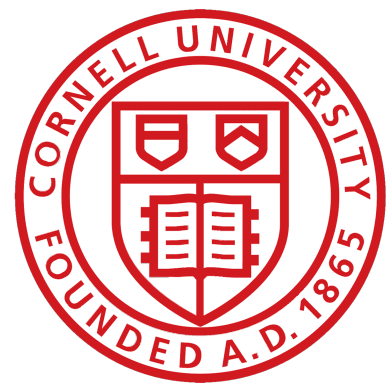




Cornell University®

On the Provable Suboptimality of Momentum SGD in Nonstationary Stochastic Optimization



Sharan Sahu*

ss4329@cornell.edu

Cameron J. Hogan*

cjh337@cornell.edu

Martin T. Wells

mtw1@cornell.edu

Department of Statistics and Data Science

Cornell University

Forty-third International Conference on Machine Learning (ICML 2026)

COEX Convention & Exhibition Center Seoul, South Korea

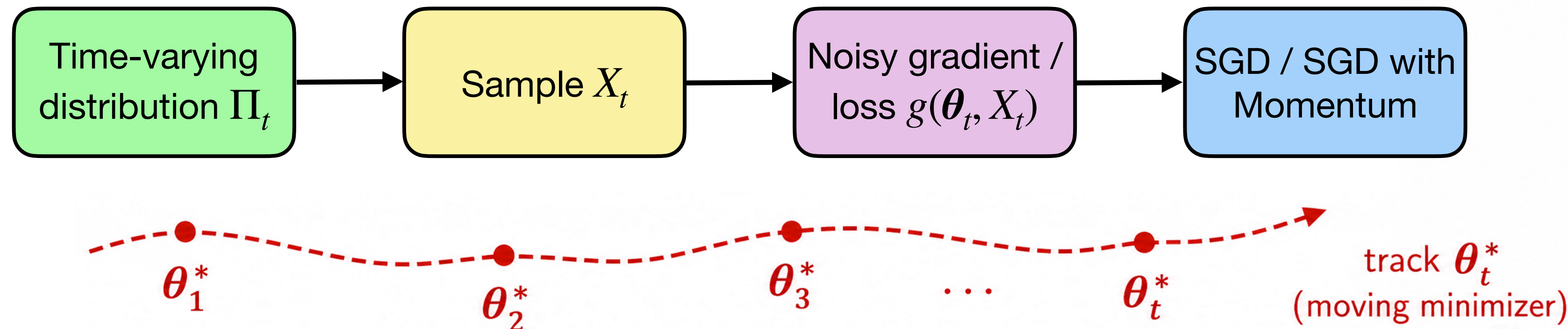
[July 6-11, 2026]

Motivation

- **Problem Setting:** We study stochastic optimization when the data distribution changes over time. At time t , a μ -strongly convex, L -smooth objective is

$$\theta_t^* = \arg \min_{\theta \in \mathbb{R}^d} G_t(\theta), \quad G_t(\theta) = \mathbb{E}_{X_t \sim \Pi_t} [g(\theta, X_t)]$$

Here $g(\theta, X_t)$ is a noisy sample loss and $X_t \sim \Pi_t$ where Π_t is a time-varying data distribution.



- **Goal:** Track the moving minimizer sequences $\{\theta_t^*\}_{t=1}^T$ for finite $T < \infty$.
- **Practical relevance:** Time-varying distribution Π_t appears in stochastic tracking and concept drift and captures many relevant settings such as policy optimization, online recommendation, continual learning, and federated learning [Kushner & Yin, 1997; Sayed, 2003; Gama et al., 2014; Hsu et al., 2021; Kakade, 2002; Schulman et al., 2015; Li et al., 2010; Parisi et al., 2019; Kairouz et al., 2021].

