Asymptotic behavior of stochastic algorithms with statistical applications
Part 2

Bernard Bercu

University of Bordeaux, France

ETICS Annual Research School, Fréjus, 2019

Outline

Outline

Stochastic approximation.

Herbert Robbins

Stochastic approximation.

Let f be an **unknown function** from $\mathbb{R}^d$ to $\mathbb{R}^d$.

Goal

$\longrightarrow$ For a **given vector** α of $\mathbb{R}^d$, find a vector $\boldsymbol{x^*}$ which satisfies

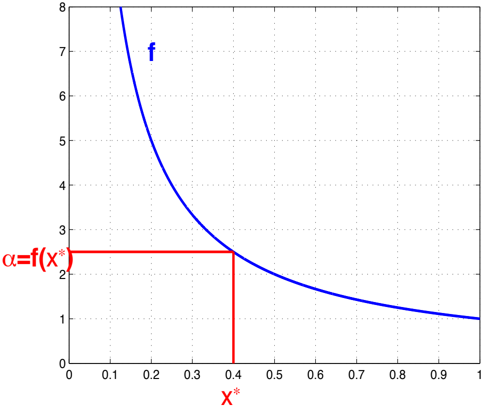
$$\boldsymbol{f(x^*) = \alpha.}$$

We will assume in all the sequel that for all $n \geqslant 1$, we can compute $X_1, \ldots, X_n$ of $\mathbb{R}^d$ and we can find Y_{n+1} of $\mathbb{R}^d$ such that

$$\mathbb{E}[Y_{n+1}|\mathcal{F}_n] = f(X_n)$$

where $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$.

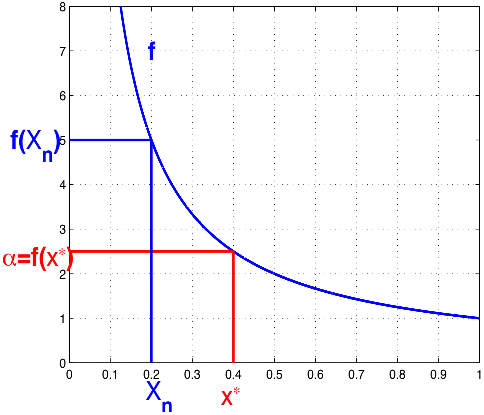
Stochastic approximation for $d = 1$.



Goal

$\longrightarrow$ Find the value x^* with very few knowledge on f.

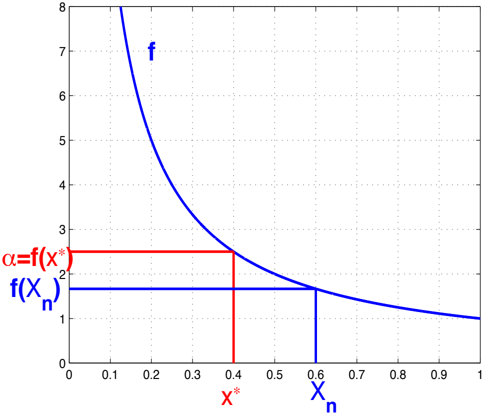
Stochastic approximation.



Basic Idea

If you are able to say that $f(X_n) > \alpha$, then increase the value of X_n.

Stochastic approximation.



Basic Idea

If you are able to say that $\boldsymbol{f(X_n)} < \alpha$, then decrease the value of $\boldsymbol{X_n}$.

Outline

The Robbins-Monro algorithm.

Let (γ_n) be a sequence of positive real numbers decreasing to zero

$$\sum_{n=1}^{\infty} \gamma_n = +\infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} \gamma_n^2 < +\infty.$$

For the sake of simplicity, we shall make use of

$$\gamma_n = \frac{1}{n}.$$

The Robbins-Monro algorithm

$$X_{n+1} = X_n + \gamma_n \Big(Y_{n+1} - \alpha \Big)$$

where the initial state X_0 is a square integrable random vector of $\mathbb{R}^d$ which can be arbitrarily chosen.

Almost sure convergence.

Let g be the positive function defined on $\mathbb{R}^d$ by

$$g(X_n) = \mathbb{E}[||Y_{n+1}||^2|\mathcal{F}_n].$$

Theorem (Robbins-Monro)

Assume that the function f is continuous from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that $f(x^*) = \alpha$, and for all x different from x^*,

$$\langle \boldsymbol{x} - \boldsymbol{x}^*, \boldsymbol{f}(\boldsymbol{x}) - \alpha \rangle < \mathbf{0}.$$

Assume that for $K > 0$ and for all $x \in \mathbb{R}^d$,

$$g(x) \leqslant K\big(1 + ||x||^2\big).$$

Then, we have the almost sure convergence

$$\lim_{n \to \infty} \boldsymbol{X}_n = \boldsymbol{x}^* \qquad \text{a.s.}$$

Proof of the almost sure convergence.

