REGULARIZED SMOOTHING FOR SOLUTION MAPPINGS OF CONVEX PROBLEMS WITH APPLICATIONS TO 2-STAGE STOCHASTIC PROGRAMMING AND SOME HIERARCHICAL PROBLEMS

Mikhail Solodov

(IMPA, Rio)

joint work with

Pedro Borges (IMPA)

Claudia Sagastizábal (UNICAMP, CeMEAI)

One World Optimization Seminar, May 2021

Outline

- Optimization problems that depend on solutions of other optimization problems
  - MOPECs, GNEPs, Multi-level optim., 2-stage stochastic progr., Hierarchical games
- Parametric optimization. Why smoothing?
- Even in "good" cases (smooth, convex, CQs)
  - solution mapping set-valued, nonsmooth
  - value-function nonsmooth, nonconvex
  - both given implicitly (compute derivatives?)
- Computationally tractable regularized smoothing (approximating solutions and value-function)

Multi-Level Optimization

E.g., (optimistic) bilevel problem:

$$\min_{(x,p)} \; f_u(x,p) \quad \text{s.t.} \quad g_u(x,p) \le 0, \; x \in SOL_l(p),$$

where

$$SOL_l(p) = \operatorname*{argmin}_x \; f_l(x,p) \quad \text{s.t.} \quad g_l(x,p) \le 0.$$

Def. value-function of the lower parametric problem:

$$V(p) = \min_x \; f_l(x,p) \quad \text{s.t.} \quad g_l(x,p) \le 0.$$

The value-function formulation of bilevel problem:

$$\min_{(x,p)} \; f_u(x,p) \quad \text{s.t.} \quad g_u(x,p) \le 0, \; g_l(x,p) \le 0, \; f_l(x,p) \le V(p).$$

Two-Stage Stochastic Programming

E.g., two-stage stochastic problem with linear 2nd stage:

$$\min_p \; c(p) + \sum_{s=1}^{K} \xi_s V_s(p), \quad \text{s.t.} \quad p \in \Pi,$$

where

$$V_s(p) = \min_x \langle q_s, x \rangle \quad \text{s.t.} \quad Wx = h_s - T_s p, \; x \geq 0.$$

Here, randomness has known probability distribution, with finite support described by scenarios $s = 1, \ldots, K$, with probabilities $\xi_s \in (0, 1)$.

Single-Leader Multi-Follower Games

An example of MOPEC:

for a given parameter p, agents $a \in A$ determine their decisions independently:

$$x_a(p) = \underset{x}{\operatorname{argmin}} \; f_a(x,p) \quad \text{s.t.} \quad g_a(x,p) \leq 0.$$

Then, some criterion F is optimized, coupling all the agents' decisions: $X_A(p) = (x_a(p), a \in A)$,

$$\min_{p} \; F(X_A(p)) \quad \text{s.t.} \quad p \in \Pi.$$

The leader's problem involves solution mappings of the followers.

Generalized Nash Equilibrium Problems

There are $a \in A$ agents with conflicting interests, each solving

$$\min_{p_a} \; f_a(p_a, p_{-a}) \quad \text{s.t.} \quad g_a(p_a) \leq 0, \; h(p_a, p_{-a}) \leq 0.$$

In particular, p_{-a} are variables not controled by agent a, i.e., parameters for agent's a problem.

On the other hand, p_a is a variable in this problem, and a parameter in all others.

The h-constraint is "shared", same for all agents.

A point $\bar{p}$ is an equilibrium if no agent can improve unilaterally.

Fully Parametrized Convex Problems

In the applications discussed above, appear parametrized problems like

$$\min_x \ f(x,p) \quad \text{s.t.} \quad B(p)x = b(p),\ g(x,p) \leq 0.$$

- All functions are suffiently smooth in x and p
- Functions f and g are convex in x for each p
- $B(p)$ full rank $\forall\, p$
- For every p Slater CQ is satisfied
  (possibly with different Slater points $\hat{x}(p)$)
- For every p the problem has at least one solution
  (need not be unique, $SOL(p)$ can be unbounded)

Solution Mapping and Value-Function

Even so, under "perfect assumptions", consider

$$\min_x \; px \quad \text{s.t.} \quad x \in [-1,1], \quad x, p \in \mathbf{R}.$$

All the assumptions above are satisfied.

$$SOL(p) = \begin{cases} 1, & p < 0 \\ [-1,1], & p = 0, \\ -1, & p > 0 \end{cases} \qquad V(p) = -|p|.$$

- Solution mapping $SOL(p)$ is multi-valued
- The value-function $V(p)$ is nonsmooth and nonconvex
- $SOL(p)$ and/or $V(p)$ enter other problem(s)!

