# Stochastic Approximation Schemes with Decision Dependent Data

Hoi-To Wai

Department of Systems Engineering & Engineering Management,
The Chinese University of Hong Kong (CUHK), Hong Kong

May 30, 2022
IORA Seminar@NUS

# Stochastic Approximation (SA) Scheme: Background

- SA scheme [Robbins and Monro, 1951] is a stochastic process:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1} H(\boldsymbol{\theta}_t; X_{t+1}), \quad t \in \mathbb{N}$$

  where $\boldsymbol{\theta}_t \in \mathbb{R}^d$ is the $t$-th iterate, $\gamma_t > 0$ is the step size, $H(\boldsymbol{\theta}_t; X_{t+1})$ is the drift term and $X_{t+1}$ represents the **data** drawn.

- **Application**: SGD – take $H(\boldsymbol{\theta}_t; X_{t+1}) = \nabla \ell(\boldsymbol{\theta}_t; X_{t+1})$ for stochastic optimization $\min_{\boldsymbol{\theta}} \mathbb{E}[\ell(\boldsymbol{\theta}; X)]$ [Bottou et al., 2018].

- *Drift* $H(\theta_n; X_{t+1})$ relies on **i.i.d. data** $X_{t+1} \Rightarrow$ *mean-field*:

$$h(\theta_t) = \mathbb{E}\big[H(\theta_t; X_{t+1}) | \mathcal{F}_t\big] =: \mathbb{E}_t\big[H(\theta_t; X_{t+1})\big],$$

  where $\mathcal{F}_t$ is the filtration generated by $\{\theta_0, \{X_m\}_{m \le t}\}$.

- **Fact**: $\theta_t \to \bar{\theta}$ such that $h(\bar{\theta}) = 0$ (+ appropriate step size).

# Stochastic Approximation (SA) Scheme: Background

- SA scheme [Robbins and Monro, 1951] is a stochastic process:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1} H(\boldsymbol{\theta}_t; X_{t+1}), \quad t \in \mathbb{N}$$

  where $\boldsymbol{\theta}_t \in \mathbb{R}^d$ is the $t$-th iterate, $\gamma_t > 0$ is the step size, $H(\boldsymbol{\theta}_t; X_{t+1})$ is the drift term and $X_{t+1}$ represents the **data** drawn.

- **Application**: SGD – take $H(\boldsymbol{\theta}_t; X_{t+1}) = \nabla \ell(\boldsymbol{\theta}_t; X_{t+1})$ for stochastic optimization $\min_{\boldsymbol{\theta}} \mathbb{E}[\ell(\boldsymbol{\theta}; X)]$ [Bottou et al., 2018].

- *Drift* $H(\boldsymbol{\theta}_n; X_{t+1})$ relies on **i.i.d. data** $X_{t+1}$ $\Rightarrow$ *mean-field*:

$$h(\boldsymbol{\theta}_t) = \mathbb{E}\big[H(\boldsymbol{\theta}_t; X_{t+1}) | \mathcal{F}_t\big] =: \mathbb{E}_t\big[H(\boldsymbol{\theta}_t; X_{t+1})\big],$$

  where $\mathcal{F}_t$ is the filtration generated by $\{\boldsymbol{\theta}_0, \{X_m\}_{m \le t}\}$.

- **Fact**: $\boldsymbol{\theta}_t \to \bar{\boldsymbol{\theta}}$ such that $h(\bar{\boldsymbol{\theta}}) = 0$ (+ appropriate step size).

# SA with Decision-Dependent Data: Motivation

- What if data $X_{t+1}$ is not i.i.d., and depends on $\boldsymbol{\theta}_t$?

$$\underline{\text{SA}}: \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1} H(\boldsymbol{\theta}_t; X_{t+1}).$$

- *Example 1*: in reinforcement learning (RL),

$$X_t = (S_t, A_t) \text{ -- state/action}, \quad \theta_t \text{ -- policy}.$$

  A policy describes conditional probability for selecting $A_t$.
- **Online policy gradient** –

$$\cdots \to A_t \to S_t \underbrace{\longrightarrow}_{\text{SA step}} \theta_t \underbrace{\longrightarrow}_{\text{Use Policy}} \underbrace{A_{t+1} \to S_{t+1}}_{\text{Calc. Reward}} \to \cdots$$

# SA with Decision-Dependent Data: Motivation

- What if data $X_{t+1}$ is not i.i.d., and depends on $\boldsymbol{\theta}_t$?

$$\underline{\text{SA}}: \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1} H(\boldsymbol{\theta}_t; X_{t+1}).$$

- *Example 1*: in reinforcement learning (RL),

$$X_t = (S_t, A_t) \text{ – state/action}, \quad \boldsymbol{\theta}_t \text{ – policy}.$$

  A policy describes conditional probability for selecting $A_t$.

- **Online policy gradient –**

$$\cdots \to A_t \to S_t \underbrace{\longrightarrow}_{\text{SA step}} \boldsymbol{\theta}_t \underbrace{\longrightarrow}_{\text{Use Policy}} \underbrace{A_{t+1} \to S_{t+1}}_{\text{Calc. Reward}} \to \cdots$$

# SA with Decision-Dependent Data: Motivation

- What if data $\color{red}{X_{t+1}}$ is not i.i.d., and depends on $\boldsymbol{\theta}_t$?

$$\underline{\text{SA}}: \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1} H(\boldsymbol{\theta}_t; \color{red}{X_{t+1}}).$$

- *Example 2*: in Strategic Classification, data may react to your decision,

$$\boldsymbol{\theta}_t \text{ -- classifier,} \quad X_{t+1} \sim \mathcal{D}(\boldsymbol{\theta}_t) \text{ -- observed data}$$

  such as in loan application, spam email classification, etc.
- **Greedy Deployment** [Perdomo et al., 2020]:

$$\cdots \to X_t \underbrace{\longrightarrow}_{\text{SA step}} \boldsymbol{\theta}_t \quad \underbrace{\longrightarrow}_{\text{Adopt/deploy the decision}} \quad X_{t+1} \to \cdots$$

# SA with Decision-Dependent Data: Challenges

**Key Q**: *Will SA with decision-dependent data converge to $h(\bar{\boldsymbol{\theta}}) = 0$ (or other meaningful point)? Under what condition? How fast?*

**Challenges —**

- The drift term is *biased*, i.e., $\mathbb{E}\big[H(\boldsymbol{\theta}_t; X_{t+1})|\mathcal{F}_t\big] \neq h(\boldsymbol{\theta}_t)$.
- If $X_{t+1}$ is *too sensitive* to $\boldsymbol{\theta}_t$, it may not converge.

**This Talk —**

- Recent results on convergence to **stationary or stable** solution with decision dependent data SA.
- Applications to online policy gradient, performative prediction, two-timescale SA, etc.

# Overview of This Talk

- General Convergence for SA with Decision-dependent Data[1]
  - Focus on a **non-convex** (but smooth) setting.
  - Expected convergence at $\mathbb{E}[\|h(\boldsymbol{\theta}_T)\|^2] = \mathcal{O}(1/\sqrt{T})$.
  - Application: Online Policy Gradient.
- State-dependent Performative Prediction[2]
  - Refined analysis on a **'strongly convex'** setting.
  - Expected convergence at $\mathbb{E}[\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_{PS}\|^2] = \mathcal{O}(1/T)$.
- Two Timescale SA and Application to Actor-Critic[3]
  - Bi-level optimization where *lower level* gives decision-dependent data.

---

[1] B. Karimi B. Miasojedow, É. Moulines, H.-T. Wai, "Non-asymptotic Analysis of Biased Stochastic Approximation Scheme", in COLT 2019.

