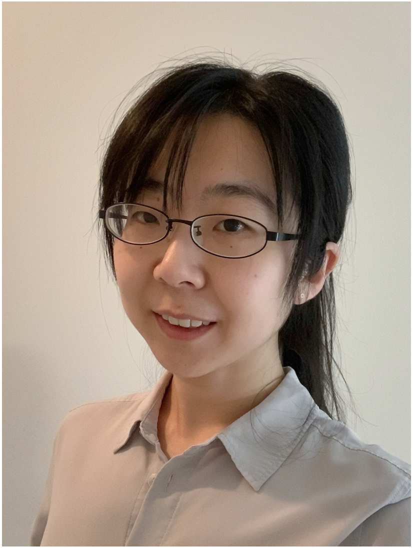
Stochastic composition optimization in the absence of Lipschitz continuous gradient

Sam Davanloo
Department of Integrated Systems Engineering
The Ohio State University

Joint work with Yin Liu

TDAI Summer School
June 3, 2022

Yin Liu, PhD Candidate, ISE, OSU



Optimization of compositional functions

- We consider the two-level stochastic compositional optimization problem of the form

$$\min_{\mathbf{x}\in\mathcal{X}} \quad F(\mathbf{x}) \triangleq f(g(\mathbf{x}))$$

$$f(\mathbf{u}) \triangleq \mathbb{E}_{\varphi}[f_{\varphi}(\mathbf{u})] \quad \text{and} \quad g(\mathbf{x}) \triangleq \mathbb{E}_{\xi}[g_{\xi}(\mathbf{x})]$$

where $f_{\varphi} : \mathbb{R}^d \to \mathbb{R}$ and $g_{\xi} : \mathbb{R}^n \to \mathbb{R}^d$ are differentiable functions, and φ, ξ are independent random variables.

Application: policy evaluation for Markov decision process

- Consider a Markov chain $\{Y_0, Y_1, \cdots\} \subset \mathcal{Y}$, unknown transition operator P, reward function $r : \mathcal{Y} \to \mathbb{R}$, discount factor $\gamma \in (0, 1)$
- We want to estimate the value function $V : \mathcal{Y} \to \mathbb{R}$ as $V(y) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(Y_t) | Y_0 = y]$
- For finite space $\mathcal{Y}$, value function (in vector form) satisfies Bellman equation $V = r + \gamma PV$
- As P is not known and $|\mathcal{Y}|$ large, system cannot be solved
- Linear model for the value function $V_x(y) \approx \sum_{i=1}^{d} x_i \phi_i(y)$ [Sutton'09]

$$\min_{\mathbf{x} \in \mathbb{R}^d} \operatorname{dist}((I - \gamma \mathbb{E}[\hat{P}])\Phi \mathbf{x}, \mathbb{E}[\hat{\mathbf{r}}])$$

Application: Model Agnostic Meta Learning (MAML)

- The goal is to find a common initialization for a set of agents $\mathcal{M} = \{1, ..., M\}$ from which they can adapt to a desired model
- Adapting involves taking one (or several) gradient step(s)
- One-step MAML

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) \triangleq \frac{1}{M} \sum_{m=1}^{M} f_m(\mathbf{x} - \alpha \nabla f_m(\mathbf{x}))$$

where $f_m(\mathbf{x}) = \mathbb{E}_{\xi_m}[f(\mathbf{x}; \xi_m)]$ [FinnAbbeelLevine'17]

Outline of the talk

- Prelimiaries
  - Main challenge with stochastic compositional optimization
  - Smoothness and the role it plays in optimization
  - Relative smoothness
- Compositional scenarios
  - Smooth of Relatively Smooth (SoR)
  - Relatively Smooth of Relatively Smooth (RoR)
  - Relatively Smooth of Smooth (RoS)
- Some numerical results
- Concluding remarks

Challenge of stochastic compositional algorithms

- Consider the main problem in an unconstrained setting

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) \triangleq f(g(\mathbf{x})) = \mathbb{E}_{\varphi}[f(\mathbb{E}_{\xi}[g(\mathbf{x};\xi)];\varphi)]$$

- Optimizing using SGD

$$x^{k+1} = x^k - \alpha \nabla g(x^k;\xi^k)^\top \nabla f(\mathbb{E}_{\xi}[g(x^k;\xi)];\varphi^k)$$

- Obtaining unbiased stochastic gradient is **costly**. Note that

$$\mathbb{E}_{\xi,\varphi}[\nabla g(x^k;\xi^k)^\top \nabla f(g(x^k;\xi);\varphi^k)] \neq \mathbb{E}_{\xi,\varphi}[\nabla g(x^k;\xi^k)^\top \nabla f(\mathbb{E}_{\xi}[g(x^k;\xi)];\varphi^k)]$$

- Is it possible to avoid the inner expectation?