Regularized Log-Barrier Smoothing

We approximate the solution mapping $SOL(p)$ and the value-function $V(p)$ of

$$\min_x \ f(x,p) \quad \text{s.t.} \quad B(p)x = b(p),\ g(x,p) \leq 0,$$

by the solution $x^\varepsilon(p)$ of Tikhonov-regularized interior penalty approximation of this problem, given by

$$\min_x \ f(x,p) - \varepsilon \sum \log(-g_i(x,p)) + \varepsilon \frac{r}{2}\|x\|^2 \quad \text{s.t.} \quad B(p)x = b(p).$$

The solution $x^\varepsilon(p)$ is unique for each p ($\varepsilon > 0$, $r \geq 0$), and $x^\varepsilon(\cdot)$ is continuously differentiable. Then, so is $V^\varepsilon(\cdot)$, where

$$V^\varepsilon(p) = f(x^\varepsilon(p), p).$$

Some Comments

- There is vast literature on continuity and generalized differentiability of value-functions
- But not much on computing (or approximating) (generalized) derivatives, except special cases
- That said, log-barrier had been used before for this purpose (Fiacco, Ishizuka, 1990), under strong assumptions (solutions unique, etc.)
- The combination of log-barrier and regularization appears to be new/important
- Approximations inside of other problems is new.

Smoothness of the Approximating Solution Mapping

Approximating subproblem

$$\min_x \; f(x,p) - \varepsilon \sum \log(-g_i(x,p)) + \varepsilon \frac{r}{2}\|x\|^2 \quad \text{s.t.} \quad B(p)x = b(p).$$

Lagrange optimality conditions (hold automatically):

$$\nabla_x f(x^\varepsilon(p),p) - \varepsilon \sum \frac{\nabla_x g_i(x^\varepsilon(p),p)}{g_i(x^\varepsilon(p),p)} + \varepsilon r x^\varepsilon(p) - (B(p))^\top \lambda^\varepsilon(p) = 0,$$

$$B(p)x^\varepsilon(p) - b(p) = 0.$$

The Jacobian of this system of nonlinear equations is

$$J^\varepsilon_{(x,\lambda)}((x^\varepsilon(p),\lambda^\varepsilon(p))p) = \begin{pmatrix} \nabla^2_{xx} f(x^\varepsilon(p),p) + \cdots + \varepsilon r I & -(B(p))^\top \\ B(p) & 0 \end{pmatrix}$$

Smoothness of the Approximations

This Jacobian is non-singular. Hence,

The Implicit Function Theorem

$\Downarrow$

continuous differentiabilty of $x^\varepsilon(\cdot)$ and of $\lambda^\varepsilon(\cdot)$,

and then also of $V^\varepsilon(\cdot) = f(x^\varepsilon(\cdot), p)$.

Moreover, their derivatives are computable by solving systems of linear equations (also via IFT).

For $\Phi(z,p) = 0$, $(\Phi'_x(z(p),p))z'(p) = -\Phi'_p(z(p),p)$.

Value-Function Bounds

For any $r \geq 0$ and $\varepsilon > 0$, if $x^\varepsilon(p)$ exists then

$$V(p) \leq V^\varepsilon(p) \leq V(p) + m\varepsilon + \varepsilon \frac{r}{2} \min_{x \in SOL(p)} \|x\|^2,$$

where m is the number of inequality constraints.

If $r > 0$, then $x^\varepsilon(p)$ exists for every $\varepsilon > 0$, and it holds in addition that

$$\frac{r}{2} \min_{x \in SOL(p)} \|x\|^2 + m \geq \frac{r}{2} \|x^\varepsilon(p)\|^2.$$

(for $r = 0$, reduces to a bound known in the interior penalties literature)

Asymptotics for Value-Function and Its Smoothing

$$V(p) \leq V^{\varepsilon}(p) \leq V(p) + m\varepsilon + \varepsilon \frac{r}{2} \min_{x \in SOL(p)} \|x\|^2.$$

We want:

$$\lim_{\varepsilon \searrow 0, p \to \bar{p}} V^{\varepsilon}(p) = V(\bar{p}).$$

This is automatic if $r = 0$. Otherwise, assume

$$\limsup_{p \to \bar{p}} \left\{ \min_{x \in SOL(p)} \|x\|^2 \right\} < +\infty.$$

- $SOL(\bar{p})$ is locally bounded (around $\bar{p}$), which is automatic if feasible set is locally bounded.
- Minimal-norm solutions are locally bounded ($SOL(p)$ can be unbounded).

