Stable Robbins-Monro approximations through stochastic proximal updates

Thibaut Horel, Panos Toulis, Edoardo Airoldi

Stochastic approximation

Problem: function $g : \mathbb{R}^n \to \mathbb{R}^n$

- estimate θ^* such that $g(\theta^*) = 0$
- g is unknown, but we observe random $G(\theta)$ s.t. $\mathbb{E}\big[G(\theta)\big] = g(\theta)$

Stochastic approximation

Problem: function $g : \mathbb{R}^n \to \mathbb{R}^n$

- estimate θ^* such that $g(\theta^*) = 0$
- g is unknown, but we observe random $G(\theta)$ s.t. $\mathbb{E}\big[G(\theta)\big] = g(\theta)$

In this talk: $g = \nabla f$ for convex function f

$$g(\theta^*) = 0 \quad \Leftrightarrow \quad f(\theta^*) = \min_\theta f(\theta)$$

Stochastic approximation

Problem: function $g : \mathbb{R}^n \to \mathbb{R}^n$

- estimate θ^* such that $g(\theta^*) = 0$
- g is unknown, but we observe random $G(\theta)$ s.t. $\mathbb{E}\big[G(\theta)\big] = g(\theta)$

In this talk: $g = \nabla f$ for convex function f

$$g(\theta^*) = 0 \quad \Leftrightarrow \quad f(\theta^*) = \min_\theta f(\theta)$$

Example: maximum likelihood estimation

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i \,|\, \theta)$$

Robbins-Monro algorithm (1951)

Iterative estimation procedure (θ_n):

$$\theta_n = \theta_{n-1} - \gamma_n G(\theta_{n-1})$$

Robbins-Monro algorithm (1951)

Iterative estimation procedure (θ_n):

$$\theta_n = \theta_{n-1} - \gamma_n G(\theta_{n-1})$$

This is stochastic gradient descent!

Robbins-Monro algorithm (1951)

Iterative estimation procedure (θ_n):

$$\theta_n = \theta_{n-1} - \gamma_n G(\theta_{n-1})$$

This is stochastic gradient descent!

If $\sum \gamma_n = \infty$ and $\sum \gamma_n^2 < \infty$:

- $\mathbb{E}\big(\|\theta_n - \theta^*\|\big) \to 0$
- asymptotic normality

Robbins-Monro algorithm (1951)

Iterative estimation procedure (θ_n):

$$\theta_n = \theta_{n-1} - \gamma_n G(\theta_{n-1})$$

This is stochastic gradient descent!

If $\sum \gamma_n = \infty$ and $\sum \gamma_n^2 < \infty$:

- $\mathbb{E}\big(\|\theta_n - \theta^*\|\big) \to 0$
- asymptotic normality

For $\gamma_n = \frac{\gamma_1}{n}$, μ-strictly convex f, if $\mu\gamma_1 > 2$:

$$\mathbb{E}\big(\|\theta_n - \theta^*\|^2\big) \leq \frac{C_1}{n} + \frac{C_2 \cdot \exp(\gamma_1^2)\|\theta_0 - \theta^*\|^2}{n^2}$$

Robbins-Monro algorithm (1951)

Iterative estimation procedure (θ_n):

$$\theta_n = \theta_{n-1} - \gamma_n G(\theta_{n-1})$$

This is stochastic gradient descent!

If $\sum \gamma_n = \infty$ and $\sum \gamma_n^2 < \infty$:

- $\mathbb{E}\big(\|\theta_n - \theta^*\|\big) \to 0$
- asymptotic normality

For $\gamma_n = \frac{\gamma_1}{n}$, μ-strictly convex f, if $\mu\gamma_1 > 2$:

$$\mathbb{E}\big(\|\theta_n - \theta^*\|^2\big) \leq \frac{C_1}{n} + \frac{C_2 \cdot \exp(\gamma_1^2)\|\theta_0 - \theta^*\|^2}{n^2}$$

$\Rightarrow$ numerical instability!

