Statistical machine learning and convex optimization **2016**

Lecture 3 — March 17th

Lecturer: Francis Bach *Scribe: Yichen Liu, Rui Li*

3.1 Motivation

In this lecture we introduce stochastic approximation methods that attempt to find zeros of functions which can be hardly computed directly. See [2] for more references.

Suppose we wish to find the root of a function h, which does not have a closed-form solution. What we can do is to do experiments or simulations to sample h at some particular values of θ. Generally, these samples are noisy, but we can make an easier assumption that experiment results are the sum of true values and noise: $y = h(\theta) + \varepsilon$, where ε denotes the noise, which is assumed to be random in each simulation.

If we try to solve such a problem with traditional algorithms for the deterministic situation, for example, Newton's procedure, then at each iterate θ_n, we need to make an estimate for h. Even if we assume that noises are zero-mean and identically distributed and independent, a reasonable estimator seems to be the empirical mean:

$$h(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} y_n$$

since the law of large numbers ensures the convergence as we take samples for large enough times. However, such a solution is unstable and can easily become time-consuming and low-efficient because no one can ensure we won't spend time taking samples at some points which are far from the root.

So the key point of this problem is that we only have access to the sample value y, and we have no way of removing the noise from it, i.e., of isolating the exact value of $h(\theta)$. We will try instead to use every new observation in the algorithm.

Link with fixed point iterations. Note that when there is no noise, and h is observed, we may consider the fixed point iteration $\theta_n = \theta_{n-1} - \gamma h(\theta_{n-1})$, which is not convergent in general.

3.2 Robbins-Monro algorithm

A classical methodology, studied by Robbins and Monro [1], is

$$\theta_n = \theta_{n-1} - \gamma_n [h(\theta_{n-1}) + \varepsilon_n]$$

where $h : \mathbb{R}^d \to \mathbb{R}^d$, $(\varepsilon_n)_{n\geq 1}$ are i.i.d random variables taking values in $\mathbb{R}^d$ denote the noise term usually having zero mean. A typical example is Gaussian noise (but we will consider more general situations, in particular in the context of machine learning).

The main challenge of Robbins-Monro algorithm is to:

- Find general sufficient conditions for iterates to converge to the root;
- Compare different types of convergence of θ_n and try to make the analysis;
- Compute the rate of convergence and decide the choice of step-sizes;
- Study asymptotical behavior.

3.2.1 Example of mean estimation

We start from a simple but inspiring example in which h is a linear function.

Let $h(\theta) = \theta - x$, obviously $\theta_* = x$ is the unique root of h; we have

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

For the choice of γ_n, if we start from $\theta_0 = 0$, a direct computation shows that

- $\theta_n = \frac{1}{n}\sum_{k=1}^n x_k$ when $\gamma_n = 1/n$
- $\theta_n = \frac{2}{n(n+1)}\sum_{k=1}^n kx_k$ when $\gamma_n = 2/(n+1)$,

that is, empirical means are instances of the Robbins-Monro algorithm, and for all cases by recursion we see

$$\begin{aligned}\theta_n - x &= (1-\gamma_n)(\theta_{n-1} - x) + \gamma_n(x_n - x),\\ &= \underbrace{\prod_{k=1}^n (1-\gamma_k)(\theta_0 - x)}_{\textit{deterministic error}} + \underbrace{\sum_{i=1}^n \prod_{k=i+1}^n (1-\gamma_k)\gamma_i(x_i - x)}_{\textit{random error}}.\end{aligned}$$

Since we have $\mathbb{E}[x_n] = x$ and $\mathbb{E}[\|x_n - x\|^2] = \sigma^2 > 0$, which follows the general assumption noises $(\varepsilon_n)_{n\geq 1}$ are zero-mean and i.i.d, so we have

$$\mathbb{E}[\|\theta_n - x\|^2] = \prod_{k=1}^n (1-\gamma_k)^2\|\theta_0 - x\|^2 + \sum_{i=1}^n \gamma_i^2 \prod_{k=i+1}^n (1-\gamma_k)^2\sigma^2.$$

If we hope θ_n converges to x in quadratic mean, it's sufficient to have

$$\lim_{n\to+\infty}\prod_{k=1}^n (1-\gamma_k)^2 = 0 \qquad \lim_{n\to+\infty}\sum_{i=1}^n \gamma_i^2 \prod_{k=i+1}^n (1-\gamma_k)^2 = 0.$$

For the deterministic error, if $\gamma_n = o(1)$,

$$\log\prod_{k=1}^n (1-\gamma_k)^2 = 2\sum_{k=1}^n \log(1-\gamma_k) \sim -2\sum_{k=1}^n \gamma_k$$

which is supposed to tend to $-\infty$ since we need $\lim_{n\to+\infty}\prod_{k=1}^{n}(1-\gamma_k)^2 = 0$.

