Reading group
Stochastic Approximation

Antoine Bambade

Willow and Sierra teams

Contents

01

Introduction

Context

- Quasi Stochastic Approximation (chap. 4),
- Broad range of applications.

Setup

$$\bar{f}(\theta^*) \stackrel{\text{def}}{=} E(f(\theta^*, \phi)) = 0, \theta^* \in \mathbb{R}^d, \phi \in \Omega \tag{1}$$

Setup

$$\bar{f}(\theta^*) \stackrel{\text{def}}{=} E(f(\theta^*, \psi)) = 0, \theta^* \in \mathbb{R}^d, \phi \in \Omega \tag{2}$$

Plan

- Step 1: Refine $\bar{f}$ (ex: s.t. globally asymptotically stable ODE):

$$\frac{d}{dt}\vartheta = \bar{f}(\vartheta) \tag{3}$$

- Step 2: Design appropriate approximation:

$$\theta_{n+1} = \theta_n + \alpha_{n+1}\bar{f}(\theta_n) \tag{4}$$

- Step 3: Design appropriate Stochastic Approximation:

$$\theta_{n+1} = \theta_n + \alpha_{n+1}(\bar{f}(\theta_n) + \Delta_{n+1}) \tag{5}$$

Plan

- Remarks on ODE design
- ODE approximation
  - > sufficient conditions for convergence,
  - > optimizing the covariance,
  - > trade-off (transient time vs optimal rate),
  - > some solutions (PJR, Zap algorithms),
  - > limits,

Measure typically $\Sigma_n \stackrel{\text{def}}{=} E[(\theta_n - \theta^*)(\theta_n - \theta^*)^T]$, but also need $\|\theta_n - \theta^*\|$

02

Remarks on ODE design

Remarks

- $\bar{f}$, f: impose transient time,
- Examples
  - > Newton-Raphson Flow

$$\theta_{n+1} = \theta_n - \alpha_{n+1}[A(\theta_n)]^{-1}\bar{f}(\theta_n), A(\theta_n) \stackrel{\text{def}}{=} [\partial_\theta \bar{f}(\theta)]_{\theta=\theta_n} \quad (6)$$

  - > Newton-Raphson algorithm ($\alpha_n = 1$),
  - > Runge-Kutta methods,
  - > etc.

03

ODE approximation

Setting

$$\theta_{n+1} = \theta_n + \alpha_{n+1}(\bar{f}(\theta_n) + \Delta_{n+1}) \tag{7}$$

with $\Delta_{n+1} = f(\theta_n, \Phi_{n+1}) - \bar{f}(\theta_n), \Phi_{n+1} \sim \Phi$

What do we compare?

$\|\vartheta_t^{(n)} - \Theta_t\|$ on time intervals T, paved by $\tau_{k+1} = \tau_k + \alpha_k, k \geq 0$

- The Original ODE $\vartheta_t^{(n)}$:

$$\frac{d}{dt}\vartheta_t^{(n)} = \bar{f}(\vartheta_t^{(n)}), t \geq \tau_n, \vartheta_{\tau_n}^{(n)} = \theta_n \tag{8}$$

- Its Stochastic Approximation Θ_t:

$$\Theta_t = \theta_n, \text{ if } t = \tau_n, \forall n \geq 0, \text{ linear interpolation otherwise} \tag{9}$$

What do we compare?

$\|\vartheta_t^{(n)} - \Theta_t\|$ on time intervals T, paved by $\tau_{k+1} = \tau_k + \alpha_k, k \geq 0$

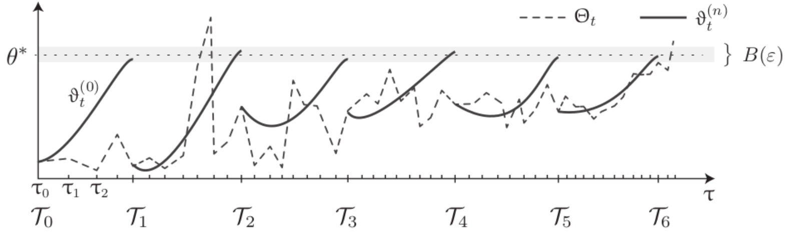

Figure 8.1: ODE approximation on time intervals $[\mathcal{T}_k, \mathcal{T}_{k+1})$ of width approximately T.

