Stochastic Gradient Method: Convergence

January 20, 2015

Lecture Outline

Problem under Consideration

Consider the following convex differentiable optimization problem:

$$\min_{u \in U^{\mathrm{ad}}} J(u) \, .$$

Let $u^\sharp \in U^{\mathrm{ad}}$ be a solution of this problem (assume that such a solution exists). The associated optimality condition writes:

$$\left\langle \nabla J(u^\sharp) \, , u - u^\sharp \right\rangle \geq 0 \, , \quad \forall u \in U^{\mathrm{ad}} \, .$$

In the deterministic framework, the Auxiliary Problem Principle (**APP**) consists in replacing the original problem by a sequence of auxiliary problems indexed by $k \in \mathbb{N}$, and without too much perturbing the optimality conditions...

APP Framework

First idea: replace $J(u)$ by its first order approximation at $u^{(k)}$:

$$J(u) \approx J(u^{(k)}) + \left\langle \nabla J(u^{(k)}) \, , u - u^{(k)} \right\rangle \ .$$

But then the criterion is no more coercive...

Second idea: add a strongly convex term:

$$\frac{1}{\epsilon} \left(K(u) - K(u^{(k)}) - \left\langle \nabla K(u^{(k)}) \, , u - u^{(k)} \right\rangle \right) \ ,$$

K being a real-valued differentiable function defined on $\mathbb{U}$ and ϵ being a positive constant. At iteration k, given $u^{(k)} \in U^{\mathrm{ad}}$, consider the following auxiliary problem:

$$\min_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \epsilon \nabla J(u^{(k)}) - \nabla K(u^{(k)}) \, , u \right\rangle .$$

The solution $u^{(k+1)}$ of the auxiliary problem at iteration k is used to formulate a new auxiliary problem at iteration $k+1$.

APP Algorithm

1. Choose a core function K and a coefficient $\epsilon > 0$.
2. Choose $u^{(0)} \in U^{\mathrm{ad}}$ and a tolerance $\sigma > 0$. Set $k = 0$.
3. Obtain the solution $u^{(k+1)}$ of the auxiliary problem

$$\min_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \epsilon \nabla J(u^{(k)}) - \nabla K(u^{(k)}) \, , u \right\rangle \ .$$

4. Set $k = k + 1$ and go to step 3 until $\left\| u^{(k+1)} - u^{(k)} \right\| < \sigma$.

Note that the optimality condition of the auxiliary problem writes

$$\left\langle \nabla K(u^{(k+1)}) + \epsilon \nabla J(u^{(k)}) - \nabla K(u^{(k)}), u - u^{(k+1)} \right\rangle \geq 0 \ , \quad \forall u \in U^{\mathrm{ad}} \ ,$$

and coincides with the optimality condition of the initial problem in case where the sequence $\{u^{(k)}\}_{k \in \mathbb{N}}$ converges.

Convergence Theorem (1)

Make the following assumptions.

H1 U^{ad} is a nonempty, closed and convex subset of an Hilbert space $\mathbb{U}$.

H2 J is a proper l.s.c. convex function, coercive on U^{ad} and differentiable, ∇J being Lipschitz with constant A.

H3 K is a proper l.s.c. function, strongly convex with modulus b and differentiable, ∇K being Lipschitz with constant B.

H4 ϵ is such that $0 < \epsilon < \frac{2b}{A}$.

Convergence Theorem (2)

Then the following conclusions hold true.

R1 The initial problem admits at least a solution $u^\sharp$, and each auxiliary problem admits an unique solution $u^{(k+1)}$.

R2 The sequence $\{J(u^{(k)})\}_{k\in\mathbb{N}}$ is strictly decreasing and converges towards $J(u^\sharp)$.

R3 The sequence $\{u^{(k)}\}_{k\in\mathbb{N}}$ is bounded, and every cluster point of this sequence is a solution of the initial problem.

Assume moreover that

H5 J is strongly convex with modulus a.

Then we obtain that

R4 the sequence $\{u^{(k)}\}_{k\in\mathbb{N}}$ converges towards the unique solution $u^\sharp$ of the initial problem.

Sketch of Proof

The proof of the first statement is based on classical theorems.

The proof of the last two statements involves four steps.

1. Select a Lyapunov function Λ.
2. Prove that $\{\Lambda(u^{(k)})\}_{k\in\mathbb{N}}$ is a decreasing sequence.
   Then it converges, and $\{u^{(k)}\}_{k\in\mathbb{N}}$ is a bounded sequence.
3. Characterize the limit of the sequence $\{\Lambda(u^{(k)})\}_{k\in\mathbb{N}}$.
4. Extract a converging subsequence of $\{u^{(k)}\}_{k\in\mathbb{N}}$ and characterize its limit.

The result holds true if $\mathbb{U}$ is an infinite dimensional Hilbert space, and may be extended to problems with explicit constraints.

Some Features of APP (1)

One can take advantage of a proper choice of K in order to obtain many special features for the auxiliary subproblems. The reader is referred to [Cohen, 2004] for a detailed description of the APP. Two of its main properties are examined hereafter.

- APP encompasses "classical" optimization algorithms. Choosing $K(u) = \|u\|^2/2$, the auxiliary problem writes

$$\min_{u \in U^{\mathrm{ad}}} \frac{1}{2}\|u\|^2 + \left\langle \epsilon \nabla J(u^{(k)}) - u^{(k)}, u \right\rangle ,$$

and its solution has the following closed-form expression:

$$u^{(k+1)} = \mathrm{proj}_{U^{\mathrm{ad}}}\left(u^{(k)} - \epsilon \nabla J(u^{(k)})\right) .$$

We obtain the well-known projected gradient algorithm.

Some Features of APP (2)

- APP allows for decomposition. Assume that the space $\mathbb{U}$ is a Cartesian product of N spaces: $\mathbb{U} = \mathbb{U}_1 \times \cdots \times \mathbb{U}_N$, and that $U^{\mathrm{ad}} = U_1^{\mathrm{ad}} \times \cdots \times U_N^{\mathrm{ad}}$, with $U_i^{\mathrm{ad}} \subset \mathbb{U}_i$. Choosing a function K additive according to that decomposition of u, that is,

$$K(u_1, \ldots, u_N) = K_1(u_1) + \ldots + K_N(u_N) \ ,$$

the auxiliary subproblem becomes

$$\min_{u_1 \in U_1^{\mathrm{ad}}, \ldots, u_N \in U_N^{\mathrm{ad}}} \quad \sum_{i=1}^{N} \left(K_i(u_i) + \left\langle \epsilon \nabla_{u_i} J(u^{(k)}) - \nabla K_i(u_i^{(k)}) \, , u_i \right\rangle \right) \ .$$

This subproblem splits up into N independent subproblems, the i-th subproblem being

$$\min_{u_i \in U_i^{\mathrm{ad}}} K_i(u_i) + \left\langle \epsilon \nabla_{u_i} J(u^{(k)}) - \nabla K_i(u_i^{(k)}) \, , u_i \right\rangle \ .$$

Standard Stochastic Gradient Method (1)

Consider the following open-loop stochastic optimization problem:

$$\min_{u \in U^{\mathrm{ad}}} J(u) \, ,$$

with $J(u) = \mathbb{E}\big(j(u, \boldsymbol{W})\big)$.