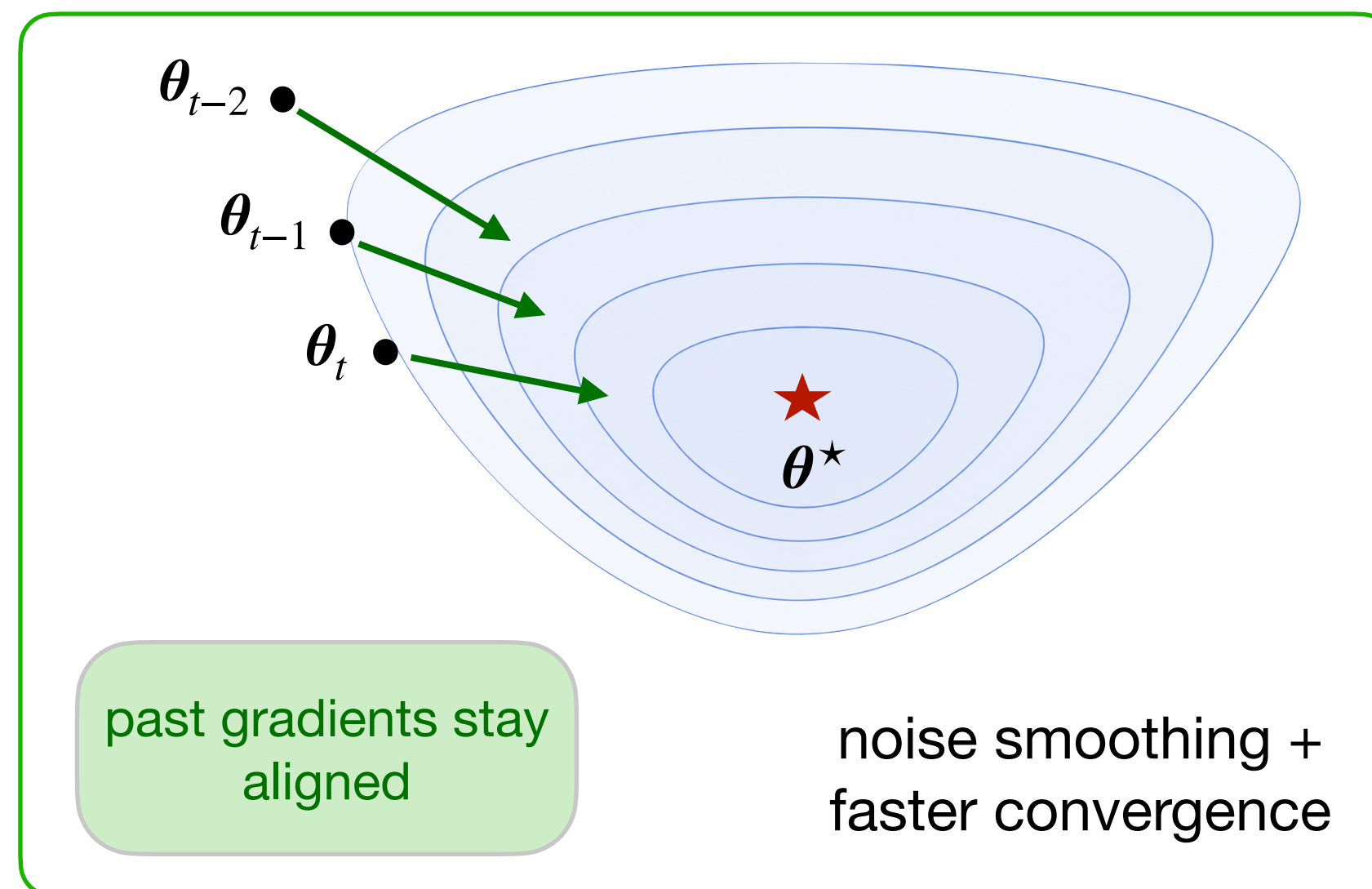
Motivation

- **Momentum in SGD:** Momentum methods average past gradients to reduce noise and often speed up optimization

$$\mathbf{v}_t = \beta \mathbf{v}_{t-1} + \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, X_t), \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma \mathbf{v}_t$$