[2] Q. Li, H.-T. Wai, "State Dependent Performative Prediction with Stochastic Approximation", in AISTATS 2022.

[3] M. Hong, H.-T. Wai, Z. Wang, Z. Yang, "A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic", in ArXiv, 2020.

# Roadmap

1. General Convergence of SA
   Application to Policy Gradient
2. Performative Prediction
3. Two-timescale SA for Bilevel Problem
4. Conclusions and Perspectives

# Roadmap

1. General Convergence of SA
   Application to Policy Gradient

2. Performative Prediction

3. Two-timescale SA for Bilevel Problem

4. Conclusions and Perspectives

# SGD Method as an SA Scheme

Consider a possibly non-convex optimization problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \; V(\boldsymbol{\theta}), \tag{1}$$

where $V : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a smooth (Lyapunov) function. Our goal is to find a stationary point of (1) by SA:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1} H(\boldsymbol{\theta}_t; X_{t+1}).$$

- **Special case – SGD**: draw i.i.d. samples $X_{t+1}$ such that $H(\boldsymbol{\theta}_t; X_{t+1})$ is *unbiased* estimate of gradient, i.e., $\mathbb{E}\big[H(\boldsymbol{\theta}_t; X_{t+1}) | \mathcal{F}_t\big] = \nabla V(\boldsymbol{\theta}_t)$.

**This Part**: *We analyze the decision-dependent relaxation to SA scheme for tackling* (1).

# Biased SA Scheme

We relax **two** restrictions in classical SA/SGD. Consider:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1} H(\boldsymbol{\theta}_t; X_{t+1}). \tag{2}$$

- The mean field is not a gradient

  $\Longrightarrow$ relevant to *non-gradient* method where the gradient is hard to compute, e.g., expectation-maximization, policy gradient.

- $\{X_t\}_{t\geq 1}$ is not i.i.d. and form a **decision-dependent Markov chain**:

  $$\mathbb{E}[H(\boldsymbol{\theta}_t; X_{t+1}) | \mathcal{F}_t] = P_{\boldsymbol{\theta}_t} H(\boldsymbol{\theta}_t; X_t) = \int H(\boldsymbol{\theta}_t; x) P_{\boldsymbol{\theta}_t}(X_t, \mathrm{d}x),$$

  where $P_{\boldsymbol{\theta}_t} : \mathsf{X} \times \mathcal{X} \to \mathbb{R}_+$ is Markov kernel with a unique stationary distribution $\pi_{\boldsymbol{\theta}_t}$, and the mean field $h(\boldsymbol{\theta}) = \int H(\boldsymbol{\theta}; x) \pi_{\boldsymbol{\theta}}(\mathrm{d}x)$.

  $\Longrightarrow$ relevant to *policy gradient*.

# Biased SA Scheme

We relax **two** restrictions in classical SA/SGD. Consider:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1} H(\boldsymbol{\theta}_t; X_{t+1}). \tag{2}$$

**Prior Works —**

- *Asymptotic Analysis*: studied with $h(\boldsymbol{\theta}) = \nabla V(\boldsymbol{\theta})$ in [Kushner and Yin, 2003], similar biased SA setting in [Tadić and Doucet, 2017].
- *Non-asymptotic Analysis*:
  - Sun et al. [2018] and Duchi et al. [2012] assumed $h(\boldsymbol{\theta}) = \nabla V(\boldsymbol{\theta})$ & **decision-independent** Markov chain.
  - Bhandari et al. [2018] studied a similar setting but focuses on linear SA with convex Lyapunov function.
  - Recent works [Chen et al., 2020, Mou et al., 2020, Durmus et al., 2021a,b] provided tight bounds for linear SA.

# Assumptions

**(A1)** For all $\boldsymbol{\theta}$, there exists $c_0 \geq 0, c_1 > 0$, $d_0 \geq 0, d_1 > 0$ such that

$$c_0 + c_1 \langle \nabla V(\boldsymbol{\theta}) \,|\, h(\boldsymbol{\theta}) \rangle \geq \|h(\boldsymbol{\theta})\|^2, \quad d_0 + d_1 \|h(\boldsymbol{\theta})\| \geq \|\nabla V(\boldsymbol{\theta})\|$$

Moreover, the Lyapunov function $V$ is $L$-smooth,

$$\|\nabla V(\boldsymbol{\theta}) - \nabla V(\boldsymbol{\theta}')\| \leq L \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \ \forall\, \boldsymbol{\theta}, \boldsymbol{\theta}'.$$

- Mean field $h(\boldsymbol{\theta})$ can be *indirectly related* to $\nabla V(\boldsymbol{\theta})$.
- Requires smooth Lyapunov function yet $V(\boldsymbol{\theta})$ is possibly *non-convex*.

**(A2)** It holds that $\sup_{\boldsymbol{\theta} \in \mathbb{R}^d, x \in \mathsf{X}} \|H(\boldsymbol{\theta}; x) - h(\boldsymbol{\theta})\| \leq \sigma$.

*Remark*: **(A2)** requires noise is *uniformly bounded* for all $x \in \mathsf{X}$.

# Assumptions (on the Markov Chain)

**(A3)** There exists a bounded measurable function $\hat{H} : \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}^d$ s.t.

$$\hat{H}_{\boldsymbol{\theta}}(x) - P_{\boldsymbol{\theta}} \hat{H}_{\boldsymbol{\theta}}(x) = H(\boldsymbol{\theta}; x) - h(\boldsymbol{\theta}), \; \forall \, \boldsymbol{\theta} \in \mathbb{R}^d, x \in \mathsf{X},$$

and $\quad \sup_{x \in \mathsf{X}} \| P_{\boldsymbol{\theta}} \hat{H}_{\boldsymbol{\theta}}(x) - P_{\boldsymbol{\theta}'} \hat{H}_{\boldsymbol{\theta}'}(x) \| \leq L_{PH}^{(1)} \| \boldsymbol{\theta} - \boldsymbol{\theta}' \|, \; \forall \, (\boldsymbol{\theta}, \boldsymbol{\theta}').$

- $\hat{H}_{\boldsymbol{\theta}}(.)$ exists if MC $P_{\boldsymbol{\theta}}$ is uniformly geometric ergodic [Douc et al., 2018].
- **Consequence**: allows *error decomposition* of

$$\begin{aligned} H(\boldsymbol{\theta}_t; X_{t+1}) - h(\boldsymbol{\theta}_t) &= \hat{H}_{\boldsymbol{\theta}_t}(X_{t+1}) - P_{\boldsymbol{\theta}_t} \hat{H}_{\boldsymbol{\theta}_t}(X_{t+1}) \\ &= \underbrace{\hat{H}_{\boldsymbol{\theta}_t}(X_{t+1}) - P_{\boldsymbol{\theta}_t} \hat{H}_{\boldsymbol{\theta}_t}(X_t)}_{\text{Martingale with conditional 0-mean}} + P_{\boldsymbol{\theta}_t} \hat{H}_{\boldsymbol{\theta}_t}(X_t) - P_{\boldsymbol{\theta}_t} \hat{H}_{\boldsymbol{\theta}_t}(X_{t+1}) \end{aligned}$$

# Main Results

**Theorem**

*Let A1–A3 hold. Suppose that the step sizes satisfy*

$$\gamma_{n+1} \le \gamma_n,\ \gamma_n \le a\gamma_{n+1},\ \gamma_n - \gamma_{n+1} \le a'\gamma_n^2,\ \gamma_1 \le 0.5\big(c_1(L + C_h)\big)^{-1},$$

*for $a, a' > 0$ and all $t \ge 0$, then*

$$\mathbb{E}[h(\boldsymbol{\theta}_T)\|^2] \le \frac{2c_1\big(V_{0,t} + C_{0,t} + (\sigma^2 L + C_\gamma)\sum_{k=0}^{t}\gamma_{k+1}^2\big)}{\sum_{k=0}^{t}\gamma_{k+1}} + 2c_0,$$

*where $C_h$, $C_\gamma$, $C_{0,t}$, $V_{0,t}$ are $\mathcal{O}(1)$ constants.*

- *Stopping Criterion*: fix any $t \ge 1$ and $T \in \{0, \ldots, t\}$ is a discrete random variable with (see [Ghadimi and Lan, 2013])

$$\mathbb{P}(T = \ell) = \big(\textstyle\sum_{t=0}^{n}\gamma_{t+1}\big)^{-1}\gamma_{\ell+1}.$$

- If $\gamma_t = (2c_1 L(1 + C_h)\sqrt{t})^{-1}$, then $\mathbb{E}[\|h(\boldsymbol{\theta}_T)\|^2] = \mathcal{O}(c_0 + \log t/\sqrt{t})$.

