Stochastic Approximation from Finance to Data Science

Gilles Pagès

LPSM-Sorbonne-Université

(Labo. Proba., Stat. et Modélisation)

M2 Probabilités & Finance

Novembre 2020

Table of Contents

Table of Contents

Abstract Learning

- Huge dataset $(z_k)_{k=1:N}$ with of possibly high dimension d: $N \simeq 10^6$, even 10^9, and $d \simeq 10^3$.
  [Image, profile, text, ...]

- Set of parameters $\theta \in \Theta \subset \mathbb{R}^K$, K large (see later on).

- There exists a smooth local loss function/local predictor

$$v(\theta, z).$$

- Global loss function: $V(\theta) = \dfrac{1}{N}\displaystyle\sum_{k=1}^{N} v(\theta, z_k)$

  with gradient $\nabla V(\theta) = \dfrac{1}{N}\displaystyle\sum_{k=1}^{N} \nabla_\theta v(\theta, z_k)$.

- Solving the minimization problem

$$\min_{\theta \in \Theta} V(\theta).$$

- Suggests a (GD) i.e. $h = \nabla V$ [or others... if $\nabla^2_\theta v(\theta, z)$ exists]:

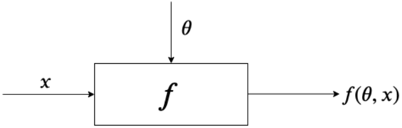
$$\begin{aligned}\theta_{n+1} &= \theta_n - \gamma_{n+1} \nabla V(\theta_n) \\ &= \theta_n - \frac{\gamma_{n+1}}{N} \sum_{k=1}^{N} \nabla_\theta v(\theta, z_k),\ n \geq 0,\end{aligned}$$

with the step sequence satisfying the (DS) assumption.

Table of Contents

Supervised learning

- Input x_k, output y_k. Data $z_k = (x_k, y_k) \in \mathbb{R}^{d_x+d_y}$, $k = 1 : N$.
- Transfer function $f : \Theta \times \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$



- Prediction/loss function (local) $v(\theta, z) = \frac{1}{2}\big|f(\theta, x_k) - y_k\big|^2$, $k = 1 : N$ so that

$$\nabla_\theta v(\theta, z) = \nabla_\theta f(\theta, x)^\top \big(f(\theta, x) - y\big).$$

- Resulting loss function gradient

$$\nabla V(\theta) = \frac{1}{N}\sum_{k=1}^{N} \nabla_\theta f(\theta, x_k)^\top \big(f(\theta, x_k) - y_k\big).$$

Table of Contents

Unsupervised learning (clustering)

- Only input $z_k = x_k \in \mathbb{R}^d$, $k = 1 : N$.
- Prototype parameter set: $\theta := (\theta^1, \ldots, \theta^r) \in \Theta = (\mathbb{R}^d)^r$, $r \in \mathbb{N}$.
- (An example of) Local loss function: nearest neighbor among the prototypes: $x \in \mathbb{R}^d$, $\theta \in \Theta$.

$$v(\theta, x) = \tfrac{1}{2} \min_{i=1:r} |\theta^i - x|^2 = \tfrac{1}{2} \mathrm{dist}\big(x, \{\theta^1, \ldots, \theta^r\}\big)^2$$

  (minimal distance to prototypes).
- $v(\theta, x)$ is not convex in θ!
- Global loss function (Distortion):

$$V(\theta) = \tfrac{1}{2N} \sum_{k=1}^{N} \min_{i=1:r} |\theta^i - x_k|^2 \quad \text{(mean minimal distance to prototypes).}$$

- Searching for the best prototypes: $\min_{\theta \in (\mathbb{R}^d)^r} V(\theta)$

Batch k-means/Forgy's algorithm

- Gradient at θ s.t. $\theta^i \neq \theta^j$: $\quad \nabla V(\theta) = \frac{1}{2N}\sum_{k=1}^{N} \nabla_\theta v(\theta, x_k)$

  with,

$$\forall\, i = 1:r, \quad \partial_{\theta^i} v(\theta, x_k) = \big(\theta^i - x_k\big)\mathbf{1}_{\{|x_k-\theta^i|<\min_{j\neq i}|x_k-\theta^j|\}} \in \mathbb{R}^d.$$

- Compute the vector of $(\mathbb{R}^d)^r$: $\mathbf{1}_{\{|x_k-\theta^i|<\min_{j\neq i}|x_k-\theta^j|\}}$ = nearest neighbour search.
- Compute $\nabla V(\theta) = \frac{1}{2N}\sum_{k=1}^{N} \nabla_\theta v(\theta, x_k)$.
- $\Longrightarrow N\times$ nearest neighbour searches among r prototypes of dim d!
- Forgy's algorithm = GD algorithm (or batch GD algorithm):

$$\theta_{n+1} = \theta_n - \gamma_{n+1}\nabla V(\theta_n).$$

Table of Contents

Pros and cons: toward stochastic algorithm I

- Numerical Probability (for Finance): we do not know how to compute $h(\theta)$.
  - h always has a probabilistic presentation in our examples:

$$h(\theta) = \mathbb{E}\, H(\theta, Z) = \int_{\mathbb{R}^q} H(\theta, z)\mathbb{P}_Z(dz) = \int_{\mathbb{R}^q} H(\theta, z) f_Z(z) dz$$

    where $H : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^d$ is Borel and q often large...
  - Cons: ...which requires the computation of (often) high dimensional integrals on $\mathbb{R}^q$ at a reasonable computational cost (complexity): impossible if $q \geq 3$!
  - Pros: The random vector Z can be simulated.
  - Pros: The function H is computable at a reasonable cost.
  - Pros: Regularizing effect of $\mathbb{E}$: h smoother than the functions $H(., z)$. (Think to $F_X(\xi) = \mathbb{E}\,\mathbf{1}_{\{X \leq \xi\}}$.)

