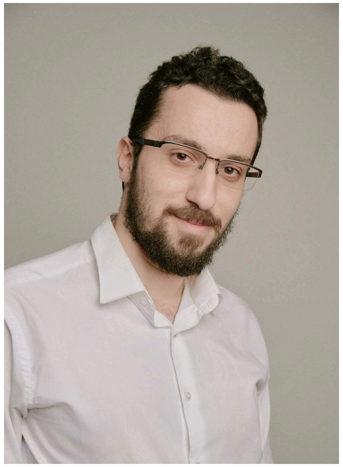
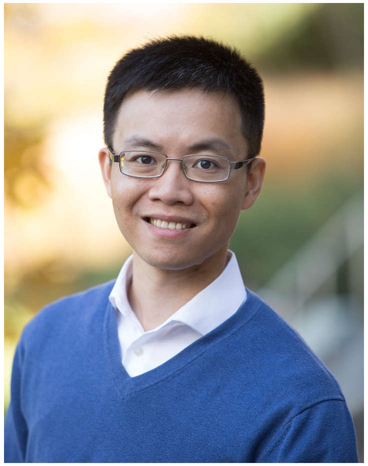
# Stochastic Methods in Variational Inequalities: Ergodicity, Bias and Refinements




Emmanouil V. Vlatakis (UC Berkeley)
**On job market**

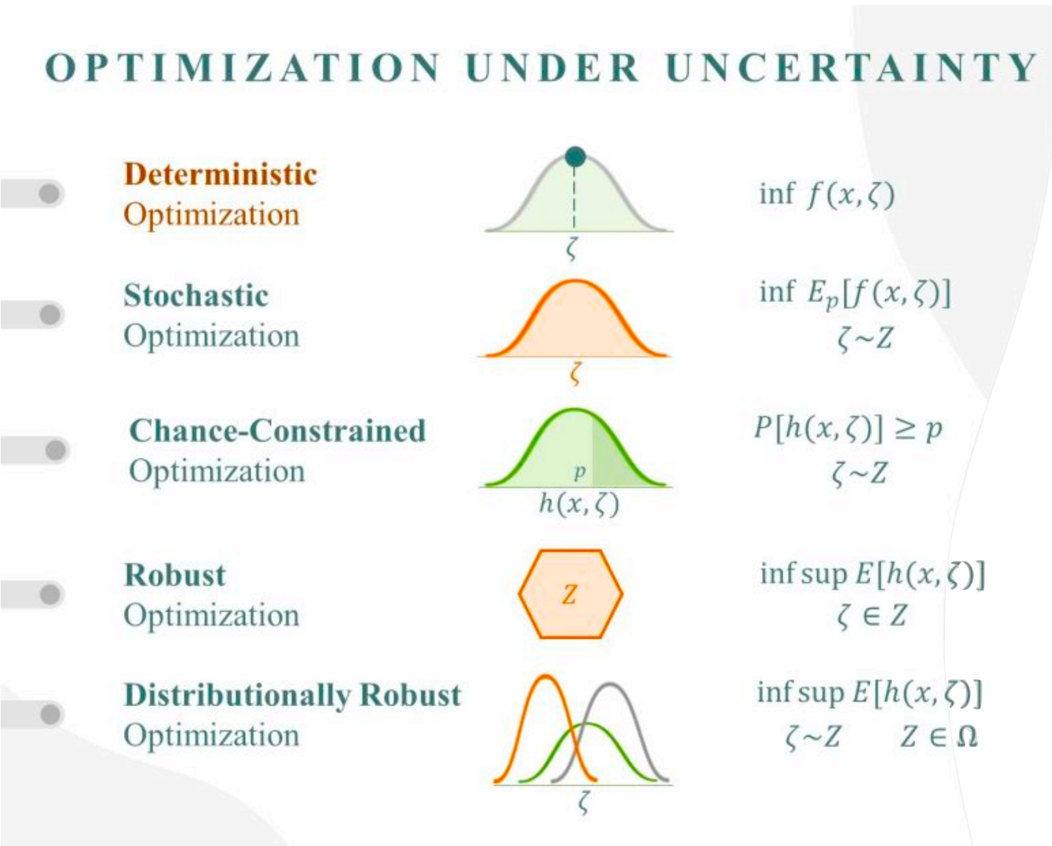
Angeliki Giannou (UW-Madison)