# Policy Optimization: Setup

- Consider a Markov Decision Process (MDP) $(\mathsf{S}, \mathsf{A}, \mathsf{R}, \mathsf{P})$:
  - $\mathsf{S}$, $\mathsf{A}$ is a finite set of state (state-space) / action (action-space)
  - $\mathsf{R} : \mathsf{S} \times \mathsf{A} \to [0, \mathsf{R}_{\max}]$ is a reward function
  - $\mathsf{P}$ is the transition model, *i.e.*, given an action $a \in \mathsf{A}$, $\mathsf{P}^a = \{\mathsf{P}^a_{s,s'}\}$ is a matrix, $\mathsf{P}^a_{s,s'}$ is the probability of transiting from the $s$th state to the $s'$th state upon taking action $a$.
- A **policy** is parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$ as:

$$\Pi_{\boldsymbol{\theta}}(a'; s') = \text{probability of taking action } a' \text{ in state } s'$$

- $\{(S_t, A_t)\}_{t \geq 1}$ forms a MC with transition probability $(s, a) \to (s', a')$:

$$Q_{\boldsymbol{\theta}}((s, a); (s', a')) := \Pi_{\boldsymbol{\theta}}(a'; s')\, \mathsf{P}^a_{s,s'} \ ,$$

also denote its invariant distribution as $\upsilon_{\boldsymbol{\theta}}(s, a)$.

**Goal**: *optimize* $\boldsymbol{\theta}$ *such that the average reward is maximized.*

# Policy Optimization: Average Reward Maximization

- **Goal:** Find a policy $\boldsymbol{\theta}$ to maximize the average reward:

$$J(\boldsymbol{\theta}) := \sum_{s \in \mathsf{S}, a \in \mathsf{A}} \upsilon_{\boldsymbol{\theta}}(s,a)\, \mathsf{R}(s,a)\ .$$

- What is the gradient of $J(\boldsymbol{\theta})$ *w.r.t.* $\boldsymbol{\theta}$?

$$\nabla J(\boldsymbol{\theta}) = \lim_{T \to \infty} \mathbb{E}_{\boldsymbol{\theta}} \big[\, \mathsf{R}(S_T, A_T) \sum_{i=0}^{T-1} \nabla \log \Pi_{\boldsymbol{\theta}}(A_{T-i}; S_{T-i}) \big].$$

- REINFORCE algorithm [Williams, 1992] uses the sample average approximation. Let $M \gg 1$, $T \gg 1$,

$$\nabla J(\boldsymbol{\theta}) \approx (1/M) \sum_{m=1}^{M} \big\{\, \mathsf{R}(S_T^m, A_T^m) \sum_{i=0}^{T-1} \nabla \log \Pi_{\boldsymbol{\theta}}(A_{T-i}^m; S_{T-i}^m) \big\}$$

where $(S_1^m, A_1^m, \ldots, S_T^m, A_T^m) \sim \Pi_{\boldsymbol{\theta}}$ are drawn from a *roll-out* for each $m \implies$ *needs many samples and* $\boldsymbol{\theta}$ *to be static during roll-out.*

# Policy Optimization: Average Reward Maximization

- **Goal:** Find a policy $\boldsymbol{\theta}$ to maximize the average reward:

$$J(\boldsymbol{\theta}) := \textstyle\sum_{s \in \mathsf{S}, a \in \mathsf{A}} \upsilon_{\boldsymbol{\theta}}(s,a)\, \mathsf{R}(s,a)\ .$$

- What is the gradient of $J(\boldsymbol{\theta})$ *w.r.t.* $\boldsymbol{\theta}$?

$$\nabla J(\boldsymbol{\theta}) = \lim_{T \to \infty} \mathbb{E}_{\boldsymbol{\theta}} \big[ \mathsf{R}(S_T, A_T) \textstyle\sum_{i=0}^{T-1} \nabla \log \Pi_{\boldsymbol{\theta}}(A_{T-i}; S_{T-i}) \big].$$

- We use a *biased* estimate of $\nabla J(\boldsymbol{\theta})$. Let $\lambda \in [0,1)$, consider the approximation [Baxter and Bartlett, 2001]:

$$\lim_{T \to \infty} \widehat{\nabla}_T J(\boldsymbol{\theta}) := \lim_{T \to \infty} \mathsf{R}(S_T, A_T) \sum_{i=0}^{T-1} \lambda^i \, \nabla \log \Pi_{\boldsymbol{\theta}}(A_{T-i}; S_{T-i}).$$

- Online method? design a *Markov chain* that converges to the limit.

# Online Policy Gradient (PG)

**Online policy gradient** [Baxter and Bartlett, 2001, Tadić and Doucet, 2017]:

$$G_{t+1} = \lambda G_t + \nabla \log \Pi_{\boldsymbol{\theta}_n}(A_{t+1}; S_{t+1}) \,, \tag{3a}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \gamma_{t+1} G_{t+1} \, \mathsf{R}(S_{t+1}, A_{t+1}) \,. \tag{3b}$$

- Let the joint state be $X_t = (S_t, A_t, G_t) \in \mathsf{S} \times \mathsf{A} \times \mathbb{R}^d$. Eq. (3b) is SA with the drift term:

$$H(\boldsymbol{\theta}_t; X_{t+1}) = G_{t+1} \, \mathsf{R}(S_{t+1}, A_{t+1})$$

- $\{X_t\}_{t \geq 1}$ forms a Markov chain and

$$h(\boldsymbol{\theta}) = \lim_{T \to \infty} \mathbb{E}_{\tau_T \sim \Pi_{\boldsymbol{\theta}}, \; S_1 \sim \overline{\Pi}_{\boldsymbol{\theta}}} \big[ \widehat{\nabla}_T J(\boldsymbol{\theta}) \big].$$

- We shall next verify (A1)–(A3).