Approximate the inner expectation

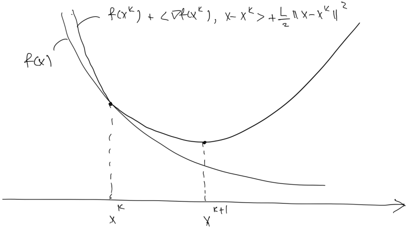
- WangFangLiu[MathProg'17] proposed to approximate the inner expectation by a running average

$$u^{k+1} = (1 - \tau_k)u^k + \tau_k g(x^k; \xi^k) \tag{1}$$

- Motivated by gradient flow ODE, ChenSunYin[NeurIPS'20] proposed an update to (1)
- GhadimiRuszczynskiWang[SIOPT'20] also proposed a running average over $\mathbf{x}^k$ and the gradient of the composition beside (1)
- The analysis of the three papers above (like many other GD-type methods) heavily depends on the Lipschitz continuity of the gradient

Gradient descent

- Upper bounding the objective function with an easy to solve quadratic function and minimize it



- First-order optimality condition (unconstrained case) to minimize the UB under ℓ_2-norm results into the gradient step

$$\nabla f(x^k) + L(x - x^k) = 0 \quad \Rightarrow \quad x = x^k - \frac{1}{L}\nabla f(x^k)$$

- Do we always have such an upper bound?

Lipschitz continuity of the gradient (smoothness)

- Existence of the UB requires Lipschitz continuity of the gradient of the function

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall x, y \in \text{dom} f$$

- For C^2 functions, Lipschitz continuity of the gradient is equivalent to

$$\nabla^2 f(x) \preceq L\mathbf{I} \quad \forall x \in \text{dom} f$$

i.e., max. eigenvalue of the hessian is bounded above by L. Hence, the quadratic form $\frac{1}{2}(x - x^k)\nabla^2 f(x^k)(x - x^k)$ is at most $\frac{L}{2}\|x - x^k\|^2$ (descent lemma)

$$f(x) \leq f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{L}{2}\|x - x^k\|_2^2$$

Absence of smoothness

- Examples of nonsmooth functions
  - Obviously, nondifferentiable functions are not smooth
  - Any (multivariate) polynomial function of degree higher than two (even convex)
  - $f(x) = -\log(x) + x^2$ with $f''(x) = \frac{1}{x^2} + 2$ on $\mathbb{R}_{++}$
  - D-optimal design $f(\mathbf{x}) = -\log\det(HXH^\top)$ with $X = \mathbf{diag}(x)$
- No global convergence for gradient descent can be established even in the convex setting
- Even if the function is smooth on some level sets, L could be huge, e.g., $f(x) = -\log(x) + x^2$ on $\{x : f(x) \leq 10\}$ has $L \approx \exp^{20}$, which results in very small step size
- In nonconvex setting, gradient descent may diverge
- BauschkeBolteTeboulle[MathOR'17] proposed a descent lemma beyond Lipschitz gradient continuity

Relative smoothness

Let h be a differentiable convex function. The Bregman distance between x, y under h is defined as

$$D_h(x,y) \triangleq h(x) - h(y) - \langle \nabla h(y), x - y \rangle \quad \forall x, y \in \text{int dom} h$$

Definition (Relative smoothness)

The function f is smooth relative to h on $\mathcal{X}$ if for any $x, \bar{x} \in \mathcal{X}$, the exists L s.t.

$$f(x) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + L D_h(x, \bar{x})$$

Proposition (LuFreundNesterov[SIOPT'17])

f is smooth relative to h on $\mathcal{X}$ iff

- *$Lh(\cdot) - f(\cdot)$ is convex on $\mathcal{X}$*
- *If twice differentiable, $\nabla^2 f(x) \preceq L \nabla^2 h(x) \quad \forall x \in \mathcal{X}$*

Note that smoothness is a special case of Relative smoothness with $h(x) = \|x\|_2^2/2$

Nonconvex setting?