Qiaomin Xie (UW-Madison)

Yudong Chen (UW-Madison)

# Motivation

- Probably, any interesting ML problem can be categorized in one of the following classical frameworks:





**Optimizers' Task**

For a problem Π:

*Find an algorithm/method*

*to compute efficiently*

$$\left\{ \text{SOL}(\Pi, Z) \right\}$$

# Motivation

- In this work, we aim to answer a question in the intersection of these three worlds:


**Our Task**

For a problem Π & a method Alg :

*Does* Alg *produce an unbiased estimator for the solution of our problem ?*

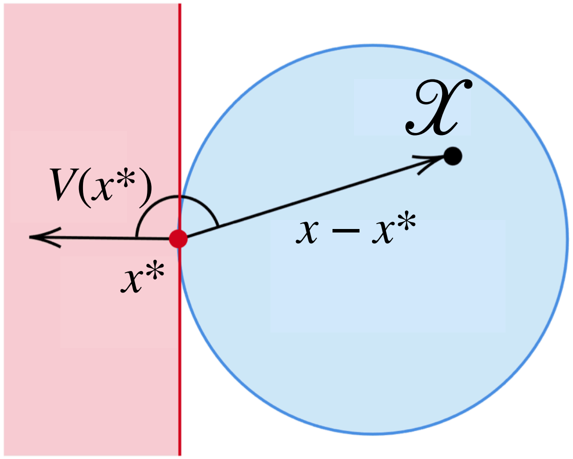
$$\Pr[\mathrm{Alg}(\Pi, Z_n) \xrightarrow[n\to\infty]{} \mathrm{SOL}(\Pi, Z)] = 1$$

# Variational Inequality Problem (VIP)

- Variational Inequality Problem:

$$\text{Find } x^* \in \mathcal{X} \subseteq \mathbb{R}^d \text{ s.t. } \langle V(x^*), x - x^* \rangle \geq 0, \text{ for all } x \in \mathcal{X}$$
$$\text{where } V : \mathcal{X} \to \mathbb{R}^d \text{ is some operator.}$$

- Example 1: **Loss minimization**
  - $V = \nabla f$: the gradient of some loss function $f: \mathbb{R}^d \to \mathbb{R}$
  - VIP: find a stationary point of $f$, i.e., $\nabla f(x^*) = 0$




- Example 2: **Fixed point problem**
  - $V(x) = F(x) - x$ for some function $F$
  - VIP: solves the fixed-point equation $F(x^*) - x^* = 0$.

- Example 3: **Saddle-point problem**
  - $L: \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$, $L(x_1, x_2)$: cost for player choosing $x_1$, payoff for player $x_2$
  - $V = (\nabla_{x_1} L, -\nabla_{x_2} L)$, VIP finds saddle point of $L$: $\min_{x_1} \max_{x_2} L(x_1, x_2)$

# Stochastic Methods for VIP

- Variational Inequality Problem:

$$\text{Find } x^* \in \mathcal{X} \subseteq \mathbb{R}^d \text{ s.t. } \langle V(x^*), x - x^* \rangle \geq 0, \text{ for all } x \in \mathcal{X}$$
$$\text{where } V : \mathcal{X} \to \mathbb{R}^d \text{ is some operator.}$$

- Typically, the exact function $V$ is **unknown, corrupted, biased**
- The optimizer has access to **stochastic** estimate of $V$: given an input $x$,

  $\hat{V}(x) = V(x) + U(x)$, where $U(\cdot)$ is any kind of noise/sample/uncertainty

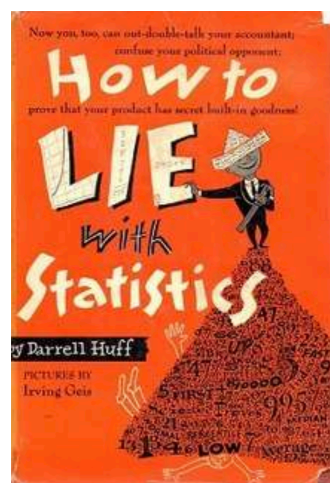
- Goal: Use stochastic estimates to find $x^*$