# Convergence Analysis

- Focus on exponential family (or soft-max) policy:

$$\Pi_{\boldsymbol{\theta}}(a; s) = \left\{ \sum_{a' \in \mathsf{A}} \exp\left( \langle \boldsymbol{\theta} \,|\, \boldsymbol{x}(s, a') \rangle \right) \right\}^{-1} \exp\left( \langle \boldsymbol{\theta} \,|\, \boldsymbol{x}(s, a) \rangle \right) .$$

- Assume that $\sup_{s,a} \|\boldsymbol{x}(s, a)\| \leq \overline{b}$ and,

**(PG1)** For any $\boldsymbol{\theta} \in \mathbb{R}^d$, $\{S_t, A_t\}_{t \geq 1}$ is geometrically ergodic. Invariant distribution $\boldsymbol{\upsilon}_{\boldsymbol{\theta}}$ and its Jacobian $\mathsf{J}^{\boldsymbol{\theta}}_{\boldsymbol{\upsilon}_{\boldsymbol{\theta}}}(\boldsymbol{\theta})$ are Lipschitz:

$$\|\boldsymbol{\upsilon}_{\boldsymbol{\theta}} - \boldsymbol{\upsilon}_{\boldsymbol{\theta}'}\| \leq L_Q \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \; \| \mathsf{J}^{\boldsymbol{\theta}}_{\boldsymbol{\upsilon}_{\boldsymbol{\theta}}}(\boldsymbol{\theta}) - \mathsf{J}^{\boldsymbol{\theta}}_{\boldsymbol{\upsilon}_{\boldsymbol{\theta}}}(\boldsymbol{\theta}')\| \leq L_\upsilon \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

- **Consequence**: $J(\boldsymbol{\theta})$ is $\mathsf{R}_{\max} |\mathsf{S}||\mathsf{A}|$-smooth w.r.t. $\boldsymbol{\theta}$,

$$(1 - \lambda)^2 \Gamma^2 + 2 \langle \nabla J(\boldsymbol{\theta}) \,|\, h(\boldsymbol{\theta}) \rangle \geq \|h(\boldsymbol{\theta})\|^2,$$

where $\Gamma := 2\overline{b}\, \mathsf{R}_{\max}\, K_R \frac{1}{(1-\rho)^2}$. Other required assumptions are satisfied too [Karimi et al., 2019].

# Convergence Analysis (cont'd)

**Corollary**

*Under PG1 and set* $\gamma_t = (2c_1 L(1 + C_h)\sqrt{t})^{-1}$. *For any* $t \in \mathbb{N}$, *the algorithm* (3) *finds a policy* $\boldsymbol{\theta}_T$ *with*

$$\mathbb{E}\left[\|\nabla J(\boldsymbol{\theta}_T)\|^2\right] = \mathcal{O}\Big((1-\lambda)^2\Gamma^2 + c(\lambda)\log t/\sqrt{t}\Big), \tag{4}$$

*where* $c(\lambda) = \mathcal{O}(\frac{1}{(1-\max\{\rho,\lambda\}})^2)$. *Expectation taken* w.r.t. $T$, $(A_t, S_t)$.

- It shows the *first convergence rate* for the online PG method.
- *Variance-bias trade-off* with $\lambda \in (0, 1)$: $\lambda \to 1$ reduces the bias, but increases the variance in static term as $c(\lambda) = \mathcal{O}((1-\lambda)^{-2})$.

# Roadmap

1. General Convergence of SA
   Application to Policy Gradient

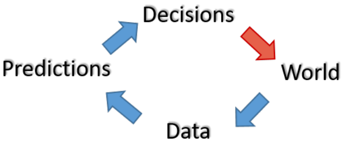
2. Performative Prediction

3. Two-timescale SA for Bilevel Problem

4. Conclusions and Perspectives

# Prediction/Machine Learning in Practice

- Prediction $\in$ a broader system.
- When predictions are used to support decisions, distribution of future observations can be altered.




- **Classical Supervised Learning**: *static world* with i.i.d. data.
- But *decision* (classifier) can cause distribution shift in the *world*.

- **Performative Prediction**: stochastic optimization problem whose data distribution depends on the decision variable[4].

**This Part**: *Performative prediction using SA comes naturally with decision-dependent distribution. Is it stable?*

[4]Thanks to Qiang Li for preparing the slides in this part.

# From Practice to Model

- **Supervised learning:**
  Data $Z = (x, y) \sim \mathcal{D}$.
- **Goal:** minimize *risk*
  $$\min_{\boldsymbol{\theta}} \ \mathbb{E}_{Z \sim \mathcal{D}}[\ell(\boldsymbol{\theta}; Z)]$$

- **Performative Prediction:**
  Data $Z = (x, y) \sim \mathcal{D}(\boldsymbol{\theta})$.
- **Goal:** minimize *performative risk*
  $$\min_{\boldsymbol{\theta}} \ \mathcal{L}(\boldsymbol{\theta}) := \mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta}; Z)]$$

- Perdomo et al. [2020] uses $\mathcal{D}(\boldsymbol{\theta})$ to capture distribution shift (agents' response) of $Z$ due to learner's state $\boldsymbol{\theta}$.
- *How should the learner deal with performativity?*
  - *Agnostic Setting:* SGD/GD on $\ell(z; \boldsymbol{\theta})$ with $z \sim \mathcal{D}(\boldsymbol{\theta})$, e.g., Perdomo et al. [2020], Mendler-Dünner et al. [2020].
  - ✓ Requires no extra knowledge on $\mathcal{L}(\boldsymbol{\theta})$ and agents...
  - *Proactive Setting:* Estimate true gradient of $\nabla \mathcal{L}(\boldsymbol{\theta})$, e.g., Izzo et al. [2021], Miller et al. [2021].
  - ✗ Needs extra knowledge on $\mathcal{L}(\boldsymbol{\theta})$ and agents...

# Greedy Deployment [Mendler-Dünner et al., 2020]

- Two different solutions to performative prediction:

$$\theta_{PO} \in \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{Z \sim \mathcal{D}(\theta)}[\ell(\theta; Z)], \quad \theta_{PS} = \arg\min_{\theta' \in \mathbb{R}^d} \mathbb{E}_{Z \sim \mathcal{D}(\theta_{PS})}[\ell(\theta'; Z)].$$

- In **agnostic setting**, our aim is to get $\theta_{PS}$, e.g., by fixed point iteration. *Can we find it in an online fashion?*

**Greedy deployment** scheme [Mendler-Dünner et al., 2020]:

$$\begin{aligned} \underline{\textit{Learner}}: &\quad \theta_{t+1} = \theta_t - \gamma_{k+1} \nabla \ell(\theta_t; Z_{t+1}), \\ \underline{\textit{Agent}}: &\quad Z_{t+1} \sim \mathcal{D}(\theta_t). \end{aligned}$$

Essentially = SGD but with *data from shifted distribution*.

- **Fact**: $\ell(\cdot; Z)$ is strongly-convex + $\mathcal{D}(\theta)$ is 'insensitive' to $\theta$, then

$$\mathbb{E}[\|\theta_t - \theta_{PS}\|^2] = \mathcal{O}(1/t).$$

# State-dependent Performative Prediction

- *Issue:* Greedy deployment in Mendler-Dünner et al. [2020]:

$$\underline{\textit{Learner}}: \quad \theta_{t+1} = \theta_k - \gamma_{t+1} \nabla \ell(\theta_t; z_{t+1}),$$
$$\underline{\textit{Agent}}: \quad z_{t+1} \sim \mathcal{D}(\theta_t) \quad \leftarrow \text{req. immediate adaptation}$$

- Example: Loan applicants may take months to build up credit history to adapt to changes in classifier of bank.
- **Our Work**: consider *stateful* (or unforgetful) agents[5].
- In other words, both *learner* and *agents* are slow adapters ⇒ fully state dependent performative prediction.

*How to model it? Can the learner still find $\boldsymbol{\theta}_{PS}$?*

[5] Brown et al. [2022] has similar setting but w/o sampling at learner.

# SA for Performative Prediction

- **Idea**: models agents' adaptation via a *controlled Markov Chain*.
  - $\mathrm{P}_{\boldsymbol{\theta}} : \mathsf{Z} \times \mathcal{Z} \to \mathbb{R}_+ =$ Markov kernel w/ stationary dist. $\mathcal{D}(\boldsymbol{\theta})$.