The standard stochastic gradient algorithm reads as follows.

1. Let $u^{(0)} \in U^{\mathrm{ad}}$ and choose a positive real sequence $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$.
2. At iteration k, draw a realization $w^{(k+1)}$ of the r.v. $\boldsymbol{W}$.
3. Compute the gradient of j and update $u^{(k+1)}$ by the formula:
$$u^{(k+1)} = \mathrm{proj}_{U^{\mathrm{ad}}} \left(u^{(k)} - \epsilon^{(k)} \nabla_u j\big(u^{(k)}, w^{(k+1)}\big) \right) \, .$$
4. Set $k = k + 1$ and go to step 2.

Standard Stochastic Gradient Method (2)

This algorithm in fact involves random variables on $(\Omega, \mathcal{A}, \mathbb{P})$:

$$\boldsymbol{U}^{(k+1)} = \mathrm{proj}_{U^{\mathrm{ad}}} \left(\boldsymbol{U}^{(k)} - \epsilon^{(k)} \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) \right) ,$$

where $\{\boldsymbol{W}^{(k)}\}_{k\in\mathbb{N}}$ is a infinite-dimensional sample of $\boldsymbol{W}$.

Recall that a sequence $\{\epsilon^{(k)}\}_{k\in\mathbb{N}}$ is called a σ-sequence if

$$\sum_{k\in\mathbb{N}} \epsilon^{(k)} = +\infty , \quad \sum_{k\in\mathbb{N}} \left(\epsilon^{(k)}\right)^2 < +\infty .$$

Robbins-Monro Theorem

Under various assumptions, the sequence $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ of random variables generated by the stochastic gradient algorithm almost surely converges to $u^\sharp$.

Moving APP to the Stochastic Framework

In order to mix the ideas of the Auxiliary Problem Principle and of the Stochastic Gradient Method, we replace the initial problem by the associated sequence of auxiliary problems, namely

$$\min_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \epsilon \nabla J(u^{(k)}) - \nabla K(u^{(k)}) \, , u \right\rangle .$$

Then, in each auxiliary problem, we replace the gradient of J by the gradient of j evaluated at sampled realizations of $\boldsymbol{W}$. Note that the "large" (constant) step size ϵ has to be replaced by "small" (going to zero as index k goes to infinity) steps $\epsilon^{(k)}$. The k-th instance of the stochastic auxiliary problem is thus

$$\min_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \epsilon^{(k)} \nabla_u j(u^{(k)}, w^{(k+1)}) - \nabla K(u^{(k)}) \, , u \right\rangle \, ,$$

$w^{(k+1)}$ being a realization of the random variable $\boldsymbol{W}$.

Stochastic APP Algorithm

We thus obtain a generalized stochastic gradient algorithm.

Stochastic APP Algorithm

1. Let $u^{(0)} \in U^{\mathrm{ad}}$ and choose a positive real sequence $\{\epsilon^{(k)}\}_{k\in\mathbb{N}}$.
2. At iteration k, draw a realization $w^{(k+1)}$ of the r.v. $\boldsymbol{W}$.
3. Update $u^{(k+1)}$ by solving the auxiliary problem:
$$u^{(k+1)} \in \underset{u\in U^{\mathrm{ad}}}{\arg\min}\, K(u) + \left\langle \epsilon^{(k)} \nabla_u j(u^{(k)}, w^{(k+1)}) - \nabla K(u^{(k)}) \,,\, u \right\rangle.$$
4. Set $k = k+1$ and go to step 2.

As usual, the algorithm is casted in the probabilistic framework:

$$\boldsymbol{U}^{(k+1)} \in \underset{u\in U^{\mathrm{ad}}}{\arg\min}\, K(u) + \left\langle \epsilon^{(k)} \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) - \nabla K(\boldsymbol{U}^{(k)}) \,,\, u \right\rangle \,.$$

The fact that the solution $\boldsymbol{U}^{(k+1)}$ of this problem corresponds to a random variable, that is, a mesurable function has to be justified.

Example

With the choice

$$K(u) = \frac{1}{2}\|u\|^2 \ ,$$

the auxiliary problem becomes

$$\min_{u \in U^{\mathrm{ad}}} \frac{1}{2}\|u\|^2 + \left\langle \epsilon^{(k)} \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) - \boldsymbol{U}^{(k)} \ , u \right\rangle .$$

The set of solutions of this problem (an unique solution per ω) forms an unique random variable $\boldsymbol{U}^{(k+1)}$, whose expression is

$$\boldsymbol{U}^{(k+1)} = \mathrm{proj}_{U^{\mathrm{ad}}} \left(\boldsymbol{U}^{(k)} - \epsilon^{(k)} \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) \right) .$$

It corresponds to the standard stochastic gradient iteration.

Convergence Theorem (1)

Make the following assumptions.

H1 U^{ad} is a nonempty closed convex subset of a Hilbert space $\mathbb{U}$.

H2 $j : \mathbb{U} \times \mathbb{W} \to \mathbb{R}$ is a normal integrand, and $\mathbb{E}\big(j(u, \boldsymbol{W})\big)$ exists for all $u \in U^{\mathrm{ad}}$.

H3 $j(\cdot, w) : \mathbb{U} \to \mathbb{R}$ is a proper convex differentiable function for all $w \in \mathbb{W}$ ($j(\cdot, w)$ is l.s.c. thanks to Assumption H2).

H4 $j(\cdot, w)$ has linearly bounded gradients (LBG):

$$\exists c_1, c_2 > 0,\ \forall (u, w) \in U^{\mathrm{ad}} \times \mathbb{W},\ \|\nabla_u j(u, w)\| \leq c_1 \|u\| + c_2 \ .$$

H5 J is Lipschitz continuous and coercive on U^{ad}.

H6 K is a proper l.s.c. function, strongly convex with modulus b and differentiable.

H7 $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ is a σ-sequence.

Convergence Theorem (2)

Then the following conclusions hold true.

R1 The initial problem has a non empty set of solutions $U^\sharp$.

R2 Each auxiliary problem has a unique solution $\boldsymbol{U}^{(k+1)}$.

R3 The sequence of random variables $\{J(\boldsymbol{U}^{(k)})\}_{k\in\mathbb{N}}$ almost surely converges to $J^\sharp = \min_{u\in U^{\mathrm{ad}}} J(u)$.

R4 The sequence of random variables $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ is almost surely bounded, and every cluster point of a realization of this sequence almost surely belongs to the optimal set $U^\sharp$.