For the noise term, we firstly also have $\sum_{i=1}^{n}\gamma_i^2\prod_{k=i+1}^{n}(1-\gamma_k)^2 \sim \sum_{i=1}^{n}\gamma_i^2\prod_{k=i+1}^{n}(1-2\gamma_k)$. We thus need a general decomposition of the noise term.

3.2.2 Forgetting of initial conditions

Now we assume $(\gamma_n)_{n\geq 1}$ is decreasing and bounded by an arbitrary $1/\mu > 0$. We have

$$\prod_{k=1}^{n}(1-\mu\gamma_k) \;\leqslant\; \prod_{k=1}^{n}\exp(-\mu\gamma_k),$$

which is going to zero exponentially fast for $\gamma_n = C/n^\alpha$ and $\alpha \in (0,1)$. If $\gamma_n = C/n$, we get

$$\prod_{k=1}^{n}\exp(-\mu\gamma_k) \approx \exp(-\mu C\log n) = 1/n^{\mu C}.$$

3.2.3 Decomposition of the noise term

Now we assume $(\gamma_n)_{n\geq 1}$ is decreasing and bounded by an arbitrary $1/\mu > 0$.

Then for any $k \leq l$, it's clear that

$$\prod_{i=k+1}^{n}(1-\mu\gamma_i) \leq \prod_{j=l+1}^{n}(1-\mu\gamma_j) \leq 1$$

With the basic inequality $\log(1-x) \leq -x$, for any $m \in \{1,...,n\}$, we have

$$\begin{aligned}
\sum_{k=1}^{n}\prod_{i=k+1}^{n}(1-\mu\gamma_i)\gamma_k^2 =\; & \sum_{k=1}^{m}\prod_{i=k+1}^{n}(1-\mu\gamma_i)\gamma_k^2 + \sum_{k=m+1}^{n}\prod_{i=k+1}^{n}(1-\mu\gamma_i)\gamma_k^2 \\
\leqslant\; & \prod_{i=m+1}^{n}(1-\mu\gamma_i)\sum_{k=1}^{m}\gamma_k^2 + \gamma_m\sum_{k=m+1}^{n}\prod_{i=k+1}^{n}(1-\mu\gamma_i)\gamma_k \\
\leqslant\; & \exp\left(-\mu\sum_{i=m+1}^{n}\gamma_i\right)\sum_{k=1}^{m}\gamma_k^2 + \frac{\gamma_m}{\mu}\sum_{k=m+1}^{n}\left[\prod_{i=k+1}^{n}(1-\mu\gamma_i) - \prod_{i=k}^{n}(1-\mu\gamma_i)\right] \\
\leqslant\; & \exp\left(-\mu\sum_{i=m+1}^{n}\gamma_i\right)\sum_{k=1}^{m}\gamma_k^2 + \frac{\gamma_m}{\mu}\left[1 - \prod_{i=m+1}^{n}(1-\mu\gamma_i)\right] \\
\leqslant\; & \exp\left(-\mu\sum_{i=m+1}^{n}\gamma_i\right)\sum_{k=1}^{n}\gamma_k^2 + \frac{\gamma_m}{\mu}.
\end{aligned}$$

Since the inequality above is true for any $n \in \mathbb{N}$ and $m \in \{1,...n\}$, and by the preceding assumption $\gamma_n \to 0$, we therefore only need

$$\lim_{n\to+\infty}\exp\left(-\mu\sum_{i=m+1}^{n}\gamma_i\right)\sum_{k=1}^{n}\gamma_k^2 = 0$$

for any fixed m. We will typically take $m = n/2$ (except for $\alpha = 1$, see below).

Typically, we take γ_n in form of $Cn^{-\alpha}$ where $\alpha \in \mathbb{R}^*$ and $C \leqslant 1/\mu$ (note that if this is not true for $n = 1$, this is true for n large enough):

- $\alpha > 1$:

  Since $\sum_{i=1}^n 1/i^\alpha = \text{ Cst } + O(1/n^{\alpha-1})$ and $\sum_{k=1}^n \gamma_k^2 < +\infty$, it's clear that

  $$\lim_{n\to+\infty} \exp\left(-\mu \sum_{i=m+1}^n \gamma_i\right) \sum_{k=1}^n \gamma_k^2 > 0.$$

  The bound does not go to zero (the step-sizes are too small).