State-dependent Performative Prediction with SA

$$\begin{aligned} &\underline{\textit{Agent}}: \quad Z_{t+1} \sim \mathrm{P}_{\boldsymbol{\theta}_t}(Z_t, \cdot) \quad (\leftarrow \text{allows slow adaptation}) \\ &\underline{\textit{Learner}}: \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1} \nabla \ell(\boldsymbol{\theta}_t; Z_{t+1}) \ \text{ \& deploys } \boldsymbol{\theta}_{t+1}. \end{aligned} \tag{5}$$

- **Example:** agents running SGD to adapt to $z \sim \mathcal{D}(\boldsymbol{\theta})$:

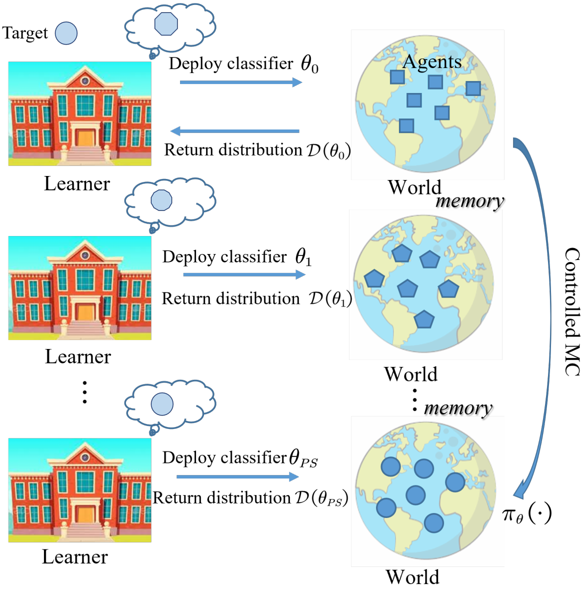
$$Z_{t+1} = Z_t + \alpha \nabla_z U(Z_t; \boldsymbol{\theta}_t, \zeta_{t+1}), \quad \leftarrow U = \text{utility fct.}$$

**Observation**: Learner's updates (5) is biased SA in Part 1 w/

$$H(\boldsymbol{\theta}_t; X_{t+1}) = \nabla \ell(\boldsymbol{\theta}_t; Z_{t+1})$$

Previous result only finds stationary point $\Leftarrow$ stronger guaranteee?

# Illustration - State-dependent Performative Prediction



# Assumptions

**(PP1)**. We assume that $\ell(\boldsymbol{\theta}; Z)$ is $\mu$-strongly convex, $L$-smooth, and the distribution $\mathcal{D}(\boldsymbol{\theta})$ satisfies $\epsilon-$sensitivity ($W_1$ denotes Wasserstein-1 distance)

$$W_1(\mathcal{D}(\boldsymbol{\theta}), \mathcal{D}(\boldsymbol{\theta}')) \leq \epsilon \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \ \forall \ \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d,$$

- PP1 specifies **sensitivity** of $\mathcal{D}(\boldsymbol{\theta})$ to $\boldsymbol{\theta}$ [Mendler-Dünner et al., 2020].
- When agents are *strategic* with a linear utility function, $Z \sim \mathcal{D}(\boldsymbol{\theta})$ if

$$Z = Z_0 + \epsilon \, \boldsymbol{\theta}, \quad Z_0 \sim \mathcal{D}_0 \text{ - some base distribution}$$

**(PP2)**. $\sigma$-perturbation with sampled gradient

$$\sup_{z \in Z} \|\nabla \ell(\boldsymbol{\theta}; z) - \nabla f(\boldsymbol{\theta}; \boldsymbol{\theta}_{PS})\| \leq \sigma \left(1 + \|\boldsymbol{\theta} - \boldsymbol{\theta}_{PS}\|\right).$$

- PP2 allows $\nabla \ell(\boldsymbol{\theta}; z) = \mathcal{O}(1 + \|\boldsymbol{\theta} - \boldsymbol{\theta}_{PS}\|)$ - compatible with strongly convex loss.

# Convergence of SA for Performative Prediction

**Theorem**

Under PP1–PP2, $\mathrm{P}_{\boldsymbol{\theta}}$ satisfies A3. Let $\epsilon < \frac{\mu}{L}$, non-increasing step sizes

$$\frac{\gamma_{t-1}}{\gamma_t} \leq 1 + \frac{\gamma_t(\mu - L\epsilon)}{4}, \quad \gamma_t \leq \min\left\{ \frac{\mu - L\epsilon}{2L^2}, \frac{\mu - L\epsilon}{2\mathcal{C}_2}, \frac{\min\{(\mu - L\epsilon)/3, 3\widehat{L}_P\}}{\mathcal{C}_3 + 3\widehat{L}_P(\mu - L\epsilon)}, \frac{1}{6\widehat{L}_P} \right\}. \quad (6)$$

For any $k \geq 1$, there exists $\mathbb{C}$ where it holds

$$\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{PS}\|^2] \leq \underbrace{\prod_{i=1}^{t}\left(1 - \gamma_i \frac{\mu - L\epsilon}{2}\right)\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{PS}\|^2}_{\text{Transient}} + \underbrace{\mathbb{C}\,\gamma_t}_{\text{Fluctuation}} .$$

- Convergence needs $\epsilon < \mu/L$ (similar to [Mendler-Dünner et al., 2020]) + Step size constrained by mixing time of MC.
- Oscillation of stochastic gradient $\sigma$, mixing time of MC $\widehat{L}$ appear in fluctuation term $\mathbb{C}$.
- Can be extended to general SA with strongly convex objective.
- (in the paper) Convergence to near-stationary point of $\mathcal{L}(\boldsymbol{\theta})$.

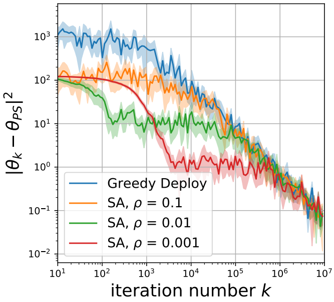
# Simulation – Gausssian Mean Estimation

- Consider the toy problem:

$$\min_{\theta \in \mathbb{R}} \ \mathbb{E}_{z \sim \mathcal{D}(\theta)} \left[ (z-\theta)^2/2 \right], \quad \mathcal{D}(\theta) \equiv \mathcal{N}\left(\bar{z} + \epsilon\theta; \sigma^2\right).$$

- **Agents:** AR model $z_{k+1} = (1-\rho) z_k + \rho \tilde{z}_{k+1}$ with independent r.v. $\tilde{z}_{k+1} \sim \mathcal{N}\left(\bar{z} + \epsilon\theta_k; \sigma^2\right)$ and $\rho \in (0,1)$.

- **Goal**: compare state dependent SA and [Mendler-Dünner et al., 2020].
- Both converge at $\mathcal{O}(1/t)$ to $\theta_{PS}$.
- As $\rho \downarrow 0$, SA is more stable and has a smaller error as the AR has stationary distribution with **lower variance.**



# Simulation – Logistic Regression

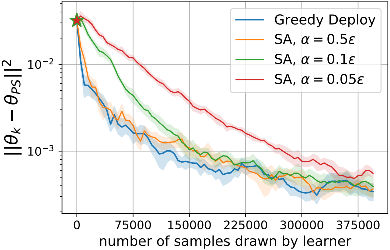
- With *Synthetic Data* for SVM problem:

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{z \sim \mathcal{D}(\theta)} \left[ \frac{\beta}{2} \|\theta\|^2 + \log(1 + \exp(\langle \theta \,|\, x \rangle)) - y \langle \theta \,|\, x \rangle \right]$$

- **Agent Response**: $\mathcal{D}(\theta)$ obtained by the best response, i.e.

$$z_{k+1} \in \arg\max_{z' \in Z} U(z'; \tilde{z}_{k+1} \sim \mathcal{D}_0), \quad U_q(z'; z, \theta) = \langle \theta \,|\, x' \rangle - \frac{\|x' - x\|}{2\epsilon}$$

- **Goal**: the impact of agents' response rate ($\alpha$) on SA.
- As $\alpha \uparrow 1\epsilon$, state-dependent SA $\rightarrow$ greedy deployment [Mendler-Dünner et al., 2020].
- $\alpha \uparrow \Rightarrow$ fast Markov chain $\Rightarrow \widehat{L} \downarrow$.



