

Table of Contents

- 1 Zero search, Optimization (deterministic, the origins)
- 2 Examples from Finance
 - Implicitation
 - Minimization
- 3 Learning procedures
 - Abstract Learning
 - Supervised Learning
 - Unsupervised Learning (clustering)
- 4 Stochastic algorithms/Approximation
 - Paradigm of stochastic approximation
 - From Robbins-Monro to Robbins-Siegmund
 - Application: Stochastic Gradient Descent (SGD) and $S(\text{pseudo-})GD$
- 5 Examples revisited by SGD
 - Numerical Probability
 - Learning theory
- 6 Application to Neural Networks and deep learning
 - Linear neural network
 - One hidden layer feedforward perceptron
 - Universal approximation property
 - Toward deep learning
 - Multilayer feedforward perceptron and Backpropagation
- 7 Unsupervised learning

Table of Contents

- 1 Zero search, Optimization (deterministic, the origins)
- 2 Examples from Finance
 - Implicitation
 - Minimization
- 3 **Learning procedures**
 - **Abstract Learning**
 - Supervised Learning
 - Unsupervised Learning (clustering)
- 4 Stochastic algorithms/Approximation
 - Paradigm of stochastic approximation
 - From Robbins-Monro to Robbins-Siegmund
 - Application: Stochastic Gradient Descent (SGD) and $S(\text{pseudo-})GD$
- 5 Examples revisited by SGD
 - Numerical Probability
 - Learning theory
- 6 Application to Neural Networks and deep learning
 - Linear neural network
 - One hidden layer feedforward perceptron
 - Universal approximation property
 - Toward deep learning
 - Multilayer feedforward perceptron and Backpropagation
- 7 Unsupervised learning

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: $V(\theta) = \frac{1}{N} \sum_{k=1}^N v(\theta, z_k)$

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

- Solving the minimization problem

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

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

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

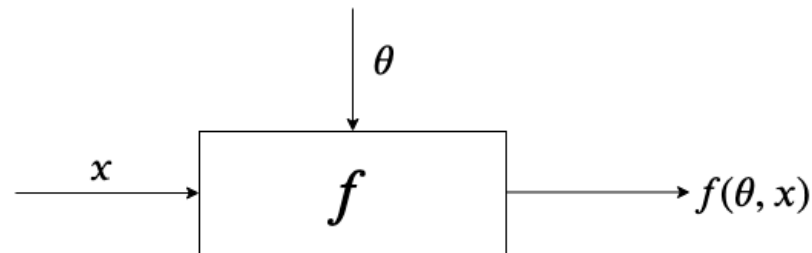
with the step sequence satisfying the *(DS)* assumption.

Table of Contents

- 1 Zero search, Optimization (deterministic, the origins)
- 2 Examples from Finance
 - Implicitation
 - Minimization
- 3 **Learning procedures**
 - Abstract Learning
 - **Supervised Learning**
 - Unsupervised Learning (clustering)
- 4 Stochastic algorithms/Approximation
 - Paradigm of stochastic approximation
 - From Robbins-Monro to Robbins-Siegmund
 - Application: Stochastic Gradient Descent (SGD) and $S(\text{pseudo-})GD$
- 5 Examples revisited by SGD
 - Numerical Probability
 - Learning theory
- 6 Application to Neural Networks and deep learning
 - Linear neural network
 - One hidden layer feedforward perceptron
 - Universal approximation property
 - Toward deep learning
 - Multilayer feedforward perceptron and Backpropagation
- 7 Unsupervised learning

Supervised learning

- **Input** x_k , **output** y_k . Data $z_k = (x_k, y_k) \in \mathbb{R}^{d_x + d_y}$, $k = 1 : N$.
 $\Theta \subset \mathbb{R}^{d_\theta}$ the set of (hyper-)parameters.
- **Transfer function** $f : \Theta \times \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y}$



- Prediction/loss function (local) $v(\theta, z) = \frac{1}{2} |f(\theta, x) - y|^2$ so that

$$\nabla_{\theta} v(\theta, z) = \underbrace{\nabla_{\theta} f(\theta, x)^{\top}}_{d_{\theta} \times d_y \text{-matrix}} (f(\theta, x) - y).$$

