Lecture 5: Algorithmic models of human behavior

Yurii Nesterov, CORE/INMA (UCL)

March 14, 2012

Main problem with the Rational Choice

- Rational choice assumption is introduced for better understanding and predicting the human behavior.
- It forms the basis of Neoclassical Economics (1900).
- The player (*Homo Economicus* $\equiv$ HE) wants to maximize his *utility function* by an appropriate adjustment of the consumption pattern.
- As a consequence, we can speak about *equilibrium* in economical systems.
- Existing literature is immense. It concentrates also on ethical, moral, religious, social, and other consequences of rationality.

  (HE = super-powerful aggressively selfish immoral individualist.)

NB: The only missing topic is the *Algorithmic Aspects* of rationality.

What do we know now?

- Starting from 1977 (Complexity Theory, Nemirovski & Yudin), we know that optimization problems in general are *unsolvable*.
- They are very difficult (and will be always difficult) for computers, independently on their speed.
- How they can be solved by us, taking into account our natural weakness in arithmetics?

NB: Mathematical consequences of unreasonable assumptions can be disastrous.

Perron paradox: The maximal integer is equal to one.
Proof: Denote by N the maximal integer. Then

$$1 \leq N \leq N^2 \leq N.$$

Hence, $N = 1$. □

What we do not know

- In which sense the human beings can solve the optimization problems?
- What is the accuracy of the solution?
- What is the convergence rate?

Main question: What are the optimization *methods*?

NB:

- Forget about Simplex Algorithm and Interior Point Methods!
- Be careful with gradients (dimension, non-smoothness).

Outline

1. Intuitive optimization (Random Search)
2. Rational activity in stochastic environment (Stochastic Optimization)
3. Models and algorithms of rational behavior

Intuitive Optimization

Problem: $\min_{x \in R^n} f(x)$, where x is the *consumption pattern.*

Main difficulties:

- High dimension of x (difficult to evaluate/observe).
- Possible non-smoothness of $f(x)$.

Theoretical advice: apply gradient method

$$x_{k+1} = x_k - hf'(x_k).$$

(In the space of all available products!)

Hint: we live in an uncertain world.

Gaussian smoothing

Let $f : E \to R$ be differentiable along any direction at any $x \in E$. Let us form its *Gaussian approximation*

$$f_\mu(x) \;=\; \frac{1}{\kappa} \int\limits_E f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du,$$

where $\kappa \stackrel{\mathrm{def}}{=} \int\limits_E e^{-\frac{1}{2}\|u\|^2} du \;=\; (2\pi)^{n/2}$.

In this definition, $\mu \geq 0$ plays a role of the *smoothing parameter*.

Why this is interesting? Define $y = x + \mu u$. Then

$$f_\mu(x) = \frac{1}{\mu^n \kappa} \int\limits_E f(y) e^{-\frac{1}{2\mu^2}\|y-x\|^2} dy. \quad \text{Hence,}$$

$$\nabla f_\mu(x) \;=\; \frac{1}{\mu^{n+2}\kappa} \int\limits_E f(y) e^{-\frac{1}{2\mu^2}\|y-x\|^2} (y - x)\, dy$$

$$= \; \frac{1}{\mu\kappa} \int\limits_E f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} u \, du \;\stackrel{(!)}{=}\; \frac{1}{\kappa} \int\limits_E \frac{f(x+\mu u) - f(x)}{\mu} e^{-\frac{1}{2}\|u\|^2} u \, du.$$

Properties of Gaussian smoothing

- If f is convex, then f_μ is convex and $f_\mu(x) \geq f(x)$.
- If $f \in C^{0,0}$, then $f_\mu \in C^{0,0}$ and $L_0(f_\mu) \leq L_0(f)$.
- If $f \in C^{0,0}(E)$, then, $|f_\mu(x) - f(x)| \leq \mu L_0(f) n^{1/2}$.

Random gradient-free oracle:

- Generate random $u \in E$.
- Return $g_\mu(x) = \frac{f(x+\mu u) - f(x)}{\mu} \cdot u$.

If $f \in C^{0,0}(E)$, then $E_u(\|g_\mu(x)\|_*^2) \leq L_0^2(f)(n+4)^2$.