Pros and Cons: toward stochastic algorithm I

- Data Science (usually V is given as well as $h = \nabla V$): but we cannot compute $h(\theta)$.
  - h still has probabilistic representation using the empirical measure.

$$h(\theta) = \frac{1}{N}\sum_{k=1}^{N} \nabla_\theta v(\theta, z_k) = \int_{\mathbb{R}^q} \nabla_\theta v(\theta, z_k)\mu_{_N}(dz) \text{ with } \mu_{_N} = \frac{1}{N}\sum_{k=1}^{N}\delta_{z_k}$$

  - Cons: But N huge $\Longrightarrow h(\theta)$ cannot be computed at a reasonable cost.
  - Pros:

$$h(\theta) = \mathbb{E}\left[\nabla_\theta v(\theta, Z)\right]$$

    where Z can be simulated by picking up a datum (uniformly) at random since

$$Z \sim z_I, \quad I \sim \mathcal{U}\big(\{1,\ldots,N\}\big).$$

  - $v(\theta, z)$ and $\nabla_\theta v(\theta, z)$ both computable hence V and $h = \nabla V$ too.
  - Cons: No regularizing effect of $\mathbb{E}$: smoothness of $[h = \nabla V]$= smoothness of $H(.,z)$.
  - Cons: Transfer of convexity in θ from $v(\cdot, z)$ to V.

Toward stochastic algorithm II

- Zero search of $h(\theta) = \mathbb{E}\, H(\theta, Z)$ as above.
- Idea 1: Use Monte Carlo simulation

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \widehat{h}_{M_{n+1}}(\theta_n)$$

$$\widehat{h}_{M_{n+1}}(\theta_n) = \frac{1}{M_n} \sum_{k=1}^{M_n} H(\theta_n, Z_k^{(n+1)}), \quad \big(Z_k^{(n+1)}\big)_{k,n}\ i.i.d. \sim Z$$

- Idea 2: Robbins-Monro, 1951 ([3]). Set

$$\forall\, n \geq 1, \quad M_n = 1\ !!$$

- Idea 1.5: Mini-batch i.e. $M_n = M > 2$. "Recently" became successful among practitioners.

[3] H. Robbins, S. Monro (1951). A stochastic approximation method, *Ann. Math. Stat.*, **22**:400–407.

Table of Contents

Robbins-Monro framework (1951)

▷ Pactitioner's corner: Replace $h(\theta_n)$ by a $H(\theta_n, Z_{n+1})$.

- Let $(\theta_n)_{n\geq 0}$ be a sequence of $\mathbb{R}^d$-valued random vectors recursively defined on $(\Omega, \mathcal{A}, \mathbb{P})$ by

$$\theta_{n+1} = \theta_n - \gamma_{n+1} H(\theta_n, Z_{n+1}), \quad \theta_0 \in L^2(\mathbb{P}, \mathcal{A})$$

with

$$\begin{aligned} &(i)\ (Z_n)_{n\geq 1} \text{ is i.i.d.} \sim Z, \text{ independent of } \theta_0 \\ &(ii)\ \big\|H(\theta, Z)\big\|_2 \leq C(1+|\theta|) \qquad (\Rightarrow h \text{ linear growth}) \\ &(iii)\ (\gamma_n)_{n\geq 1} \text{ is a } (0,+\infty)\text{-valued deterministic step sequence} \end{aligned}$$

so that $(\theta_n)_{n\geq 0}$ is $(\mathcal{F}_n)$-adapted with $\mathcal{F}_n = \sigma(\theta_0, Z_1, \ldots, Z_n)$. Then

$$\theta_{n+1} = \theta_n - \gamma_{n+1} h(\theta_n) - \gamma_{n+1} \Delta M_{n+1}, \ n \geq 0 \ldots$$

- where $\gamma_{n+1}\Delta M_{n+1}$ is a perturbation of the deterministic procedure with

$$\Delta M_{n+1} = H(\theta_n, Z_{n+1}) - h(\theta_n).$$

Interpretations

- Idea 1 (Robbins-Monro 1951, Robbins-Siegmund 1971):
  - Perturbed zero search procedure with decreasing step for h.
  - The perturbation is a martingale increment since

$$\mathbb{E}\big(H(\theta_n, Z_{n+1})\,|\,\mathcal{F}_n\big) \underbrace{=}_{Z_{n+1}\perp\!\!\!\perp\mathcal{F}_n} \Big[\mathbb{E}\,H(\theta, Z_{n+1})\Big]_{|\theta=\theta_n} = h(\theta_n),\ n \geq 0.$$

- Idea 2 (Ljung, 1977): Perturbed Euler scheme with decreasing step of the ODE

$$\dot{\theta} = -h(\theta).$$

(nice theory in connection with perturbed dynamical systems but no time time be exploited here).