The New York Times

Opinion

OP-ED CONTRIBUTOR

When an Algorithm Helps Send You to Prison




# Stochastic Methods for VIP: SGDA and SEG

- Stochastic Gradient Descent Ascent (SGDA) [Nemirovski et al '09]:

$$x_{t+1} = x_t - \gamma_t \Big( V(x_t) + U_t(x_t) \Big)$$

  - $\gamma_t > 0$: stepsize
  - For loss minimization problems: SGDA reduces to SGD

- Stochastic Extra Gradient (SEG) [Korpelevich '76]: at each iteration *t*

$$x_{t+1/2} = x_t - \gamma_t \Big( V(x_t) + U_{t+1/2}(x_t) \Big), \qquad \text{\%Extra look-ahead step}$$

$$x_{t+1} = x_t - \eta_t \Big( V(x_{t+1/2}) + U_t(x_{t+1/2}) \Big) \qquad \text{\%update}$$

  - $\gamma_t > 0, \ \eta_t > 0$: stepsizes

- Classical asymptotic convergence results with **diminishing** stepsizes

SGDA: $\sum_t \gamma_t = \infty, \ and \sum_t \gamma_t^2 < \infty$

SEG: $\sum_t \gamma_t \eta_t = \infty, \ \sum_t \gamma_t^2 \eta_t < \infty \ and \ \sum_t \eta_t^2 < \infty$ [Hsieh '20]

Standard Example: $\gamma_t = 1/\sqrt{t}$

For Simplicity $\mathcal{X} = \mathbb{R}^d$

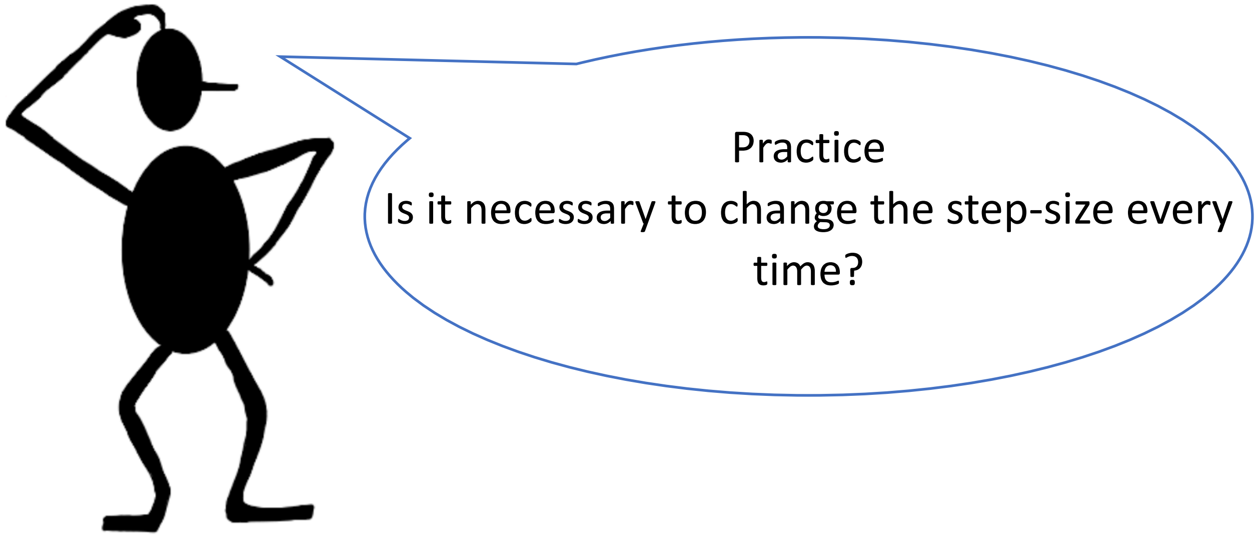
# Our Focus: SGDA/SEG with Constant Stepsizes

- St

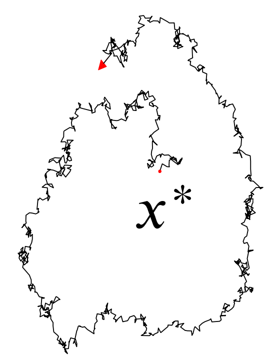
$x_{t+1}$

- St

p




- Using constant stepsizes:
  - Might be non-convergent
  - But faster converges to the neighborhood
- Goal: A fine-grained characterization of ***distributional*** behaviors of SEG/SGDA with constant stepsize