At last, if J is strongly convex, then $U^\sharp$ reduces to a singleton $\{u^\sharp\}$ and the sequence $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ almost surely converges to the unique solution $u^\sharp$ of the initial problem.

Sketch of Proof

The proof of the 1st statement is based on optimization theorems.

The proof of the 2nd statement involves measurability arguments.

The proof of the last two statements consists of three steps.

1. **Select a Lyapunov function**.
   Here we choose $\Lambda(u) = K(u^\sharp) - K(u) - \left\langle \nabla K(u) \, , u^\sharp - u \right\rangle$.
2. **Bound from above the variation of Λ.**
   Using assumptions and writing optimality conditions, we get:
   $$\mathbb{E}\left(\Lambda(\boldsymbol{U}^{(k+1)}) \mid \mathcal{F}^{(k)}\right) \leq (1+\alpha^{(k)})\Lambda(\boldsymbol{U}^{(k)}) + \beta^{(k)} - \epsilon^{(k)}\left(J(\boldsymbol{U}^{(k)}) - J(u^\sharp)\right) .$$
3. **Prove the convergence of the sequences.**
   Using two technical lemmas, we obtain that $\{\Lambda(\boldsymbol{U}^{(k)})\}_{k\in\mathbb{N}}$ almost surely converges to a finite random variable $\boldsymbol{\Lambda}^\infty$, and that $\{J(\boldsymbol{U}^{(k)})\}_{k\in\mathbb{N}}$ almost surely converges to $J(u^\sharp)$. Using a compactness argument, it exists subsequences of $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ converging almost surely to elements belonging to the set $U^\sharp$.

Two Useful Lemmas (1)

Robbins-Siegmund Theorem

Let $\{\boldsymbol{\Lambda}^{(k)}\}_{k\in\mathbb{N}}$, $\{\boldsymbol{\alpha}^{(k)}\}_{k\in\mathbb{N}}$, $\{\boldsymbol{\beta}^{(k)}\}_{k\in\mathbb{N}}$ and $\{\boldsymbol{\eta}^{(k)}\}_{k\in\mathbb{N}}$ be four positive sequences of real-valued random variables adapted to the filtration $\{\mathcal{F}^{(k)}\}_{k\in\mathbb{N}}$. Assume that

$$\mathbb{E}\big(\boldsymbol{\Lambda}^{(k+1)} \,\big|\, \mathcal{F}^{(k)}\big) \leq \big(1+\boldsymbol{\alpha}^{(k)}\big)\boldsymbol{\Lambda}^{(k)} + \boldsymbol{\beta}^{(k)} - \boldsymbol{\eta}^{(k)} \;,\quad \forall k\in\mathbb{N}\;,$$

and that

$$\sum_{k\in\mathbb{N}} \boldsymbol{\alpha}^{(k)} < +\infty \quad \text{and} \quad \sum_{k\in\mathbb{N}} \boldsymbol{\beta}^{(k)} < +\infty \;,\ \mathbb{P}\text{-a.s.}\;.$$

Then, the sequence $\{\boldsymbol{\Lambda}^{(k)}\}_{k\in\mathbb{N}}$ almost surely converges to a finite[a] random variable $\boldsymbol{\Lambda}^{\infty}$, and we have that $\sum_{k\in\mathbb{N}} \boldsymbol{\eta}^{(k)} < +\infty$, $\mathbb{P}$-a.s..

[a] A random variable $\boldsymbol{X}$ is finite if $\mathbb{P}\big(\{\omega\in\Omega \mid \boldsymbol{X}(\omega) = +\infty\}\big) = 0$.

Two Useful Lemmas (2)

Technical Lemma

Let J be a real-valued function defined on a Hilbert space $\mathbb{U}$.
We assume that J is Lipschitz continuous with constant L.
Let $\{u^{(k)}\}_{k\in\mathbb{N}}$ be a sequence of elements of $\mathbb{U}$ and let $\{\epsilon^{(k)}\}_{k\in\mathbb{N}}$ be a sequence of positive real numbers such that

$$\text{(a)} \sum_{k\in\mathbb{N}} \epsilon^{(k)} = +\infty,$$

$$\text{(b)} \exists\, \mu \in \mathbb{R},\ \sum_{k\in\mathbb{N}} \epsilon^{(k)} \left| J(u^{(k)}) - \mu \right| < +\infty,$$

$$\text{(c)} \exists\, \delta > 0,\ \forall k \in \mathbb{N},\ \left\| u^{(k+1)} - u^{(k)} \right\| \leq \delta \epsilon^{(k)}.$$

Then the sequence $\left\{J(u^{(k)})\right\}_{k\in\mathbb{N}}$ converges to μ.

Proof of Convergence (1)

The proof of the first statement is based on standard theorems in the field of convex optimization ensuring the existence of solutions in a general Hilbert space. Let $u^\sharp \in U^\sharp$ be a solution of the initial problem.

The existence of a r.v. $\boldsymbol{U}^{(k+1)}$ solution of the auxiliary problem

$$\min_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \epsilon^{(k)} \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) - \nabla K(\boldsymbol{U}^{(k)}) \, , u \right\rangle \, ,$$

is a consequence of the fact that the criterion to be minimized is a normal integrand. The arg min is a closed-valued and measurable multifunction and thus at least admits a measurable selection (see [Rockafellar & Wets, 1998, Theorem 14.37] for further details).

The solution $\boldsymbol{U}^{(k+1)}$ is unique because K is strongly convex.

Proof of Convergence (2)

- Let $\Lambda(u) = K(u^\sharp) - K(u) - \left\langle \nabla K(u) \, , u^\sharp - u \right\rangle$. We have

$$\Lambda(u) \geq \frac{b}{2} \left\| u - u^\sharp \right\|_{\mathbb{U}}^2 , \tag{7}$$

(strong convexity of K) so that Λ is bounded from below.

- Consider the variation of Λ during the algorithm:

$$\begin{aligned} \boldsymbol{\Delta}^{(k)} &= \Lambda(\boldsymbol{U}^{(k+1)}) - \Lambda(\boldsymbol{U}^{(k)}) \\ &= \underbrace{K(\boldsymbol{U}^{(k)}) - K(\boldsymbol{U}^{(k+1)}) - \left\langle \nabla K(\boldsymbol{U}^{(k)}) \, , \boldsymbol{U}^{(k)} - \boldsymbol{U}^{(k+1)} \right\rangle}_{T_1} \\ &\qquad + \underbrace{\left\langle \nabla K(\boldsymbol{U}^{(k)}) - \nabla K(\boldsymbol{U}^{(k+1)}) \, , u^\sharp - \boldsymbol{U}^{(k+1)} \right\rangle}_{T_2} . \end{aligned}$$

From the convexity of K, we have that $T_1 \leq 0$.