Random intuitive optimization

Problem: $f^* \stackrel{\text{def}}{=} \min_{x \in Q} f(x)$, where $Q \subseteq E$ is a closed convex set, and f is a nonsmooth convex function.

Let us choose a sequence of positive steps $\{h_k\}_{k \geq 0}$.

Method $\mathcal{RS}_\mu$: Choose $x_0 \in Q$.

For $k \geq 0$: a). Generate u_k.

b). Compute $\Delta_k = \frac{1}{\mu}[f(x_k + \mu u_k) - f(x_k)]$.

c). Compute $x_{k+1} = \pi_Q(x_k - h_k \Delta_k u_k)$.

NB: μ can be arbitrary small.

Convergence results

This method generates random $\{x_k\}_{k\geq 0}$. Denote $S_N = \sum\limits_{k=0}^{N} h_k$,

$\mathcal{U}_k = (u_0, \ldots, u_k)$, $\phi_0 = f(x_0)$, and $\phi_k \stackrel{\mathrm{def}}{=} E_{\mathcal{U}_{k-1}}(f(x_k))$, $k \geq 1$.

Theorem: Let $\{x_k\}_{k\geq 0}$ be generated by $\mathcal{RS}_\mu$ with $\mu > 0$. Then,

$$\sum_{k=0}^{N} \frac{h_k}{S_N}(\phi_k - f^*) \leq \mu L_0(f) n^{1/2} + \frac{1}{2S_N}\|x_0 - x^*\|^2 + \frac{(n+4)^2}{2S_N} L_0^2(f) \sum_{k=0}^{N} h_k^2.$$

In order to guarantee $E_{\mathcal{U}_{N-1}}(f(\hat{x}_N)) - f^* \leq \epsilon$, we choose

$$\mu = \frac{\epsilon}{2L_0(f)n^{1/2}}, \quad h_k = \frac{R}{(n+4)(N+1)^{1/2}L_0(f)}, \quad N = \frac{4(n+4)^2}{\epsilon^2} L_0^2(f) R^2.$$

Interpretation

- Disturbance μu_k may be caused by external random factors.
- For small μ, the sign and the value of Δ_k can be treated as an *intuition*.
- We use a random experience accumulated by a very small shift along a random direction.
- The reaction steps h_k are big. (Emotions?)
- The dimension of x slows down the convergence.

Main ability: to implement an action, which is absolutely opposite to the proposed one. (Needs training.)

NB: Optimization method has a form of emotional reaction.

It is efficient in the absence of stable coordinate system.

Optimization in Stochastic Environment

Problem: $\min_{x \in Q} \left[\phi(x) = E(f(x,\xi)) \equiv \int_{\Omega} f(x,\xi)\, p(\xi)\, d\xi \right]$, where

- $f(x,\xi)$ is convex in x for any $\xi \in \Omega \subseteq R^m$,
- Q is a closed convex set in R^n,
- $p(\xi)$ is the density of random variable $\xi \in \Omega$.

Assumption: We can generate a sequence of random events $\{\xi_i\}$:

$$\frac{1}{N} \sum_{i=1}^{N} f(x,\xi_i) \quad \stackrel{N \to \infty}{\to} \quad E(f(x,\xi)), \quad x \in Q.$$

Goal: For $\epsilon > 0$ and $\phi^* = \min_{x \in Q} \phi(x)$ find $\bar{x} \in Q$: $\phi(\bar{x}) - \phi^* \leq \epsilon$.

Main trouble: For finding δ-approximation to $\phi(x)$, we need $O\left(\left(\frac{1}{\delta}\right)^m\right)$ computations of $f(x,\xi)$.

Stochastic subgradients (Ermoliev, Wetz, 70's)

Method: Fix some $x_0 \in Q$ and $h > 0$. For $k \geq 0$, repeat:

$$\text{generate } \xi_k \text{ and update } x_{k+1} = \pi_Q\left(x_k - h \cdot f'(x_k, \xi_k)\right).$$

Output: $\bar{x} = \frac{1}{N+1} \sum\limits_{k=0}^{N} x_k$.

Interpretation: Learning process in stochastic environment.

Theorem: For $h = \frac{R}{L\sqrt{N+1}}$ we get $\boxed{E(\phi(\bar{x})) - \phi^* \leq \frac{LR}{\sqrt{N+1}}}$.