# Roadmap

# Bilevel Optimization

- Many problems can be described as **bilevel optimization**:

$$\min_{x \in X \subseteq \mathbb{R}^{d_1}} \quad \ell(x) := f(x, y^\star(x)) \qquad \text{s.t.} \quad y^\star(x) \in \arg\min_{y \in \mathbb{R}^{d_2}} \; g(x, y), \tag{Bi}$$

- **Upper-level** = *leader* / decision maker, **lower-level** = *follower*.
- Related to *mathematical program with equilibrium constraint (MPEC)* Luo et al. [1996], *stackelberg game* Stackelberg [1952].
- **Applications**: meta learning, policy optimization, etc..

**This Part**: $f, g$ are *stochastic functions* – $f(x, y) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, y, \xi)]$.
$\Rightarrow$ Consider tackling upper level by SA: samples $y^\star(x)$ are **decision-dependent**: there are more structure than previous part.

# Motivation: Policy Optimization via Actor Critic

- Consider **tabular policy** given by $\pi : \mathsf{S} \times \mathsf{A} \to \mathbb{R}_+$ with $|\mathsf{S}|, |\mathsf{A}| < \infty$.
- Let $\rho_0$ be init. distribution, the $\gamma$-discounted reward[6] of $\pi$ is:

$$\mathbb{E}_\pi[Q^\pi(S,A)] = \mathbb{E}_{S\sim\rho_0}\left[\langle Q^\pi(S,\cdot) \,|\, \pi(\cdot|S)\rangle\right],$$

$$\text{with} \quad Q^\pi(S,A) = \mathbb{E}_\pi\left[\sum_{t\geq 0} \gamma^t R(S_t,A_t) | S_0 = S, A_0 = A\right]$$

- Note $Q^\pi(S,A)$ is $\gamma$-discounted reward (Q-function) given init. $(S,A)$.
- With **fixed** $\pi$, $Q^\pi(S,A)$ can be evaluated by solving Bellman equation; or through linear approximation $Q^\pi(S,A) \approx \langle \theta^\star(\pi) \,|\, \phi(S,A)\rangle$.

A **Bilevel Optimization** problem:

$$\min_{\pi \in X \subseteq \mathbb{R}^{|S|\times|A|}} \ell(\pi) = -\langle Q_{\theta^\star(\pi)}, \pi \rangle_{\rho_0} \qquad \text{(Actor)}$$

$$\text{s.t.} \quad \theta^\star(\pi) \in \arg\min_{\theta\in\mathbb{R}^d} \ \tfrac{1}{2}\| Q_\theta - R - \gamma P^\pi Q_\theta \|^2_{\mu^\pi \otimes \pi}. \qquad \text{(Critic)}$$

[6] In Part 1, we have considered average reward with paramterized policy.

# Tackling the Bilevel Problem (Bi)

- Recall the bi-level optimization problem:

$$\min_{x \in X \subseteq \mathbb{R}^{d_1}} \ell(x) \iff \begin{array}{rl} \min\limits_{x \in X \subseteq \mathbb{R}^{d_1}} & \ell(x) := f(x, y^\star(x)) \\ \text{s.t.} & y^\star(x) \in \arg\min\limits_{y \in \mathbb{R}^{d_2}} g(x, y), \end{array}$$

- The gradient of $\ell(x)$ is:

$$\nabla_x \ell(x) = \nabla_x f(x, y^\star) - \nabla^2_{xy} g(x, y^\star)[\nabla^2_{yy} g(x, y^\star)]^{-1} \nabla_y f(x, y^\star)$$

**Stationary Condition**: (Bi) can be tackled by finding $(x^\star, y^\star)$ s.t.

$$F(x, y) = 0, \quad G(x, y) = \nabla_y g(x, y) = 0$$

where $F(x, y) = \nabla_x f(x, y) - \nabla^2_{xy} g(x, y)[\nabla^2_{yy} g(x, y)]^{-1} \nabla_y f(x, y)$

# Finding Fixed Points with Stochastic Samples

- We only have **stochastic samples** and the problems are **coupled**.
- Let $\xi_{k+1}$ denotes the random 'seed' at iteration $k$, and $F(\cdot;\xi_{k+1})$, $G(\cdot;\xi_{k+1})$ denote the stochastic samples of $F, G$, respectively.
- If $x$ **is fixed** and under suitable conditions, the recursion

$$y_{k+1} = y_k + \beta_k G(x, y_k; \xi_{k+1}) \stackrel{k\to\infty}{\longrightarrow} y^\star(x) \text{ s.t. } G(x, y^\star(x)) = 0.$$

- Furthermore, the recursion

$$x_{k+1} = x_k + \alpha_k F(x_k, y^\star(x_k); \xi_{k+1}) \stackrel{k\to\infty}{\longrightarrow} x^\star \text{ s.t. } F(x^\star, y^\star(x^\star)) = 0.$$

- If one could run the two recursions $\Rightarrow$ fixed point, but the $y_k$ recursion requires $x$ to be fixed; and $x_k$ recursion requires $y^\star(x_k)$.

*Suggesting a double-loop algorithm? e.g., [Ghadimi and Wang, 2018].*

# Two Timescale Stochastic Approximation (TTSA)

- Consider a **single-loop, two timescale** algorithm [Borkar, 1997]:

$$x_{k+1} = x_k + \alpha_k F(x_k, y_k; \xi_{k+1})$$
$$y_{k+1} = y_k + \beta_k G(x_k, y_k; \xi_{k+1})$$

- We require that

$$\lim_{k \to \infty} \frac{\alpha_k}{\beta_k} = 0$$

  $x$-update is at slow timescale; while $y$-update is at fast timescale.

- **Intuition**: when updating $y_k$, as $\alpha_k \ll \beta_k$, then $x_k$ is *almost static*; when updating $x_k$, the used $y_k$ have *almost converged* to $y^\star(x_k)$.

# TTSA for Tackling (Bi): The Algorithm

## TTSA Algorithm for (Bi)

Follow the recursion:

$$x_{k+1} = x_k - \alpha_k h_f^k \qquad [h_f^k \approx F(x^k, y^k)]$$
$$y_{k+1} = y_k - \beta_k \nabla_y g(x_k, y_k; \zeta_{k+1}) \qquad \text{(TTSA-Bi)}$$

- $x_k$ update uses **decision-dependent data** via $y_k$ driven by $x_{k-1}$.
- **Two timescale** step sizes to *balance* upper and lower level updates.

- **Challenge**: *easy* to estimate $G(\cdot) = \nabla_y g(\cdot)$, but $F(\cdot)$ is non-trivial since

$$F(x,y) = \nabla_x f(x,y) - \nabla^2_{xy} g(x,y) \underbrace{[\nabla^2_{yy} g(x,y)]^{-1}}_{\text{can't replace by } \nabla^2_{yy} g(x,y;\zeta)} \nabla_y f(x,y)$$

*Biased* estimate is possible; see details in the paper.