- Resulting loss function gradient

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

Table of Contents

- 1 Zero search, Optimization (deterministic, the origins)
- 2 Examples from Finance
 - Implicitation
 - Minimization
- 3 **Learning procedures**
 - Abstract Learning
 - Supervised Learning
 - **Unsupervised Learning (clustering)**
- 4 Stochastic algorithms/Approximation
 - Paradigm of stochastic approximation
 - From Robbins-Monro to Robbins-Siegmund
 - Application: Stochastic Gradient Descent (SGD) and $S(\text{pseudo-})GD$
- 5 Examples revisited by SGD
 - Numerical Probability
 - Learning theory
- 6 Application to Neural Networks and deep learning
 - Linear neural network
 - One hidden layer feedforward perceptron
 - Universal approximation property
 - Toward deep learning
 - Multilayer feedforward perceptron and Backpropagation
- 7 Unsupervised learning

Unsupervised learning (clustering)

- Only input $z_k = x_k \in \mathbb{R}^d$, $k = 1 : N$.
- **Prototypes/codewords** parameter set: $\theta := (\theta^1, \dots, \theta^r) \in \Theta = (\mathbb{R}^d)^r$, $r \in \mathbb{N}$.
- **Similarity measure** $\phi : (x, \theta) \mapsto \phi : (x, \theta) \geq 0$.
- (An example of) Local loss function: nearest neighbor among the prototypes: $x \in \mathbb{R}^d$, $\theta \in \Theta$.

$$v(\theta, x) = \frac{1}{2} \min_{i=1:r} \phi(\theta^i, x)$$

(**best similarity to prototypes**).

- $v(\theta, x)$ is **not convex** in θ (due to min...)!
- Global loss function (**Distortion**):

$$V_\phi(\theta) = \frac{1}{2N} \sum_{k=1}^N \min_{i=1:r} \phi(\theta^i, x_k) \quad (\text{mean best similarity to prototypes}).$$

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

Examples of similarity measures

- $\phi(\theta^i, x) = \|x - \theta^i\|^2$ (Euclidean norm)
- $\phi(\theta^i, x) = \|x - \theta^i\|^r$ ($\|\cdot\|$ norm on $\mathbb{R}^d$ norm, $r > 0$).
- $\phi(\theta^i, x) = F(x) - F(\theta^i) - (\nabla F(\theta^i) | x - \theta^i) > 0$ (Bregman divergence induced by a **strictly convex function** $F : \mathbb{R}^d \rightarrow \mathbb{R}$) where
 - $F(x) = -\log(x)$ on $\mathbb{R}_+$,
 - $F(x) = x \log(x)$ on $\mathbb{R}_+$,
 - $F(x) = \log(1 + e^{\lambda x})/\lambda$, $\lambda > 0$ on $\mathbb{R}$, etc.

Batch k -means/Forgey's algorithm (squared Eucl. norm)

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

with

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

- 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 $\nabla V(\theta) = \frac{1}{2N} \sum_{k=1}^N \nabla_{\theta} v(\theta, x_k)$.
- $\implies N \times$ nearest neighbour searches among r prototypes of dim d !
- Forgey's algorithm = (GD) algorithm (or batch (GD) algorithm):