- Bolt et al. [SIOPT'18] extended the Bregman descent lemma

Definition

f is smooth and/or weakly-convex relative to h on $\mathcal{X}$ if there exists $L_\ell > 0$ and $L_u > 0$ s.t.

$$-L_\ell D_h(x,\bar{x}) \leq f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), x - \bar{x} \rangle \leq L_u D_h(x,\bar{x})$$

- The LHS inequality is equivalent to $f + L_\ell h$ is convex
- Authors also showed global convergence of their Bregman Proximal Gradient algorithm to first-order stationary point

Contribution of this work

- We developed stochastic optimization algorithms to solve the constrained compositional problem with nonconvex components in the absence of smoothness
- This consists of three algorithms:
  - Smooth of Relatively smooth (SoR)
  - Relatively smooth of Relatively smooth (RoR)
  - Relatively smooth of Smooth (RoS)
- We establish conditions for (relatively) smoothness of the composition
- Establish (sample) iteration complexity of the proposed algorithms

Smooth of Relatively smooth (SoR) composition

Lemma (Stationarity Measure)

Given a μ_h-strongly convex function h, define

$$\hat{\mathbf{x}}^+ \triangleq \operatorname*{argmin}_{\mathbf{y}\in\mathcal{X}} \langle \nabla F(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{1}{\tau} D_h(\mathbf{y},\mathbf{x}),$$

where $\tau > 0$. Then $\hat{\mathbf{x}}^+ = \mathbf{x}$ if and only if $-\nabla F(\mathbf{x}) \in \mathcal{N}_{\mathcal{X}}(\mathbf{x})$.

The Lemma shows $\mathbf{dist}(\hat{\mathbf{x}}^+, \mathbf{x})$ is a suitable measure for stationarity.

Assumption

i. *The function f_φ is average L_f-smooth*
ii. *The function g_ξ is average L_g-smooth relative to 1-strongly convex function h_g*
iii. *The stochastic gradients of f_φ and g_ξ are bounded in expectation*
iv. *The variance of g_ξ is bounded*

SoR Algorithm

Require: $\mathbf{x}^0, \mathbf{u}^0, \mathbf{w}^0, \tau_k, \beta_k$

for $k = 0, 1, 2, \ldots, K$ **do**

Solve

$$\mathbf{x}^{k+1} = \operatorname*{argmin}_{\mathbf{y} \in \mathcal{X}} \left\langle \mathbf{w}^k, \mathbf{y} - \mathbf{x}^k \right\rangle + \frac{1}{\tau_k} D_h(\mathbf{y}, \mathbf{x}) \tag{2}$$

with $h(\mathbf{x}) = \frac{C_g^2 L_f}{2} \|\mathbf{x}\|^2 + C_f L_g h_g(\mathbf{x})$

Take i.i.d. samples $\{\xi_j^k\}_{j=1}^n$ and update

$$\mathbf{u}^{k+1} = \frac{1}{n} \sum_{j=1}^{n} \left[(1 - \beta_k)(\mathbf{u}^k + g_{\xi_j^k}(\mathbf{x}^{k+1}) - g_{\xi_j^k}(\mathbf{x}^k)) + \beta_k g_{\xi_j^k}(\mathbf{x}^{k+1}) \right] \tag{3}$$

Take i.i.d. samples $\{\varphi_i^k\}_{i=1}^n, \{\xi_i^k\}_{i=1}^n$ and calculate

$$\mathbf{w}^{k+1} = \frac{1}{n} \sum_{i=1}^{n} \nabla g_{\xi_i^k}(\mathbf{x}^{k+1})^{\mathsf{T}} \nabla f_{\varphi_i^k}(\mathbf{u}^{k+1}) \tag{4}$$

end for

Convergence rate of the algorithm

Lemma

Under Assumption above, $F(\mathbf{x}) = f(g(\mathbf{x}))$ *is 1-smooth relative to* $C_g^2 L_f$*-strongly convex function* $h(\mathbf{x}) = \frac{C_g^2 L_f}{2}\|\mathbf{x}\|^2 + C_f L_g h_g(\mathbf{x})$.

Corollary

Setting $\tau_k = \tau < \min\{1/2, L_f/(L_f+8), 1/L_f\}$ *and* $\beta_k = L_f\tau$*, we have*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{D_h(\hat{\mathbf{x}}^{k+1}, \mathbf{x}^k)}{\tau^2}\right] \le \frac{V^0}{\eta K} + \frac{\sigma_F^2\tau}{C_g^2 L_f \eta n} + \frac{2L_f^2\tau^2\sigma_g^2}{n},$$

where $\eta \triangleq \tau - 2\tau^2$.