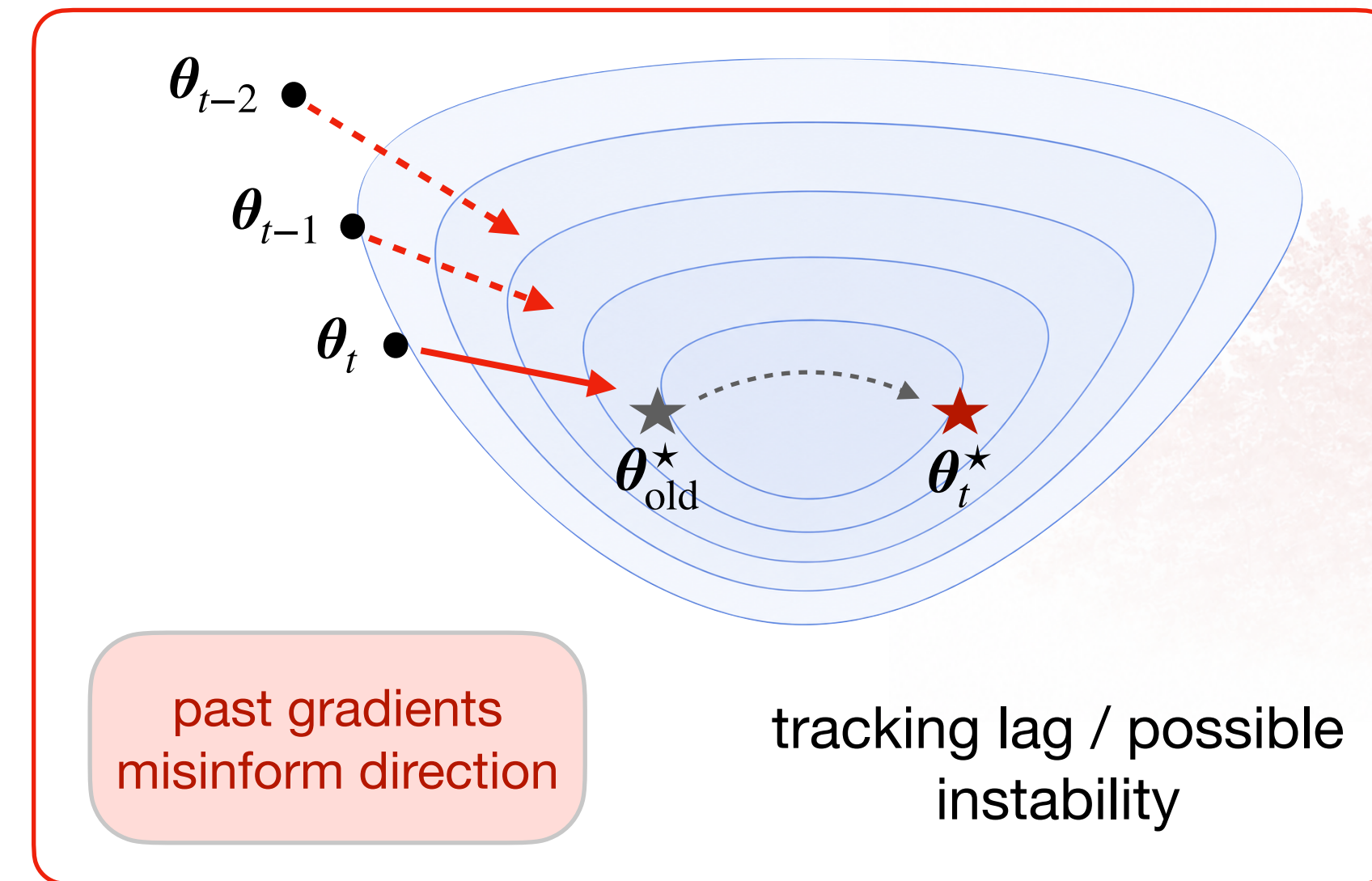
Stationary objective

fixed distribution / fixed minimizer



Nonstationary objective

drifting distribution / moving minimizer



past gradients from Π_{t-k} for $k = 0, \dots, t$ may misalign with the current objective G_t .

- **Main question:** *When does momentum accelerate tracking, and when does uninformative gradient information create unavoidable lag in nonstationary stochastic optimization?*

Main Results (I) - Expectation Tracking Bound

► **A1:** Second-moment bounds

- **Drift:** $\Delta_t = \theta_t^\star - \theta_{t+1}^\star$, $\mathbb{E}[\|\Delta_t\|_2^2] \leq \Delta^2$
- **Noise:** $\mathbb{E}[\|\xi_{t+1}(\theta_t)\|^2] \leq \sigma^2$ and $\mathbb{E}[\|\xi_{t+1}(\psi_t)\|^2] \leq \sigma^2$