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

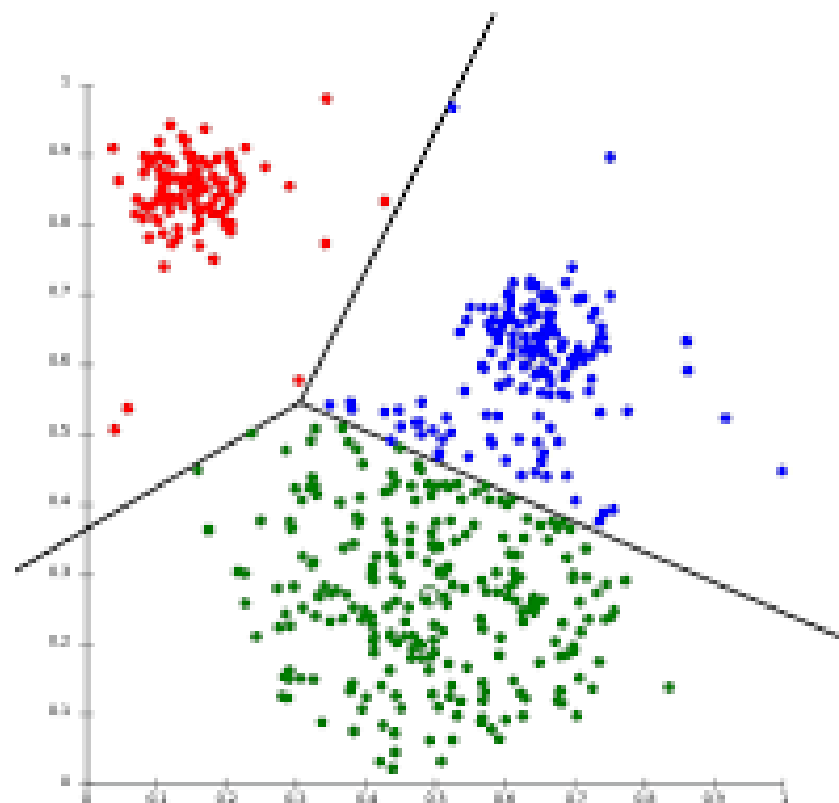


Table of Contents

- 1 Zero search, Optimization (deterministic, the origins)
- 2 Examples from Finance
 - Implicitation
 - Minimization
- 3 Learning procedures
 - Abstract Learning
 - Supervised Learning
 - Unsupervised Learning (clustering)
- 4 Stochastic algorithms/Approximation
 - Paradigm of stochastic approximation
 - From Robbins-Monro to Robbins-Siegmund
 - Application: Stochastic Gradient Descent (SGD) and $S(\text{pseudo-})GD$
- 5 Examples revisited by SGD
 - Numerical Probability
 - Learning theory
- 6 Application to Neural Networks and deep learning
 - Linear neural network
 - One hidden layer feedforward perceptron
 - Universal approximation property
 - Toward deep learning
 - Multilayer feedforward perceptron and Backpropagation
- 7 Unsupervised learning

Table of Contents

- 1 Zero search, Optimization (deterministic, the origins)
- 2 Examples from Finance
 - Implicitation
 - Minimization
- 3 Learning procedures
 - Abstract Learning
 - Supervised Learning
 - Unsupervised Learning (clustering)
- 4 Stochastic algorithms/Approximation
 - Paradigm of stochastic approximation
 - From Robbins-Monro to Robbins-Siegmund
 - Application: Stochastic Gradient Descent (SGD) and $S(\text{pseudo-})GD$
- 5 Examples revisited by SGD
 - Numerical Probability
 - Learning theory
- 6 Application to Neural Networks and deep learning
 - Linear neural network
 - One hidden layer feedforward perceptron
 - Universal approximation property
 - Toward deep learning
 - Multilayer feedforward perceptron and Backpropagation
- 7 Unsupervised learning

Pros and cons: toward stochastic algorithm I

- Numerical Probability (for Fin.): 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 $H : \mathbb{R}^d \times \mathbb{R}^q \rightarrow \mathbb{R}^d$ 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 e.g. to $F_X(\xi) = \mathbb{E} \mathbf{1}_{\{X \leq \xi\}}$.)
- **Pros:** Can use **stochastic analysis**! E.g. if $V(x_0) = \mathbb{E} f(X_T^{x_0})$, (i.e. $\theta = x_0$) in a B-S model. Then $V'(x_0) = \mathbb{E} \left[\underbrace{f'(X_T^{x_0}) \frac{X_T^{x_0}}{x_0}}_{\text{pathwise derivative}} \right] = \mathbb{E} \left[\underbrace{f(X_T^{x_0}) \frac{W_T}{x_0 \sigma T}}_{\text{substitution}} \right].$

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**.

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

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

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

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

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

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

Toward stochastic algorithm II

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

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \hat{h}_{[n+1]}(\theta_n)$$
$$\hat{h}_{[n+1]}(\theta_n) = \frac{1}{M_{n+1}} \sum_{k=1}^{M_{n+1}} H(\theta_n, Z_k^{(n+1)}), \quad (Z_k^{(n+1)})_{k,n} \text{ i.i.d. } \sim Z$$

- **Idea 2:** Robbins-Monro, 1951 ⁽³⁾. Set: $\forall n \geq 1, \quad M_n = 1 !!$

$$\theta_{n+1} = \theta_n - \gamma_{n+1} H(\theta_n, Z_{n+1}), \quad n \geq 0.$$

- **Idea 1.5:** Mini-batch i.e. $M_n = M > 2$ medium size. “Recently” became successful among practitioners.