# Recent Non-asymptotic Results (Incomplete List)

- SGDA/SEG and variants: constant or diminishing stepsizes
  - **Upper-bound** on *mean-squared error (MSE)* $\mathbb{E}\left\|x_t - x^*\right\|^2$ *or*

    *vector-field amplitude* $\mathbb{E}\left\|V(x_t)\right\|^2$ *or other metrics …*

  [Gorbuno-Berard-Gidel-Loizou, '22] [Gorbunov-Loizou-Gidel '22] [Hsie-Iutzeler-Malick-Mertikopoulos, '20] [Beznosikov-Gorbunov-Berard-Loizou, '23]…

- Special case of VI: Constant stepsize SGD and Stochastic approximation
  - Study $\{x_t\}$ from the lens of Markov chain
    - Distributional convergence, characterization of stationary distribution
  - SGD for strongly convex objectives: [Dieuleveut-Durmus-Bach '20]
  - SGD for non-smooth non-convex functions: [Yu-Balasubramania-Volgushev-Erdogdu, '21]
  - Linear stochastic approximation with Markovian data: [Huo-Chen-Xie, '23]

*Today Question…What is the distribution of* $x_t$?

# This Talk: Weak Quasi Strongly Monotone Operator V

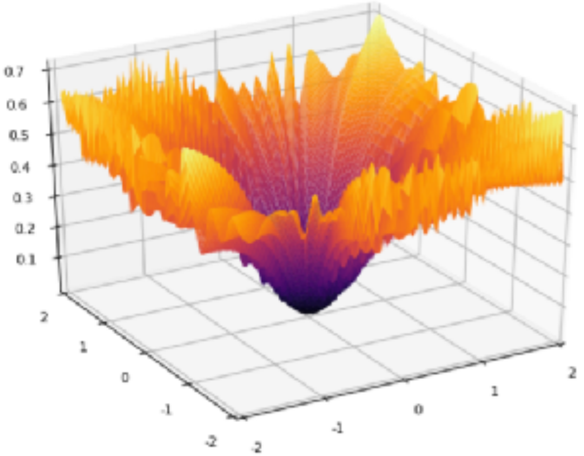
- The operator V is **$\lambda$-weak** **$\mu$-quasi** strongly monotone with $\lambda \geq 0,\ \mu > 0$

$$\langle V(x), x - x^* \rangle \geq \mu \left\| x - x^* \right\|^2 - \lambda, \qquad \forall x \in \mathbb{R}^d.$$

- (Quasi-)strong **monotonicity**
  - Resemble the notion of (quasi-)strong **convexity** in optimization literature
- **$\mu$-quasi** strongly monotone: relaxation of $\mu$-strong monotone:

$$\langle V(x) - V(x'), x - x' \rangle \geq \mu \left\| x - x' \right\|^2, \qquad \forall x,\ x' \in \mathbb{R}^d.$$

- **$\lambda$-weak**
  - Resemble the notion of weak convex optimization




- Assume the operator $V$ is at most $L$-linear growth, i.e.,
  $\left\| V(x) \right\| \leq L(1 + \|x\|), \qquad \forall x \in \mathbb{R}^d.$

For Simplicity $\mathcal{X} = \mathbb{R}^d$

# Our Analytical Approach: the Lens of Markov Chain

- Stochastic Gradient Descent Ascent (SGDA):

$$x_{t+1} = x_t - \gamma\Big(V(x_t) + U_t(x_t)\Big)$$

- Stochastic Extra Gradient (SEG) [Korpelevich '76]: at iteration *t*

$$x_{t+1/2} = x_t - \gamma\Big(V(x_t) + U_{t+1/2}(x_t)\Big), \qquad \text{\%Extra look-ahead step}$$