Proof.

First of all, denote

$$V_n = ||X_n - x^*||^2.$$

For all $n \geqslant 0$, we clearly have

$$\begin{aligned} V_{n+1} &= ||X_{n+1} - x^*||^2, \\ &= ||X_n + \gamma_n\big(Y_{n+1} - \alpha\big) - x^*||^2, \\ &= ||X_n - x^*||^2 + 2\gamma_n\langle X_n - x^*, Y_{n+1} - \alpha\rangle + \gamma_n^2||Y_{n+1} - \alpha||^2, \end{aligned}$$

which leads to

$$V_{n+1} = V_n + \gamma_n^2||Y_{n+1} - \alpha||^2 + 2\gamma_n\langle X_n - x^*, f(X_n) + \varepsilon_{n+1} - \alpha\rangle$$

where $\varepsilon_{n+1} = Y_{n+1} - f(X_n)$. □

Proof of the almost sure convergence, continued

Proof.

Since $\mathbb{E}[Y_{n+1}|\mathcal{F}_n]=f(X_n)$, $\mathbb{E}[\varepsilon_{n+1}|\mathcal{F}_n]=0$. It means that (ε_n) is a **martingale difference sequence.** Consequently,

$$\mathbb{E}[V_{n+1}|\mathcal{F}_n] = V_n + \gamma_n^2\mathbb{E}[||Y_{n+1} - \alpha||^2|\mathcal{F}_n] - B_n$$

where (B_n) is the positive sequence given by

$$B_n = -2\gamma_n\langle X_n - x^*, f(X_n) - \alpha\rangle$$

Moreover,

$$\mathbb{E}[||Y_{n+1}||^2|\mathcal{F}_n] \leqslant K(1 + ||X_n||^2) \leqslant L(1 + V_n)$$

where $L = 2K(1 + ||x^*||^2)$. □

Proof of the almost sure convergence, continued

Proof.

Therefore, we obtain that

$$\mathbb{E}[V_{n+1}|\mathcal{F}_n] \leqslant V_n(1 + a_n) + A_n - B_n$$

where $a_n = 2L\gamma_n^2$ and $A_n = 2(L + ||\alpha||^2)\gamma_n^2$. The assumption

$$\sum_{n=1}^{\infty} \gamma_n^2 < +\infty$$

clearly implies that

$$\sum_{n=1}^{\infty} a_n < +\infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} A_n < +\infty \qquad \text{a.s.}$$

□

Proof of the almost sure convergence, continued

Proof.

Hence, it follows from Robbins-Siegmund theorem that (V_n) converges a.s. to a random variable V_∞ and

$$\sum_{n=1}^{\infty} B_n < +\infty \qquad \text{a.s.}$$

It remains to prove that $V_\infty = 0$. Assume by contradiction that $V_\infty > 0$. Then, we can find two finite constants $0 < a < b$ such that, for n large enough, $a \leqslant ||X_n - x^*|| \leqslant b$. Denote by Δ the annulus of $\mathbb{R}^d$,

$$\Delta = \Big\{ x \in \mathbb{R}^d \ \text{ such that } \ a \leqslant ||x - x^*|| \leqslant b \Big\}.$$

Let F be the continuous negative function defined, for all $x \in \mathbb{R}^d$, by

$$F(x) = \langle x - x^*, f(x) - \alpha \rangle.$$

Proof.

One can find $c > 0$ such that, for all $x \in \Delta$,

$$F(x) \leqslant -c.$$

However, for n large enough $X_n \in \Delta$, which implies that $F(X_n) \leqslant -c$. Consequently, for n large enough,

$$B_n = -2\gamma_n F(X_n) \geqslant 2c\gamma_n.$$

Finally, the assumption

$$\sum_{n=1}^{\infty} \gamma_n = +\infty \qquad \Longrightarrow \qquad \sum_{n=1}^{\infty} B_n = +\infty$$

leading to a contradiction. It means that $V_\infty = 0$ so $X_n \to x^*$ a.s. □

Outline

Asymptotic normality.