# This Work

We characterize the **rate of convergence** for TTSA when:

- the inner objective $g(x, y)$ is strongly convex in $y$, and
- the outer objective $\ell(x)$ is *smooth, convex, strongly convex*.

| $\ell(x)$ | CONSTRAINT | STEP SIZE $(\alpha_k, \beta_k)$ | RATE (OUTER) | RATE (INNER) |
|---|---|---|---|---|
| SC | $X \subseteq \mathbb{R}^{d_1}$ | $\mathcal{O}(k^{-1})$, $\mathcal{O}(k^{-2/3})$ | $\mathcal{O}(K^{-2/3})$ | $\mathcal{O}(K^{-2/3})$ |
| WC | $X \subseteq \mathbb{R}^{d_1}$ | $\mathcal{O}(K^{-3/5})$, $\mathcal{O}(K^{-2/5})$ | $\mathcal{O}(K^{-2/5})$ | $\mathcal{O}(K^{-2/5})$ |

**Prior Works** — many and

- *Linear TTSA* $\approx$ *solving* **quadratic** *upper/lower level*
  - Dalal et al. [2018, 2019] obtained high probability bounds with a **projection** step, recent work [Kaledin et al., 2020].
- *Finite-time Analysis of Bilevel Stochastic Optimization*
  - [Couellan and Wang, 2016, Ghadimi and Wang, 2018] – double loop SA & recently [Yang et al., 2021, Chen et al., 2021, Guo and Yang, 2021].

# General Assumptions (Informal)

**(TT1).** Consider the upper-level function $f(x,y)$, $\ell(x) = f(x, y^\star(x))$:

1. $\nabla_y f(x,y)$ is Lipschitz in $(x,y)$ + bounded; $\nabla_x f(x,y)$ is Lipschitz in $y$.
2. The objective function is **$\mu_\ell$-weakly convex**

$$\ell(w) \geq \ell(v) + \langle \nabla \ell(v), w - v \rangle + \mu_\ell \|w - v\|^2, \ \forall\ w, v \in X.$$

**(TT2).** Consider the lower-level function $g(x,y)$:

1. For any $x \in X$, $g(x,y)$ is **strongly convex** in $y$.
2. The Jacobian/Hessian $\nabla^2_{xy} g(x,y), \nabla^2_{yy} g(x,y)$ are Lipschitz in $(x,y)$. Moreover, $\nabla^2_{xy} g(x,y)$ is bounded.

## Key Consequence

Under TT1–TT2, the following holds:

$$\|F(x,y) - \nabla \ell(x)\| \leq L\|y^\star(x) - y\|, \ \|y^\star(x_1) - y^\star(x_2)\| \leq L_y \|x_1 - x_2\|$$

# Main Results (Strongly Convex $\ell$)

**Theorem**

*Under TT1, TT2, suppose that $\mu_\ell > 0$, then*

$$\mathbb{E}[\|x^k - x^\star\|^2] \lesssim \underbrace{\prod_{i=0}^{k-1}\Big(1 - \mu_\ell \alpha_i\Big)\mathrm{V}_0}_{\text{transient term - decay exponentially}} + \underbrace{\alpha_k^{2/3}}_{\text{steady state term}}$$

$$\mathbb{E}[\|y^k - y^\star(x^{k-1})\|^2] \lesssim \prod_{i=0}^{k-1}\Big(1 - \beta_i \mu_g / 4\Big)\mathrm{V}_0 + \beta_k$$

*where $\mathrm{V}_0$ depends on the initialization, the inequality is up to constants not depending on k (exact expressions can be found in the paper)*

- **Consequence**: if we set $\alpha_k = c_\alpha/(k + k_\alpha), \beta_k = c_\beta/(k + k_\beta)^{2/3}$,

$$\Delta_x^k = \mathcal{O}(1/k^{2/3}), \quad \Delta_y^k = \mathcal{O}(1/k^{2/3}).$$

# Main Results (Weakly Convex $\ell$)

**Theorem**

*Under TT1, TT2, suppose that $\mu_\ell \in \mathbb{R}$. Set $\mathsf{K} \sim \mathcal{U}\{0, \ldots, K-1\}$ and $\alpha_k \asymp K^{-3/5}, \beta_k \asymp K^{-2/5}$. For sufficiently large $K \geq 1$, it holds*

$$\mathbb{E}[\|\nabla \ell(x^{\mathsf{K}})\|^2] \lesssim \left[ L^2 \left( \Delta^0 + \frac{\sigma^2}{\mu_g^2} \right) + \mu_g \sigma^2 \right] \frac{K^{-\frac{2}{5}}}{|\mu_\ell|^2},$$

$$\mathbb{E}[\|y^{\mathsf{K}} - y^\star(x^{\mathsf{K}-1})\|^2] \lesssim \left[ \frac{\Delta^0}{\mu_g} + \frac{\sigma^2}{\mu_g^2} + \frac{\mu_g \sigma^2}{L^2} \right] K^{-\frac{2}{5}},$$

*where $\Delta^0$ depends on the initialization, the inequality is up to constants not depending on $k$ (exact expressions can be found in the paper)*

- **Consequence**: we get $\mathbb{E}[\tilde{\Delta}_x^{\mathsf{K}}] = \mathcal{O}(1/K^{2/5})$, $\mathbb{E}[\Delta_y^{\mathsf{K}}] = \mathcal{O}(1/K^{2/5})$.
- **Note**: $\tilde{\Delta}_x^{\mathsf{K}}$ is a stationarity measure for $x^{\mathsf{K}}$ related to Moreau envelope.
- Actor-critic requires slightly different algorithm than (TTSA-Bi); but similar analysis applies.

# Reflection: why two-timescale?

- In the upper-level update,

$$x_{k+1} = x_k - \alpha_k h_f^k \leftarrow \text{note } h_f^k \approx F(x,y) \neq \nabla \ell(x)$$

- We recall that $\|F(x,y) - \nabla \ell(x)\| = \mathcal{O}(\|y - y^\star(x)\|)$.
- Need $\alpha_k \leq c_0 \beta_k^{3/2}$ to balance the errors, leading to the step sizes

$$\text{strongly convex } \ell(x): \quad \alpha_k \asymp k^{-1}, \beta_k \asymp k^{-2/3}$$
$$\text{weakly convex } \ell(x): \quad \alpha_k \asymp K^{-3/5}, \beta_k \asymp K^{-2/5}$$

- Ultimately, the convergence rate is **limited** by the 'faster' timescale; also see [Kaledin et al., 2020].
- For 1-level problem, even naive SGD achieves $\mathbb{E}[\tilde{\Delta}_x^K] = \mathcal{O}(1/K^{1/2})$.

*Accelerated bilevel optimization? Yes, [Khanduri et al., 2021].*

# Agenda

# Summary

We have studied variants of SA with decision dependent data:

$$\underline{\text{SA}}: \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1} H(\boldsymbol{\theta}_t; X_{t+1}),$$

where $X_{t+1}$ is not i.i.d., and depends on $\boldsymbol{\theta}_t$ (via a controlled MC).

- *General SA* with possibly non-gradient $H(\boldsymbol{\theta}; X)$:
  $\Rightarrow$ convergence to stationary point $\mathbb{E}[\|h(\boldsymbol{\theta}_T)\|^2] = \mathcal{O}(\log T/\sqrt{T})$.
  $\Rightarrow$ application to online policy gradient.
- *Performative Prediction* through SA:
  $\Rightarrow$ modelling stateful agents through controlled MC.
  $\Rightarrow$ convergence to PS solution $\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{PS}\|^2] = \mathcal{O}(1/t)$.
- *Bilevel optimization* via TTSA:
  $\Rightarrow$ utilizes two timescales for coupled SAs & application to actor-critic.
  $\Rightarrow$ convergence rates to stationary solution.