NB: This is an estimate for the *average* performance.

Hint: For us, it is enough to ensure a *Confidence Level* $\beta \in (0, 1]$:

$$\textbf{Prob}\ [\ \phi(\bar{x}) \geq \phi^* + \epsilon V_\phi\ ] \leq 1 - \beta,$$

where $V_\phi = \max\limits_{x \in Q} \phi(x) - \phi^*$.

In the real world, we *always* apply solutions with $\beta < 1$.

What do we have now?

After N-steps we observe a *single* implementation of the random variable $\bar{x}$ with $\quad E(\phi(\bar{x})) - \phi^* \leq \frac{LR}{\sqrt{N+1}}$.

What about the level of confidence?

1. For random $\psi \geq 0$ and $T > 0$ we have

$$E(\psi) = \int \psi = \int\limits_{\psi \geq T} \psi + \int\limits_{\psi < T} \psi \geq T \cdot \mathbf{Prob}\,[\psi \geq T].$$

2. With $\psi = \phi(\bar{x}) - \phi^*$ and $T = \epsilon V_\phi$ we need

$$\frac{1}{\epsilon V_\phi}[E(\phi(\bar{x})) - \phi^*] \leq \frac{LR}{\epsilon V_\phi \sqrt{N+1}} \leq 1 - \beta.$$

Thus, we can take $\quad N + 1 = \frac{1}{\epsilon^2 (1-\beta)^2} \left(\frac{LR}{V_\phi} \right)^2$.

NB: 1. For personal needs, this may be OK. What about $\beta \to 1$?

2. How we increase the confidence level in our life?

Ask for advice as many persons as we can!

Pooling the experience

Individual learning process (Forms opinion of one expert)

Choose $x_0 \in Q$ and $h > 0$. For $k = 0, \ldots, N$ repeat

$$\boxed{\textit{generate } \xi_k, \textit{ and set } x_{k+1} = \pi_Q(x_k - hf'(x_k, \xi_k)).}$$

Compute $\bar{x} = \frac{1}{N+1} \sum_{k=0}^{N} x_k$.

Pool the experience:

For $j = 1, \ldots, K$ compute $\bar{x}_j$. Generate the output $\hat{x} = \frac{1}{K} \sum_{j=1}^{K} \bar{x}_j$.

Note: All learning processes start from the same x_0.

Probabilistic analysis

Theorem. Let $Z_j \in [0, V]$, $j = 1, \ldots, K$ be independent random variables with the same average μ. Then for $\hat{Z}_K = \frac{1}{K} \sum_{j=1}^{K} Z_j$

$$\textbf{Prob} \left[\hat{Z}_k \geq \mu + \hat{\epsilon} \right] \leq \exp \left(- \frac{2\hat{\epsilon}^2 K}{V^2} \right).$$

Corollary.

Let us choose $K = \frac{2}{\epsilon^2} \ln \frac{1}{1-\beta}$, $N = \frac{4}{\epsilon^2} \left(\frac{LR}{V_\phi} \right)^2$, and $h = \frac{R}{L\sqrt{N+1}}$.
Then the pooling process implements an (ϵ, β)-solution.

Note: Each 9 in $\beta = 0.9 \cdots 9$ costs $\frac{4.6}{\epsilon^2}$ experts.

Comparison (ϵ is not too small $\equiv$ Q is reasonable)

Denote $\rho = \frac{LR}{V_\phi}$	Single Expert (SE)	Pooling Experience (PE)
Number of experts	1	$\frac{2}{\epsilon^2} \ln \frac{1}{1-\beta}$
Length of life	$\frac{\rho^2}{\epsilon^2(1-\beta)^2}$	$\frac{4\rho^2}{\epsilon^2}$
Computational efforts	$\frac{\rho^2}{\epsilon^2(1-\beta)^2}$	$\frac{8\rho^2}{\epsilon^4} \ln \frac{1}{1-\beta}$

- Reasonable computational expenses (for Multi-D Integrals)
- Number of experts does not depend on dimension.

Differences

- For low level of confidence, SE may be enough.
- High level of confidence needs independent expertise.
- Average experience of young population has much higher level of confidence than the experience of a long-life wizard.
- In PE, the confidence level of "experts" is only $\frac{1}{2}$ (!).

Why this can be useful?