Hence to achieve ϵ-stationarity, the algorithm needs $K = O(\epsilon^{-1})$ and $n = O(\epsilon^{-1})$, i.e., the number of calls to the g_ξ , ∇g_ξ, and ∇f_φ oracles are $O(\epsilon^{-2})$.

Relatively smooth of Relatively smooth regime (RoS)

Assumption

i. *The function f_φ is average L_f-smooth relative to h_f*
ii. *The function g_ξ is average L_g-smooth.*
iii. *The stochastic gradients of f_φ, g_ξ are bounded in expectation*
iv. *The variance of g_ξ is bounded*

Lemma

Under Assumption above, $F(\mathbf{x})$ is 1-smooth relative to $h(\mathbf{x}) = \frac{C_f L_g + C_{h_f} L_g L_f}{2} \|\mathbf{x}\|^2 + L_f h_f(g(\mathbf{x}))$, which is shown to be $C_f L_g$-strongly convex.

Relatively smooth of Relatively smooth regime (RoR)

Assumption

i. *The function f_φ is average L_f-smooth relative to h_f*

ii. *The function g_ξ is average L_g-smooth relative to 1-strongly convex function h_g*

iii. *The stochastic gradients of f_φ, g_ξ are bounded in expectation*

iv. *The variance of g_ξ is bounded*

Lemma

Under Assumption above, $F(\mathbf{x})$ is 1-smooth relative to $h(\mathbf{x}) = (C_f L_g + C_{h_f} L_g L_f) h_g(\mathbf{x}) + L_f h_f(g(\mathbf{x}))$, which is shown to be convex. Furthermore, if $h_g(\mathbf{x})$ is 1-strongly convex, then $h(\mathbf{x})$ is $C_f L_g$-strongly convex.

Algorithm RoS and RoR

Require: $\mathbf{x}^0, \tau_k \leq 1, \lambda \triangleq C_f L_g + 2C_{h_f} L_g L_f$

1: **for** $k = 0, 1, 2, \ldots, K$ **do**

2: Take i.i.d. samples $\{\xi_i^k\}_{i=1}^{n_k}$

$$\mathbf{u}^k = \frac{1}{n_k} \sum_{i=1}^{n_k} g_{\xi_i^k}(\mathbf{x}^k) \tag{5}$$

3: Take i.i.d. samples $\{\varphi_i^k\}_{i=1}^{m_k}, \{\xi_i^k\}_{i=1}^{m_k}$ and calculate

$$\mathbf{v}^k = \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla g_{\xi_i^k}(\mathbf{x}^k), \quad \mathbf{s}^k = \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla f_{\varphi_i^k}(\mathbf{u}^k) \tag{6}$$

$$\mathbf{w}^k = \mathbf{v}^k \mathbf{s}^k \tag{7}$$

4: Solve

$$\mathbf{x}^{k+1} = \operatorname*{argmin}_{\mathbf{y} \in \mathcal{X}} \left\langle \mathbf{w}^k, \mathbf{y} - \mathbf{x}^k \right\rangle + \frac{L_f}{\tau_k} D_{h_f}(\mathbf{u}^k + (\mathbf{v}^k)^\mathsf{T}(\mathbf{y} - \mathbf{x}^k), \mathbf{u}^k) + \frac{\lambda}{\tau_k} D_{h_g}(\mathbf{y}, \mathbf{x}^k)$$

with $h_g(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$ in the RoS case.

5: **end for**

Stationarity measure (RoS and RoR)

Lemma

Define

$$\tilde{\mathbf{x}}_\tau \triangleq \operatorname*{argmin}_{y \in \mathcal{X}} \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L_f}{\tau} D_{h_f}(g(\mathbf{x}) + \nabla g(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}), g(\mathbf{x})) + \frac{\lambda}{\tau} D_{h_g}(\mathbf{y}, \mathbf{x}),$$

where $\lambda \triangleq C_f L_g + 2C_{h_f} L_g L_f$*, then* $\tilde{\mathbf{x}}_\tau = \mathbf{x}$ *if and only if* $-\nabla F(\mathbf{x}) \in \mathcal{N}_{\mathcal{X}}(\mathbf{x})$*.*

- The above Lemma motivates the use of $\mathbf{dist}(\tilde{\mathbf{x}}_\tau, \mathbf{x})$ to measure stationarity error.