- $\alpha \in (0, 1)$:

  Because $\sum_{i=1}^n 1/i^\alpha = \text{ Cst} \times n^{1-\alpha} + O(1)$, so whatever $\sum_{k=1}^n \gamma_k^2 < \infty$ or $= \infty$, we always have

  $$\lim_{n\to+\infty} \exp\left(-\mu \sum_{i=m+1}^n \gamma_i\right) \sum_{k=1}^n \gamma_k^2 = 0 \text{ exponentially fast.}$$

  This leads to a rate in $O(n^{-\alpha})$.

- $\alpha = 1$:

  We know

  $$\sum_{i=1}^n 1/i = \log(n) + \gamma + O(1/n),$$

  where $\gamma > 0$ is Euler-Mascheroni constant, and $\sum_{k=1}^n \gamma_k^2 < +\infty$.

  Thus, we have, with $\gamma = C/n$, and $C \leqslant 1/\mu$:

  $$\begin{aligned}
  \sum_{k=1}^n \prod_{i=k+1}^n (1 - \mu\gamma_i)\gamma_k^2 &\leqslant \sum_{k=1}^n \prod_{i=k+1}^n \exp(-\mu\gamma_i)\gamma_k^2 \\
  &= \sum_{k=1}^n \exp(-\mu \sum_{i=k+1}^n \gamma_i)\gamma_k^2 \\
  &\approx \sum_{k=1}^n \exp(-\mu \log n + \mu \log k)\frac{C^2}{k^2} \\
  &\approx \frac{C^2}{n^{\mu C}} \sum_{k=1}^n \frac{1}{k^{2-C\mu}} \\
  &\approx O(1/n).
  \end{aligned}$$

  Initial conditions are forgotten as $1/n^{\mu C}$ and hence if C is too small, convergence is slow.

Thus a sufficient condition for convergence in quadratic mean is $\alpha \in (0,1)$ when γ_n has the form of $Cn^{-\alpha}$.

In practice, we hope that the iterates will converge to the root $\theta_* = x$ in the strongest sense, i.e. $\theta_n \to \theta_*$ almost surely, since a random noise exists all the time. From the example above we see that for some specific case, it's quite feasible to realize convergence in quadratic mean, which implies convergence in probability (see below).

3.3 Recall: convergence of random variables

We review several convergences in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 3.1 *A sequence of random variables $(X_n)_{n\geq 1}$ is said to converge to X almost-surely if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$.*

Definition 3.2 *A sequence of random variables $(X_n)_{n\geq 1}$ is said to converge to X in probability if for any $\varepsilon, \delta > 0$, there exists a $N \in \mathbb{N}$, such that for $n \geq N$, $\mathbb{P}(|X_n - X| \geq \varepsilon) \leq \delta$.*

Definition 3.3 *A sequence of random variables $(X_n)_{n\geq 1}$ is said to converge in r-th ($r \geq 1$) mean to X if $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] = 0$.*

Relationship between convergences

- It's obvious that convergence a.s. implies convergence in probability.
- Convergence in r-th mean implies convergence in probability by using Markov's inequality
$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}[|X_n - X|^r]}{\varepsilon^r}$$
- Convergence a.s. and dominant convergence theorem implies convergence in mean.
- To get convergence a.s. from convergence in probability, recall that

Lemma 3.4 (Borel-Cantelli) *Let $(E_n)_{n\geq 1}$ be a sequence of sets such that $\sum_{i=1}^{\infty} \mathbb{P}(E_n) < \infty$, then $\mathbb{P}(\limsup_{n\to\infty} E_n) = 0$, where $\limsup_{n\to\infty} E_n := \cap_{N=1}^{\infty} \cup_{n\geq N} E_n$.*

If for any $\varepsilon > 0$, we propose $A_n(\varepsilon) := \{|X_n - X| > \varepsilon\}$, and $\sum_{n=1}^{\infty} \mathbb{P}(A_n(\varepsilon)) < \infty$, then by the lemma above we have
$$\mathbb{P}(\limsup_{n\to\infty} A_n(\varepsilon)) = 0$$
this is to say there exists $N = N(\varepsilon)$ such that for $n \geq N$, $\mathbb{P}(|X_n - X| \leq \varepsilon) = 1$, which means $X_n \to X$ *a.s.* (note that this will apply to our convergence in high-probability in the next lecture).

3.4 Need for Lyapunov function

Robbins-Monro algorithm cannot converge all the time, we need to introduce the Lyapunov function.