Typical assumptions (Th. 8.1)

- $\bar{f}$ Lipschitz continuous,
- $\{\theta_n\}$ bounded a.s.,
- Cumulative disturbance M_K vanishes for each T,

Then:

$$\forall T > 0, \lim_{n \longrightarrow \infty} \sup_{\tau_n \geq t \leq \tau_n + T} \|\vartheta_t^{(n)} - \Theta_t\| = 0 \tag{10}$$

If also the ODE is globally asymptotically stable with unique minimum θ^*:

$$\lim_{t \longrightarrow \infty} \Theta_t = \lim_{n \longrightarrow \infty} \theta_n = \theta^* \tag{11}$$

Step-size choice

Sufficient conditions (Robbins–Monro) for matching vanishing disturbance of Th. 8.1

- $\sum_k \alpha_k = \infty$,
- $\sum_k \alpha_k^2 < \infty$.

Hence, typical choice: $\alpha_n = \frac{g}{(n+n_0)^\rho}, \rho \in (0,1], g > 0$

Cumulative disturbance

$$M_K^{(n)} = \sum_{i=n+1}^{K} \alpha_i \Delta_i \tag{12}$$

Vanishing in the sense (cf. Th 8.1):

$$\lim_{n \longrightarrow \infty} \sup_{K > n, \tau_K - \tau_n \leq T} \| M_K^{(n)} \| = 0 \tag{13}$$

$\Sigma_n \stackrel{\text{def}}{=} E[(\theta_n - \theta^*)(\theta_n - \theta^*)^T]$, assume $n^\rho \Sigma_n \to \Sigma_\theta$

Scalar gain

- $\alpha_n = \frac{g}{n+n_0}$: If Real$(\lambda(gA)) < -1/2$ then Σ_θ must solve

$$(gA + \frac{1}{2}I)\Sigma_\theta + \Sigma_\theta(gA + \frac{1}{2}I)^T + g^2\Sigma_\Delta = 0 \tag{14}$$

- $\alpha_n = \frac{g}{(n+n_0)^\rho}, \rho \in (1/2, 1)$: If Real$(\lambda(A)) < 0$ then Σ_θ must solve

$$A\Sigma_\theta + \Sigma_\theta A^T + g\Sigma_\Delta = 0 \tag{15}$$

with

- $A \stackrel{\text{def}}{=} \partial_\theta \bar{f}(\theta^*)$,
- $\Sigma_\Delta \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n} E[M_n M_n^T]$, with $M_n = \sum_{k=1}^n f_k(\theta^*)$

Matrix gain

- $\alpha_n = \frac{G}{n+n_0}$: If Real$(\lambda(GA)) < -1/2$ then Σ_θ^G must solve

$$(GA + \frac{1}{2}I)\Sigma_\theta^G + \Sigma_\theta^G(GA + \frac{1}{2}I)^T + G\Sigma_\Delta G^T = 0 \tag{16}$$

with

- $A \overset{\text{def}}{=} \partial_\theta \bar{f}(\theta^*)$,
- $\Sigma_\Delta \overset{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n} E[M_n M_n^T]$, with $M_n = \sum_{k=1}^{n} f_k(\theta^*)$

Optimal choice

The choice $G^* = -A^{-1}$ results in:

$$\Sigma_\theta^* \overset{\text{def}}{=} A^{-1}\Sigma_\Delta (A^{-1})^T, \Sigma_\theta^G - \Sigma_\theta^* \geq 0 \tag{17}$$

Naive recap

Suppose

- $\bar{f}$ Lipsichitz continuous,
- ODE Globally asymptotically stable,
- $\alpha_n = -\frac{A^{-1}}{n+n_0}$ (i.e., Newton-Raphson flow type approximation)
- other mild necessary assumptions

Then

- Optimal rate : $n\Sigma_n \to \Sigma_\theta^*$
- Optimal covariance Σ_θ^*

Are we done ?

