Asymptotic Behaviors of Projected Stochastic Approximation: A Jump Diffusion Perspective

Jiadong Liang; Yuze Han; Xiang Li; Zhihua Zhang

Loopless Projection Stochastic Approximation

- We aim to solve the following problem
$$\min_{\boldsymbol{x}} \mathbb{E}_{\zeta\sim\mathscr{D}} f(\boldsymbol{x},\zeta) \text{ subject to } \boldsymbol{A}^\top \boldsymbol{x} = \mathbf{0}$$

- The LPSA first performs as $\boldsymbol{x}_{n+\frac{1}{2}} = \boldsymbol{x}_n - \eta_n \nabla f\left(\boldsymbol{x}_n\right) + \eta_n \xi_n$ with a martingale difference sequence $\{\xi_n\}$.

- Then we independently cast a coin with the head probability p_n and obtain $\omega_n \sim \text{Bernoulli}\left(p_n\right)$. If $\omega_n = 1$, we perform one step of projection onto the null space of $\boldsymbol{A}^\top$. Otherwise, we let $\boldsymbol{x}_{n+1} = \boldsymbol{x}_{n+\frac{1}{2}}$.

- It's obvious that Local SGD is a specialized case of LPSA under federated learning scenario.

Convergence Rate Analysis

- Let $\eta_n = \eta_0 n^{-\alpha};\ p_n = \min\{\eta_n^{\beta}, 1\}$ with $\beta \in [0,1)$.
- Theorem 3.1

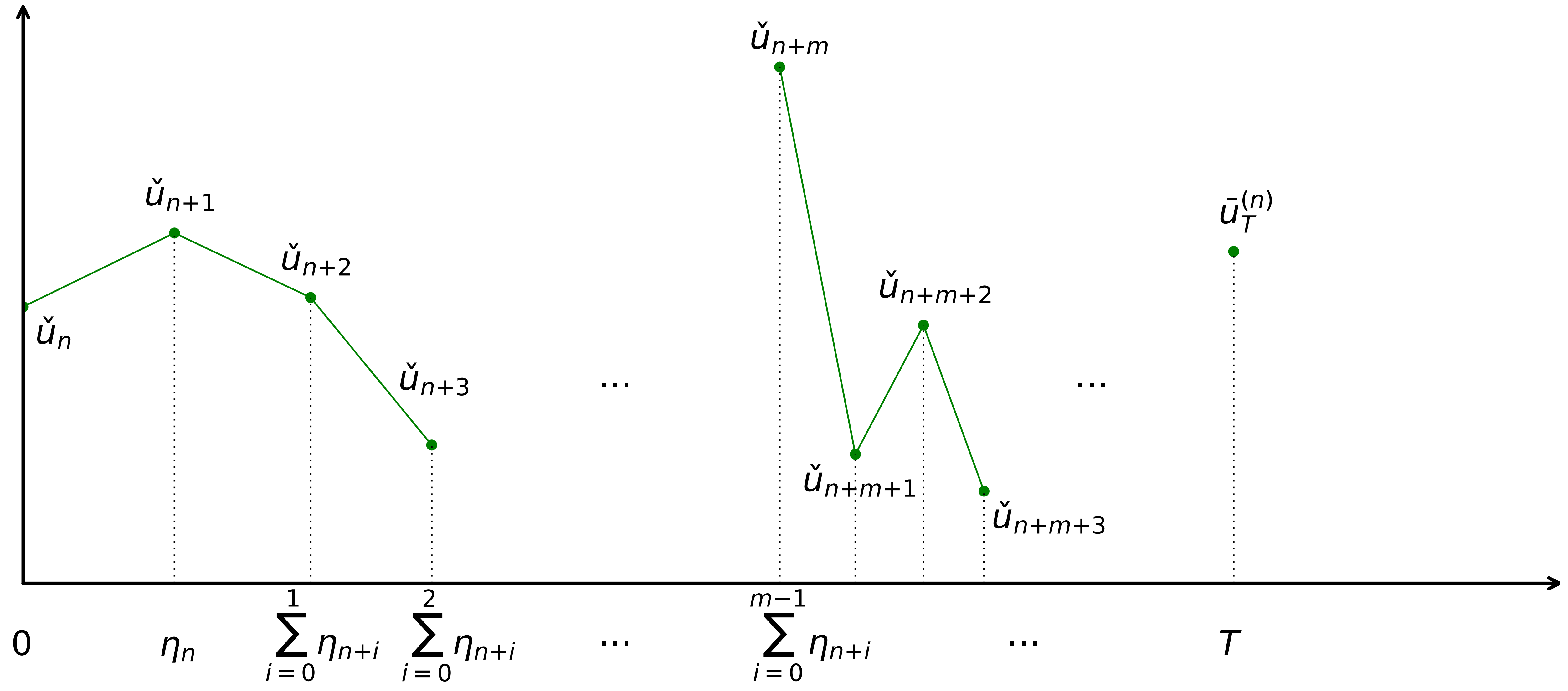
  Under appropriate assumptions, for (i) $0 < \alpha < 1$ or (ii) $\alpha = 1$ with $\eta_0 > 2/\mu$ (μ is the strong convexity parameter), we have

  $$\mathbb{E}\left\|\boldsymbol{u}_n - \boldsymbol{x}^\star\right\|^2 = \mathcal{O}\left(n^{-\alpha\min\{1,2-2\beta\}}\right)$$