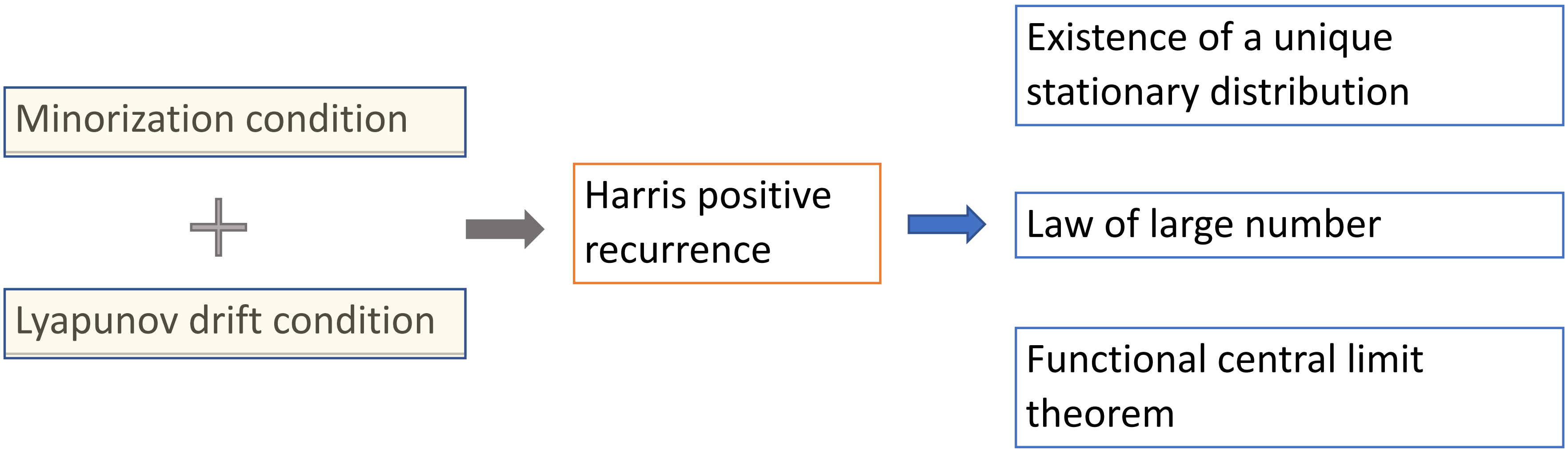
$$x_{t+1} = x_t - \eta\Big(V(x_{t+1/2}) + U_t(x_{t+1/2})\Big) \qquad \text{\%update}$$

- Assumptions on noise:
  - Zero-mean: $\|\mathbb{E}[U_t(x_t) \mid \mathcal{F}_t]\| \le b_{\text{bias}}$;
  - Bounded variance: $\mathbb{E}[\|U_t(x_t)\|^2 \mid \mathcal{F}_t] \le \sigma^2_{\text{variance}} + \rho^2 d(x_t, \mathcal{X}^*)$

- Key observations: with **constant stepsizes**,
  - the iterates $\{x_t\}_{t\ge 0}$ of SGDA/SEG forms a **homogeneous Markov chain in** $\mathbb{R}^d$.

# Roadmap for Understanding Distributional Properties

- For a homogeneous Markov chain $\{x_t\}_{t\geq 0}$ in **continuous** state space $\mathbb{R}^d$:




[Meyn-Tweedie, '09]

# First Result: Convergence up to Constant Factors

**Theorem 1**

Under previous assumptions, for SGDA with $\gamma$ satisfies $\gamma < \frac{\mu}{L^2}$, then for any initial point $x_0 \in \mathbb{R}^d$,

$$\mathbb{E}[\|x_t - x^*\|^2] \leq (1 - c_1)^t \|x_0 - x^*\|^2 + c_2,$$

with $c_1 \gtrsim \mu\gamma$, $c_2 \lesssim \frac{\lambda + \gamma\sigma^2}{\mu}$.

- Similar guarantee for SEG
- Byproduct of the proof: **Geometric** Lyapunov drift condition

$$\mathbb{E}\left[W(x_{t+1}) - W(x_t) \mid \mathcal{F}_t\right] \leq -\beta W(x_t) + b\mathbb{1}_C(x)$$

where $W(x) := \|x - x^*\|^2 + 1$, and C is bounded set.

# Main Results: Harris Positive Recurrence of Markov Chain

**Theorem 2**

Under previous assumptions, the iterates $\{x_t\}_{t\geq 0}$ of SGDA/SEG is a Harris positive recurrent Markov chain.

1. It admits a unique stationary distribution $\pi_\gamma$ ;
2. For each test function $\phi: \mathbb{R}^d \to \mathbb{R}$ with $\|\phi(x)\| \leq L_\phi(1 + \|x\|)$ ,

$$\left|\mathbb{E}[\phi(x_t)] - \mathbb{E}_{\pi_\gamma}[\phi(x)]\right| \leq c\rho^t,$$

where $\rho \in (0,1)$;

# Main Results: LLN and CLT of Averaged Iterates

**Theorem 3**

Under previous assumptions, for any function $\phi$ with $\pi_\gamma(|\phi|) < \infty$,

1. **(LLN)** $\frac{1}{T}\sum_{t=0}^{T-1} \phi(x_t) \to \mathbb{E}_{\pi_\gamma}[\phi(x)]$, a.s.;
2. **(CLT)** $\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1} \left[\phi(x_t) - \mathbb{E}_{\pi_\gamma}[\phi(x)]\right] \xrightarrow{d} N(0, Var_{\pi_\gamma}(\phi))$.