The aymptotic normality requires more assumption on the function f. We now assume that f is twice differentiable. It follows from **Taylor's formula** that

$$f(x) = \alpha + H(x - x^*) + O\big(||x - x^*||^2\big)$$

where H is the **Jacobian matrix** of f at x^*. We also assume that H is an **Hurwitz matrix.** It means that the real parts of all the eigenvalues of H are negative. Let $\lambda_{max}(H)$ be the eigenvalue of H with the largest real part and denote

$$\rho = -\text{Re}(\lambda_{max}(H)).$$

In dimension $d = 1$, we have $f(x^*) = \alpha$, $H = f'(x^*)$ and $\rho = -f'(x^*)$.

Asymptotic normality, continued.

Let (ε_n) be the **martingale difference sequence** given by

$$\varepsilon_{n+1} = Y_{n+1} - \mathbb{E}[Y_{n+1}|\mathcal{F}_n] = Y_{n+1} - f(X_n).$$

Theorem (Robbins-Monro, continued)

Assume that the function f is twice differentiable from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that $f(x^*) = \alpha$. Suppose that f and g satisfy the same assumptions as in Robbins-Monro Theorem. Moreover, assume that

$$\lim_{n\to\infty} \mathbb{E}[\varepsilon_{n+1}\varepsilon_{n+1}^T|\mathcal{F}_n] = \Gamma \qquad a.s.$$

where Γ is a symmetric definite positive matrix and that (ε_n) has a conditional moment of order $> \mathbf{2}$. *If* $\boldsymbol{\rho > 1/2}$, *we have the aymptotic normality*

$$\sqrt{n}(X_n - x^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

Asymptotic normality for $d = 1$.

The limiting covariance matrix Σ is the unique solution of the **Lyapunov equation**

$$\left(H + \frac{1}{2} I_d\right)\Sigma + \Sigma\left(H^T + \frac{1}{2} I_d\right) = -\Gamma.$$

It is quite complicated to evaluate Σ. However, in the special case $d = 1$ and $\Gamma = \sigma^2$, we have $\rho = -H = -f'(x^*)$,

$$\Sigma = \frac{\sigma^2}{2\rho - 1}.$$

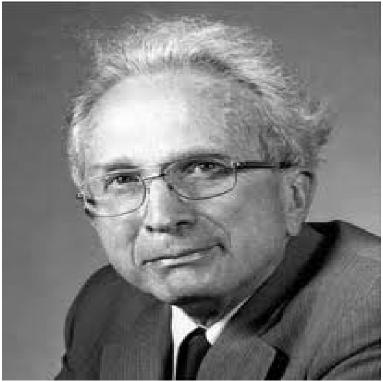
Consequently, as soon as $\boldsymbol{\rho > 1/2}$, we have

$$\boldsymbol{\sqrt{n}(X_n - x^*) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2}{2\rho - 1}\right).}$$

Outline

Stochastic approximation.

Jack Kiefer



Jacob Wolfowitz

Stochastic approximation.

Let f be an **unknown differentiable function** from $\mathbb{R}^d$ to $\mathbb{R}$.

Goal

$\longrightarrow$ Find a vector $\boldsymbol{x^*}$ of $\mathbb{R}^d$ which satisfies

$$\nabla f(x^*) = 0.$$

We will make use of the **directional derivative** of f at $x \in \mathbb{R}^d$ along the vector $y \in \mathbb{R}^d$, given by

$$\langle \nabla f(x), y \rangle = \lim_{t \to 0} \frac{f(x+ty) - f(x-ty)}{2t}.$$

Stochastic approximation.

We will assume in all the sequel that for all $n \geqslant 1$, we can compute $X_1, \ldots, X_n$ of $\mathbb{R}^d$ and we can find Y_{n+1} and Z_{n+1} of $\mathbb{R}^d$ such that

$$\mathbb{E}[Y_{n+1}|\mathcal{F}_n] = \begin{pmatrix} f(X_n + c_n e_1) \\ \vdots \\ f(X_n + c_n e_d) \end{pmatrix} = \Phi(X_n)$$

and

$$\mathbb{E}[Z_{n+1}|\mathcal{F}_n] = \begin{pmatrix} f(X_n - c_n e_1) \\ \vdots \\ f(X_n - c_n e_d) \end{pmatrix} = \Psi(X_n)$$

where $(e_1, \ldots, e_d)$ is the canonical basis of $\mathbb{R}^d$, (c_n) is a sequence of positive real numbers decreasing to zero, and $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. In dimension $d = 1$, $\mathbb{E}[Y_{n+1}|\mathcal{F}_n] = f(X_n + c_n)$, $\mathbb{E}[Z_{n+1}|\mathcal{F}_n] = f(X_n - c_n)$.