Proof of Convergence (3)

Let $\boldsymbol{G}^{(k)} = \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)})$. From the optimality conditions of the auxiliary problem evaluated at $u^\sharp$, we have that

$$T_2 \leq \epsilon^{(k)} \underbrace{\left\langle \boldsymbol{G}^{(k)}, u^\sharp - \boldsymbol{U}^{(k)} \right\rangle}_{T_3} + \epsilon^{(k)} \underbrace{\left\langle \boldsymbol{G}^{(k)}, \boldsymbol{U}^{(k)} - \boldsymbol{U}^{(k+1)} \right\rangle}_{T_4} .$$

- From the convexity of $j(\cdot, w)$, we have that
$$T_3 \leq j(u^\sharp, \boldsymbol{W}^{(k+1)}) - j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) .$$
- The optimality condition at $\boldsymbol{U}^{(k)}$ and the strong convexity of K lead to: $\epsilon^{(k)} \left\langle \boldsymbol{G}^{(k)}, \boldsymbol{U}^{(k)} - \boldsymbol{U}^{(k+1)} \right\rangle \geq b \left\| \boldsymbol{U}^{(k+1)} - \boldsymbol{U}^{(k)} \right\|_{\mathbb{U}}^2$.

  Using the Schwartz inequality, we obtain: $T_4 \leq \frac{\epsilon^{(k)}}{b} \left\| \boldsymbol{G}^{(k)} \right\|_{\mathbb{U}}^2$.

  The LBG assumption and the majoration (7) of Λ yield:
$$T_4 \leq \frac{\epsilon^{(k)}}{b} \left(\alpha \Lambda(\boldsymbol{U}^{(k)}) + \beta \right) .$$

Proof of Convergence (4)

Collecting the upper bounds obtained for T_1, T_3 and T_4 leads to

$$\boldsymbol{\Delta}^{(k)} \leq \epsilon^{(k)} \big(j(u^\sharp, \boldsymbol{W}^{(k+1)}) - j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) \big) + \frac{\big(\epsilon^{(k)}\big)^2}{b} \big(\alpha \Lambda(\boldsymbol{U}^{(k)}) + \beta \big) \, .$$

Taking the conditional expectation w.r.t. the σ-field $\mathcal{F}^{(k)}$ generated by $(\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(k)})$, we obtain that[a]

$$\mathbb{E}\big(\Lambda(\boldsymbol{U}^{(k+1)}) \mid \mathcal{F}^{(k)} \big) \leq \big(1 + \alpha^{(k)}\big) \Lambda(\boldsymbol{U}^{(k)}) + \beta^{(k)} + \epsilon^{(k)} \big(J(u^\sharp) - J(\boldsymbol{U}^{(k)}) \big) \, ,$$

with $\alpha^{(k)} = (\alpha/b)(\epsilon^{(k)})^2$ and $\beta^{(k)} = (\beta/b)(\epsilon^{(k)})^2$.

[a] Recall that $\boldsymbol{W}^{(k+1)}$ is independent of $\mathcal{F}^{(k)}$ and that $\boldsymbol{U}^{(k)}$ is $\mathcal{F}^{(k)}$-measurable.

Reminder. We have also obtained the two following inequalities:

$$\Lambda(\boldsymbol{U}^{(k)}) \geq \frac{b}{2} \big\| \boldsymbol{U}^{(k)} - u^\sharp \big\|_{\mathbb{U}}^2 \quad \text{and} \quad \big\| \boldsymbol{U}^{(k)} - \boldsymbol{U}^{(k+1)} \big\|_{\mathbb{U}} \leq \frac{\epsilon^{(k)}}{b} \big\| \boldsymbol{G}^{(k)} \big\|_{\mathbb{U}} \, .$$

Proof of Convergence (5)

- From the Robbins-Siegmund theorem, $\left\{\Lambda(\boldsymbol{U}^{(k)})\right\}_{k\in\mathbb{N}}$ almost surely converges to a finite random variable $\boldsymbol{\Lambda}^{\infty}$ and we have

$$\sum_{k=0}^{+\infty} \epsilon^{(k)} \left(J(\boldsymbol{U}^{(k)}) - J(u^{\sharp})\right) < +\infty \ , \ \ \mathbb{P}\text{-a.s. .}$$

  Let Ω_0 denote the subset of Ω such that the two almost sure properties mentioned above are fulfilled: $\mathbb{P}(\Omega_0) = 1$.
- We deduce that both sequences $\left\{\boldsymbol{U}^{(k)}\right\}_{k\in\mathbb{N}}$ and $\left\{\boldsymbol{G}^{(k)}\right\}_{k\in\mathbb{N}}$ are a.s. bounded, so that the same holds true for the sequence $\left\{\frac{1}{\epsilon^{(k)}}\|\boldsymbol{U}^{(k+1)} - \boldsymbol{U}^{(k)}\|_{\mathbb{U}}\right\}_{k\in\mathbb{N}}$. This makes it possible to use the second technical lemma and claim that $\left\{J(\boldsymbol{U}^{(k)})\right\}_{k\in\mathbb{N}}$ almost surely converges to $J(u^{\sharp})$.

Proof of Convergence (6)

- Pick some $\omega \in \Omega_0$. The sequence of realizations $\{u^{(k)}\}_{k\in\mathbb{N}}$ of $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ associated with ω is bounded, and $u^{(k)} \in U^{\mathrm{ad}}$. By a compactness argument,[6] it exists a convergent subsequence $\{u^{(\Phi(k))}\}_{k\in\mathbb{N}}$, with limit $\bar{u}$. Using the lower semi-continuity of function J, we have that

$$J(\bar{u}) \leq \liminf_{k\to+\infty} J(u^{(\Phi(k))}) = J(u^\sharp) \; .$$

Since $\bar{u} \in U^{\mathrm{ad}}$, we deduce that $\bar{u} \in U^\sharp$.

[6] A subset $U^{\mathrm{ad}} \subset \mathbb{U}$ is compact if it is closed and bounded, provided that $\mathbb{U}$ is a finite-dimensional Hilbert space. If U^{ad} is an infinite-dimensional Hilbert space, such a property remains true only in the weak topology. If U^{ad} is closed in the strong topology and is convex, then it is also closed in the weak topology, and hence compact if bounded. In the same vein, the l.s.c. property of J is preserved in the weak topology if J is convex (see [Ekeland & Temam, 1999]).

Proof of Convergence (7)

We ultimately consider the case when J is strongly convex with modulus a. Then the initial problem has a unique solution $u^\sharp$. Thanks to the strong convexity property of J, we have

$$J(\boldsymbol{U}^{(k)}) - J(u^\sharp) \geq \left\langle \nabla J(u^\sharp) \, , \boldsymbol{U}^{(k)} - u^\sharp \right\rangle + \frac{a}{2} \| \boldsymbol{U}^{(k)} - u^\sharp \|_{\mathbb{U}}^2$$
$$\geq \frac{a}{2} \| \boldsymbol{U}^{(k)} - u^\sharp \|_{\mathbb{U}}^2 \, .$$

Since $J(\boldsymbol{U}^{(k)})$ converges almost surely to $J(u^\sharp)$, we deduce that $\| \boldsymbol{U}^{(k)} - u^\sharp \|_{\mathbb{U}}$ almost surely converges to zero.