- Implication: Statistical inference
  - CLT results can be used for constructing confidence intervals.
- But how far $\mathbb{E}_{\pi_\gamma}[x]$ is away from $x^*$?

# Main Results: Bias Characterization w.r.t. step-size

**Theorem 4**

Under previous assumptions, for SGDA with stepsize $\gamma < \bar{\gamma}$,

$$\mathbb{E}_{\pi_\gamma}[x] - x^* = \boldsymbol{\gamma \Delta(x^*)} + O(\gamma^2),$$

with $\boldsymbol{\Delta(x^*)}$ being **independent** of the stepsize $\boldsymbol{\gamma}$.

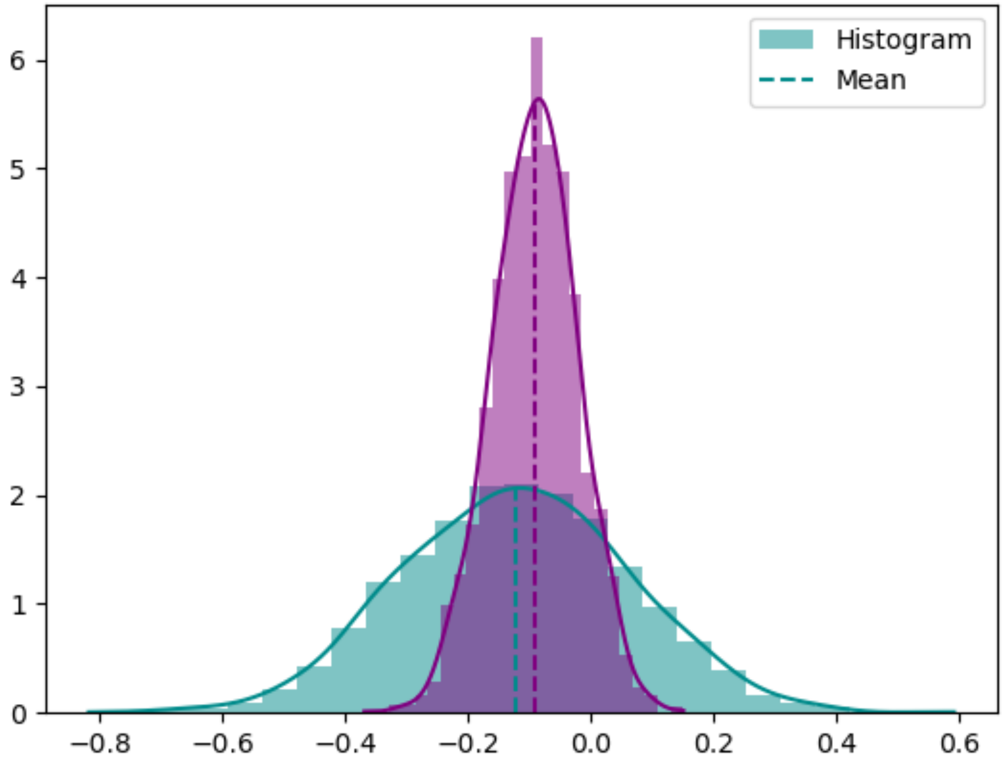
- Implication: Richardson-Romberg (RR) extrapolation for bias reduction
  - Run SGDA with two stepsizes $\gamma$ and $2\gamma$ in parallel
    - Let $\{\bar{x}_t^{(\gamma)}\}, \{\bar{x}_t^{(2\gamma)}\}$ be the averaging iterates
  - Richardson-Romberg (RR)-extrapolated iterate:

$$\begin{aligned} \hat{x}_t &:= 2\bar{x}_t^{(\alpha)} - \bar{x}_t^{(2\alpha)} \\ &\to 2\mathbb{E}_{\pi_\gamma}[x] - \mathbb{E}_{\pi_{2\gamma}}[x] \quad \textbf{(LLN)} \\ &= x^* + O(\gamma^2) \end{aligned}$$