Tracking error bound in expectation for SGD

Under **A1**, for any $t \geq 0$ and constant stepsize $\gamma \leq \min\{\mu/L^2, 1/L\}$,

$$\mathbb{E} \left[\left\| \theta_t - \theta_t^\star \right\|^2 \right] \lesssim \underbrace{\exp \left(-\frac{\gamma \mu t}{2} \right) \left\| \theta_0 - \theta_0^\star \right\|^2}_{\text{optimization error / transient decay}} + \underbrace{\frac{\Delta^2}{\gamma^2 \mu^2}}_{\text{tracking floor from drift}} + \underbrace{\frac{\sigma^2 \gamma}{\mu}}_{\text{noise floor}}.$$

- **Takeaway:** The error decomposes into initialization error, a floor due to nonstationary drift, and a floor due to stochastic gradient noise.

Main Results (I) - Expectation Tracking Bound

Tracking error bound in expectation for SGDM

Under **A1**, for any $t \geq 0$ and constant stepsize $\gamma \leq \mu(1 - \beta)^2/4L^2$,

$$\mathbb{E} \left[\left\| \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^* \right\|^2 \right] \lesssim \underbrace{\frac{1}{(1 - \beta)^2} \exp \left(-\frac{\gamma \mu t}{1 - \beta} \right) \left\| \boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^* \right\|^2}_{\text{sensitivity to initialization}} + \frac{(2 + \beta)^2}{\gamma^2 \mu^2} \Delta^2 + \underbrace{\frac{\sigma^2 \gamma}{\mu(1 - \beta)}}_{\text{inflates noise floor}}.$$

- ▶ Tracking error admits the same three-term decomposition as SGD but with explicit momentum penalties.
- ▶ Letting $\kappa := L/\mu$ be the condition number, $\gamma\mu = \mathcal{O}(1/\kappa)$ so ill-conditioned problems can have a long burn-in time before reaching steady state.
- ▶ Classical Nesterov acceleration improves deterministic GD from $\mathcal{O}(\kappa)$ to $\mathcal{O}(\sqrt{\kappa})$ [Nesterov, 1983; Nesterov, 2014].

Main Results (II) - High Probability Tracking Bounds

- **A2:** Conditional sub-Gaussian noise. There exists a constant $\sigma > 0$ such that for all $t \geq 0$,
 $\left\| \xi_{t+1}(\theta_t) \mid \mathcal{F}_t \right\|_{\Psi_2} \leq \sigma$ and $\left\| \xi_{t+1}(\psi_t) \mid \mathcal{F}_t \right\|_{\Psi_2} \leq \sigma$

High probability tracking error bound for SGD

Under **A2**, for all $t \in [T]$, $\gamma \leq \min \{ \mu/L^2, 1/L \}$, and $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\left\| \theta_t - \theta_t^\star \right\|^2 \lesssim \exp \left(-\frac{\gamma \mu t}{2} \right) \left\| \theta_0 - \theta_0^\star \right\|^2 + \frac{\mathfrak{D}_t}{\gamma \mu} + \frac{d\sigma^2\gamma}{\mu} + \left(d\sigma^2\gamma^2 + \frac{\sigma^2\gamma}{\mu} + \gamma^2\sigma^2\mathfrak{D}_t^{(2)} \right) \log \frac{2T}{\delta},$$

$$\text{where } \mathfrak{D}_t := \sum_{\ell=0}^{t-1} (1 - \gamma\mu/2)^{t-\ell-1} \left\| \Delta_\ell \right\|^2 \text{ and } \mathfrak{D}_t^{(2)} := \sum_{\ell=0}^{t-1} (1 - \gamma\mu/2)^{2(t-\ell-1)} \left\| \Delta_\ell \right\|^2.$$

- Our bound is *drift-adaptive and time-resolved* and captures that **(i)** only recent nonstationarity drift affects the guarantee and **(ii)** drift amplifies stochastic fluctuations by increasing the tracking mismatch.