- Understanding of the actual role of existing social an political phenomena (education, medias, books, movies, theater, elections, etc.)
- Future changes (Internet, telecommunications)
- Development of new averaging instruments (Theory of expertise: mixing opinion of different experts, competitions, etc.)

Conscious versus subconscious

NB: Conscious behavior can be irrational.

Subconscious behavior is often rational.

- Animals.
- Children education: First level of knowledge is subconscious.
- Training in sport (optimal technique $\Rightarrow$ subconscious level).

Examples of subconscious estimates:

- Mental "image processing".
- Tracking the position of your body in space.
- Regular checking of your status in the society (?)

Our model: Conscious behavior based on dynamically updated subconscious estimates.

Model of consumer: What is easy for us?

Question 1: $123 * 456 = ?$

Question 2: How often it rains in Belgium?

Easy questions:

- average salary,
- average gas consumption of your car,
- average consumption of different food,
- average commuting time,

and many other (survey-type) questions.

Main abilities of anybody:

1. Remember the past experience (often by *averages*).

2. Estimate *probabilities* of some future events, taking into account their *frequencies* in the past.

Guess: We are *Statistical* Homo Economicus? (SHE)

Main features of SHE

Main passion: Observations.

Main abilities:

- Can select the best variant from several possibilities.
- Can compute average characteristics for some actions.
- Can compute frequencies of some events in the past.
- Can estimate the “faire” prices for products.

As compared with HE: A huge step back in the computational power and informational support.

Theorem: SHE can be rational.

(The proof is constructive.)

Consumption model

Market

- There are n products with unitary prices p_j.
- Each product is described by the *vector of qualities* $a_j \in R^m$. Thus, $a_j^{(i)}$ is the *volume* of quality i in the unit of product j.

Consumer SHE

- Forms and updates the *personal prices* $y \in R^m$ for qualities.
- Can estimate the personal quality/price ratio for product j:

$$\pi_j(y) = \frac{1}{p_j}\langle a_j, y\rangle.$$

- Has standard σ_i for consumption of quality i, $\sum\limits_{i=1}^{m} \sigma_i y_i = 1$.

Denote $A = (a_1, \ldots, a_n)$, $\sigma = (\sigma_1, \ldots, \sigma_m)^T$, $\pi(y) = \max\limits_{1 \leq j \leq n} \pi_j(y)$.

Consumption algorithm (CA) for kth weekend

For Friday night, SHE has personal prices y_k, budget λ_k, and cumulative consumption vector of qualities $s_k \in R^m$, $s_0 = 0$.

1. Define the set $J_k = \{j : \pi_j(y_k) = \pi(y_k)\}$, containing the products with the best quality/price ratio.
2. Form partition $x_k \geq 0$: $\sum\limits_{j=1}^{n} x_k^{(j)} = 1$, and $x_k^{(j)} = 0$ for $j \notin J_k$.
3. Buy all products in volumes $X_k^{(j)} = \lambda_k \cdot x_k^{(j)} / p_j$, $j = 1, \dots, n$.
4. Consume the bought products: $s_{k+1} = s_k + AX_k$.
5. During the next week, SHE watches the results and forms the personal prices for the next shopping.

NB: Only Item 5 is not defined.

Updating the personal prices for qualities

Define $\xi_i = \sigma_i y_k^{(i)}$, the *relative importance* of quality i, $\sum_{i=1}^{m} \xi_i = 1$.

Denote by $\hat{s}_k = \frac{1}{k} s_k$ the average consumption.

Assumption. 1. During the week, SHE performs regular detections of the most deficient quality by computing $\psi_k = \min_{1 \le i \le m} \hat{s}_k^{(i)} / \sigma_i$.

2. This detection is done with random additive errors. Hence, we observe

$$E_\epsilon \left(\min_{1 \le i \le m} \left\{ \frac{\hat{s}_k^{(i)}}{\sigma_i} + \epsilon_i \right\} \right).$$

Thus, any quality has a chance to be detected as the worst one.

3. We define ξ_i as the frequency of detecting the quality i as the most deficient one with respect to $\hat{s}_k$.

This is it. Where is Optimization? Objective Function, etc.?