Assumption (Extra Assumption for RoS and RoR)

Similar to Bolt et al. [SIOPT'18], we also assume the function h_f *is* L_{h_f} *Lipschitz smooth on any bounded subset of* $\mathbb{R}^d$*.*

Convergence rate for the RoR setting

Corollary

Setting $\tau_k = \tau$ *such that* $\lambda/\tau - C_f - 1 - C_f L_g > 0$, $n_k = n$, $m_k = m$ *and define* $A \triangleq \left(\frac{\lambda}{\tau} - C_f - 1 - C_f L_g\right) / \left(C_f L_g + \frac{5 C_{h_f} L_g L_f}{4}\right)$, *then under the RoR and extra Assumptions, we have*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2}{\tau^2}\right]$$

$$\leq \frac{1}{\tau^2 A (C_f L_g + \frac{C_{h_f} L_g L_f}{2})}\left(\frac{f(g(\mathbf{x}^0)) - F^*}{K} + \frac{\sigma_g}{\sqrt{n}}(2C_f + 4AC_{h_f}L_f) + \frac{2\sigma_g^2}{n}\left(\frac{AC_g^2 L_f^2 L_{h_f}^2 \tau^2}{C_{h_f} L_g} + C_g^2 L_f^2 L_{h_f}^2\right)\right.$$
$$\left. + \frac{1}{m}\left(\frac{C_g^2 C_f + C_g^2 C_f^2}{2} + \frac{6AC_{h_f} C_g^2 L_f}{L_g} + \frac{2A\tau^2\sigma_F^2}{C_{h_f} L_g L_f}\right)\right).$$

- Hence to achieve ϵ-stationary solution, the algorithm needs $K = O(\epsilon^{-1})$, $n = O(\epsilon^{-2})$, and $m = O(\epsilon^{-1})$, i.e., the number of calls to the g_ξ , ∇g_ξ, and ∇f_φ oracles are $O(\epsilon^{-3})$, $O(\epsilon^{-2})$, and $O(\epsilon^{-2})$, respectively.
- For the **RoS composition** and through **variance reduction**, we obtained $O(\epsilon^{-5/2})$, $O(\epsilon^{-3/2})$, $O(\epsilon^{-2})$ sample complexities for the g_ξ , ∇g_ξ, and ∇f_φ oracles, respectively.

Numerical experiments (SoR)

- Risk-averse portfolio optimization

$$\min_{\mathbf{x}\in\mathcal{X}} -\mathbb{E}[r_\xi(\mathbf{x})] + \lambda\left(\mathbb{E}[(r_\xi(\mathbf{x}))^2] - \mathbb{E}[r_\xi(\mathbf{x})]^2\right)$$

- This problem is a stochastic compositional problem with

$$g(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^2 = \left[\mathbb{E}[r_\xi(\mathbf{x})], \mathbb{E}[(r_\xi(\mathbf{x}))^2]\right]$$
$$f(\mathbf{u}) : \mathbb{R}^2 \to \mathbb{R} = -u_1 + \lambda u_2 - \lambda u_1^2$$

- The random reward is modeled as $r_\xi(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\intercal A_\xi \mathbf{x}$ with a symmetric matrix A_ξ
- It can be shown that g is 1-smooth relative to $h_g = \frac{\|A\|}{2}\|\mathbf{x}\|^2 + \frac{3\mathbb{E}[\|A_\xi\|^2]}{4}\|\mathbf{x}\|^4$, while f is smooth (quadratic)
- Also, $f \circ g$ is 1-smooth relative to

$$h(\mathbf{x}) = \frac{C_g^2 L_f + C_f L_g \|A\|}{2}\|\mathbf{x}\|^2 + \frac{3 C_f L_g \mathbb{E}[\|A_\xi\|^2]}{4}\|\mathbf{x}\|^4$$

Numerical experiments (RoS)

- Policy evaluation for MDP with $V^{\pi}(i) \approx \langle \Phi_i, \mathbf{x} \rangle$ results in minimization $\mathbf{dist}(\mathbf{r}, (\mathbf{I} - \gamma P^{\pi})\Phi\mathbf{x})$
- Under KL divergence $D_{KL}(\mathbf{a}, \mathbf{b}) = \sum_i a_i \log(a_i/b_i) + b_i - a_i$, $\forall a_i, b_i > 0$, the problem can be written as

$$\min_{\mathbf{x}} \quad D_{KL}\big(\mathbf{r}, (\mathbf{I} - \gamma \mathbb{E}[P])\Phi\mathbf{x}\big) \quad s.t. \quad (\mathbf{I} - \gamma \mathbb{E}[P])\Phi\mathbf{x} \geq \epsilon \mathbf{1}$$