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

Table of Contents

- 1 Zero search, Optimization (deterministic, the origins)
- 2 Examples from Finance
 - Implicitation
 - Minimization
- 3 Learning procedures
 - Abstract Learning
 - Supervised Learning
 - Unsupervised Learning (clustering)
- 4 Stochastic algorithms/Approximation
 - Paradigm of stochastic approximation
 - From Robbins-Monro to Robbins-Siegmund
 - Application: Stochastic Gradient Descent (SGD) and $S(\text{pseudo-})GD$
- 5 Examples revisited by SGD
 - Numerical Probability
 - Learning theory
- 6 Application to Neural Networks and deep learning
 - Linear neural network
 - One hidden layer feedforward perceptron
 - Universal approximation property
 - Toward deep learning
 - Multilayer feedforward perceptron and Backpropagation
- 7 Unsupervised learning

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

- (i) $(Z_n)_{n \geq 1}$ is i.i.d. $\sim Z$, independent of θ_0 ,
- (ii) $\|H(\theta, Z)\|_2 \leq C(1 + |\theta|)$ ($\Rightarrow |h(\theta)| \leq C(1 + |\theta|)$ linear growth),
- (iii) $(\gamma_n)_{n \geq 1}$ is a $(0, +\infty)$ -valued **deterministic step** sequence,

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

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

- 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

$$\mathbb{E}(H(\theta_n, Z_{n+1}) | \mathcal{F}_n) \underbrace{=}_{Z_{n+1} \perp\!\!\!\perp \mathcal{F}_n \& i.i.d \sim Z} \left[\mathbb{E} H(\theta, Z) \right]_{\theta=\theta_n} = h(\theta_n), \quad n \geq 0.$$

- 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*: $V : \mathbb{R}^d \rightarrow \mathbb{R}_+, \mathcal{C}^1, \lim_{\infty} V = +\infty, \nabla V$ Lipschitz, $|\nabla V|^2 \leq c(1 + V)$ and
- *Mean-reversion*: $(\nabla V|h) \geq 0$.
- *L^2 -Growth control*: $\|H(\theta, Z)\|_2 \leq C\sqrt{1 + V(\theta)}$.
- *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) $V(\theta_n) \xrightarrow{\text{a.s.}} V_\infty \in L^1 [\Rightarrow (\theta_n)_{n \geq 0} \text{ pathwise bounded}]$ 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 $\theta_n - \theta_{n-1} \rightarrow 0$ a.s.).
- (iv) $\sum_n \gamma_n \Delta M_n$ converges a.s. and in L^2 .

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

Proof.

- Set $\mathcal{F}_n := \sigma(\theta_0, Z_1, \dots, Z_n)$, $n \geq 1$ 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.

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

where

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

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

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

- 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: $|(a|b)| \leq |a| \cdot |b| \leq \frac{1}{2}(|a|^2 + |b|^2)$ so that

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

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

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

- So that $(\Delta M_n)_{n \geq 1}$ is a sequence of L^2 $(\mathcal{F}_n)$ -martingale increments satisfying

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

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

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

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

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

- Conditioning with respect to $\mathcal{F}_n$ yields, as $\nabla V(\theta_n)$ is $\mathcal{F}_n$ -measurable,

- $\mathbb{E}[(\nabla V(\theta_n)|\Delta M_{n+1}) | \mathcal{F}_n] = (\nabla V(\theta_n)|\mathbb{E}[\Delta M_{n+1} | \mathcal{F}_n]) = 0$

- and, other terms in the RHS being (also) $\mathcal{F}_n$ -measurable,

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

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

- 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,

- $(1 + C_v \gamma_{n+1}^2) \times [\text{this positive term}]$ 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)

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

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

- Hence

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

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

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

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

$(S_n)_{n \geq 0}$ is $L^1(\mathbb{P})$ -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 $\sum_{0 \leq k \leq n-1} \gamma_{k+1}(\nabla V|h)(\theta_k)$ satisfies for every $n \geq 1$,