Definition 3.5 *A Lyapunov function V associated to h verifies the following properties:*

- *Non-negative: $V \geq 0$;*
- *Continuously-differentiable with L-Lipschitz-continuous gradients;*
- *Control of h: $\forall \theta$, $\|h(\theta)\|^2 \leqslant C(1 + V(\theta))$;*
- *Gradient condition: there exists $\alpha > 0$, $\forall \theta$, $h(\theta)^\top V'(\theta) \geqslant \alpha \|V'(\theta)\|^2$.*

If $h = f'$, then $V(\theta) = f(\theta) - \inf f$ is the default (but not only) choice for Lyapunov function, even if f is not convex. However, this usually requires some additional condition:

- $\|V'(\theta)\|^2 \geqslant 2\mu V(\theta)$ (which is satisfied for V a μ-strongly-convex function, or more generally for Polyak-Lojasiewicz conditions).

Generalized noise sequence In the rest part of this section, we no longer restrict ourselves to the case that $(\varepsilon_n)_{n\geq 1}$ are i.i.d., and we assume that for each $n \in \mathbb{N}$, ε_n is $\mathcal{F}_n$-measurable where $(\mathcal{F}_n)_{n\geq 1}$ is a filtration on the probability space such that

$$\mathbb{E}[\varepsilon_n | \mathcal{F}_{n-1}] = 0 \quad \mathbb{E}[\|\varepsilon_n\|^2 | \mathcal{F}_{n-1}] \leq \sigma^2$$

almost surely.

Since $\theta_n = \theta_{n-1} - \gamma_n \big[h(\theta_{n-1}) + \varepsilon_n\big]$, we see θ_n is also $\mathcal{F}_n$-measurable.

3.4.1 Convergence of the Lyapunov function

Applying Robbins-Monro algorithm to h and use the regularity (or other properties) of Lyapunov function V, we have

$$\begin{aligned} V(\theta_n) &\leq V(\theta_{n-1}) + V'(\theta_{n-1})^\top (\theta_n - \theta_{n-1}) + \frac{L}{2}\|\theta_n - \theta_{n-1}\|^2 \\ &= V(\theta_{n-1}) - \gamma_n V'(\theta_{n-1})^\top (h(\theta_{n-1}) + \varepsilon_n) + \frac{L\gamma_n^2}{2}\|h(\theta_{n-1}) + \varepsilon_n\|^2, \end{aligned}$$

and by the properties above and assumptions on noise,

$$\begin{aligned} \mathbb{E}[V(\theta_n)|\mathcal{F}_{n-1}] &\leq V(\theta_{n-1}) - \gamma_n V'(\theta_{n-1})^\top h(\theta_{n-1}) + \frac{L\gamma_n^2}{2}\|h(\theta_{n-1})\|^2 + \frac{L\gamma_n^2}{2}\sigma^2 \\ &\leq V(\theta_{n-1}) - \alpha\gamma_n \|V'(\theta_{n-1})\|^2 + \frac{LC\gamma_n^2}{2}\big[1 + V(\theta_{n-1})\big] + \frac{L\gamma_n^2}{2}\sigma^2 \\ &\leq V(\theta_{n-1})[1 + \frac{LC\gamma_n^2}{2}] - \alpha\gamma_n \|V'(\theta_{n-1})\|^2 + \frac{L\gamma_n^2}{2}(C + \sigma^2) \end{aligned}$$

If the additional condition $\|V'(\theta)\|^2 \geqslant 2\mu V(\theta)$ holds, with $\gamma_n \leq 2\alpha\mu/LC$ for n large enough, we have

$$\begin{aligned}\mathbb{E}[V(\theta_n)|\mathcal{F}_{n-1}] &\leq V(\theta_{n-1})[1 + \frac{LC\gamma_n}{2}\frac{2\alpha\mu}{LC}] - 2\alpha\gamma_n\mu V(\theta_{n-1}) + \frac{L\gamma_n^2}{2}(C+\sigma^2) \\ &\leq V(\theta_{n-1})[1-\alpha\mu\gamma_n] + M\gamma_n^2\end{aligned}$$

which implies $\mathbb{E}V(\theta_n) \leq \mathbb{E}V(\theta_{n-1})[1-\alpha\mu\gamma_n] + M\gamma_n^2$ where $M := L(C+\sigma^2)/2$.