The proof is complete.

Conclusions

The stochastic APP algorithm encompasses the stochastic gradient algorithm (obtained using $K(u) = \|u\|^2/2$), as well as the so-called matrix-gain algorithm (K being in this case $K(u) = \langle u\,, Au\rangle\,/2$ and A being a positive definite matrix).

From a theoretical point of view, the convergence theorem has been proved under natural assumptions. As a matter of fact, the convexity and differentiability assumptions are standard in the framework of convex optimization. Note that, even if an explicit convexity property is not required in the Robbins-Monro theorem, another assumption playing a very similar role is used.

As far as decomposition is concerned, the stochastic APP algorithm opens this possibility as a way to solve large stochastic optimization problems. Of course, the convergence remains slow because it is driven by the σ-sequence $\{\epsilon^{(k)}\}_{k\in\mathbb{N}}$.

A Constrained Stochastic Optimization Problem

Recall that in our stochastic optimization setting, the probability space is denoted $(\Omega, \mathcal{A}, \mathbb{P})$, and $\boldsymbol{W}$ is a random variable defined on the space $(\mathbb{W}, \mathcal{W})$.

- We are interested in the case where the criterion J is defined as $J(u) = \mathbb{E}\big(j(u, \boldsymbol{W})\big)$, with $j : \mathbb{U} \times \mathbb{W} \to \overline{\mathbb{R}}$. This is the standard framework when studying open-loop stochastic optimization problems.
- We will hereafter consider only constraints Θ which are of a deterministic nature: $\Theta : \mathbb{U} \to \mathbb{V}$.
- The problem we deal with has thus the following expression:

$$\min_{u \in U^{\mathrm{ad}}} \mathbb{E}\big(j(u, \boldsymbol{W})\big) \quad \text{subject to} \quad \Theta(u) \in -C \ .$$

Nature of Constraints in Stochastic Optimization

Constraints in stochastic optimization arise in different ways, have different meanings and need various mathematical treatments.

- A constraint may be deterministic: $\Theta(u) \in -C$.
- A constraint may be formulated in the almost sure sense: $\theta(u, \boldsymbol{W}) \in -C \ \ \mathbb{P}$-a.s.. It is generally used to express hard constraints (physical laws, ...).
- Another (more realistic) way is to formulate stochastic constraints in probability: $\mathbb{P}\big(\theta(u, \boldsymbol{W}) \in -C\big) \geq \pi$, which means that the constraints can sometimes be violated.
- Another possibility is to have a constraint in expectation: $\mathbb{E}\big(\theta(u, \boldsymbol{W})\big) \in -C$. Although usually non intuitive, such a formulation proves useful in some specific problems, as it will be illustrated in the lecture next week.

Problem under Consideration

Consider the following convex optimization problem:

$$\min_{u \in U^{\mathrm{ad}} \subset \mathbb{U}} J(u) \quad \text{subject to} \quad \Theta(u) \in -C \subset \mathbb{V} \,,$$

where U^{ad} is a closed convex subset of an Hilbert space $\mathbb{U}$, and where C is a closed convex salient cone of another Hilbert space $\mathbb{V}$.

Let $C^\star$ be the dual cone[a] of C. We introduce the Lagrangian L of the constrained optimization problem, defined on $U^{\mathrm{ad}} \times C^\star$:

$$L(u,p) = J(u) + \big\langle p \,, \Theta(u) \big\rangle \,.$$

Under standard convexity and continuity assumptions, and under a Constraint Qualification Condition, solving the initial problem is equivalent to determining a saddle point of the Lagrangian L.

[a] defined as $C^\star = \big\{ p \in \mathbb{V},\ \langle p \,, v \rangle \geq 0 \ \ \forall v \in C \big\}$.

Uzawa and Arrow-Hurwicz Algorithms

Assuming that a saddle point of L exists, the initial problem is equivalent to the following dual problem:

$$\max_{p \in C^\star} \left(\min_{u \in U^{\mathrm{ad}}} L(u,p) \right) .$$

This problem can be solved by using the Uzawa algorithm:

$$u^{(k+1)} \in \arg\min_{u \in U^{\mathrm{ad}}} J(u) + \left\langle p^{(k)} , \Theta(u) \right\rangle ,$$
$$p^{(k+1)} = \mathrm{proj}_{C^\star} \left(p^{(k)} + \rho \, \Theta(u^{(k+1)}) \right) .$$

Another possibility is to use the Arrow-Hurwicz algorithm:

$$u^{(k+1)} = \mathrm{proj}_{U^{\mathrm{ad}}} \left(u^{(k)} - \epsilon \left(\nabla J(u^{(k)}) + \left(\Theta'(u^{(k)}) \right)^\top p^{(k)} \right) \right) ,$$
$$p^{(k+1)} = \mathrm{proj}_{C^\star} \left(p^{(k)} + \rho \, \Theta(u^{(k+1)}) \right) .$$

APP with Explicit Constraints

A "natural" extension of the APP to constrained optimization problems consists in choosing a function $K : \mathbb{U} \to \mathbb{R}$ and then replacing the resolution of the initial problem by the resolution of the following sequence of auxiliary problems:[7]

$$u^{(k+1)} = \underset{u \in U^{\mathrm{ad}}}{\arg\min}\, K(u) + \left\langle \epsilon \nabla J(u^{(k)}) - \nabla K(u^{(k)}) \,, u \right\rangle + \epsilon \left\langle p^{(k)} \,, \Theta(u) \right\rangle ,$$

$$p^{(k+1)} = \mathrm{proj}_{C^\star} \left(p^{(k)} + \rho\, \Theta(u^{(k+1)}) \right) .$$

It is not difficult to show that this APP framework encompasses

- the Uzawa algorithm (using $K(u) = J(u)$ and $\epsilon = 1$),
- the Arrow-Hurwicz algorithm (using $K(u) = \|u\|^2 / 2$).

Moreover, choosing an additive core K allows for decomposition in the minimization stage of the APP algorithm.

[7] Note that the term $\epsilon \left\langle p^{(k)} \,, \Theta(u) \right\rangle$ may be replaced by $\epsilon \left\langle p^{(k)} \,, \Theta'(u^{(k)}).u \right\rangle$.

Lagrangian Stability

The standard Lagrangian approach has the following drawback. Assume that $p^\sharp$ is a solution of the dual problem, and consider the set $\widehat{U}(p^\sharp)$ of solutions associated to the primal minimization:

$$\widehat{U}(p^\sharp) = \arg\min_{u \in U^{\mathrm{ad}}} L(u, p^\sharp) \ .$$

Then the set $U^\sharp$ of solutions of the initial problem may be strictly included in $\widehat{U}(p^\sharp)$, as illustrated by the following (linear) example:

$$\min_{u \in [-1,1]} \ -u \quad \text{s.t.} \quad u = 0 \ ,$$

whose unique saddle point is $\{0\} \times \{1\}$ whereas $\widehat{U}(1) = [-1, 1]$.