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

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

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

so that, in particular,

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

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 $n \rightarrow +\infty$. $V_\infty \in L^1$ by Fatou's Lemma since $(V(\theta_n))_{n \geq 0}$ is L^1 -bounded.
- (iii) Again by Beppo Levi's monotone convergence Theorem for series with non-negative terms,

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

so that

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

which in turns yields

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

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

$$\begin{aligned} \langle M^\gamma \rangle_n &= \sum_{k=1}^n \gamma_k^2 \mathbb{E} (|\Delta M_k|^2 | \mathcal{F}_{k-1}) \leq \sum_{k=1}^n \gamma_k^2 \mathbb{E} (|H(\theta_{k-1}, Z_k)|^2 | \mathcal{F}_{k-1}) \\ &\leq C \sum_{k=1}^n \gamma_k^2 (1 + V(\theta_{k-1})) \end{aligned}$$

so that, owing to (i)_a,

$$\mathbb{E} \langle M^\gamma \rangle_\infty \leq C \sum_{n \geq 1} \gamma_n^2 (1 + \mathbb{E} V(\theta_{n-1})) \leq C' \sum_{n \geq 1} \gamma_n^2 < +\infty.$$

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

- Which completes the proof !

□

Table of Contents

- 1 Zero search, Optimization (deterministic, the origins)
- 2 Examples from Finance
 - Implicitation
 - Minimization
- 3 Learning procedures
 - Abstract Learning
 - Supervised Learning
 - Unsupervised Learning (clustering)
- 4 Stochastic algorithms/Approximation
 - Paradigm of stochastic approximation
 - From Robbins-Monro to Robbins-Siegmund
 - Application: Stochastic Gradient Descent (SGD) and $S(\text{pseudo-})GD$
- 5 Examples revisited by SGD
 - Numerical Probability
 - Learning theory
- 6 Application to Neural Networks and deep learning
 - Linear neural network
 - One hidden layer feedforward perceptron
 - Universal approximation property
 - Toward deep learning
 - Multilayer feedforward perceptron and Backpropagation
- 7 Unsupervised learning

Robbins-Monro (pathwise reasoning): zero search algorithm

Theorem (Robbins-Monro algorithm)

- Assume the mean function h is continuous and satisfies

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

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

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

- Finally, assume $(\gamma_n)_{n \geq 1}$ satisfies (DS).

Then

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

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

- If $H(\theta, z) = h(\theta)$, back to a deterministic zero search procedure!!!
- if $h \uparrow\uparrow$ in the sense $(h(\theta) - h(\theta') | \theta - \theta') > 0$, $\theta \neq \theta'$ and $h(\theta_*) = 0$, (i) holds.

Proof

- Let $\theta_\lambda = \theta^* - \lambda h(\theta^*)$, $\lambda > 0$. If $h(\theta^*) \neq 0$, then $-(h(\theta_\lambda)|h(\theta^*)) > 0$. Letting $\lambda \rightarrow 0$ 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 $\sum_{n \geq 1} \gamma_n (h(\theta_{n-1})|\theta_{n-1} - \theta_*) < +\infty \quad \mathbb{P}\text{-a.s.}$
 - 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.,

$$\lim_n (\theta_{n-1}(\omega) - \theta_* |h(\theta_{n-1}(\omega))) = 0.$$

- In fact if $\lim_n (\theta_{n-1}(\omega) - \theta_* |h(\theta_{n-1}(\omega))) > 0$, the above convergence induces a contradiction with $\sum_{n \geq 1} \gamma_n = +\infty$.

- Let $(\phi(n, \omega))_{n \geq 1}$ be a subsequence such that

$$(\theta_{\phi(n, \omega)}(\omega) - \theta_* |h(\theta_{\phi(n, \omega)}(\omega))) \longrightarrow 0 \quad \text{as} \quad n \rightarrow +\infty.$$

- 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,

$$\theta_{\phi(n, \omega)}(\omega) \rightarrow \theta_\infty = \theta_\infty(\omega).$$