3.4.2 Convergence of expectation

Let $\delta_n := \mathbb{E}[V(\theta_n)] \geq 0$, so, $\delta_n \leqslant \delta_{n-1}[1-\alpha\mu\gamma_n] + M\gamma_n^2$. By recursion,

$$\delta_n \leqslant \prod_{k=1}^{n}(1-\alpha\mu\gamma_k)\delta_0 + M\sum_{i=1}^{n}\gamma_i^2\prod_{k=i+1}^{n}(1-\alpha\mu\gamma_k).$$

By an analogous argument to mean estimation (analysis of the noise term), the sufficient conditions for convergence in mean of the Lyapunov function is $\sum_n \gamma_n = +\infty$ and $\gamma_n \to 0$. Particularly, when γ_n is in the form $Cn^{-\alpha}$ where $C > 0$, $\alpha \in (0,1)$ works, and for $\alpha = 1$, only for C large enough.

3.4.3 Convergence almost surely

To establish almost-sure convergence from the recurrence inequality derived above, we firstly present the Robbins-Siegmund theorem [3].

Theorem 3.6 (Robbins-Siegmund) *Let $(V_n)_{n\geq1}, (\beta_n)_{n\geq1}, (\chi_n)_{n\geq1}, (\eta_n)_{n\geq1}$ be four non-negative $(\mathcal{F}_n)_{n\geq1}$-adapted processes such that $\sum_n \beta_n < \infty$ and $\sum_n \chi_n < \infty$ almost surely. If for each $n \in \mathbb{N}$,*

$$\mathbb{E}[V_n|\mathcal{F}_{n-1}] \leqslant V_{n-1}(1+\beta_{n-1}) + \chi_{n-1} - \eta_{n-1}$$

then $(V_n)_{n\geq1}$ converges almost surely to a random variable V_∞ and $\sum_{n=1}^{\infty}\eta_n$ is finite almost surely.

Proof Define $\alpha_n := \prod_{k=1}^{n}(1+\beta_k)$, then $\alpha_n(\geq 1)$ is $\mathcal{F}_n$-measurable, and converges almost surely by the assumption on $(\beta_n)_{n\geq1}$. Since $(1+\beta_n) = \alpha_n/\alpha_{n-1}$,

$$\begin{aligned}&\mathbb{E}[V_n|\mathcal{F}_{n-1}] \leqslant V_{n-1}\frac{\alpha_{n-1}}{\alpha_{n-2}} + \chi_{n-1} - \eta_{n-1} \\ \Rightarrow \quad &\mathbb{E}[\frac{V_n}{\alpha_{n-1}}|\mathcal{F}_{n-1}] \leqslant \frac{V_{n-1}}{\alpha_{n-2}} + \frac{\chi_{n-1}}{\alpha_{n-1}} - \frac{\eta_{n-1}}{\alpha_{n-1}}\end{aligned}$$

Define

$$V_n' := \frac{V_n}{\alpha_{n-1}}, \quad \chi_n' := \frac{\chi_n}{\alpha_n}, \quad \eta_n' := \frac{\eta_n}{\alpha_n}$$

we can rewrite the inequality as

$$\mathbb{E}[V_n'|\mathcal{F}_{n-1}] \leqslant V_{n-1}' + \chi_{n-1}' - \eta_{n-1}'$$

Define $Y_n := V'_n - \sum_{k=1}^{n-1}(\chi'_k - \eta'_k)$, then

$$\mathbb{E}[Y_n|\mathcal{F}_{n-1}] \leqslant V'_{n-1} + \chi'_{n-1} - \eta'_{n-1} - \sum_{k=1}^{n-1}(\chi'_k - \eta'_k) = Y_{n-1}$$

which verifies $(Y_n)_{n\geq 1}$ is a supermartingale. Let $\tau_a = \inf\{n \geq 1, \sum_{k=1}^{n} \chi'_k > a\}$ be a stopping time. Then

$$\begin{aligned}\mathbb{E}[Y_{n\wedge\tau_a}|\mathcal{F}_{n-1}] &= \mathbb{E}[Y_{\tau_a}\mathbb{1}_{\{\tau_a\leq n-1\}} + Y_n\mathbb{1}_{\{\tau_a>n-1\}}|\mathcal{F}_{n-1}] \\ &\leq Y_{\tau_a}\mathbb{1}_{\{\tau_a\leq n-1\}} + Y_{n-1}\mathbb{1}_{\{\tau_a>n-1\}} \\ &= Y_{(n-1)\wedge\tau_a}\end{aligned}$$

where we use the fact $Y_{\tau_a}\mathbb{1}_{\{\tau_a\leq n-1\}}$ is $\mathcal{F}_{n-1}-$ measurable and $(Y_n)_{n\geq 1}$ is a supermartingale.