Suggests to use tools from ODE theory like ... Lyapunov functions.

Idea 1: Robbins-Siegmund Lemma, 1971

Theorem (Robbins-Siegmund Lemma, 1971)

- *Lyapunov function:* $V : \mathbb{R}^d \to \mathbb{R}_+, \mathcal{C}^1$, $\lim_{\infty} V = +\infty$, ∇V *Lipschitz*, $|\nabla V|^2 \le c(1+V)$ *and*
- *Mean-reversion:* $(\nabla V | h) \ge 0$.
- *L^2-Growth control:* $\|H(\theta, Z)\|_2 \le C\sqrt{1+V(\theta)}$.
- *Decreasing Step assumption (DS):* $\sum_n \gamma_n = +\infty$ *and* $\sum_n \gamma_n^2 < +\infty$.

Then, if $V(\theta_0) \in L^1$,

(*i*) $V(\theta_n) \stackrel{a.s.}{\to} V_\infty \in L^1$ [$\Rightarrow (\theta_n)_{n\ge 0}$ *pathwise bounded*] *and* L^1-*bounded*.

(*ii*) $\sum_n \gamma_n (\nabla V | h)(\theta_{n-1}) \in L^1$ *a.s.* , *hence* $< +\infty$ *a.s.*

(*iii*) $\sum_n |\Delta\theta_n|^2 < +\infty$ *a.s.* (*so that* $\theta_n - \theta_{n-1} \to 0$ *a.s.*).

(*iv*) $\sum_n \gamma_n \Delta M_n$ *converges a.s. and in* L^2.

- Note $|\nabla V|^2 \le c(1+V) \Rightarrow V$ sub-quadratic: $V(\theta) \le \kappa(1+|\theta|^2)$ and h sublinear.

Proof.

- Set $\mathcal{F}_n := \sigma(\theta_0, Z_1, \ldots, Z_n)$, $n \geq 1$ and $\Delta\theta_n := \theta_n - \theta_{n-1}$, $n \geq 1$.
- There exists $\xi_{n+1} \in (\theta_n, \theta_{n+1})$ s.t.

$$\begin{aligned}
V(\theta_{n+1}) &= V(\theta_n) + (\nabla V(\xi_{n+1})|\Delta\theta_{n+1}) \\
(\star) \quad &\leq V(\theta_n) + (\nabla V(\theta_n)|\Delta\theta_{n+1}) + [\nabla V]_{\rm Lip}|\Delta\theta_{n+1}|^2 \\
(\star\star) \quad &= V(\theta_n) - \gamma_{n+1}(\nabla V(\theta_n)|H(\theta_n, Z_{n+1})) \\
&\qquad + [\nabla V]_{\rm Lip}\gamma_{n+1}^2|H(\theta_n, Z_{n+1})|^2 \\
(\star\star\star) \quad &= V(\theta_n) - \gamma_{n+1}(\nabla V(\theta_n)|h(\theta_n)) - \gamma_{n+1}(\nabla V(\theta_n)|\Delta M_{n+1}) \\
&\qquad + [\nabla V]_{\rm Lip}\gamma_{n+1}^2|H(\theta_n, Z_{n+1})|^2,
\end{aligned}$$

where

$$\Delta M_{n+1} = H(\theta_n, Z_{n+1}) - h(\theta_n).$$

and where we used in $(\star)$:

$|\nabla V(\xi_{n+1}) - \nabla V(\theta_n)| \leq [\nabla V]_{\rm Lip}|\xi_{n+1} - \theta_n| \leq [\nabla V]_{\rm Lip}|\theta_{n+1} - \theta_n|$.

- Show by induction that $V(\theta_n)\in L^1(\mathbb{P})$, given that $V(\theta_0)\in L^1(\mathbb{P})$ via $(\star\star)$.
- Technical key: $|(a|b)|\le |a|\cdot|b|\le \frac{1}{2}(|a|^2+|b|^2)$ so that

$$\mathbb{E}\left|\big(\nabla V(\theta_n)|H(\theta_n,Z_{n+1})\big)\right|\le \frac{1}{2}\big(\mathbb{E}\,|\nabla V(\theta_n)|^2+\mathbb{E}\,|H(\theta_n,Z_{n+1})|^2\big).$$

and, still using that $Z_{n+1}\perp\!\!\!\perp \mathcal{F}_n$ and $\theta_n \underline{\in} \mathcal{F}_n$,

$$\mathbb{E}\,|H(\theta_n,Z_{n+1})|^2=\mathbb{E}\big[\mathbb{E}\,(|H(\theta_n,Z_{n+1})|^2|\mathcal{F}_n)\big]\le C\big(1+\mathbb{E}\,V(\theta_n)\big).$$

- So that $(\Delta M_n)_{n\ge1}$ is a sequence of L^2 $(\mathcal{F}_n)$-martingale increments satisfying

$$\begin{aligned}\mathbb{E}\,(|\Delta M_{n+1}|^2\,|\,\mathcal{F}_n)&=\mathbb{E}\,(|H(\theta_n,Z_{n+1})|^2\,|\,\mathcal{F}_n)-|h(\theta_n|)^2\\&\le \mathbb{E}\,(|H(\theta_n,Z_{n+1})|^2\,|\,\mathcal{F}_n)\le C(1+V(\theta_n)).\end{aligned}$$