- To simplify the problem, given $A_\xi \in \mathbb{R}_+^{S \times n}$, we solve

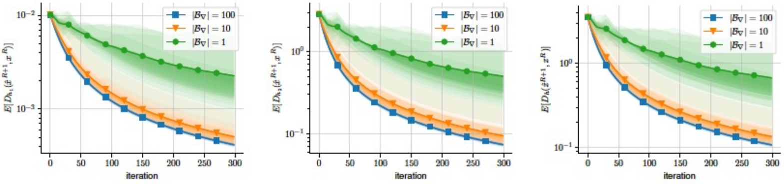
$$\min_{\mathbf{x} \in \mathbb{R}_+^n} \quad D_{KL}\big(\mathbf{r}, \mathbb{E}[A_\xi])\mathbf{x}\big)$$

- This problem is the stochastic compositional problem with

$$g(\mathbf{x}) = \mathbb{E}[A_\xi]\mathbf{x}, \quad f(\mathbf{u}) = \sum_{i=1}^{S} u_i - r_i \log u_i.$$

g is smooth and f is smooth relative to $h_f(\mathbf{u}) = -\sum_{j=1}^{n} \log u_j$

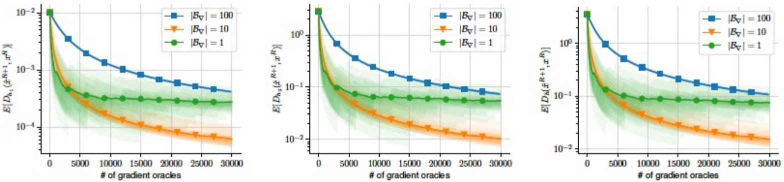
SoR: Risk-averse optimization



(a) $h_1(\mathrm{x}) = \frac{1}{2}\|\mathrm{x}\|^2$

(b) $h_2(\mathrm{x}) = \frac{1}{4}\|\mathrm{x}\|^4$

(c) $h(\mathrm{x}) = \frac{100}{2}\|\mathrm{x}\|^2 + \frac{0.9}{4}\|\mathrm{x}\|^4$

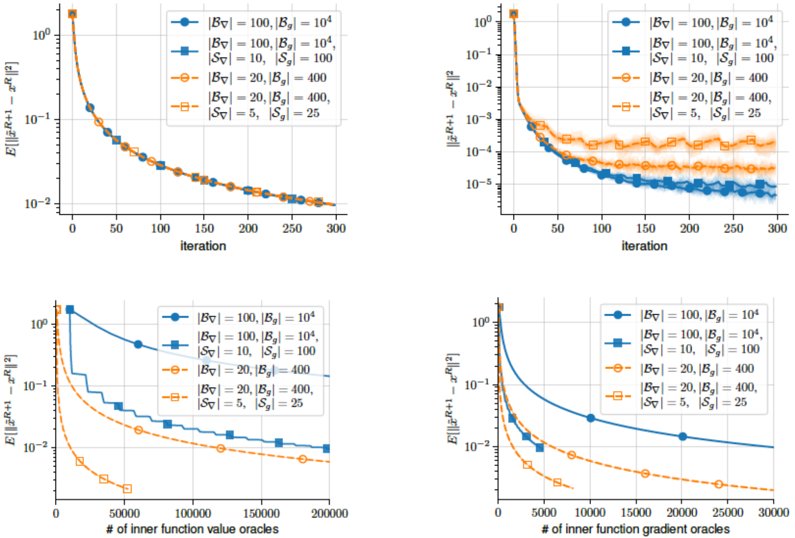


(d) $h_1(\mathrm{x}) = \frac{1}{2}\|\mathrm{x}\|^2$

(e) $h_2(\mathrm{x}) = \frac{1}{4}\|\mathrm{x}\|^4$

(f) $h_2(\mathrm{x}) = \frac{100}{2}\|\mathrm{x}\|^2 + \frac{0.9}{4}\|\mathrm{x}\|^4$

RoS: Policy evaluation for MDP



Concluding remarks

- This works looks into two-level stochastic compositional optimization problem in the absence of smoothness
- It considers the notion of "relative smoothness" for the inner, outer, or both functions
- The work proposes two Bregman-based algorithms to solve SoR, RoR, and RoR compositions over closed convex sets
- Iteration/oracle complexity of the proposed algorithms to obtain first-order stationarity solutions are established

Thank you for listening