So $(Y_{n\wedge\tau_a})_{n\geq 1}$ is also a supermartingale and

$$Y_{n\wedge\tau_a} \geq \sum_{k=1}^{(n-1)\wedge\tau_a} \chi'_k \geq -a$$

for all n. It follows from the Doob convergence theorem (see [2]) that

$$\lim_{n\to\infty} Y_{n\wedge\tau_a} \text{ exists}$$

and is finite a.s., i.e. $\lim_{n\to\infty} Y_n$ exists and finite on

$$\{\tau_a = \infty\} = \{\sum_{n=1}^{\infty} \chi'_k \leq a\}$$

Since $\sum_{n=1}^{\infty} \chi'_k \leq \sum_{n=1}^{\infty} \chi_k < \infty$ a.s., let $a \to \infty$, we see that $\lim_{n\to\infty} Y_n$ exists and is finite almost surely.

Hence by $Y_n = V'_n - \sum_{k=1}^{n-1}(\chi'_k - \eta'_k)$ and V', η' are non-negative,

$$\lim_{n\to\infty} V'_n \text{ exists} \quad \text{and} \quad \sum_{k=1}^{\infty} \eta'_k < \infty$$

and it follows from

$$V'_n := \frac{V_n}{\alpha_{n-1}}, \quad \eta_n := \eta'_n\alpha_n \leq \eta'_n \prod_{k=1}^{\infty}(1+\beta_k)$$

that we conclude $\lim_{n\to\infty} V_n$ exists and is finite, and $\sum_{n=1}^{\infty} \eta_n < \infty$ almost surely. ■

Applying Robbins-Siegmund theorem to Lyapunov function V, where $V(\theta_n)_{n\geq 1}$ satisfies:

$$\mathbb{E}[V(\theta_n)|\mathcal{F}_{n-1}] = V(\theta_{n-1})[1 + \frac{LC\gamma_n^2}{2}] - \alpha\gamma_n\|V'(\theta_{n-1})\|^2 + \frac{L\gamma_n^2}{2}(C+\sigma^2)$$

and set

$$V_n = V(\theta_n), \quad \beta_n = \frac{LC\gamma_{n+1}^2}{2}, \quad \chi_n = \frac{L\gamma_{n+1}^2}{2}(C+\sigma^2), \quad \eta_n = \alpha\gamma_{n+1}\|V'(\theta_n)\|^2$$

It's clear that these four sequence are non-negative and adapted, so we conclude $\lim_{n\to\infty} V(\theta_n)$ exists and is finite almost surely. This requires that $\sum_n \gamma_n^2 < \infty$, i.e., that the step-sizes are squared summable.

3.5 Robbins-Monro analysis - asymptotic normality

We consider the asymptotic normality of the output of Robbins-Monro algorithm. We emphasize here intuitive results with simple (simplistic) assumptions. For more precise statements and proofs, see [7].

We consider the traditional step-size $\gamma = C/n$, and provide a proof sketch for the differential A of h at unique θ_* symmetric:

$$\begin{aligned}
\theta_n &= \theta_{n-1} - \gamma_n h(\theta_{n-1}) - \gamma_n \varepsilon_n \\
&\approx \theta_{n-1} - \gamma_n \big[h'(\theta_*)(\theta_{n-1} - \theta_*) \big] - \gamma_n \varepsilon_n + \gamma_n O(\|\theta_n - \theta_*\|^2) \\
&\approx \theta_{n-1} - \gamma_n A(\theta_{n-1} - \theta_*) - \gamma_n \varepsilon_n \\
\theta_n - \theta_* &\approx (I - \gamma_n A) \cdots (I - \gamma_1 A)(\theta_0 - \theta_*) - \sum_{k=1}^{n} (I - \gamma_n A) \cdots (I - \gamma_{k+1} A) \gamma_k \varepsilon_k \\
\theta_n - \theta_* &\approx \exp\big[- (\gamma_n + \cdots + \gamma_1) A \big] (\theta_0 - \theta_*) - \sum_{k=1}^{n} \exp\big[- (\gamma_n + \cdots + \gamma_{k+1}) A \big] \gamma_k \varepsilon_k \\
&\approx \exp\big[- CA \log n \big] (\theta_0 - \theta_*) - \sum_{k=1}^{n} \exp\big[- C(\log n - \log k) A \big] \frac{C}{k} \varepsilon_k .
\end{aligned}$$

We have used above the approximation of the harmonic series by the logarithm. We obtain asymptotic normality by averaging zero mean random variables.