  Where $\boldsymbol{u}_n = \mathcal{P}_{\boldsymbol{A}^\perp}\left(\boldsymbol{x}_n\right)$.
- As β decreases, i.e. the projection happens more frequently, $\mathbb{E}\left\|\boldsymbol{u}_n - \boldsymbol{x}^\star\right\|^2$ converges faster. What's more, we can find a phase transition when β goes cross 0.5.

Asymptotic Behavior via Diffusion Approximation

- Frequent Projection $\beta \in [0,1/2)$
- Let $\check{\boldsymbol{u}}_n := \dfrac{\boldsymbol{u}_n - \boldsymbol{x}^*}{\sqrt{\eta_{n-1}}}$. And let $\overline{\boldsymbol{u}}_t^{(n)}$ be the continuous random process which starts at $\check{\boldsymbol{u}}_n$ and takes value $\check{\boldsymbol{u}}_{n+m}$ at time point $\eta_n + \cdots + \eta_{n+m-1}$.
- The trajectory is presented in the form shown on the right.

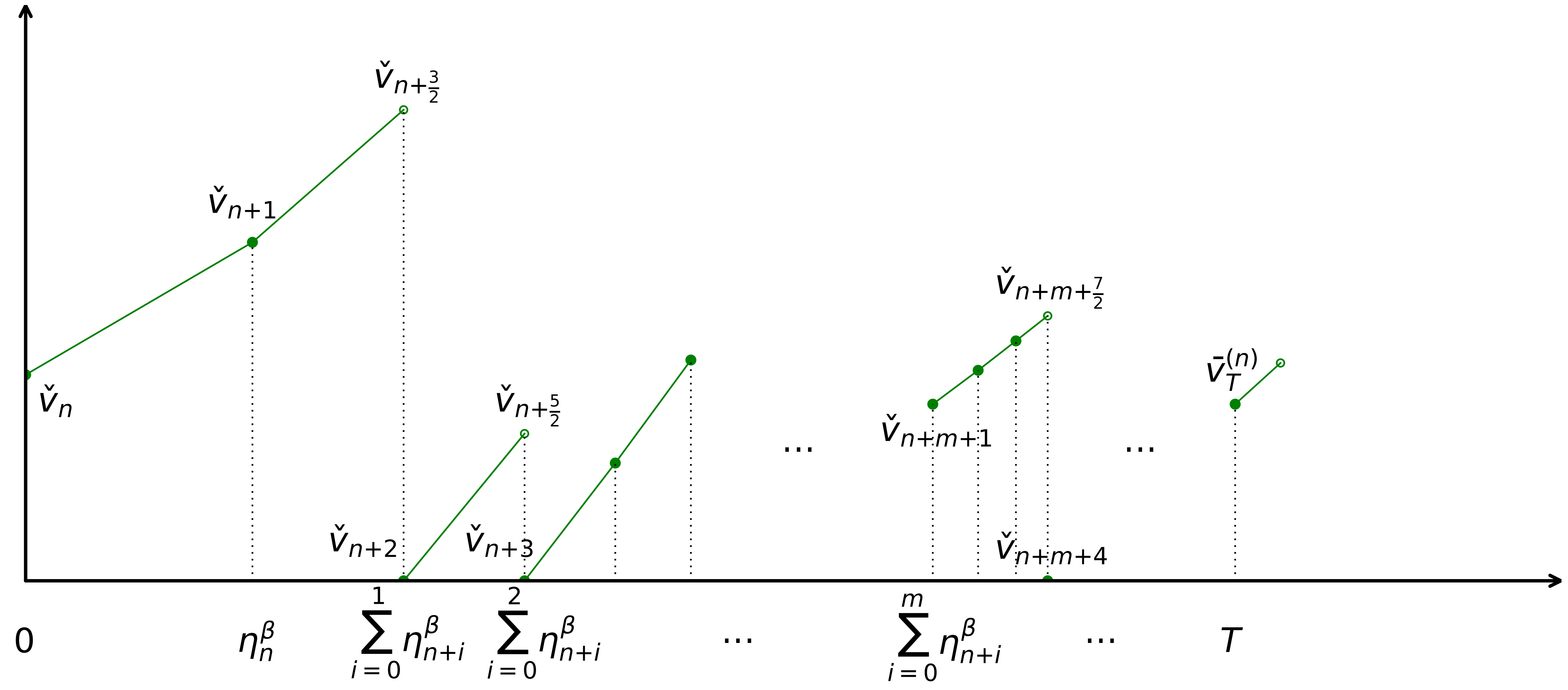


Asymptotic Behavior via Diffusion Approximation

- Theorem 3.3.
  Let regular assumptions hold. Then the sequence of random processes $\left\{\bar{\boldsymbol{u}}_t^{(n)} : t \geq 0\right\}_{n=1}^{\infty}$ converges weakly to the stationary weak solution of the following SDE:
  $$d\mathbf{X}_t = -\mathscr{P}_{\boldsymbol{A}^\perp}\left(\nabla^2 f\left(\boldsymbol{x}^\star\right) - \frac{1}{2\eta_0}\mathbf{1}_{\{\alpha=1\}}\mathbf{I}_d\right)\mathbf{X}_t dt + \mathscr{P}_{\boldsymbol{A}^\perp}\Sigma\left(\boldsymbol{x}^\star\right)^{\frac{1}{2}} d\mathbf{W}_t.$$
  Further, the rescaled sequence $\left\{\check{\boldsymbol{u}}_n\right\}_{n=1}^{\infty}$ converges weakly to the invariant distribution of this dynamics.

Asymptotic Behavior via Jump Approximation

- Occasional Projection $\beta \in (1/2,1)$.
- Let $\boldsymbol{v}_n = \mathscr{P}_{\boldsymbol{A}}\left(\boldsymbol{x}_n\right)$ and $\check{\boldsymbol{v}}_n = \eta_{n-1}^{\beta-1}\boldsymbol{v}_n$. And we define $\overline{\boldsymbol{v}}_t^{(n)}$ as the cadlag process which starts at $\check{\boldsymbol{v}}_n$ and take values $\check{\boldsymbol{v}}_{n+m+\frac{1}{2}}$, $\check{\boldsymbol{v}}_{n+m+1}$ at time points $(\eta_n^\beta + \cdots + \eta_{n+m-1}^\beta)-$ and $\eta_n^\beta + \cdots + \eta_{n+m-1}^\beta$ respectively.