Outline

The Kiefer-Wolfowitz algorithm.

Let (γ_n) be a sequence of positive real numbers decreasing to zero

$$\sum_{n=1}^{\infty} \gamma_n = +\infty, \qquad \sum_{n=1}^{\infty} \left(\frac{\gamma_n}{c_n}\right)^2 < +\infty, \qquad \sum_{n=1}^{\infty} \gamma_n c_n < +\infty.$$

For the sake of simplicity, we can choose $0 < c < 1/2$,

$$\gamma_n = \frac{1}{n} \qquad \text{and} \qquad c_n = \frac{1}{n^c}.$$

The Kiefer-Wolfowitz algorithm

$$X_{n+1} = X_n + \frac{\gamma_n}{2c_n}\Big(Y_{n+1} - Z_{n+1}\Big)$$

where the initial state X_0 is a square integrable random vector of $\mathbb{R}^d$ which can be arbitrarily chosen.

Almost sure convergence.

Let g and h be the two positive functions defined on $\mathbb{R}^d$ by

$$g(X_n) = \mathbb{E}[||Y_{n+1}||^2|\mathcal{F}_n] \qquad \text{and} \qquad h(X_n) = \mathbb{E}[||Z_{n+1}||^2|\mathcal{F}_n].$$

Theorem (Kiefer-Wolfowitz)

Assume that the function f is twice continuously differentiable from $\mathbb{R}^d$ to $\mathbb{R}$ such that $\nabla f(x^*) = 0$, and for all x different from x^*,

$$\langle \boldsymbol{x} - \boldsymbol{x}^*, \nabla \boldsymbol{f}(\boldsymbol{x}) \rangle < \mathbf{0}.$$

Assume that for $L > 0$ and for all $x \in \mathbb{R}^d$,

$$||\nabla^2 f(x)|| \leqslant L\big(1 + ||x||\big).$$

Almost sure convergence, continued

Theorem (Kiefer-Wolfowitz, continued)

Moreover, assume that for $K_g > 0$, $K_h > 0$ *and for all* $x \in \mathbb{R}^d$,

$$g(x) \leqslant K_g\big(1 + ||x||^2\big) \qquad \text{and} \qquad h(x) \leqslant K_h\big(1 + ||x||^2\big).$$

Then, we have the almost sure convergence

$$\lim_{n \to \infty} X_n = x^* \qquad \text{a.s.}$$

Outline

Asymptotic normality.

The asymptotic normality requires more assumption on the function f. We now assume that $f \in \mathcal{C}^3(\mathbb{R}^d)$ with $\nabla f(x^*) = 0$. As $\nabla f \in \mathcal{C}^2(\mathbb{R}^d)$, it follows from **Taylor's formula** that

$$\nabla f(x) = H(x - x^*) + O\big(||x - x^*||^2\big)$$

where $H = \nabla^2 f(x^*)$ is the **Hessian matrix** of f at x^*. We also assume that H is a **negative definite matrix.** It means that all the eigenvalues of H are negative. Denote

$$\rho = -\lambda_{max}(H).$$

In dimension $d = 1$, we have $f'(x^*) = 0$, $H = f''(x^*)$ and $\rho = -f''(x^*)$.

Asymptotic normality, continued

Let (ε_n) and (ξ_n) be the **martingale difference sequences** given by

$$\varepsilon_{n+1} = Y_{n+1} - \mathbb{E}[Y_{n+1}|\mathcal{F}_n] = Y_{n+1} - \Phi(X_n),$$
$$\xi_{n+1} = Z_{n+1} - \mathbb{E}[Z_{n+1}|\mathcal{F}_n] = Z_{n+1} - \Psi(X_n).$$

Theorem (Kiefer-Wolfowitz, continued)

Assume that the function $f \in \mathcal{C}^3(\mathbb{R}^d)$ such that $\nabla f(x^*) = 0$. Suppose that f and g satisfy the same assumptions as in Kiefer-Wolfowitz Theorem. Moreover, assume that

$$\lim_{n\to\infty} \mathbb{E}[\varepsilon_{n+1}\varepsilon_{n+1}^T|\mathcal{F}_n] = \Gamma_g \qquad a.s.$$
$$\lim_{n\to\infty} \mathbb{E}[\xi_{n+1}\xi_{n+1}^T|\mathcal{F}_n] = \Gamma_h \qquad a.s.$$

where Γ_g and Γ_h are symmetric definite positive matrices and that (ε_n) and (ξ_n) have conditional moments of order $>$ **2.**