Assuming A, $(\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top$ and $\mathbb{E}(\varepsilon_k \varepsilon_k^\top) = \Sigma$ commute, we may compute the expected covariance matrix as follows (note that we allow ourselves to take powers and logarithms of matrices, which can be done formally by taking functions of eigenvalues while preserving eigenvectors):

$$\begin{aligned}
\mathbb{E}(\theta_n - \theta_*)(\theta_n - \theta_*)^\top &\approx \exp\big[- 2CA \log n \big] (\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top \\
&\qquad + \sum_{k=1}^{n} \exp\big[- 2C(\log n - \log k) A \big] \frac{C^2}{k^2} \mathbb{E}(\varepsilon_k \varepsilon_k^\top) \\
&\approx n^{-2CA} (\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top + n^{-2CA} \sum_{k=1}^{n} C^2 k^{2CA - 2} \Sigma \\
&\approx n^{-2CA} (\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top + n^{-2CA} C^2 \frac{n^{2CA-1}}{2CA - 1} \Sigma .
\end{aligned}$$

With the step-size $\gamma = C/n$, we need $2C\lambda_{\min}(A) \geqslant 1$ for convergence, which implies that the first term depending on initial condition $\theta_* - \theta_0$ is negligible.

Moreover, we see the difficulty in setting the constant C: if C too small, we may have no convergence, while when C is too large, we obtain a large variance.

Finally, there is a strong dependence on the conditioning of the problem, that is, if $\lambda_{\min}(A)$ is small, then C has to be large. One way to improve the conditioning is to "choose"

A proportional to identity for optimal behavior (by premultiplying A by a conditioning matrix that make A close to a constant times identity).

3.6 Polyak-Ruppert averaging

3.6.1 Problems with Robbins-Monro algorithm

As we have illustrated before, we can see the convergence of the Robbins-Monro algorithm heavily depends on the choice of step size $(\gamma_n)_{n\geq 1}$ which not only determines the efficiency, i.e., convergence rate, but also decides whether the iterates affiliated to $(\gamma_n)_{n\geq 1}$ will truly converge to result we desire or not. Besides, uncontrolled unknown conditions of the problem also affects the algorithm behavior since we are supposed to know nothing about the noise.

3.6.2 Cesaro means

An alternative way, proposed by Polyak and Juditsky [4] and Ruppert [5], is to estimate θ_* by $(\bar{\theta}_n)_{n\geq 1}$ instead of $(\theta_n)_{n\geq 1}$, where $\bar{\theta}_n$ is defined

$$\bar{\theta}_n = \frac{1}{n}\sum_{k=1}^{n}\theta_k$$

named Cesaro mean of $(\theta_n)_{n\geq 1}$ and can be computed recursively as

$$\bar{\theta}_n = (1-\frac{1}{n})\bar{\theta}_{n-1} + \frac{1}{n}\theta_n$$

The idea is inspired by Cesaro's theorem.

Theorem 3.7 (Cesaro) *Assume* $\lim_{n\to\infty}\theta_n = \theta_*$ *with convergence rate* $\|\theta_n - \theta_*\| \leq \alpha_n$ *where* $\alpha_n \to 0$*, then* $\lim_{n\to\infty}\bar{\theta}_n = \theta_*$ *with convergence rate* $\bar{\alpha}_n := \frac{1}{n}\sum_{k=1}^{n}\alpha_k$.

Proof The convergence $\bar{\theta}_n \to \theta_*$ is well-known, and

$$\|\bar{\theta}_n - \theta_*\| \leq \frac{1}{n}\sum_{k=1}^{n}\|\theta_k - \theta_*\| \leq \frac{1}{n}\sum_{k=1}^{n}\alpha_k = \bar{\alpha}_n$$

By applying the result above to α_n, we see $\bar{\alpha}_n \to 0$. Thus $\theta_n \to \theta_*$ with the convergence rate $\bar{\alpha}_n$. ■

The convergence rate of $\bar{\theta}_n \to \theta_*$ depends on the rate of original sequence. However, in the case $\sum_{k=1}^{\infty}\alpha_n < \infty$, the rate is equivalent to $1/n$ (and we lose potential exponential convergence).

Note that there are many counterexamples in which the sequence of Cesaro means converges, but the original sequence does not.

3.6.3 Asymptotic distribution of $\bar{\theta}_n - \theta_*$

In this section, we only give intuitive arguments, for more details, see [3, 4].