Asymptotic Behavior via Jump Approximation

- Theorem 3.4
  Let regular assumptions hold. Then the sequence of cadlag stochastic processes $\left\{\bar{\boldsymbol{v}}_t^{(n)} : t \geq 0\right\}_{n=1}^{\infty}$ weakly converges to the stationary weak solution of the following Jump-SDE:
  $$d\mathbf{Y}_t = -\nabla f\left(\boldsymbol{x}^\star\right) dt - \mathbf{Y}_{t-} \cdot \mathbf{N}_\gamma(dt).$$
  Further, the rescaled sequence $\left\{\check{\boldsymbol{v}}_n\right\}_{n=1}^{\infty}$ weakly converges to the invariant distribution of this dynamics, i.e., $-\frac{\nabla f\left(\boldsymbol{x}^\star\right)}{\left\|\nabla f\left(\boldsymbol{x}^\star\right)\right\|} \cdot \mathscr{E}\left(\frac{\left\|\nabla f\left(\boldsymbol{x}^\star\right)\right\|}{\gamma}\right)$

Asymptotic Behavior via Jump Approximation

- Corollary 1

  Let regular assumptions hold. Then for $\beta \in (1/2, 1)$, $\hat{\boldsymbol{u}}_n := \eta_{n-1}^{\beta-1}\left(\boldsymbol{u}_n - \boldsymbol{x}^\star\right)$ converges to a non-zero vector

  $$\frac{1}{\gamma}\left\{\mathscr{P}_{\boldsymbol{A}^\perp}\left(\nabla^2 f\left(\boldsymbol{x}^\star\right) - \frac{1-\beta}{\eta_0}\mathbf{1}_{\{\alpha=1\}}\mathbf{I}\right)\mathscr{P}_{\boldsymbol{A}^\perp}\right\}^\dagger\left(\mathscr{P}_{\boldsymbol{A}^\perp}\nabla^2 f\left(\boldsymbol{x}^\star\right)\nabla f\left(\boldsymbol{x}^\star\right)\right)$$

- Remark: From the above derivation, for the choice $p_n \propto \eta_n^\beta$, when β varies, our algorithm has an interesting bias-variance tradeoff.

Interesting Bias-Variance Tradeoff

- Order of fluctuation: $\mathcal{O}(\eta_n^{1/2})$
- Order of Bias: $\mathcal{O}(\eta_n^{1-\beta})$
- When $\beta \in [0,1/2)$
  The fluctuation caused by the randomness of gradient queries in every iteration dominates the optimization accuracy.
- When $\beta \in [1/2,1)$
  Manipulated by the biases formed by the accumulation of skewed updates in the unconstrained state within each 'inner loop'

The End