- Hence, $(\nabla V(\theta_n)|\Delta M_{n+1})$ (is a true martingale increment and) satisfies

$$\|(\nabla V(\theta_n)|\Delta M_{n+1})\|_1\le \|\nabla V(\theta_n)\|_2\|\Delta M_{n+1}\|_2\le C\big(1+\mathbb{E}\,V(\theta_n)\big).$$

- Let us come back to $(\star\star\star)$:

$$V(\theta_{n+1}) \leq V(\theta_n) - \gamma_{n+1}(\nabla V|h)(\theta_n) - \gamma_{n+1}(\nabla V(\theta_n)|\Delta M_{n+1}) \\ + [\nabla V]_{\mathrm{Lip}}\gamma_{n+1}^2|H(\theta_n, Z_{n+1})|^2.$$

- Conditioning with respect to $\mathcal{F}_n$ yields, as $\nabla V(\theta_n)$ is $\mathcal{F}_n$-measurable,
  - $\mathbb{E}\big[(\nabla V(\theta_n)|\Delta M_{n+1})\,|\,\mathcal{F}_n\big] = \big(\nabla V(\theta_n)|\mathbb{E}\big[\Delta M_{n+1}\,|\,\mathcal{F}_n\big]\big) = 0$
  - and, other terms in the RHS being (also) $\mathcal{F}_n$-measurable,

$$\mathbb{E}\left(V(\theta_{n+1})\,|\,\mathcal{F}_n\right) + \gamma_{n+1}(\nabla V|h)(\theta_n) \leq V(\theta_n) + C_V\gamma_{n+1}^2\big(1 + V(\theta_n)\big) \\ = V(\theta_n)(1 + C_V\gamma_{n+1}^2) + C_V\gamma_{n+1}^2$$

with $C_V = C^2[\nabla V]_{\mathrm{Lip}} > 0$.

- Add
  - the [positive term] $= \sum_{k=1}^{n}\gamma_k \underbrace{(\nabla V|h)(\theta_{k-1})}_{\geq 0 \text{ by (mean-reversion)}} + C_V\sum_{k\geq n+2}\gamma_k^2$ on the left-hand side of the above inequality,
  - $(1 + C_V\gamma_{n+1}^2)\times$[this positive term] on the right-hand side .

- Divide the resulting inequality by $\prod_{k=1}^{n+1}(1+C_V\gamma_k^2)$ shows that (the $\mathcal{F}_n$-adapted sequence)

$$S_n = \frac{V(\theta_n) + \sum_{k=0}^{n-1}\gamma_{k+1}(\nabla V|h)(\theta_k) + C_V\sum_{k\ge n+1}\gamma_k^2}{\prod_{k=1}^{n}(1+C_V\gamma_k^2)},\ n\ge 0,$$

is a non-negative super-martingale with $S_0 = V(\theta_0) \in L^1(\mathbb{P})$.

- Hence

$$S_n \xrightarrow{a.s.} S_\infty \in L^1_{\mathbb{R}_+}(\mathbb{P}).$$

- Consequently, using that $\sum_{k\geq n+1}\gamma_k^2 \to 0$ (by (DS)), we get

$$V(\theta_n) + \sum_{k=0}^{n-1}\gamma_{k+1}(\nabla V|h)(\theta_k) \xrightarrow{a.s.} \widetilde{S}_\infty = S_\infty \prod_{n\geq 1}(1 + C_V\gamma_n^2) \in L^1(\mathbb{P}).$$

- $(i)_a$ The non-negative super-martingale

$$(S_n)_{n\geq 0} \;\text{ is } L^1(\mathbb{P})\text{-bounded by }\; \mathbb{E}\, S_0 = \mathbb{E}\, V(\theta_0) < +\infty,$$

hence $(V(\theta_n))_{n\geq 0}$ is L^1-bounded since

$$V(\theta_n) \leq \left(\prod_{k=1}^{n}(1 + C_V\gamma_k^2)\right) S_n, \quad n \geq 0,$$

and $\prod_{k\geq 1}(1 + C_V\gamma_k^2) < +\infty$ by the (DS) assumption on $(\gamma_n)_{n\geq 1}$.

- *(ii)* Now, for the same reason, the series with non-negative terms $\sum_{0\le k\le n-1}\gamma_{k+1}(\nabla V|h)(\theta_k)$ satisfies for every $n\ge 1$,

$$\mathbb{E}\left(\sum_{k=0}^{n-1}\gamma_{k+1}(\nabla V|h)(\theta_k)\right)\le\prod_{k=1}^{n}(1+C_V\gamma_k^2)\mathbb{E}\,S_0$$

so that, by the Beppo Levi monotone convergence Theorem for series with non-negative terms,

$$\mathbb{E}\left(\sum_{n\ge 0}\gamma_{n+1}(\nabla V|h)(\theta_n)\right)<+\infty$$

so that, in particular,

$$\sum_{n\ge 0}\gamma_{n+1}(\nabla V|h)(\theta_n)<+\infty\qquad \mathbb{P}\text{-}a.s.$$

and the series converges in L^1 to its *a.s.* limit.

- $(i)_b$ It follows that $V(\theta_n) \longrightarrow V_\infty$ a.s. as $n \to +\infty$. $V_\infty \in L^1$ by Fatou's Lemma since $\big(V(\theta_n)\big)_{n\geq 0}$ is L^1-bounded.