- By continuity of h , $(\theta_\infty - \theta_* |h(\theta_\infty)) = 0$ which implies $\theta_\infty = \theta_*$.
Now, since we know that $V(\theta_n(\omega)) = \frac{1}{2}|\theta_n(\omega) - \theta_*|^2$ converges,

$$\lim_n |\theta_n(\omega) - \theta_*|^2 = \lim_n |\theta_{\phi(n, \omega)}(\omega) - \theta_*|^2 = 0.$$

- Convergence in L^p , $p \in (0, 2)$ follows by uniform integrability. □

Theorem (Stochastic optimization: Stochastic Gradient Descent)

- Let $V : \mathbb{R}^d \rightarrow \mathbb{R}_+$ be a *Lyapunov function for h* i.e. differentiable,
 $\lim_{|\theta| \rightarrow +\infty} V(\theta) = +\infty$, ∇V Lip, $|\nabla V|^2 \leq C(1 + V)$, $\|H(\theta, Z)\|_2 \leq C\sqrt{1 + V(\theta)}$.
- Assume $\{\nabla V = 0\} = \{\theta_*\}$.
- Let $h(\theta) = \nabla V(\theta) = \mathbb{E} H(\theta, Z)$ with H s.t. $\|H(\theta, Z)\|_2 \leq C\sqrt{1 + V(\theta)}$,
 $V(\theta_0) \in L^1(\mathbb{P})$ and $\theta_0 \perp\!\!\!\perp (Z_n)_{n \geq 1}$.
- Assume $(\gamma_n)_{n \geq 1}$ satisfies (DS).

Then

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

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

- *Proof.* Use (almost) the same arguments as above but with $(\nabla V | h)(\theta) = |\nabla V(\theta)|^2 > 0$, $\theta \neq \theta^*$, instead of $(\theta - \theta_* | h(\theta))$. 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 $\theta_n \rightarrow \theta^*$ (details left as an exercise). $\square$
- **Remark !** If $H(\theta, z) = h(\theta) = \nabla V(\theta)$: **Convergence thm for regular (GD) !!**

Stochastic Pseudo-Gradient Descent

- This time h is given and we search for V such that $(\nabla V | h) \geq 0$ (mean-reversion).

Theorem (Zero search: Stochastic Pseudo-Gradient Descent)

- ▶ Let $h(\theta) = \mathbb{E} H(\theta, Z)$ and $(Z_n)_{n \geq 1}$ an i.i.d. sequence with $Z_1 \stackrel{d}{\sim} Z$.
- ▶ Let $V : \mathbb{R}^d \rightarrow \mathbb{R}_+$ be a *Lyapunov function for h* : differentiable,
 $\lim_{|\theta| \rightarrow +\infty} V(\theta) = +\infty$, ∇V Lip, $|\nabla V|^2 \leq C(1 + V)$, $\|H(\theta, Z)\|_2 \leq C\sqrt{1 + V(\theta)}$
- ▶ and (**NEW !**)

$$\forall K \subset \mathbb{R}^d, \text{ compact, } (h | \nabla V) \geq a_K |\nabla V|^2, a_K > 0 \text{ and } \{\nabla V = 0\} = \{\theta_*\}.$$

- ▶ Assume $(\gamma_n)_{n \geq 1}$ satisfies (DS), $V(\theta_0) \in L^1(\mathbb{P})$ and $\theta_0 \perp\!\!\!\perp (Z_n)_{n \geq 1}$.

Then

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

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

Proof

- Proof

- V and h satisfy the assumptions of the Robbins-Siegmund Lemma.
- Reasoning on a (somewhat hidden) generic ω .
- Hence $V(\theta_n) \rightarrow V_\infty$ and $\sum_{n \geq 1} \gamma_n (h|\nabla V)(\theta_{n-1}) < +\infty$.
- As $\lim_{|\theta| \rightarrow +\infty} V(\theta) = +\infty$, $(\theta_n)_{n \geq 1}$ is bounded so that $K = \overline{\{\theta_n, n \geq 0\}}$ is compact.
- Hence $\sum_{n \geq 1} \gamma_n |\nabla V|^2(\theta_{n-1}) < +\infty$ since $a_K > 0$.
- Etc (like for the *SGD*). □

- Question:

What happens when $\{\nabla V = 0\}$ and or $\{h = 0\}$ are not singleton?

- Multi-target framework !