Main Results (II) - High Probability Tracking Bounds

High probability tracking error bound for SGDM

Under **A2**, for all $t \in [T]$, $\gamma \leq \min \{1/L, \mu(1 - \beta)^2/4L^2\}$, and $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\| \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^* \|^2 \lesssim \frac{1}{(1 - \beta)^2} \exp \left(-\frac{\gamma \mu t}{4(1 - \beta)} \right) \| \boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^* \|^2 + \frac{\mathfrak{D}_t^{\text{lag}}}{\gamma \mu (1 - \beta)} + \frac{d\sigma^2 \gamma}{\mu (1 - \beta)} + \left(\frac{d\sigma^2 \gamma^2}{(1 - \beta)^2} + \frac{\sigma^2 \gamma}{\mu (1 - \beta)} + \frac{\sigma^2 \gamma^2 \mathfrak{D}_t^{\text{lag},(2)}}{(1 - \beta)^4} \right) \log \frac{2T}{\delta},$$

where $\mathfrak{D}_t^{\text{lag}} := \sum_{\ell=0}^{t-1} \tilde{\rho}^{t-\ell-1} \| \mathbf{b}_\ell \|^2$ and $\mathfrak{D}_t^{\text{lag},(2)} := \sum_{\ell=0}^{t-1} \tilde{\rho}^{2(t-\ell-1)} \| \mathbf{b}_\ell \|^2$ with $\tilde{\rho} := 1 - \frac{\gamma \mu}{4(1 - \beta)}$ and $\mathbf{b}_\ell$ depends on lagged drift terms $\Delta_\ell, \Delta_{\ell-1}$.

- Momentum adds an inertia horizon of order $\mathcal{O}(1/(1 - \beta))$, amplifies the drift-noise coupling by $\mathcal{O}(1/(1 - \beta)^2)$ and slows transient decay and prevents SGDM from adapting to drift in ill-conditioned regimes ($\kappa \gg 1$).
- Explains why SGD can be strictly more robust in nonstationary, ill-conditioned settings: it avoids the compounded inertia and variance penalties induced by momentum.

Main Results (III) - Minimax Lower Bounds

► **Setting:** Dynamic regret for non stationary stochastic optimization under distribution shift.

• **Dynamic regret:** $\mathcal{R}_T^\pi(G) := \mathbb{E}^\pi \left[\sum_{t=0}^{T-1} \left(G_{t+1}(\theta_t) - G_{t+1}(\theta_{t+1}^*) \right) \right]$, **Minimax risk:** $\mathfrak{M}_T(\Pi_\beta, \mathbb{V}_T) := \inf_{\pi \in \Pi_\beta} \sup_{G: \text{GVar}_{p,q}(G) \leq \mathbb{V}_T} \mathcal{R}_T^\pi(G)$.

► **Function class (gradient variation budget):** for $1 \leq p, q \leq \infty$, the $L_{p,q}$ gradient-variational functional of

$g = (g_1, \dots, g_T)$ is:

$$\text{GVar}_{p,q}(g) := \begin{cases} \left(\frac{1}{T} \sum_{t=1}^{T-1} \left\| \nabla g_{t+1} - \nabla g_t \right\|_p^q \right)^{1/q} & q < \infty \\ \max_{1 \leq t \leq T-1} \left\| \nabla g_{t+1} - \nabla g_t \right\|_p & q = \infty, \end{cases} \quad \|g\|_p := \begin{cases} \left(\int_{\Theta} \|g(\theta)\|_2^p d\theta \right)^{1/p} & p < \infty \\ \sup_{\theta \in \Theta} \|g(\theta)\|_2 & p = \infty. \end{cases}$$

Minimax lower bound for strongly-convex function sequences using SGDM

Fix arbitrary $1 \leq p, q \leq \infty$. Consider the class Π_β of SGDM(β) policies with constant step size $\gamma \leq c_0(1 - \beta)^2/L$. Then there exists a class $\mathcal{G}_{p,q}(\mathbb{V}_T)$ of μ -strongly convex, L -smooth function sequences whose gradient-variational functional budget satisfies $\text{GVar}_{p,q}(G) \leq \mathbb{V}_T$ such that

$$\mathfrak{M}_T(\Pi_\beta, \mathbb{V}_T) \gtrsim \max \left\{ \underbrace{(1 - \beta)^{-2/(\alpha q + 2)} \sigma^{4/(\alpha q + 2)} \mu^{(\alpha q - 2q - 2)/(\alpha q + 2)} \mathbb{V}_T^{2q/(\alpha q + 2)} T^{\alpha q/(\alpha q + 2)}}_{\text{Noise-limited regime}}, \underbrace{(1 - \beta)^{-2/(\alpha q)} \mu^{(\alpha q - 2q - 2)/(\alpha q)} L^{2/(\alpha q)} \mathbb{V}_T^{2/\alpha} T^{1 - 2/(\alpha q)}}_{\text{Drift-limited regime}} \right\}.$$