A solution $\widehat{u} \in \widehat{U}(p^\sharp)$ induced by the dual problem is not always a solution of the initial problem (**stability of the Lagrangian**)!

Augmented Lagrangian

A remedy to this difficulty is to use a different duality theory, based on the idea of regularization. This idea leads to a new Lagrangian L_c which is called the augmented Lagrangien. The Lagrangian L_c is defined on the set $U^{\mathrm{ad}} \times \mathbb{V}$,[8] and its expression, which depends on a scalar parameter $c > 0$, is given by:

$$L_c(u,p) = J(u) + \frac{1}{2c}\left(\|\mathrm{proj}_{C^\star}(p + c\Theta(u))\|^2 - \|p\|^2 \right) .$$

The augmented Lagrangian has the two following properties.

1. The standard Lagrangian L and the augmented Lagrangian L_c have the same set of saddle points.
2. The augmented Lagrangian L_c is always stable:

$$U^\sharp = \arg\min_{u \in U^{\mathrm{ad}}} L_c(u, p^\sharp) .$$

[8] whereas the standard Lagrangian L is defined on $U^{\mathrm{ad}} \times C^\star$

APP and Augmented Lagrangian

The solution of the initial problem can be obtained by solving the "augmented" dual problem:

$$\max_{p \in \mathbb{V}} \left(\min_{u \in U^{\mathrm{ad}}} L_c(u,p) \right) .$$

The extension of the APP to that dual problem consists in solving the following sequence of auxiliary problems (see [Cohen, 2004]):

$$u^{(k+1)} = \arg\min_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \epsilon \nabla J(u^{(k)}) - \nabla K(u^{(k)}) \, , u \right\rangle + \epsilon \left\langle \mathrm{proj}_{C^\star} \left(p^{(k)} + c\Theta(u^{(k)}) \right) , \Theta(u) \right\rangle ,$$

$$p^{(k+1)} = \left(1 - \frac{\rho}{c} \right) p^{(k)} + \frac{\rho}{c} \, \mathrm{proj}_{C^\star} \left(p^{(k)} + c\Theta(u^{(k+1)}) \right) .$$

On the Agenda

1. **Extension of the Uzawa Algorithm.**
2. **Stochastic APP Algorithm with Constraints.**
3. **Stochastic APP and Augmented Lagrangian.**
4. **What happens if** $\Theta(u) = \mathbb{E}\big(\theta(u, \boldsymbol{W})\big)$?

A Useful Tool in Stochastic Approximation

In the context of Stochastic Approximation, strong connections exist between the convergence of the standard SA algorithm:

$$\boldsymbol{U}^{(k+1)} = \boldsymbol{U}^{(k)} + \epsilon^{(k)} \left(h(\boldsymbol{U}^{(k)}) + \boldsymbol{\xi}^{(k+1)} \right) ,$$

and the behavior of the ordinary differential equation (ODE) associated to this algorithm:

$$\dot{u} = h(u) ,$$

(see [Kushner & Clark, 1978]). A useful corollary is the following.

Let $\{\boldsymbol{U}^{(k)}\}_{k \in \mathbb{N}}$ be the sequence generated by the Stochastic Approximation algorithm, and assume that

$$\exists u^\sharp \in \mathbb{U}, \text{ such that } \mathbb{P}\left(\lim_{k \to +\infty} \boldsymbol{U}^{(k)} = u^\sharp \right) > 0 .$$

Then $u^\sharp$ is a stable equilibrium point of the associated ODE.

Extension of the Uzawa Algorithm to the Stochastic Case

Our first attempt for solving the stochastic constrained problem:

$$\min_{u \in U^{\mathrm{ad}}} \mathbb{E}\big(j(u, \boldsymbol{W})\big) \quad \text{subject to} \quad \Theta(u) \in -C \ ,$$

is to propose an extension of the Uzawa algorithm. More precisely, during the minimization stage w.r.t. u, we propose to replace the expectation $J(u)$ by the value $j(u, w^{(k+1)})$.[9] We thus obtain a tentative Stochastic Uzawa Algorithm:

$$\boldsymbol{U}^{(k+1)} = \arg\min_{u \in U^{\mathrm{ad}}} j(u, \boldsymbol{W}^{(k+1)}) + \left\langle \boldsymbol{P}^{(k)} \ , \Theta(u) \right\rangle \ ,$$

$$\boldsymbol{P}^{(k+1)} = \mathrm{proj}_{C^\star}\left(\boldsymbol{P}^{(k)} + \rho^{(k)}\, \Theta(\boldsymbol{U}^{(k+1)})\right) \ .$$

Question: what about the convergence of this algorithm?

[9] Note that we have replaced here the evaluation of J by the one of j, whereas ∇J is replaced by $\nabla_u j$ in the stochastic gradient method...

Stochastic Uzawa Algorithm Counter-Example (1)

Consider a constrained stochastic optimization problem with:

- $\mathbb{U} = \mathbb{R}^2$ and $U^{\mathrm{ad}} = \mathbb{U}$,
- $\mathbb{V} = \mathbb{R}$ and $C = \{0\}$ (equality constraint),
- $\mathbb{W} = \mathbb{R}^4$ and $\boldsymbol{W} = (\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{B}_1, \boldsymbol{B}_2)$,
- $j(u, w) = \frac{1}{2}\left(a_1 {u_1}^2 + a_2 {u_2}^2\right) + \left(b_1 u_1 + b_2 u_2\right)$,
- $\Theta(u) = \theta_1 u_1 + \theta_2 u_2$.

The optimality conditions (KKT) of this problem write:

$$\mathbb{E}\left(\boldsymbol{A}_1\right) u_1^\sharp + \mathbb{E}\left(\boldsymbol{B}_1\right) + \theta_1 p^\sharp = 0 \ , \ \mathbb{E}\left(\boldsymbol{A}_2\right) u_2^\sharp + \mathbb{E}\left(\boldsymbol{B}_2\right) + \theta_2 p^\sharp = 0 \ , \ \theta_1 u_1^\sharp + \theta_2 u_2^\sharp = 0 \ ,$$

so that the value of the optimal multiplier is:

$$p^\sharp = - \frac{\dfrac{\mathbb{E}\left(\boldsymbol{B}_1\right)}{\mathbb{E}\left(\boldsymbol{A}_1\right)} \theta_1 + \dfrac{\mathbb{E}\left(\boldsymbol{B}_2\right)}{\mathbb{E}\left(\boldsymbol{A}_2\right)} \theta_2}{\dfrac{\theta_1^2}{\mathbb{E}\left(\boldsymbol{A}_1\right)} + \dfrac{\theta_2^2}{\mathbb{E}\left(\boldsymbol{A}_2\right)}} \ .$$