- (iii) Again by Beppo Levi's monotone convergence Theorem for series with non-negative terms,

$$\mathbb{E}\left(\sum_{n\geq 1}|\Delta\theta_n|^2\right) = \sum_{n\geq 1}\mathbb{E}\,|\Delta\theta_n|^2 \leq \sum_{n\geq 1}\gamma_n^2\,\mathbb{E}\,|H(\theta_{n-1},Z_n)|^2$$
$$\leq C\sum_{n\geq 1}\gamma_n^2(1+\mathbb{E}\,V(\theta_{n-1})) < +\infty$$

so that

$$\sum_{n\geq 1}|\Delta\theta_n|^2 \in L^1(\mathbb{P}) \quad \text{(hence } as. \text{ finite)}$$

which in turns yields

$$\Delta\theta_n = \theta_n - \theta_{n-1} \to 0 \qquad a.s. \text{ and in } \; L^2(\mathbb{P})$$

- (iv) Set $M_n^\gamma = \sum_{k=1}^n \gamma_k \Delta M_k$. M^γ is clearly an $(\mathcal{F}_n)$-martingale. Moreover,

$$\langle M^\gamma\rangle_n = \sum_{k=1}^n \gamma_k^2\, \mathbb{E}\left(|\Delta M_k|^2\,|\mathcal{F}_{k-1}\right) \le \sum_{k=1}^n \gamma_k^2\, \mathbb{E}\left(|H(\theta_{k-1}, Z_k)|^2\,|\mathcal{F}_{k-1}\right)$$
$$\le C\sum_{k=1}^n \gamma_k^2\big(1+V(\theta_{k-1})\big)$$

so that, owing to $(i)_a$,

$$\mathbb{E}\,\langle M^\gamma\rangle_\infty \le C\sum_{n\ge 1}\gamma_n^2\big(1+\mathbb{E}\,V(\theta_{n-1})\big) \le C'\sum_{n\ge1}\gamma_n^2 < +\infty.$$

Hence M_n^γ converges *a.s.* and in L^2.

- Which completes the proof ! $\square$

Table of Contents

Robbins-Monro (pathwise reasoning): zero search algorithm

Theorem (Robbins-Monto algorithm)

Assume the mean function h is continuous and satisfies

$$\forall\,\theta\in\mathbb{R}^d,\ \theta\neq\theta_*,\quad \big(\theta-\theta_*|h(\theta)\big)>0.$$

Suppose furthermore that $\theta_0\in L^2$ and that H satisfies

$$\forall\,\theta\in\mathbb{R}^d,\quad \big\|H(\theta,Z)\big\|_2\leq C(1+|\theta|).$$

Finally, assume $(\gamma_n)_{n\geq1}$ satisfies (DS). Then

$$\{h=0\}=\{\theta_*\}\quad \text{and}\quad \theta_n\xrightarrow{a.s.}\theta_*.$$

The convergence also holds in every L^p, $p\in(0,2)$ (and $\big(|\theta_n-\theta_*|\big)_{n\geq0}$ is L^2-bounded).

Proof

- Let $\theta_\lambda = \theta^* - \lambda h(\theta^*)$, $\lambda > 0$. If $h(\theta^*) \neq 0$, then $-(h(\theta_\lambda)|h(\theta^*) > 0$. Letting $\lambda \to 0$ implies $-|h(\theta^*)|^2 \geq 0$ hence $h(\theta^*) = 0$ so that, clearly, $\{h = 0\} = \{\theta^*\}$.
- The function $V(\theta) = \frac{1}{2}|\theta - \theta_*|^2$ is a Lyapunov function since $\nabla V = \theta - \theta^*$.
- The quadratic linear growth assumption on H is clearly satisfied too.
- Robbins-Siegmund's Lemma implies

  1. $|\theta_n - \theta_*|^2 \longrightarrow 2\, V_\infty \in L^1 \quad \mathbb{P}\text{-}a.s.$,
  2. $\displaystyle\sum_{n\geq 1} \gamma_n \big(h(\theta_{n-1})|\theta_{n-1} - \theta_*\big) < +\infty \quad \mathbb{P}\text{-}a.s.$
  3. $(|\theta_n - \theta_*|^2)_{n\geq 0}$ is L^1-bounded.

- Now we keep on reasoning pathwise: let ω be generic (in the sense it satisfies both above *a.s.* properties 1. & 2.).

- On has combining (DS) $\sum_n \gamma_n = +\infty$ and the above 2.,
$$\underline{\lim}_n \big(\theta_{n-1}(\omega) - \theta_* | h(\theta_{n-1}(\omega))\big) = 0.$$
- In fact if $\underline{\lim}_n \big(\theta_{n-1}(\omega) - \theta_* | h(\theta_{n-1}(\omega))\big) > 0$, the above convergence induces a contradiction with $\displaystyle\sum_{n\geq 1} \gamma_n = +\infty$.
- Let $\big(\phi(n,\omega)\big)_{n\geq 1}$ be a subsequence such that
$$\big(\theta_{\phi(n,\omega)}(\omega) - \theta_* | h(\theta_{\phi(n,\omega)}(\omega))\big) \longrightarrow 0 \quad \text{as} \quad n \to +\infty.$$
- Now, $(\theta_n(\omega))_{n\geq 0}$ being bounded, one may assume, up to one further extraction $\phi \circ \psi(n,\omega)$ but still denoted $\phi(n,\omega)$ for convenience,
$$\theta_{\phi(n,\omega)}(\omega) \to \theta_\infty = \theta_\infty(\omega).$$
- By continuity of h, $\big(\theta_\infty - \theta_* | h(\theta_\infty)\big) = 0$ which implies $\theta_\infty = \theta_*$. Now, since we know that $V(\theta_n(\omega)) = \frac{1}{2}|\theta_n(\omega) - \theta_*|^2$ converges,
$$\lim_n \big|\theta_n(\omega) - \theta_*\big|^2 = \lim_n \big|\theta_{\phi(n,\omega)}(\omega) - \theta_*\big|^2 = 0.$$
- Convergence in L^p, $p \in (0,2)$ follows by uniform integrability. $\square$