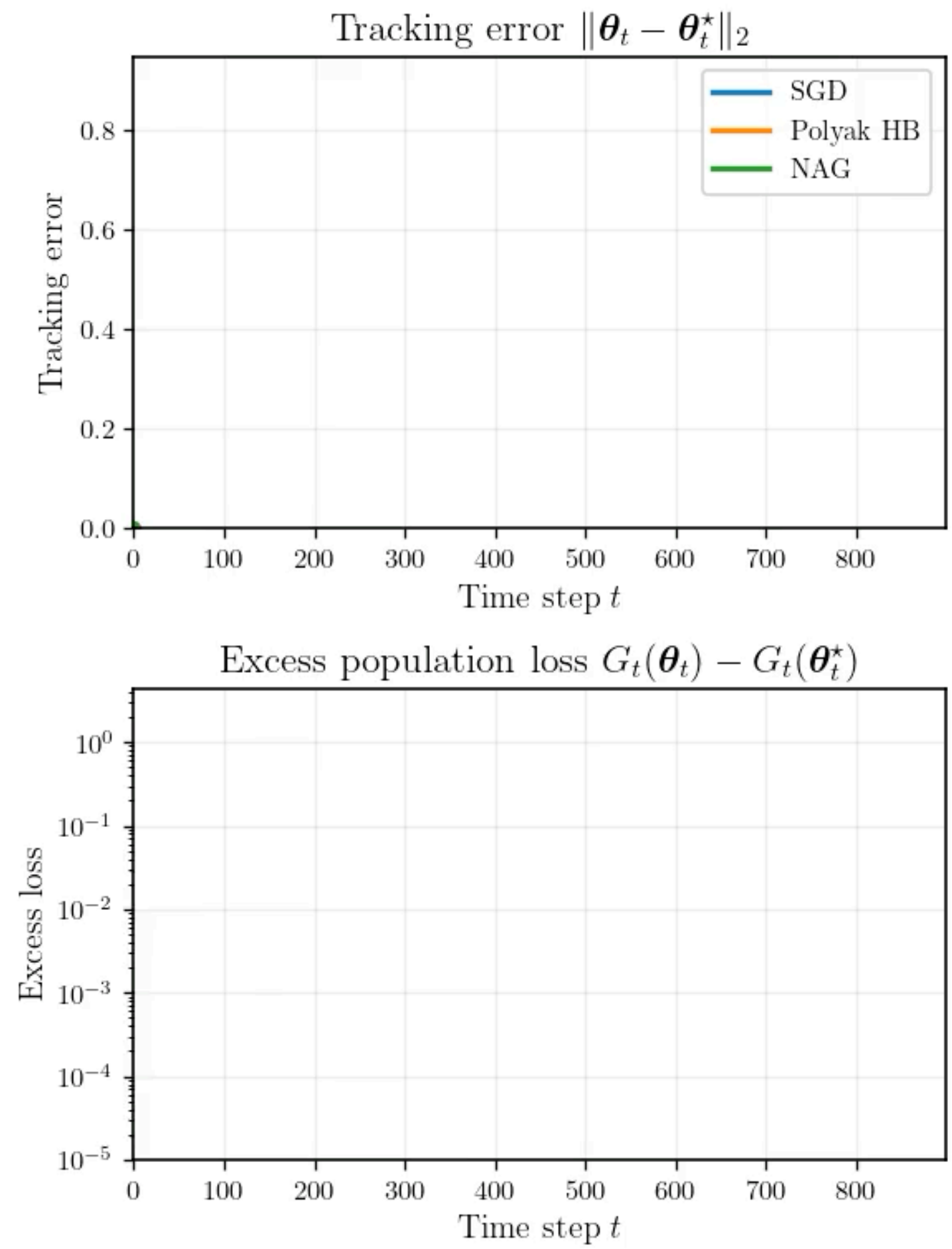