Algorithmic aspects

1. If ϵ_i are doubly-exponentially i.i.d. with variance μ, then

$$y_k^{(i)} = \frac{1}{\sigma_i} \exp\left\{ -\frac{s_k^{(i)}}{k\sigma_i\mu} \right\} / \sum_{j=1}^{m} \exp\left\{ -\frac{s_k^{(j)}}{k\sigma_j\mu} \right\}$$

Therefore, $y_k = \arg\min\limits_{\langle \sigma, y\rangle = 1} \{ \langle s_k, y\rangle + \gamma d(y) \}$,

where $\gamma = k\mu$, $d(y) = \sum\limits_{i=1}^{m} \sigma_i y^{(i)} \ln(\sigma_i y^{(i)})$ (prox-function).

2. $AX_k = \lambda_k A \left[\frac{x_k}{p} \right] \equiv \lambda_k g_k$, where $g_k \in \partial\pi(y_k)$ (subgradient).

3. Hence, s_k is an accumulated *model* of function $\pi(y)$.

Hence, CA is a *primal-dual* method for solving the (dual) problem

$$\min_{y \geq 0} \left\{ \pi(y) \equiv \max_{1 \leq i \leq m} \frac{1}{p_i} \langle a_i, y\rangle : \ \langle \sigma, y\rangle = 1 \right\}.$$

Comments

1. The primal problem is

$$\max_{u,\tau}\{\tau : \ Au \geq \tau\sigma, \ u \geq 0, \ \langle p, u\rangle = 1\}.$$

We set $u_k = [x_k/p]$ and approximate u^* by averaging $\{u_k\}$.

2. No “computation” of subgradients (we just buy).
Model is updated implicitly (we just eat).

3. CA is an example of *unintentional* optimization.
(Other examples in the nature: Fermat principle, etc.)

4. SHE does not recognize the objective. However, it exists.
SHE is rational by behavior, not by the goal (which is absent?).

5. Function $\pi(y)$ measures the positive appreciation of the market.
By minimizing it, we develop a pessimistic vision of the world.
(With time, everything becomes expensive.)

6. For a better life, allow a bit of irrationality. (Smooth objective, faster convergence.)

Conclusion

1. Optimization patterns are widely presented in the social life. Examples:

- Forming the traditions (Inaugural Lecture)
- Efficient collaboration between industry, science and government (Lecture 1)
- Local actions in problems of unlimited size (Lecture 3).

2. The winning social systems give better possibilities for rational behavior of people. (Forget about ants and bees!)

3. Our role could be the discovering of such patterns and helping to improve them by an appropriate mathematical analysis.

References

Lecture 1: Intrinsic complexity of Black-Box Optimization

- Yu. Nesterov. *Introductory Lectures on Convex Optimization*. Chapters 2, 3. Kluwer, Boston, 2004.
- Yu. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(\frac{1}{k^2})$. *Doklady AN SSSR* (translated as Soviet Math. Dokl.), 1983, v.269, No. 3, 543-547.

Lecture 2: Looking into the Black Box

- Yu. Nesterov. "Smooth minimization of non-smooth functions", *Mathematical Programming* (A), **103** (1), 127-152 (2005).
- Yu. Nesterov. "Excessive gap technique in nonsmooth convex minimization". SIAM J. Optim. **16** (1), 235-249 (2005).
- Yu.Nesterov. Gradient methods for minimizing composite functions. Accepted by *Mathematical Programming*.
-

References

Lecture 3: Huge-scale optimization problems

- Yu.Nesterov. Efficiency of coordinate descent methods on large scale optimization problems. Accepted by SIAM.
- Yu.Nesterov. Subgradient methods for huge-scale optimization problems. CORE DP 2012/02.

Lecture 4: Nonlinear analysis of combinatorial problems.

- Yu.Nesterov. Semidefinite Relaxation and Nonconvex Quadratic Optimization. *Optimization Methods and Software*, vol.9, 1998, pp.141–160.
- Yu.Nesterov. Simple bounds for boolean quadratic problems. EUROPT Newsletters, **18**, 19-23 (December 2009).

References

Lecture 5:

- Yu.Nesterov, J.-Ph.Vial. Confidence level solutions for stochastic programming. *Auromatica*, **44**(6), 1559-1568 (2008)
- Yu.Nesterov. Algorithmic justification of intuitive rationality in consumer behavior. CORE DP.

Thank you for your attention!