# Perspectives

SA with decision dependent data:

$$\underline{\text{SA}}: \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1} H(\boldsymbol{\theta}_t; X_{t+1}),$$

where $X_{t+1}$ is not i.i.d., and depends on $\boldsymbol{\theta}_t$ (via a controlled MC).

**Theory**:

- Current results require 'strong' assumptions on MC which makes sense only for finite-state space, see [Durmus et al., 2021b].
- Strong convergence, e.g., with high probability [Durmus et al., 2021a].
- Avoid saddle point in non-convex problems? [Lee et al., 2019]

**Applications/Algorithmic**:

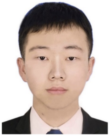
- Decentralized & federated learning; see [Wai, 2020].
- Beyond reinforcement learning & performative prediction — Langevin Monte-carlo [De Bortoli et al., 2021], search engine optimization [Avrachenkov et al., 2022], etc.

# Most importantly, thanks to ...

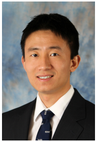
Belhal Karimi
(Baidu Research)

Blazej Miasojedow
(U of Warsaw)

Eric Moulines
(Ecole Polytechnique)



Qiang Li
(CUHK)



Mingyi Hong
(UMN)

Zhuoran Yang
(Yale)

Zhaoran Wang
(Northwesthern)

And thank you all for attending! Questions?

For more info: `http://www1.se.cuhk.edu.hk/~htwai/`

# References I


Konstantin Avrachenkov, Kishor Patil, and Gugan Thoppe. Online algorithms for estimating change rates of web pages. *Performance Evaluation*, 153:102261, 2022.

Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692, 2018.

Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5): 291–294, 1997.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *AISTATS*, 2022.

Shuhang Chen, Adithya Devraj, Ana Busic, and Sean Meyn. Explicit mean-square error bounds for monte-carlo and linear stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 4173–4183. PMLR, 2020.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*, 2021.

Nicolas Couellan and Wenjuan Wang. On the convergence of stochastic bi-level gradient methods. *Optimization*, 2016.

Gal Dalal, Gugan Thoppe, Balázs Szörényi, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*, pages 1199–1233, 2018.

# References II


Gal Dalal, Balazs Szorenyi, and Gugan Thoppe. A tale of two-timescale reinforcement learning with the tightest finite-time bound. *arXiv preprint arXiv:1911.09157*, 2019.

Valentin De Bortoli, Alain Durmus, Marcelo Pereyra, and Ana F Vidal. Efficient stochastic optimisation by unadjusted langevin monte carlo. *Statistics and Computing*, 31(3):1–18, 2021.

Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer, 2018.

John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. Tight high probability bounds for linear stochastic approximation with fixed stepsize. *Advances in Neural Information Processing Systems*, 34, 2021a.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Hoi-To Wai. On the stability of random matrix product with markovian noise: Application to linear stochastic approximation and td learning. In *Conference on Learning Theory*, pages 1711–1752. PMLR, 2021b.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Zhishuai Guo and Tianbao Yang. Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: Performative gradient descent. In *ICML*, 2021.

Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In *Conference on Learning Theory*, pages 2144–2203. PMLR, 2020.

# References III


Belhal Karimi, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR, 2019.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *arXiv preprint arXiv:2102.07367*, 2021.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1):311–337, 2019.

Zhi-Quan Luo, Jong-Shi Pang, and Daniel Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, 1996. doi: 10.1017/CBO9780511983658.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.

John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720. PMLR, 2021.

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. *arXiv preprint arXiv:2004.04719*, 2020.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *ICML*, 2020.

# References IV


Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

H. Van Stackelberg. *The theory of market economy*. Oxford University Press, 1952.

Tao Sun, Yuejiao Sun, and Wotao Yin. On Markov chain gradient descent. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9918–9927. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8195-on-markov-chain-gradient-descent.pdf.

Vladislav B Tadić and Arnaud Doucet. Asymptotic bias of stochastic gradient search. *The Annals of Applied Probability*, 27(6):3255–3304, 2017.

Hoi-To Wai. On the convergence of consensus algorithms with markovian noise and gradient bias. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 4897–4902. IEEE, 2020.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *arXiv preprint arXiv:2106.04692*, 2021.

**Proof Sketch**: From the $L$-smoothness of $V$, we have

$$\underbrace{V(\boldsymbol{\theta}_{k+1}) - V(\boldsymbol{\theta}_k)}_{\text{telescoping sum} \to \text{repeated terms are cancelled}} \leq$$

$$\underbrace{-\gamma_{k+1} \langle \nabla V(\boldsymbol{\theta}_k) \,|\, h(\boldsymbol{\theta}_k) \rangle + \frac{\gamma_{k+1}^2 L}{2} \| h(\boldsymbol{\theta}_k) + \boldsymbol{e}_{k+1} \|^2}_{\text{sum controlled by biasedness + others}} \underbrace{-\gamma_{k+1} \langle \nabla V(\boldsymbol{\theta}_k) \,|\, \boldsymbol{e}_{k+1} \rangle}_{\text{good if summable!}}$$

**Idea —** under mild conditions, there exists $\hat{H}_{\boldsymbol{\theta}}(\cdot)$ such that $\boldsymbol{e}_{k+1} = \hat{H}_{\boldsymbol{\theta}_k}(X_{k+1}) - P_{\boldsymbol{\theta}_k} \hat{H}_{\boldsymbol{\theta}_k}(X_{k+1})$ (Poisson equation), consequently,

$$\textstyle\sum_{k=0}^n \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\theta}_k) \,\middle|\, \hat{H}_{\boldsymbol{\theta}_k}(X_{k+1}) - P_{\boldsymbol{\theta}_k} \hat{H}_{\boldsymbol{\theta}_k}(X_{k+1}) \right\rangle \equiv A_1 + A_2 + A_3 + A_4 + A_5$$

$$\text{Martingale} \to A_1 = \textstyle\sum_{k=1}^n \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\theta}_k) \,\middle|\, \hat{H}_{\boldsymbol{\theta}_k}(X_{k+1}) - P_{\boldsymbol{\theta}_k} \hat{H}_{\boldsymbol{\theta}_k}(X_k) \right\rangle$$

$$\text{Smoothness} \to A_2 = \textstyle\sum_{k=1}^n \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\theta}_k) \,\middle|\, P_{\boldsymbol{\theta}_k} \hat{H}_{\boldsymbol{\theta}_k}(X_k) - P_{\boldsymbol{\theta}_{k-1}} \hat{H}_{\boldsymbol{\theta}_{k-1}}(X_k) \right\rangle$$

$$\text{Smoothness} \to A_3 = \textstyle\sum_{k=1}^n \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\theta}_k) - \nabla V(\boldsymbol{\theta}_{k-1}) \,\middle|\, P_{\boldsymbol{\theta}_{k-1}} \hat{H}_{\boldsymbol{\theta}_{k-1}}(X_k) \right\rangle$$

$$\text{Step size} \to A_4 = \textstyle\sum_{k=1}^n (\gamma_{k+1} - \gamma_k) \left\langle \nabla V(\boldsymbol{\theta}_k) \,\middle|\, P_{\boldsymbol{\theta}_{k-1}} \hat{H}_{\boldsymbol{\theta}_{k-1}}(X_k) \right\rangle$$

$$\text{Finite number} \to A_5 = \gamma_1 \left\langle \nabla V(\boldsymbol{\theta}_0) \,\middle|\, \hat{H}_{\boldsymbol{\theta}_0}(X_1) \right\rangle - \gamma_{n+1} \left\langle \nabla V(\boldsymbol{\theta}_n) \,\middle|\, P_{\boldsymbol{\theta}_n} \hat{H}_{\boldsymbol{\theta}_n}(X_{n+1}) \right\rangle$$