Asymptotic normality, continued

Theorem (Kiefer-Wolfowitz, continued)

If $\rho > 2c$ *where* $1/6 < c < 1/2$*, we have the aymptotic normality*

$$\sqrt{nc_n^2}(X_n - x^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

$\longrightarrow$ In the special case $\rho > 2c$ with $c = 1/6$, we also have

$$n^{1/3}(X_n - x^*) \xrightarrow{\mathcal{L}} \mathcal{N}(m, \Sigma)$$

where the mean m can be explicitely calculated.

Asymptotic normality for $d = 1$.

The limiting covariance matrix Σ is the unique solution of the **Lyapunov equation**

$$\Big(H + \Big(\frac{1}{2} - c\Big) I_d\Big)\Sigma + \Sigma\Big(H + \Big(\frac{1}{2} - c\Big) I_d\Big) = -\frac{1}{4}\Gamma.$$

It is quite complicated to evaluate Σ. However, in the special case $d = 1$ and $\Gamma = \sigma^2$, we have

$$\Sigma = \frac{\sigma^2}{8(\rho + c - 1/2)}.$$

Consequently, as soon as $\rho > 2c$ where $1/6 < c < 1/2$, we have

$$\sqrt{nc_n^2}(X_n - x^*) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \frac{\sigma^2}{8(\rho + c - 1/2)}\Big).$$

Outline

Stochastic averaging.

David Ruppert

Boris Polyak

Stochastic averaging.

The idea of using averaging to accelerate stochastic algorithms in due to Poliak and Ruppert. It consists in introducing a Cesaro mean over the iterations of the original stochastic algorithm

$$\overline{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

Goal

$\longrightarrow$ Improve the convergence properties of the stochastic algorithm by minimizing its asymptotic variance.
$\longrightarrow$ Substantially weaken the conditions of the previous asymptotic results on the stochastic algorithm.

Stochastic averaging on the Robbins-Monro algorithm.

Consider the slow down Robbins-Monro algorithm given by

$$X_{n+1} = X_n + \gamma_n \Big(Y_{n+1} - \alpha \Big)$$

where the initial state X_0 is a square integrable random vector of $\mathbb{R}^d$ which can be arbitrarily chosen and the step

$$\gamma_n = \frac{1}{n^c} \qquad \text{with} \qquad \frac{1}{2} < c < 1.$$

At time $n \geqslant 1$, compute de Cesaro mean

$$\overline{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

A second-order recursive equation.

We have

$$(n+1)\overline{X}_{n+1} = \sum_{k=1}^{n} X_k + X_{n+1} = n\overline{X}_n + X_{n+1},$$

which implies that

$$\boldsymbol{\overline{X}_{n+1} = \overline{X}_n + \frac{1}{n+1}\Big(X_{n+1} - \overline{X}_n\Big).}$$

However,

$$X_{n+1} = X_n + \gamma_n\Big(Y_{n+1} - \alpha\Big).$$

A second-order recursive equation.

Consequently, as $X_n = n\overline{X}_n - (n-1)\overline{X}_{n-1}$ we obtain that

$$\begin{aligned}
\overline{X}_{n+1} &= \overline{X}_n + \frac{1}{n+1}\Big(X_n - \overline{X}_n + \gamma_n\Big(Y_{n+1} - \alpha\Big)\Big), \\
&= \overline{X}_n + \frac{1}{n+1}\Big((n-1)\Big(\overline{X}_n - \overline{X}_{n-1}\Big) + \gamma_n\Big(Y_{n+1} - \alpha\Big)\Big), \\
&= \overline{X}_n + \Big(\frac{n-1}{n+1}\Big)\overline{X}_n - \Big(\frac{n-1}{n+1}\Big)\overline{X}_{n-1} + \frac{\gamma_n}{n+1}\Big(Y_{n+1} - \alpha\Big),
\end{aligned}$$

leading to **the second-order recursive equation**

$$\overline{X}_{n+1} = \Big(\frac{2n}{n+1}\Big)\overline{X}_n - \Big(\frac{n-1}{n+1}\Big)\overline{X}_{n-1} + \frac{\gamma_n}{n+1}\Big(Y_{n+1} - \alpha\Big).$$

A second-order recursive equation.