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. $\text{card}(\{\nabla V = 0\}) \geq 2$, one has *mutatis mutandis*: a.s. there exists $v_\infty \in \mathbb{R}_+$ and a connected component χ_∞ of $\{\nabla V = 0\} \cap \{V = v_\infty\}$ such that

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

(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 $(V(\theta_n))_{n \geq 0}$ is L^1 -bounded so that $(\nabla V(\theta_n))_{n \geq 0}$ is L^2 -bounded).

- Same type of result for *pseudo-stochastic gradient*.

- By the **R.-S. Lemma**, one has for free that, a.s., $(\theta_n)_{n \geq 0}$ is pathwise bounded and $\theta_n - \theta_{n-1} \rightarrow 0$ 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⁾.

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

Mountain Pass Lemma

Theorem (A. Ambrosetti & P. Rabinowitz)

- Let $V : \mathbb{R}^d \rightarrow \mathbb{R}_+$ be C^1 , coercive [and ...].
If $\text{card}(\text{argminloc}_{\mathbb{R}^d} V) \geq 2$, then

$$\text{argminloc}_{\mathbb{R}^d} V \neq \{\nabla V = 0\}.$$

- To be more precise ^(a): if V has *(at least) two strict local minima*, then it has *at least one critical point which is not a local minimum*. Moreover, if $d \geq 2$, this critical point is a (kind of) *saddle point* i.e. is *not a local extremum*.

^aAubin J.-P., Ekeland I. (2006). *Applied Nonlinear Analysis*. Dover Books

- In terms of stochastic optimization, this theorem claims the existence of *traps* for multi-target *SGD* or *SPGD*.

Avoiding (noisy) traps ?

- We consider here for simplicity the framework of *SGD* for simplicity ($D^2V(\theta)$ is diagonalisable) .

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

- Assume $\nabla V(\theta) = \mathbb{E} H(\theta, Z)[, \text{ etc}]$.
- Let $\theta_* \in \{\nabla V = 0\}$. If there exists a **negative eigen-(value,vector)** (λ, u) i.e. such that $D^2V(\theta^*)u = \lambda u$ such that

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

i.e. θ^* is a **noisy trap** (noise in a repulsive repulsive direction) then

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

- This allows to eliminate **noisy** local maxima but also saddle points, “monkey” saddle points, etc.

Rates of convergence

Later on...

Deterministic framework (exercise)

- What happens if

$$\forall \theta \in \mathbb{R}^d, \forall z \in \mathbb{R}^q, \quad H(\theta, z) = h(\theta)$$

i.e. **no randomness** ?

- In fact **corresponding deterministic optimization theorems are proved...**

Table of Contents

- 1 Zero search, Optimization (deterministic, the origins)
- 2 Examples from Finance
 - Implicitation
 - Minimization
- 3 Learning procedures
 - Abstract Learning
 - Supervised Learning
 - Unsupervised Learning (clustering)
- 4 Stochastic algorithms/Approximation
 - Paradigm of stochastic approximation
 - From Robbins-Monro to Robbins-Siegmund
 - Application: Stochastic Gradient Descent (SGD) and $S(\text{pseudo-})GD$
- 5 Examples revisited by SGD
 - Numerical Probability
 - Learning theory
- 6 Application to Neural Networks and deep learning
 - Linear neural network
 - One hidden layer feedforward perceptron
 - Universal approximation property
 - Toward deep learning
 - Multilayer feedforward perceptron and Backpropagation
- 7 Unsupervised learning

Table of Contents

- 1 Zero search, Optimization (deterministic, the origins)
- 2 Examples from Finance
 - Implicitation
 - Minimization
- 3 Learning procedures
 - Abstract Learning
 - Supervised Learning
 - Unsupervised Learning (clustering)
- 4 Stochastic algorithms/Approximation
 - Paradigm of stochastic approximation
 - From Robbins-Monro to Robbins-Siegmund
 - Application: Stochastic Gradient Descent (SGD) and $S(\text{pseudo-})GD$
- 5 Examples revisited by SGD
 - Numerical Probability
 - Learning theory
- 6 Application to Neural Networks and deep learning
 - Linear neural network
 - One hidden layer feedforward perceptron
 - Universal approximation property
 - Toward deep learning
 - Multilayer feedforward perceptron and Backpropagation
- 7 Unsupervised learning

Numerical Probability: Implied volatility II ($\theta = \sigma$)

- Let ⁽⁵⁾ $h(\sigma) = \text{Call}_{BS}(\sigma) - \text{Call}_{Mkt}(K, T) = \mathbb{E}H(\sigma, Z)$ with

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

(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 (say 1 or 2).
- **Exercise:** Write and execute a script with both Newton and Robbins-Monro algorithms. [Step calibration is the challenge].