Not the best in practice

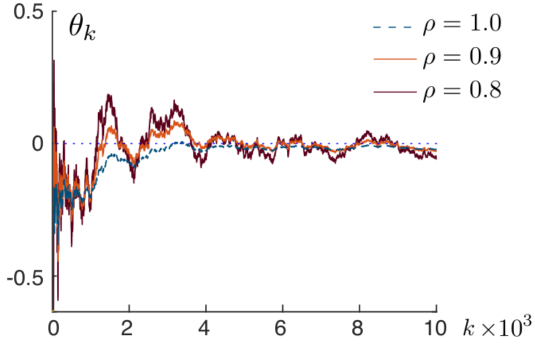

Figure 8.2: Comparison of three values of ρ for the step-size $\alpha_n = g/n^\rho$.

Transient time estimate

cf. Lemma 8.4 (scalar step size)

- If $\alpha_n = g/n$, for large $k \geq 0$

$$\|\theta_{n+k} - \theta^*\| \approx \|\vartheta^{(n)}_{\tau_{n+k}} - \theta^*\| \leq B_0 e^{\rho_0 g}\|\theta_n - \theta^*\|\left(\frac{n}{n+k}\right)^{-\rho_0 g} \tag{18}$$

- If $\alpha_n = g/n^\rho$, $\rho < 1$, for large $k \geq 0$

$$\|\theta_{n+k} - \theta^*\| \approx \|\vartheta^{(n)}_{\tau_{n+k}} - \theta^*\| \leq B_0 e^{\rho_0 g(1+\tau_n)}\|\theta_n - \theta^*\| e^{\frac{-\rho_0 g}{1-\rho}(n+k+1)^{1-\rho}} \tag{19}$$

Trade-off:

- $\alpha_n = g/n$: optimal rate ($O(1/n)$) but slower transient time,
- $\alpha_n = g/n^\rho$, $\rho < 1$, slower rate, but quicker transient time.

JPR: Simple but impactful idea

Wait until transient time "terminated" at $N_0 > 0$ ($\rho < 1$), then reduce volatility using averaging.

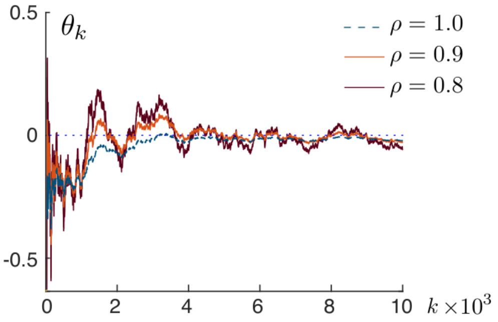

Figure 8.2: Comparison of three values of ρ for the step-size $\alpha_n = g/n^\rho$.

Polyak-Juditsky-Ruppert Averaging

Initialization: $\theta_0 \in \mathbb{R}^d$

- $\theta_{n+1} = \theta_n + \beta_{n+1} f_{n+1}(\theta_n)$, $0 \leq n \leq N-1$,
- $\theta_N^{PR} = \frac{1}{N-N_0} \sum_{k=N_0+1}^{N} \theta_k$

with $1 << N_0 << N$, β_n square summable and $\lim_n n\beta_n = \infty$ (typically $\beta_n = g/n^\rho$, $\rho < 1$)

Optimal rate

Under mild assumptions (Section 8.6.3)

$$nE[(\theta_n^{PR} - \theta^*)(\theta_n^{PR} - \theta^*)^T] \to \Sigma_\theta^* \stackrel{\text{def}}{=} A^{-1}\Sigma_\Delta (A^{-1})^T \tag{20}$$