Consequently, as $X_n = n\overline{X}_n - (n-1)\overline{X}_{n-1}$ we obtain that

$$\begin{aligned}
\overline{X}_{n+1} &= \overline{X}_n + \frac{1}{n+1}\Big(X_n - \overline{X}_n + \gamma_n\big(Y_{n+1} - \alpha\big)\Big), \\
&= \overline{X}_n + \frac{1}{n+1}\Big((n-1)\big(\overline{X}_n - \overline{X}_{n-1}\big) + \gamma_n\big(Y_{n+1} - \alpha\big)\Big), \\
&= \overline{X}_n + \Big(\frac{n-1}{n+1}\Big)\overline{X}_n - \Big(\frac{n-1}{n+1}\Big)\overline{X}_{n-1} + \frac{\gamma_n}{n+1}\big(Y_{n+1} - \alpha\big),
\end{aligned}$$

leading to **the second-order recursive equation**

$$\overline{X}_{n+1} = \Big(\frac{2n}{n+1}\Big)\overline{X}_n - \Big(\frac{n-1}{n+1}\Big)\overline{X}_{n-1} + \frac{\gamma_n}{n+1}\big(Y_{n+1} - \alpha\big).$$

A second-order recursive equation.

Consequently, as $X_n = n\overline{X}_n - (n-1)\overline{X}_{n-1}$ we obtain that

$$\begin{aligned}
\overline{X}_{n+1} &= \overline{X}_n + \frac{1}{n+1}\Big(X_n - \overline{X}_n + \gamma_n\big(Y_{n+1} - \alpha\big)\Big), \\
&= \overline{X}_n + \frac{1}{n+1}\Big((n-1)\big(\overline{X}_n - \overline{X}_{n-1}\big) + \gamma_n\big(Y_{n+1} - \alpha\big)\Big), \\
&= \overline{X}_n + \Big(\frac{n-1}{n+1}\Big)\overline{X}_n - \Big(\frac{n-1}{n+1}\Big)\overline{X}_{n-1} + \frac{\gamma_n}{n+1}\big(Y_{n+1} - \alpha\big),
\end{aligned}$$

leading to **the second-order recursive equation**

$$\overline{X}_{n+1} = \Big(\frac{2n}{n+1}\Big)\overline{X}_n - \Big(\frac{n-1}{n+1}\Big)\overline{X}_{n-1} + \frac{\gamma_n}{n+1}\big(Y_{n+1} - \alpha\big).$$

Outline

Almost sure convergence.

Theorem (Robbins-Monro averaging)

Assume that the function f is continuous from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that $f(x^*) = \alpha$, and for all x different from x^*,

$$\langle \boldsymbol{x} - \boldsymbol{x}^*, \boldsymbol{f}(\boldsymbol{x}) - \alpha \rangle < \mathbf{0}.$$

Assume that for $K > 0$ and for all $x \in \mathbb{R}^d$,

$$g(x) \leqslant K\left(1 + ||x||^2\right).$$

Then, we have the almost sure convergence

$$\lim_{n\to\infty} \overline{X}_n = x^* \qquad \text{a.s.}$$

Outline

Asymptotic normality.

Theorem (Robbins-Monro averaging)

Assume that the function f is twice differentiable from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that $f(x^*) = \alpha$. Suppose that f and g satisfy the same assumptions as in Robbins-Monro Theorem. Moreover, assume that

$$\lim_{n\to\infty} \mathbb{E}[\varepsilon_{n+1}\varepsilon_{n+1}^T | \mathcal{F}_n] = \Gamma \qquad a.s.$$

*where Γ is a symmetric definite positive matrix and that (ε_n) has a conditional moment of order $>$ **2.** Then, we have the asymptotic normality*

$$\sqrt{n}(\overline{X}_n - x^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

where the asymptotic matrix Σ is given by

$$\Sigma = H^{-1}\Gamma(H^{-1})^T.$$

Asymptotic normality for $d = 1$.

It is not necessary to assume that

$$\boldsymbol{\rho = -\mathrm{Re}(\lambda_{max}(H)) > \frac{1}{2}}.$$

In the special case $d = 1$ and $\Gamma = \sigma^2$, we have $\rho = -H = -f'(x^*)$, which means that

$$\Sigma = \frac{\sigma^2}{\big(f'(x^*)\big)^2}.$$

Consequently, the asymptotic normality reduces to

$$\boldsymbol{\sqrt{n}(\overline{X}_n - x^*) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \frac{\sigma^2}{\big(f'(x^*)\big)^2}\Big)}.$$