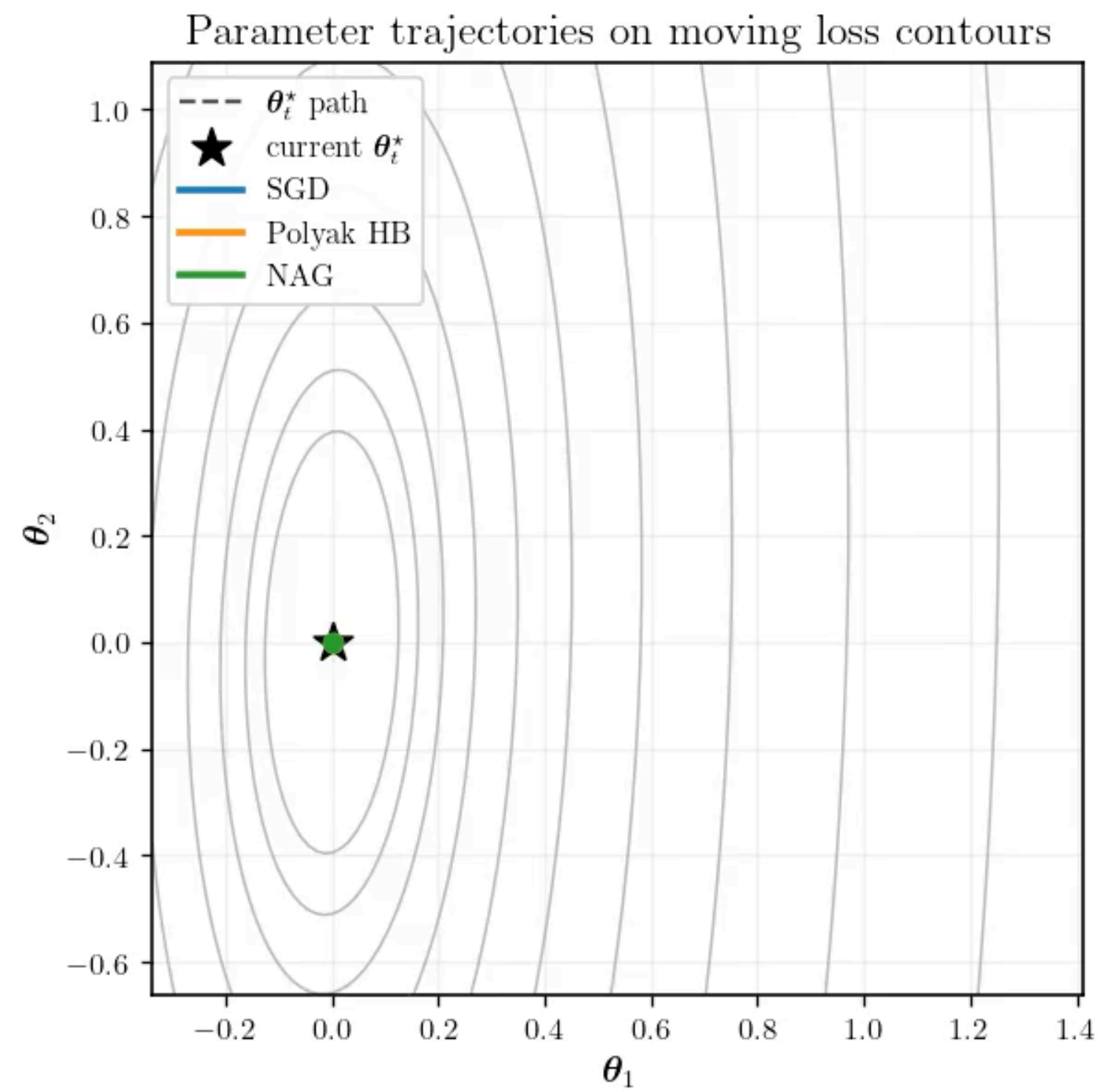
Noise-limited regime