Bias reduced from $\boldsymbol{\gamma \Delta(x^*)} + O(\gamma^2)$ to $O(\boldsymbol{\gamma^2})$

# Numerical Result: Normality and Bias

Histogram after T=1000

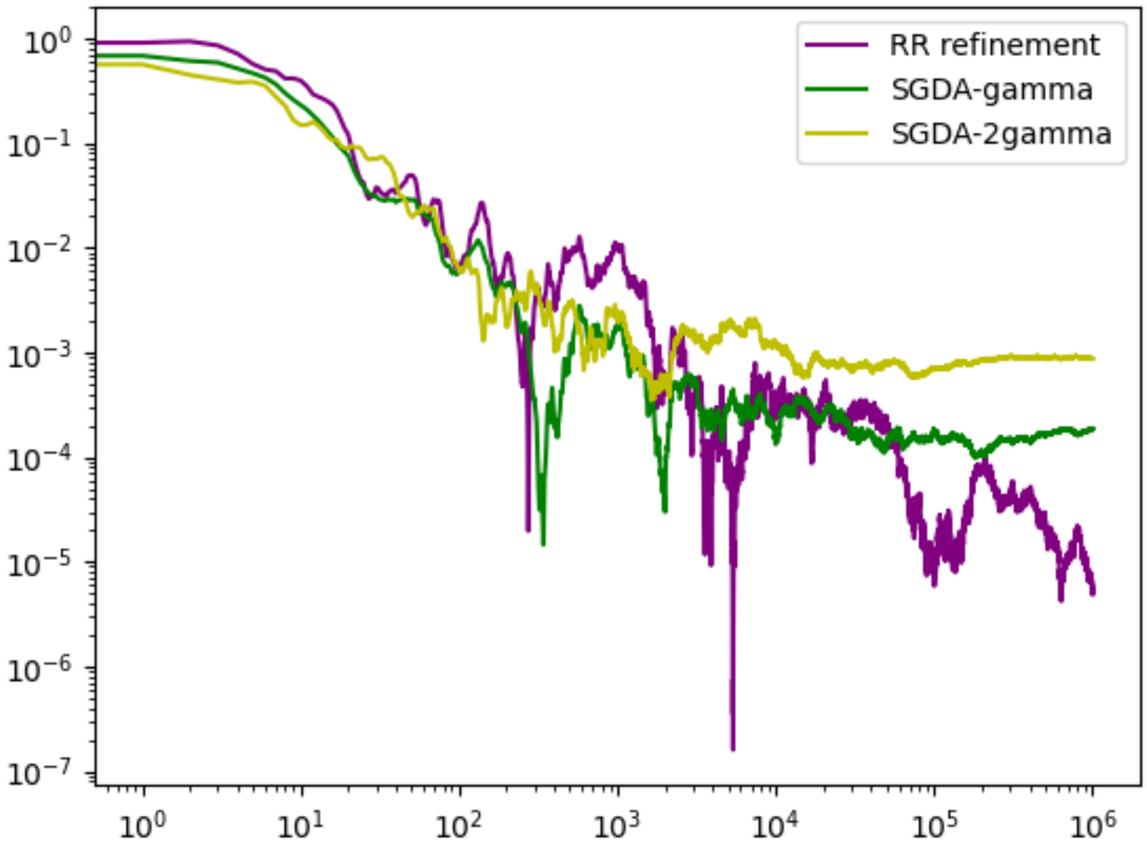



-Purple: $\gamma = 0.01$

-Green: $\gamma = 0.1$

- SEG/SGDA for min-max game with $\min_{x_1} \max_{x_2} L(x_1, x_2)$=0

# Numerical Results: RR for Bias Reduction in Zero-sum Games




- SGDA for min-max problems

# Summary

- Stochastic VIP: Constant Stepsize + Ergodicity + Bias Reduction
  - Constant stepsize: fast convergence with exponential decay rate of optimization error
  - Polyak-Ruppert average: LLN and Asymptotic normality
  - RR Extrapolation: reduce bias
- Extensions:
  - Beyond martingale noise: Markovian noise **(Goal Multi-agent RL)**
  - Statistical inference: variance estimation and CI construction
    - Constant stepsize with RR extrapolation


Thank you for
your attention