Theorem (Stochastic Gradient Descent)

◀ *Let* $V : \mathbb{R}^d \to \mathbb{R}_+$ *be a differentiable function* $\lim_{\infty} V(\theta) = +\infty$, ∇V *Lipschitz*, $|\nabla V|^2 \le C(1+V)$ *and* $\{\nabla V = 0\} = \{\theta_*\}$.

◀ *Let* $h(\theta) = \mathbb{E}\, H(\theta, Z) = \nabla V(\theta)$ *with* H *s.t.* $\|H(\theta,Z)\|_2 \le C\sqrt{1+V(\theta)}$ *and that* $V(\theta_0) \in L^1(\mathbb{P})$. *Assume* $(\gamma_n)_{n\ge1}$ *satisfies (DS).*

Then

$$V(\theta_*) = \min_{\mathbb{R}^d} V \quad \text{and} \quad \theta_n \xrightarrow{a.s.} \theta_* \quad \text{as} \quad n \to +\infty.$$

Moreover, $\nabla V(\theta_n)$ *converges to* 0 *in every* L^p, $p \in (0,2)$ *(and* $\big(V(\theta_n)\big)_{n\ge0}$ *is* L^1*-bounded so that* $\big(\nabla V(\theta_n)\big)_{n\ge0}$ *is* L^2*-bounded).*

- *Proof.* Use (almost) the same arguments as above but with $(\nabla V \mid h)(\theta) = |\nabla V(\theta)|^2 > 0$, $\theta \neq \theta^*$, instead of $(\theta - \theta_* \mid h(\theta))$. Thus, for a fixed generic scenario ω, there exists a limiting value θ_∞ such that $|\nabla V(\theta_\infty)|^2 = 0$ so that $\theta_\infty = \theta^*$ and $\lim_n V(\theta_n) = \lim_n V(\theta_{\phi(n)} = V(\theta^*) = \min_{\mathbb{R}^d} V$. Hence $\theta_n \to \theta^*$ (details left as an exercise). □

- **Remark**. If $H(\theta, z) = h(\theta) = \nabla V(\theta)$: Convergence thm for Gradient descent (GD)!!

Multi-target stochastic algorithms

Theorem (Multitarget Stochastic Gradient Descent (Fort-P., Benaïm, $\simeq$ 1990)

(a) *If the former assumption* $\{\nabla V = 0\} = \{\theta_*\}$ *IS NOT SATISFIED i.e.* $\mathrm{card}(\{\nabla V = 0\}) \geq 2$, *one has mutatis mutandis: a.s. there exists* $v_\infty \in \mathbb{R}_+$ *and a connected component* χ_∞ *of* $\big\{\nabla V = 0\big\} \cap \big\{V = v_\infty\big\}$ *such that*

$$\mathrm{dist}(\theta_n, \chi_\infty) \longrightarrow 0 \quad a.s.$$

(b) *In particular if* $\{\nabla V = 0\} \cap \{V = v\}$ *is locally finite for every* $v \geq 0$ *is finite, then there exists a r.v.* θ_∞ *such that*

$$\nabla V(\theta_\infty) = 0 \quad \text{and} \quad \theta_n \longrightarrow \theta_\infty.$$

(c) *Moreover,* $\nabla V(\theta_n)$ *converges to* 0 *in every* L^p, $p \in (0, 2)$ *(and* $\big(V(\theta_n)\big)_{n\geq 0}$ *is* L^1*-bounded so that* $\big(\nabla V(\theta_n)\big)_{n\geq 0}$ *is* L^2*-bounded).*

- By the R.-S. Lemma, one has for free that, a.s., $(\theta_n)_{n\geq 0}$ is pathwise bounded and $\theta_n - \theta_{n-1} \to 0$ pathwise. Hence its limiting values makes up a connected compact set Θ_∞, clearly included in some $\{V = v_\infty\}$.
- But this is not enough... Θ_∞ is also invariant under the flow of

$$ODE \equiv \dot{\theta} = -\nabla V(\theta)$$

  which converges toward $\{\nabla V = 0\}$.
- Still not enough : needs to make a transfer from *ODE* to algorithm.
- Needs further insights based on topology and the *ODE* method ([4]).