Remark: two time-scale ODE

PJR averaging:

$$\begin{cases} \theta_{n+1} & = \theta_n + \beta_{n+1} f_{n+1}(\theta_n) \\ \theta_{n+1}^{PR} & = \theta_n^{PR} + \alpha_{n+1}[\theta_{n+1} - \theta_n^{PR}], n \geq N_0 \end{cases} \tag{21}$$

with $\theta_{N_0}^{PR} = 0$, $\alpha_n = 1/n$, $\lim_n \frac{\beta_n}{\alpha_n} = \infty$

More generally (ex: The 8.3)

$$\begin{cases} \theta_{n+1} & = \theta_n + \beta_{n+1} f_{n+1}(\theta_n, \omega_n) \\ \omega_{n+1} & = \omega_n + \alpha_{n+1} g_{n+1}(\theta_n, \omega_n) \end{cases} \tag{22}$$

with $\lim_n \frac{\beta_n}{\alpha_n} = \infty$, which implies $\theta_n \approx \theta^s(\omega_n)$ for large n.

$$\frac{d}{dt} w_t = \bar{g}(\theta^s(w_t), w_t) \tag{23}$$

ZAP algorithm

Objective: approximating Newton-Raphson flow

$$\frac{d}{dt}\vartheta_t = -[\varepsilon I + A(\vartheta_t)^T A(\vartheta_t)]^{-1} A(\vartheta_t)^T f(\vartheta_t) \qquad (24)$$

Motivations (=gain matrix algorithm):

- ideal transient time ("$\bar{f}(\vartheta_t) = \bar{f}(\vartheta_0)e^{-t}$"),
- optimal rate (under mild assumptions),
- mild assumptions for $\bar{f}$.

ZAP Stochastic Approximation

- $\theta_0 \in \mathbb{R}^d$, $\hat{A}_0 \in \mathbb{R}^{d\times d}, \varepsilon > 0$
- For $n \geq 0$

$$\begin{cases} \hat{A}_{n+1} = \hat{A}_n + \beta_{n+1}[A_{n+1} - \hat{A}_n], \\ A_{n+1} \stackrel{\text{def}}{=} \partial_\theta f_{n+1}(\theta_n), \\ \theta_{n+1} = \theta_n + \alpha_{n+1} G_{n+1} f_{n+1}(\theta_n), \\ G_{n+1} \stackrel{\text{def}}{=} -[\varepsilon I + \hat{A}_{n+1}^T \hat{A}_{n+1}]^{-1} \hat{A}_{n+1}^T. \end{cases} \tag{25}$$

with $\lim_n \beta_n/\alpha_n = \infty$.

Remark

If $\varepsilon = 0$ and $\alpha_n = \beta_n = \frac{1}{n}$: Stochastic Newton Raphson (SNR) algorithm.

Some curse

- curse of condition number: $\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$

$$\theta_{n+1} = \theta_n + \alpha_{n+1}(A\theta_n + \Delta_{n+1}) \tag{26}$$

  Optimal covariance $\Sigma_\theta^* = (A^2)^{-1}$

- curse of Markovian memory: when noise is not i.i.d., but Markovian, like in RL.

04

Conclusion

Synthesis

- SA = tool for solving $\bar{f}(\theta^*) = 0$,
- Step 1: ODE design (transient time)
- Step 2: ODE approximation
  - > Convergence sufficient conditions ($\bar{f}$ Lipschitz continuous, ODE globally asymptotically stable, Robbins-Monro step sizes, etc.),
  - > Optimizing covariance ($G^* = -[\partial_\theta \bar{f}(\theta^*)]^{-1}$),
  - > Trade off ($\rho = 1$ optimal rate $O(1/n)$ with slow transient time vs opposite for $\rho < 1$),
  - > Solutions (PJR averaging, ZAP),
  - > Limits (condition number, Markovian memory).

Other solutions?

Matrix Momentum Algorithms ?