Stochastic Uzawa Algorithm Counter-Example (2)

Apply our Uzawa algorithm. The minimization stage leads to:

- $\boldsymbol{A}_1^{(k+1)} \boldsymbol{U}_1^{(k+1)} + \boldsymbol{B}_1^{(k+1)} + \theta_1 \boldsymbol{P}^{(k)} = 0,$
- $\boldsymbol{A}_2^{(k+1)} \boldsymbol{U}_2^{(k+1)} + \boldsymbol{B}_2^{(k+1)} + \theta_2 \boldsymbol{P}^{(k)} = 0,$

and the update of the multiplier writes:

- $\boldsymbol{P}^{(k+1)} = \boldsymbol{P}^{(k)} + \rho^{(k)} \left(\theta_1 \boldsymbol{U}_1^{(k+1)} + \theta_2 \boldsymbol{U}_2^{(k+1)} \right).$

We thus obtain

$$\boldsymbol{P}^{(k+1)} = \boldsymbol{P}^{(k)} - \rho^{(k)} \left(\frac{\theta_1^2}{\boldsymbol{A}_1^{(k+1)}} + \frac{\theta_2^2}{\boldsymbol{A}_2^{(k+1)}} \right) \boldsymbol{P}^{(k)} - \rho^{(k)} \left(\theta_1 \frac{\boldsymbol{B}_1^{(k+1)}}{\boldsymbol{A}_1^{(k+1)}} + \theta_2 \frac{\boldsymbol{B}_2^{(k+1)}}{\boldsymbol{A}_2^{(k+1)}} \right).$$

$\left\{ \boldsymbol{P}^{(k)} \right\}_{k \in \mathbb{N}}$ can only converge to a stable equilibrium point of the associated differential equation (ODE argument), that is,

$$\overline{p} = - \frac{\mathbb{E}\left(\frac{\boldsymbol{B}_1}{\boldsymbol{A}_1} \right) \theta_1 + \mathbb{E}\left(\frac{\boldsymbol{B}_2}{\boldsymbol{A}_2} \right) \theta_2}{\frac{\theta_1^2}{\mathbb{E}\left(\boldsymbol{A}_1 \right)} + \frac{\theta_2^2}{\mathbb{E}\left(\boldsymbol{A}_2 \right)}}.$$

Stochastic Uzawa Algorithm Counter-Example (3)

As soon as the random variables $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are non independent, we have

$$p^\sharp = -\frac{\frac{\mathbb{E}\left(\boldsymbol{B}_1\right)}{\mathbb{E}\left(\boldsymbol{A}_1\right)}\theta_1 + \frac{\mathbb{E}\left(\boldsymbol{B}_2\right)}{\mathbb{E}\left(\boldsymbol{A}_2\right)}\theta_2}{\frac{\theta_1^2}{\mathbb{E}\left(\boldsymbol{A}_1\right)} + \frac{\theta_2^2}{\mathbb{E}\left(\boldsymbol{A}_2\right)}} \neq -\frac{\mathbb{E}\left(\frac{\boldsymbol{B}_1}{\boldsymbol{A}_1}\right)\theta_1 + \mathbb{E}\left(\frac{\boldsymbol{B}_2}{\boldsymbol{A}_2}\right)\theta_2}{\frac{\theta_1^2}{\mathbb{E}\left(\boldsymbol{A}_1\right)} + \frac{\theta_2^2}{\mathbb{E}\left(\boldsymbol{A}_2\right)}} = \overline{p}\,.$$

The stochastic Uzawa algorithm does not solve the problem!

Remark. The standard stochastic gradient produces an averaging effect on the iterates $\boldsymbol{U}^{(k)}$ by means of the coefficients $\epsilon^{(k)}$. In the Uzawa algorithm, $\boldsymbol{U}^{(k)}$ is obtained by a minimization procedure which does not incorporate any averaging effect, hence the failure.

Stochastic APP Algorithm with Constraints

Consider the APP algorithm in the deterministic setting:

$$u^{(k+1)} = \arg\min_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \epsilon \nabla J(u^{(k)}) - \nabla K(u^{(k)}) \, , u \right\rangle + \epsilon \left\langle p^{(k)} \, , \Theta(u) \right\rangle ,$$

$$p^{(k+1)} = \mathrm{proj}_{C^\star} \left(p^{(k)} + \rho \, \Theta(u^{(k+1)}) \right) .$$

The extension to the stochastic case is obtained in a canonical way by replacing in the above minimization stage the gradient of J by the partial gradient of j w.r.t. u, evaluated at a sample $\boldsymbol{W}^{(k+1)}$ of $\boldsymbol{W}$. Using the notation $\boldsymbol{G}^{(k)} = \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)})$, we obtain:

Stochastic APP Algorithm in the Constrained Case

$$\boldsymbol{U}^{(k+1)} = \arg\min_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \epsilon^{(k)} \boldsymbol{G}^{(k)} - \nabla K(\boldsymbol{U}^{(k)}) \, , u \right\rangle + \epsilon^{(k)} \left\langle \boldsymbol{P}^{(k)} \, , \Theta(u) \right\rangle ,$$

$$\boldsymbol{P}^{(k+1)} = \mathrm{proj}_{C^\star} \left(\boldsymbol{P}^{(k)} + \epsilon^{(k)} \Theta(\boldsymbol{U}^{(k+1)}) \right) .$$

Note that this never leads to the Uzawa algorithm because $\epsilon^{(k)} \to 0$.

Convergence Theorem (1)

Make the following assumptions.

H1 U^{ad} is a nonempty closed convex subset of a Hilbert space $\mathbb{U}$, and C is a closed convex salient cone of a Hilbert space $\mathbb{V}$.

H2 $j : \mathbb{U} \times \mathbb{W} \to \mathbb{R}$ is a normal integrand, and $\mathbb{E}\big(j(u, \boldsymbol{W})\big)$ exists for all $u \in U^{\mathrm{ad}}$.

H3 $j(\cdot, w) : \mathbb{U} \to \mathbb{R}$ is a proper convex l.s.c. differentiable function with linearly bounded gradients (LBG), for all $w \in \mathbb{W}$.

H4 J is strictly convex, Lipschitz and coercive on U^{ad}.

H5 Θ is C-convex, Lipschitz with constant L_Θ.

H6 A constraint qualification condition holds true.

H7 K is a proper l.s.c. function, strongly convex with modulus b and differentiable.

H8 The sequence $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ is a σ-sequence.

Convergence Theorem (2)

Then the following conclusions hold true.

R1 The initial constrained problem has a non empty set of saddle points $\{u^\sharp\} \times P^\sharp$.

R2 Each auxiliary problem has a unique solution $\boldsymbol{U}^{(k+1)}$.

R3 The sequence of random variable $\{L(\boldsymbol{U}^{(k)}, p^\sharp)\}_{k\in\mathbb{N}}$ almost surely converges to $L(u^\sharp, p^\sharp)$ for all $p^\sharp \in P^\sharp$.