Our procedure

Iterative procedure (θ_n):

$$\theta_n = \theta_{n-1} - \gamma_n G(\theta_n^+)$$
$$\text{where } \theta_n^+ = \theta_{n-1} - \gamma_n g(\theta_n^+)$$

Idealized implicit procedure

Our procedure

Iterative procedure (θ_n):

$$\theta_n = \theta_{n-1} - \gamma_n G(\theta_n^+)$$
$$\text{where } \theta_n^+ = \theta_{n-1} - \gamma_n g(\theta_n^+)$$

Idealized implicit procedure

$$\begin{aligned}\theta_n^+ &= \arg\min_\theta \left\{ \gamma_n f(\theta) + \frac{1}{2}\|\theta - \theta_{n-1}\|^2 \right\} \\ &= \text{prox}_{\gamma_n f}(\theta_{n-1})\end{aligned}$$

Robbins-Monro + proximal updates = "stochastic proximal updates"

The best of both worlds

If $\sum \gamma_n = \infty$ and $\sum \gamma_n^2 < \infty$:

- $\mathbb{E}\big(\|\theta_n - \theta^*\|\big) \to 0$
- asymptotic normality

The best of both worlds

If $\sum \gamma_n = \infty$ and $\sum \gamma_n^2 < \infty$:

- $\mathbb{E}\big(\|\theta_n - \theta^*\|\big) \to 0$
- asymptotic normality

For $\gamma_n = \frac{\gamma_1}{n}$ and strictly convex f:

$$\mathbb{E}\big(\|\theta_n - \theta^*\|^2\big) \leq \frac{C_1 \cdot \|\theta_0 - \theta^*\|^2 + C_2 \cdot \gamma_1}{n}$$

$\Rightarrow$ initial error is not amplified by γ_1!

Approximate instantiations

- Nested procedure:

$$\theta_n = \theta_{n-1} - \gamma_n G(\theta'_n)$$

θ'_n approximates $\theta_n^+ = \mathrm{prox}_{\gamma_n f}(\theta_{n-1})$ using RM algorithm:

$$\begin{aligned} x_0 &= \theta_{n-1} \\ x_k &= x_{k-1} - a_k\big(\gamma_n G(x_k) + x_k - \theta_{n-1}\big) \\ \theta'_n &= x_K \end{aligned}$$

Approximate instantiations

- Nested procedure:

$$\theta_n = \theta_{n-1} - \gamma_n G(\theta'_n)$$

θ'_n approximates $\theta_n^+ = \mathrm{prox}_{\gamma_n f}(\theta_{n-1})$ using RM algorithm:

$$\begin{aligned} x_0 &= \theta_{n-1} \\ x_k &= x_{k-1} - a_k\big(\gamma_n G(x_k) + x_k - \theta_{n-1}\big) \\ \theta'_n &= x_K \end{aligned}$$

We analyze the convergence rate of this procedure!

Approximate instantiations

- Nested procedure:
$$\theta_n = \theta_{n-1} - \gamma_n G(\theta'_n)$$
θ'_n approximates $\theta^+_n = \mathrm{prox}_{\gamma_n f}(\theta_{n-1})$ using RM algorithm:
$$\begin{aligned} x_0 &= \theta_{n-1} \\ x_k &= x_{k-1} - a_k\big(\gamma_n G(x_k) + x_k - \theta_{n-1}\big) \\ \theta'_n &= x_K \end{aligned}$$
We analyze the convergence rate of this procedure!

- Implicit SGD [Bertsekas 2011, Toulis & Airoldi 2017]:
$$\theta_n = \theta_{n-1} - \gamma_n G(\theta_n)$$
θ_n is an unbiased estimator of $\theta^+_n = \mathrm{prox}_{\gamma_n f}(\theta_{n-1})$

Numerical evaluation

Quantile estimation:

- distribution with CDF F, estimate θ^*_α such that $F(\theta^*_\alpha) = \alpha$
- $g(\theta) = F(\theta) - \alpha$
- observations: $G(\theta) = \mathbb{I}[Z \leq \theta] - \alpha$ where $Z \sim F$