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

Numerical Probability: Implied correlation search II

- $(\theta \rightsquigarrow \rho)$. Let $h(\rho) := \text{Best-of-Call}_{Mkt}(K, T) - \text{Best-of-Call}_{BS}(\dots, \rho, \dots)$ $\uparrow\uparrow$.
 $= \mathbb{E} H(\rho, Z), \quad Z = (Z^1, Z^2) \sim \mathcal{N}(0, I_2)$

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

- 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

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

- There exists a C_{Lip}^1 -Lyapunov function $V : \mathbb{R} \rightarrow \mathbb{R}_+$ such that $V'(\theta)h(\varphi(\theta)) \geq 0$ and $= 0$ iff $h(\varphi(\theta)) = 0$ i.e. $\varphi(\theta) = \rho_*$ [Easy **exercise** !].

⁶G. Pagès (2018). *Introduction to Numerical Probability with application to Finance*, Springer-Verlag, Berlin, 576p.

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

- Set

$$\rho = \sin(\theta) =: \tilde{\varphi}(\theta)$$

- Then one checks that, as $(Z^1, Z^2) \stackrel{d}{=} (Z^1, -Z^2)$,

$$(Z^1, \sin(\theta)Z^1 + \sqrt{1 - \sin^2(\theta)}Z^2) \sim (Z^1, \sin(\theta)Z^1 + \cos(\theta)Z^2).$$

- It introduces **countably many solutions** (“half” parasitic as noisy traps) to the implicitation problem in θ . But it does not matter in practice !
- Set $\tilde{H}(\theta, z) = H(\sin(\theta), z)$ so that $\tilde{h}(\theta) = h(\sin(\theta))$ is 2π -periodic and implement

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \tilde{H}(\theta_n, Z_{n+1}), \quad n \geq 0.$$

- In fact it even improves the convergence by reducing the “exploring phase” since the algorithm is always close to a solution.
- Lyapunov fucntion ? **YES!**, but not so easy exercise...

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

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

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

- (Hyper-)spherical parametrization

$$t_{11} = 1$$

$$t_{21} = \sin(\theta_2), \quad t_{22} = \cos(\theta_2)$$

$$t_{31} = \sin(\theta_3) \sin(\phi_3), \quad t_{32} = \sin(\theta_3) \cos(\phi_3), \quad t_{33} = \cos(\theta_3)$$

$$t_{41} = \sin(\theta_4) \sin(\phi_4) \sin(\psi_4), \quad t_{42} = \sin(\theta_4) \sin(\phi_4) \cos(\psi_4)$$

$$t_{43} = \sin(\theta_4) \cos(\phi_4), \quad t_{44} = \cos(\theta_4).$$

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 $H(\xi, x) = \partial_\xi v(\xi, x) = 1 - \frac{1}{1-\alpha} \mathbf{1}_{\{x \geq \xi\}} = \frac{1}{1-\alpha} (\mathbf{1}_{\{x \leq \xi\}} - \alpha)$ 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

$$\xi_{n+1} = \xi_n - \frac{\gamma_{n+1}}{1-\alpha} (\mathbf{1}_{\{X_{n+1} \leq \xi_n\}} - \alpha) \longrightarrow \xi_\alpha = VaR_\alpha(X).$$

- What about $CVaR_\alpha(X)$? Various solutions...

$$\Xi_n = \frac{v(\xi_0, X_1) + \cdots + v(\xi_{n-1}, X_n)}{n} \longrightarrow \mathbb{E} v(\xi_\alpha, X) = CVaR_\alpha(X).$$

- Recursive form $\Xi_n = \Xi_{n-1} - \frac{1}{n} \left(\Xi_{n-1} - v(\xi_{n-1}, X_n) \right)$, $\Xi_0 = 0$.
- 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**.

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