R4 The sequences of r.v. $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ and $\{\boldsymbol{P}^{(k)}\}_{k\in\mathbb{N}}$ are almost surely bounded, and the sequence $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ almost surely converges to $u^\sharp$.

Sketch of Proof

The proof of the first two statement is based on standard theorems.

The proof of the last two statements consists of three steps.

1. **Select a Lyapunov function**.
   $\Lambda(u,p) = K(u^\sharp) - K(u) - \left\langle \nabla K(u)\,, u^\sharp - u \right\rangle + \left\| p - p^\sharp \right\|^2 /2.$
2. **Bound from above the variation of $\boldsymbol{\Lambda}$**.
   Using assumptions and writing optimality conditions, we get:
   $$\mathbb{E}\left(\Lambda(\boldsymbol{U}^{(k+1)}, \boldsymbol{P}^{(k+1)}) \mid \mathcal{F}^{(k)}\right) \leq (1+\alpha^{(k)})\Lambda(\boldsymbol{U}^{(k)}, \boldsymbol{P}^{(k)}) + \beta^{(k)} - \epsilon^{(k)}\left(L(\boldsymbol{U}^{(k)}, p^\sharp) - L(u^\sharp, p^\sharp)\right).$$
3. **Prove the convergence of the sequences.**
   Using the two lemmas, we obtain that $\{\Lambda(\boldsymbol{U}^{(k)}, \boldsymbol{P}^{(k)})\}_{k\in\mathbb{N}}$ almost surely converges to a finite random variable $\boldsymbol{\Lambda}^\infty$, and that $\{L(\boldsymbol{U}^{(k)}, p^\sharp)\}_{k\in\mathbb{N}}$ almost surely converges to $L(u^\sharp, p^\sharp)$.
   By a compactness argument and uniqueness of $u^\sharp$, the sequence $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ almost surely converges to $u^\sharp$.

Stochastic APP and Augmented Lagrangian

In order to deal with non stable problems, we extend the use of the Augmented Lagrangian to the stochastic framework. The extension is obtained by replacing the gradient of J by the partial gradient of j w.r.t. u. Using the notation $\boldsymbol{G}^{(k)} = \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)})$, we obtain:

Regularized Stochastic APP algorithm with Constraints

$$\boldsymbol{U}^{(k+1)} = \arg\min_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \epsilon^{(k)} \boldsymbol{G}^{(k)} - \nabla K(\boldsymbol{U}^{(k)}) \, , u \right\rangle + \epsilon^{(k)} \left\langle \mathrm{proj}_{C^\star}\left(\boldsymbol{P}^{(k)} + c\Theta(\boldsymbol{U}^{(k)})\right) , \Theta(u) \right\rangle ,$$

$$\boldsymbol{P}^{(k+1)} = \left(1 - \frac{\epsilon^{(k)}}{c}\right) \boldsymbol{P}^{(k)} + \frac{\epsilon^{(k)}}{c} \mathrm{proj}_{C^\star}\left(\boldsymbol{P}^{(k)} + c\Theta(\boldsymbol{U}^{(k+1)})\right) .$$

Convergence Theorem (1)

Make the following assumptions.

H1 U^{ad} is a nonempty closed convex subset of a Hilbert space $\mathbb{U}$, and C is a closed convex cone of another Hilbert space $\mathbb{V}$.

H2 $j : \mathbb{U} \times \mathbb{W} \to \mathbb{R}$ is a normal integrand, and $\mathbb{E}\big(j(u, \boldsymbol{W})\big)$ exists for all $u \in U^{\mathrm{ad}}$.

H3 $j(\cdot, w) : \mathbb{U} \to \mathbb{R}$ is a proper convex l.s.c. differentiable function with linearly bounded gradients (LBG), for all $w \in \mathbb{W}$.

H4 J is Lipschitz continuous, coercive on U^{ad}.

H5 Θ is C-convex, Lipschitz with constant L_Θ.

H6 A constraint qualification condition holds true.

H7 K is a proper l.s.c. function, strongly convex with modulus b and differentiable.

H8 The sequence $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ is a σ-sequence.

Convergence Theorem (2)

Then the following conclusions hold true.

R1 The initial constrained problem has a non empty set of saddle points $U^\sharp \times P^\sharp$.

R2 Each auxiliary problem has a unique solution $\boldsymbol{U}^{(k+1)}$.

R3 The sequence of r.v. $\{L_c(u^\sharp, \boldsymbol{P}^{(k)}) - L_c(\boldsymbol{U}^{(k)}, p^\sharp)\}_{k\in\mathbb{N}}$ almost surely converges to zero for all saddle point $(u^\sharp, p^\sharp)$.

R4 The sequences of r.v. $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ and $\{\boldsymbol{P}^{(k)}\}_{k\in\mathbb{N}}$ are almost surely bounded, and each cluster point of a realization of the sequence $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ almost surely converges to an element of $U^\sharp$.

Sketch of Proof

The proof of the first two statement is based on standard theorems.

The proof of the last two statements follows the usual scheme, with

$$\Lambda(u,p) = K(u^\sharp) - K(u) - \left\langle \nabla K(u)\,, u^\sharp - u \right\rangle + \frac{1}{2}\left\| p - p^\sharp \right\|^2 ,$$

and a substantial amount of technicalities. . .

Stochastic APP and Constraints in Expectation

We finally aim at solving the following stochastic optimization problem, in which the constraint corresponds to an expectation:

$$\min_{u \in U^{\mathrm{ad}}} \mathbb{E}\big(j(u, \boldsymbol{W})\big) \quad \text{subject to} \quad \mathbb{E}\big(\theta(u, \boldsymbol{W})\big) \in -C \ .$$

In the spirit of the stochastic gradient method, we use values of θ evaluated in realizations of $\boldsymbol{W}$ rather than expected values of Θ. With $\boldsymbol{G}^{(k)} = \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)})$ and $\boldsymbol{\vartheta}^{(k)} = \theta'_u(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)})$, the extension of the APP method is:

Stochastic APP Algorithm with Expected Constraints

$$\boldsymbol{U}^{(k+1)} \in \arg\min_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \epsilon^{(k)} \boldsymbol{G}^{(k)} - \nabla K(\boldsymbol{U}^{(k)}) \, , u \right\rangle + \epsilon^{(k)} \left\langle \boldsymbol{P}^{(k)} \, , \boldsymbol{\vartheta}^{(k)} \cdot u \right\rangle \, ,$$

$$\boldsymbol{P}^{(k+1)} = \mathrm{proj}_{C^\star} \left(\boldsymbol{P}^{(k)} + \rho^{(k)} \, \theta(\boldsymbol{U}^{(k+1)}, \boldsymbol{W}^{(k+1)}) \right) \ .$$

Convergence Theorem and Proof

Long and intricate. . . See the lecture notes.