[4] G. Pagès (2018). *Introduction to Numerical Probability with application to Finance, Springer-Verlag, Berlin, 576p.*

Avoiding traps

Theorem (Traps (Pemantle 1984, Lazarev 1989, Brandière-Duflo 1996, Fort-P. 1997, Benaïm 1998))

Assume $\nabla V(\theta) = \mathbb{E}\, H(\theta, Z)$*[, etc].*

Let $\theta_* \in \{\nabla V = 0\}$*. If there exists a negative eigen-(value,vector)* (λ, u) *such that* $D^2 V(\theta^*)u = \lambda u$ *such that*

$$\lambda < 0 \quad \textit{and} \quad \mathbb{E}\,\big(H(\theta_*, Z)|u\big)^2 > 0$$

i.e. θ^* *is a noisy trap then*

$$\mathbb{P}\big(\theta_n \to \theta_*\big) = 0.$$

- This allows to eliminate noisy local maxima, saddle points, monkey saddle points, etc.

Table of Contents

Table of Contents

- Let ([5]) $h(\sigma) = \mathrm{Call}_{BS}(\sigma) - \mathrm{Call}_{M2Mkt} = \mathbb{E}H(\sigma, Z)$ with

$$H(\sigma, z) = \left(x_0\, e^{-\frac{\sigma^{+2}}{2}T + \sigma^{+}\sqrt{T}z} - e^{-rT}K\right)_+ - \mathrm{Call}_{M2Mkt}$$

  (σ_+ to ensure that h is increasing).
- Then the recursive stochastic zero search reads

$$\sigma_{n+1} = \sigma_n - \gamma_{n+1} H(\sigma_n, Z_{n+1}), \quad \sigma_0 > 0.$$

  with $(Z_n)_{n\geq 1}$ i.i.d.,$\sim \mathcal{N}(0,1)$ and $\sum_n \gamma_n = +\infty$, $\sum_{n\geq 1} \gamma_n^2 < +\infty$
- Try with $\gamma_n = \frac{a}{b+n}$ so that $\gamma_1 \times H(\sigma_0, Z_{+1}) \simeq$ few units.
- Exercise: write and execute a script with Both Newton and Robbins-Montro algorithms.

[5] G. Pagès (2018). *Introduction to Numerical Probability with application to Finance, Springer-Verlag, Berlin, 576p.*

Numerical Probability: Implied correlation search II $(\theta \rightsquigarrow \rho)$

- Let $h(\rho) = \text{Best-ofCall}_{BS}(\ldots, \rho, \ldots) - \text{Best-of-Call}_{M2Mkt}$

$$= \mathbb{E}\, H(\rho, Z),\ Z = (Z^1, Z^2) \sim \mathcal{N}(0, I_2)$$

with $H(\rho, Z) = \left(\max\left(x_0^1 e^{-\frac{\sigma_1^2 T}{2} + \sigma_1\sqrt{T}z^1}, x_0^2 e^{-\frac{\sigma_2^2 T}{2} + \sqrt{T}\sigma_2(\rho z^1 + \sqrt{1-\rho^2}z^2)}\right) - e^{-rT}K\right)_+ - \text{Best-of-Call}_{M2Mkt}.$

- The naive algorithm (with $(\gamma_n)_{n\geq 1}$ satisfying the (DS) assumption)

$$\rho_{n+1} = \rho_n - \gamma_{n+1} H(\rho_n, Z_{n+1})$$

does not live inside $[-1, 1]$!! ...

- What to do ? Project on $[-1, 1]$ (theorems do exist) or change of variable ([6]) e.g. by an homeomorphism from $\mathbb{R}$ to $(-1, 1)$ like

$$\rho = \frac{2}{\pi}\arctan(\theta) =: \varphi(\theta)$$

so that

$$\theta_{n+1} = \theta_n - \gamma_{n+1} H\big(\varphi(\theta_n), Z_{n+1}\big),\ \theta_0 \in \mathbb{R}.$$

- There exists a C^1_{Lip}-Lyapunov function $V : \mathbb{R} \to \mathbb{R}_+$ such that

Implied correlation search II: A better choice ! ($\theta = \ldots\theta$)

- Set
$$\rho = \sin(\theta) =: \tilde{\varphi}(\theta)$$
- Then one checks that, as $(Z^1, Z^2) \stackrel{d}{=} (Z^1, -Z^2)$,
$$\left(Z^1, \sin(\theta)Z^1 + \sqrt{1-\sin^2(\theta)}Z^2\right) \sim \left(Z^1, \sin(\theta)Z^1 + \cos(\theta)Z^2\right).$$
- It introduces countably many solutions ("half" parasitic as noisy traps) to the implicitation problem in θ. But it does not matter in practice !
- Set $\tilde{H}(\theta, z) = H\big(\sin(\theta), z\big)$ so that $\tilde{h}(\theta) = h\big(\sin(\theta)\big)$ is 2π-periodic and implement
$$\theta_{n+1} = \theta_n - \gamma_{n+1}\tilde{H}(\theta_n, Z_{n+1}),\ n \geq 0.$$
- In fact it even improves the convergence by reducing the "exploring phase" since the algorithm is always close to a solution.