Asymptotics for Smoothing of the Solution Mapping

$$V(p) \le V^{\varepsilon}(p) \le V(p) + m\varepsilon + \varepsilon \frac{r}{2} \min_{x \in SOL(p)} \|x\|^2,$$

$$\text{If } r > 0, \quad \frac{r}{2} \min_{x \in SOL(p)} \|x\|^2 + m \ge \frac{r}{2} \|x^{\varepsilon}(p)\|^2.$$

$$\text{If } \limsup_{p \to \bar{p}} \{ \min_{x \in SOL(p)} \|x\|^2 \} < +\infty,$$

$$\text{then, for any } r \ge 0, \ \limsup_{\varepsilon \searrow 0, p \to \bar{p}} x^{\varepsilon}(p) \subset SOL(\bar{p}).$$

If $r > 0$, then $x^{\varepsilon}(p)$ is uniformly bounded (for small ε and p close to $\bar{p}$), even if $SOL(p)$ is not.

If $r = 0$, for $x^{\varepsilon}(p)$ uniformly bounded, have to assume local boundedness of $SOL(p)$ around $\bar{p}$, or directly $\limsup_{\varepsilon \searrow 0, p \to \bar{p}} \|x^{\varepsilon}(p)\| < +\infty$.

Boundedness of Smoothing Gradients of the Value-Function

Let the smoothing be built with $r > 0$, and assume

$$\limsup_{p \to \bar{p}} \left\{ \min_{x \in SOL(p)} \|x\|^2 \right\} < +\infty.$$

Then

- Primal approx. $x^\varepsilon(p)$ is locally uniformly bounded (for small ε and p close to $\bar{p}$)
- Dual approx. $\lambda^\varepsilon(p)$ and $\mu_i^\varepsilon(p) := -\varepsilon / g_i(x^\varepsilon(p), p)$ are locally uniformly bounded

And finally,

$$\nabla V^\varepsilon(p) \text{ is locally uniformly bounded}$$

(important for numerics)

On Lipschitz-Continuity of the Value-Function

It is interesting that, as a by-product of our algorithmic development, we (easily) recover the following result from the literature:

Under our standing assumptions and

$$\limsup_{p \to \bar{p}} \{ \min_{x \in SOL(p)} \|x\|^2 \} < +\infty,$$

the value-function $V(\cdot)$ is

locally Lipschitz-continuous around $\bar{p}$

This "$\limsup$" condition is called
"restricted inf-compactness" in
L. Guo, G.-H. Lin, J. Ye, J. Zhang (SIOPT 2014)

Proof: Under

$$\limsup_{p \to \bar{p}} \{ \min_{x \in SOL(p)} \|x\|^2 \} < +\infty,$$

Taking $r > 0$ in our smoothing algorithm, we proved that $\nabla V^{\varepsilon}(p)$ is locally bounded, and

$$\lim_{\varepsilon \searrow 0, p \to \bar{p}} V^{\varepsilon}(p) = V(\bar{p}).$$

It is known that the latter (generally) implies that

$$\partial V(\bar{p}) \subset \text{conv} \{ \limsup_{\varepsilon \searrow 0, p \to \bar{p}} \nabla V^{\varepsilon}(p) \}.$$

Hence, $\partial V(\bar{p})$ is also bounded. In this context,

$\partial V(p)$ locally bounded iff $V(p)$ is locally Lipschitz.