Drift-limited regime

► **Takeaway:** No sequence of functions $\{g_t\}_{t=1}^T \subset \mathcal{G}_{p,q}(\mathbb{V}_T)$ using SGDM with any constant stepwise can achieve smaller regret and must incur penalties depending on $(1 - \beta)^{-1}$.

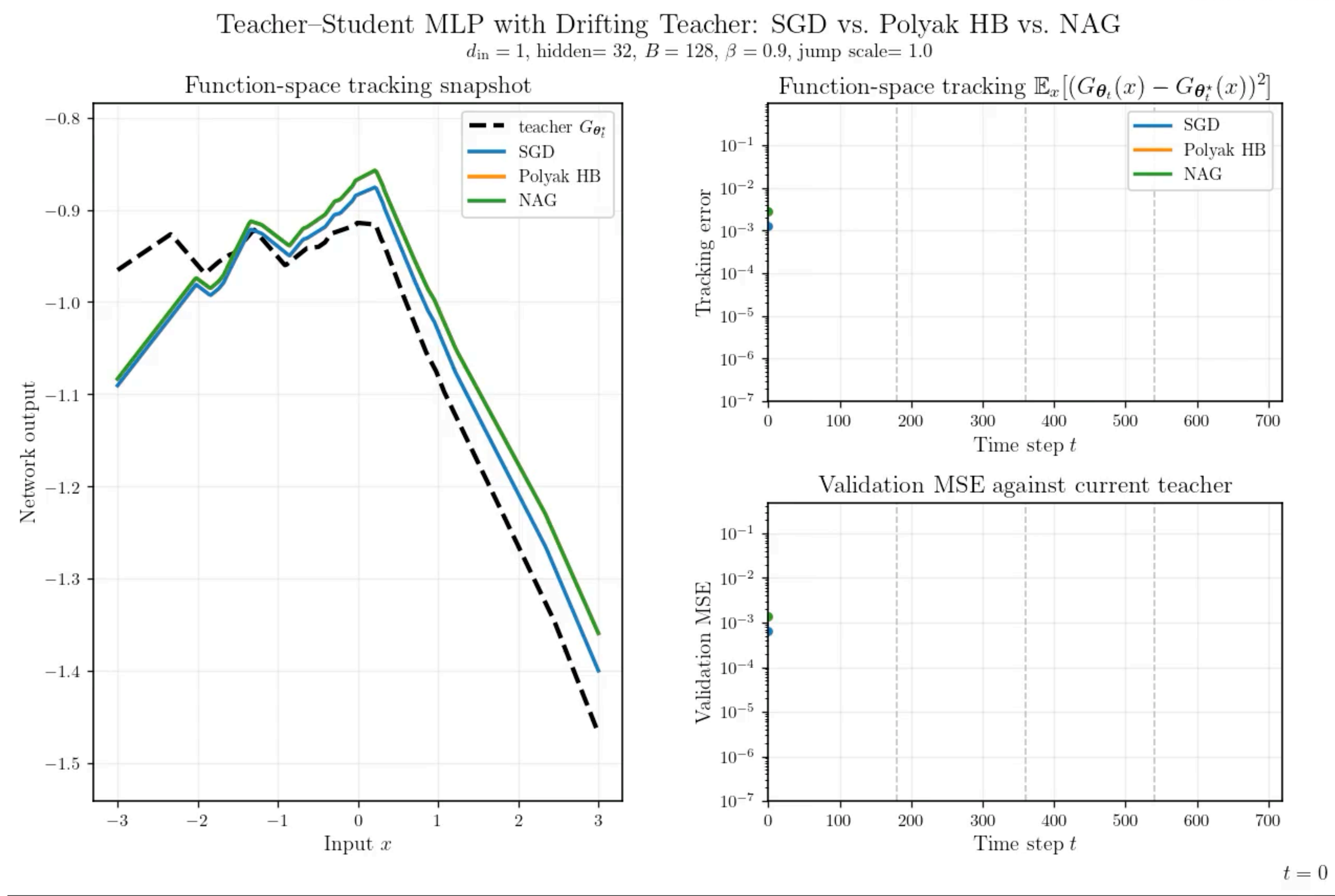
Empirical Results - Linear Regression

Linear Regression with Drifting Minimizer: SGD vs. Polyak HB vs. NAG
 $d = 2, B = 128, \kappa = 10.0, \beta = 0.9, \text{drift rate} = 0.05$



$t = 0$

Empirical Results - Teacher Student MLP



Thank you for listening!

For further details, please visit:

- ▶ Our paper: <https://openreview.net/forum?id=U1bxelQLaK>
- ▶ Our poster: <https://icml.cc/virtual/2026/poster/63752>