In higher dimension ($\theta = (\theta_1, \ldots, \theta_{d(d-1)/2})$)

- In higher dimension a correlation matrix $R = \left[\frac{(\sigma_{i.}|\sigma_{j.})}{|\sigma_{i.}||\sigma_{j.}|}\right]$ whose Cholesky decomposition reads

$$R = TT^\top \quad \text{with } T \text{ lower triangular and} \quad \sum_{j=1}^{i} t_{ij}^2 = 1.$$

- (Hyper-)spherical parametrization

$$\begin{aligned}
&t_{11} = 1\\
&t_{21} = \cos(\theta_2),\ t_{22} = \sin(\theta_2)\\
&t_{31} = \cos(\theta_3)\sin(\phi_3);\ t_{32} = \cos(\theta_3)\cos(\phi_3),\ t_{33} = \sin(\theta_3)\\
&t_{41} = \cos(\theta_4)\cos(\phi_4)\cos(\psi_4),\ t_{42} = \cos(\theta_4)\cos(\phi_4)\sin(\psi_4)\\
&t_{43} = \cos(\theta_4)\sin(\phi_4),\ t_{44} = \cos(\theta_4).
\end{aligned}$$

etc.

- Then $\theta = (\theta_2, \theta_3, \phi_3, \theta_4, \phi_4, \psi_4)$.
- More involved problem: periodicity introduces multiple solutions. Which is no longer an asset due to many saddle points...

Numerical probability: VaR_α-$CVaR_\alpha$ II ($\theta \rightsquigarrow \xi$)

- Set $H(\xi,x) = \partial_\xi v(\xi,x) = 1 - \frac{1}{1-\alpha}\mathbf{1}_{\{x\geq\xi\}} = \frac{1}{1-\alpha}\big(\mathbf{1}_{\{x\leq\xi\}} - \alpha\big)$ so that
$$V'(\xi) = \mathbb{E}\, H(\xi, X)$$
- Set $\gamma_n = \frac{1}{n}$ and let X_n i.i.d., $\sim X$, then
$$\xi_{n+1} = \xi_n - \frac{\gamma_{n+1}}{1-\alpha}\big(\mathbf{1}_{\{X_{n+1}\leq\xi_n\}} - \alpha\big) \longrightarrow \xi_\alpha = \mathrm{VaR}_\alpha(X).$$
- What about $\mathrm{CVaR}_\alpha(X)$? Various solutions...
$$\Xi_n = \frac{v(\xi_0, X_1) + \cdots + v(\xi_{n-1}, X_n)}{n} \longrightarrow \mathbb{E}\, v(\xi_\alpha, X) = \mathrm{CVaR}_\alpha(X).$$
- Recursive form $\Xi_n = \Xi_{n-1} - \frac{1}{n}\Big(\Xi_{n-1} - v(\xi_{n-1}, X_n)\Big)$, $\Xi_0 = 0$.
- Warning ! Rare events phenomenon tends to freeze the algorithm $\Rightarrow$ adaptive Importance Sampling ([7]) !
- ...and try to slowly increase $\alpha = \alpha_n$ from $\alpha_0 = \frac{1}{2}$ to the target level.

[7] O. Bardou, N. Frikha, G. Pagès (2009). Computing VaR and CVaR using Stochastic Approximation and Adaptive Unconstrained Importance Sampling, *Monte Carlo and Applications Journal*, **15**(3):173–210.

First conclusions

- Low dimensional examples selected on purpose for expository.
- Not as automatic as (linear) Monte Carlo simulation: tuning of the step is mandatory.
- Many other examples : adaptive variance reduction (see Lemaire-P. 2007, AAP).
- Central-Limit Theorem, Averaging principle (Ruppert-Polyak).
- More details and results in ([8]) if interested and the references therein.

[8] G. Pagès (2018). *Introduction to Numerical Probability with application to Finance, Springer-Verlag, Berlin, 576p.*

Table of Contents

Learning

- Database $(z_k)_{k=1:N}$, parameters $\theta \in \Theta \subset \mathbb{R}^K$ and (local) loss function/predictor $v(\theta, z)$.
- Let $(I_k)_{k\geq 1}$ be an i.i.d. sequence $\mathcal{U}\big(\{1,\ldots,N\}\big)$-distributed.
- The stochastic gradient descent reads

$$\theta_{n+1} = \theta_n - \gamma_{n+1}\nabla_\theta v\big(\theta_n, z_{I_{n+1}}\big)$$

  where $z_{I_{n+1}}$ means that a datum has been picked up at random in the database uniformly in $\{1,\ldots,N\}$.
- Check that

$$\begin{aligned}\mathbb{E}\,\nabla_\theta v\big(\theta_n, z_I\big) &= \frac{1}{N}\sum_{k=1}^{N}\nabla_\theta v\big(\theta_n, z_k\big)\\ &= \int \nabla_\theta v\big(\theta_n, z\big)\mu_{_N}(dz) = \nabla V(\theta_n)\end{aligned}$$

CLVQ/k-means (unsupervised learning)

- Aim:

$$\min_{(\theta^j)_{j=1:r}} \left[V(\theta) = \frac{1}{2} \sum_{k=1}^{N} \min_{i=1:r} |\theta^i - x_k|^2 \right]$$

(mean minimal distance to prototypes).

- Competitive Learning Vector Quantization:

$$\theta^i_{n+1} = \begin{cases} \theta^i_n - \gamma_{n+1}\big(\theta^i_n - x_{n+1}\big) & \text{if } |x_{n+1} - \theta^i_n| < \min_{j\neq i} |x_{n+1} - \theta^j_n| \\ = 0 & \text{otherwise} \end{cases}$$

- In other words: $\to n+1$ reads
  - Nearest neighbour search to the datum among r prototypes of dimension d .
  - Moving the winner by a dilatation centered at the datum with ratio $1 - \gamma_{n+1} > 0$.

CLVQ