Now we assume in the recursion

$$\theta_n = \theta_{n-1} - \gamma_n[h(\theta_{n-1}) + \varepsilon_n]$$

the function $h \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$, $\gamma_n = Cn^{-\alpha}$ and the noise $(\varepsilon_n)_{n\geq 1}$ are i.i.d. following $\mathcal{N}(0, \Sigma)$.

So we have

$$h(\theta_{n-1}) = \frac{1}{\gamma_n}\big[\theta_{n-1} - \theta_n\big] - \varepsilon_n$$

and by Taylor series expansion, we have

$$A(\theta_{n-1} - \theta_*) + O(\|\theta_{n-1} - \theta_*\|^2) = \frac{1}{\gamma_n}[\theta_{n-1} - \theta_n] - \varepsilon_n$$

where $A = h'(\theta_*) \in \mathbb{R}^{d\times d}$, and we use the fact $h(\theta_*) = 0$.

By the previous argument, we know $\|\theta_n - \theta_*\|^2 = O(n^{-\alpha})$, thus

$$\begin{aligned} A(\theta_{n-1} - \theta_*) =& \frac{1}{\gamma_n}[\theta_{n-1} - \theta_n] - \varepsilon_n + O(n^{-\alpha}) \\ \Rightarrow \quad \frac{1}{n}\sum_{k=1}^{n} A(\theta_{k-1} - \theta_*) =& \frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}\big[\theta_{k-1} - \theta_k\big] - \frac{1}{n}\sum_{k=1}^{n}\varepsilon_k + O(n^{-\alpha}) \\ =& \frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}\big[\theta_{k-1} - \theta_k\big] + \mathcal{N}(0, \Sigma/n) + O(n^{-\alpha}) \end{aligned}$$

because of the central limit theorem.

Using Abel's summation formula, we have

$$\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\theta_{k-1} - \theta_k) = \frac{1}{n}\sum_{k=1}^{n-1}(\theta_k - \theta_*)(\gamma_{k+1}^{-1} - \gamma_k^{-1}) - \frac{1}{n}(\theta_n - \theta_*)\gamma_n^{-1} + \frac{1}{n}(\theta_0 - \theta_*)\gamma_1^{-1}$$

which implies

$$\Big\|\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\theta_{k-1} - \theta_k)\Big\| \leqslant \frac{1}{n}\sum_{k=1}^{n-1}\|\theta_k - \theta_*\| \cdot |\gamma_{k+1}^{-1} - \gamma_k^{-1}| + \frac{1}{n}\|\theta_n - \theta_*\|\gamma_n^{-1} + \frac{1}{n}\|\theta_0 - \theta_*\|\gamma_1^{-1}$$

Since $\|\theta_n - \theta_*\| = O(n^{-\frac{\alpha}{2}})$ and $\gamma_n = Cn^{-\alpha}$, so

$$\frac{1}{n}\sum_{k=1}^{n-1}\|\theta_k - \theta_*\| \cdot |\gamma_{k+1}^{-1} - \gamma_k^{-1}| = O(n^{\frac{\alpha}{2}-1}), \qquad \frac{1}{n}\|\theta_n - \theta_*\|\gamma_n^{-1} = O(n^{\frac{\alpha}{2}-1})$$

Thus

$$\frac{1}{n}\sum_{k=1}^{n} A(\theta_{k-1} - \theta_*) = \mathcal{N}(0, \Sigma/n) + O(n^{-\alpha}) + O(n^{\alpha/2-1})$$

and we can conclude that $\bar{\theta}_n - \theta_*$ is asymptotically Gaussian with zero-mean and covariance $\frac{1}{n}A^{-1}\Sigma A^{-1}$.

Moreover, the asymptotic variance is independent of the step-size (if this steps-size is $\gamma_n = O(n^{-\alpha})$ for $n \in (1/2, 1)$).

Bibliography

[1] H. Robbins and S. Monro. A stochastic approximation method. Ann. Math. Statistics, 22:400-407, 1951. ISSN 0003-4851

[2] M. Duflo. Algorithmes Stochastiques. Springer-Verlag, 1996.

[3] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. Optimizing Methods in Statistics, 1971.

[4] B. T. Polyak and A. B. Juditsky. Acceleration of Stochastic Approximation by Averaging. SIAM Journal on Control and Optimization, 1992.

[5] D. Ruppert. Efficient Estimations from a Slowly Convergent Robbins-Monro Process. Cornell University Operations Research and Industrial Engineering, 1988.

[6] M. Schmidt, N. Le Roux, and F. Bach. Optimization with approximate gradients. Technical report, HAL, 2011.

[7] V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.