On Gradient Consistency

We have

$$\lim_{\varepsilon \searrow 0, p \to \bar{p}} V^{\varepsilon}(p) = V(\bar{p}),$$

and hence,

$$\partial V(\bar{p}) \subset \text{conv} \left\{ \limsup_{\varepsilon \searrow 0, p \to \bar{p}} \nabla V^{\varepsilon}(p) \right\}.$$

When V is locally Lipschitz-continuous,
gradient consistency between V^{ε} and V holds, if

$$\text{conv} \left\{ \limsup_{\varepsilon \searrow 0, p \to \bar{p}} \nabla V^{\varepsilon}(p) \right\} \subset \partial_C V(\bar{p}).$$

(When V is locally Lipschitz, $\partial_C V(\bar{p}) = \text{conv}\, \partial V(\bar{p})$)

On Gradient Consistency

Let the smoothing be built with $r > 0$, and assume

$$\limsup_{p \to \bar{p}} \{ \min_{x \in SOL(p)} \|x\|^2 \} < +\infty.$$

Define

$$W^\varepsilon(p) = V^\varepsilon(p) - \varepsilon \sum \log(-g_i(x,p)) + \varepsilon \frac{r}{2} \|x\|^2.$$

- $\lim_{\varepsilon \searrow 0, p \to \bar{p}} W^\varepsilon(p) = V(\bar{p})$.
- If V is convex, and W^ε is convex for all ε small enough, then W^ε is gradient consistent with V.
- If the problem is parametrized only on the mapping b in the RHS of constraint $Bx = b(p)$, and $b(\cdot)$ is affine, then V and W^ε are convex.

Two-Stage Stochastic Programming

Given convex 1st and 2nd stage objective functions c and q_s, risk-averse two-stage stochastic problem is

$$\min_p \; c(p) + R(q_1(x_1(p)), \ldots, q_k(x_k(p))) \quad \text{s.t.} \quad p \in \Pi,$$

where R is some risk measure, and

$x_s(p),\ s = 1, \ldots, k,$ are solutions of

$$\min_x \; q_s(x) \quad \text{s.t.} \quad Wx + T_s p = h_s,\ x \geq 0.$$

Let the risk measure R be

$$AVaR_\alpha(q_s) = \min_{t \in \mathbf{R}} \left\{ t + \frac{1}{1-\alpha} E[\max\{q_s - t, 0\}] \right\}, \ \alpha \in (0,1).$$

Re-writing "generic" min-max expression
$\min(\max\{z_1, z_2\})$ as $\min\ y$ s.t. $y \geq z_1, y \geq z_2$, we get

$$\min_{(p,t)}\ c(p) + t + \sum_{s=1}^{k} \xi_s V_s(p,t) \quad \text{s.t.} \quad p \in \Pi,\ t \in \mathbf{R},$$

where

$$V_s(p,t) = \min_{(x,y)}\ q_s(x) + \tfrac{y}{1-\alpha} \text{ s.t. } Wx = h_s - T_s p,\ q_s(x) - y \leq t,$$
$$x \geq 0,\ y \geq 0.$$

Here, have RHS parametrization in the constraints.
Hence, "recourse functions" V_s are convex.

This two-stage problem

$$\min_{(p,t)} \; c(p) + t + \sum_{s=1}^{k} \xi_s V_s(p,t) \quad \text{s.t.} \quad p \in \Pi, \; t \in \mathbf{R},$$

$$V_s(p,t) = \min_{(x,y)} \; q_s(x) + \tfrac{y}{1-\alpha} \; \text{s.t.} \quad Wx = h_s - T_s p, \; q_s(x) - y \leq t,$$

$$x \geq 0, \; y \geq 0,$$

is usually solved by L-shaped (cutting-planes) or, much better, bundle methods.

Observe: If $\nu_s(p,t)$ and $\rho_s(p,t)$ are Lagrange multipliers associated at the solution $x_s(p,t)$ to the first two constraints in the second-stage, then

$$(T_s^\top \nu_s(p,t), \, -\rho_s(p,t)) \in \partial V_s(p,t).$$

Convexity is important! (Otherwise... $\partial V_s =$???)

Two-Stage Stochastic Programming

Our approach to solving this two-stage problem

$$\min_{(p,t)} \; c(p) + t + \sum_s \xi_s V_s(p,t) \quad \text{s.t.} \quad p \in \Pi,\; t \in \mathbf{R},$$

$$V_s(p,t) = \min_{(x,y)} \; q_s(x) + \tfrac{y}{1-\alpha} \;\text{s.t.}\; Wx = h_s - T_s p,\; q_s(x) - y \le t,$$

$$x \ge 0,\; y \ge 0,$$

is different, applicable to the nonconvex case too(!):

$$\min_{(p,t)} \; c(p) + t + \sum_s \xi_s V_s^{\varepsilon}(p,t) \quad \text{s.t.} \quad p \in \Pi,\; t \in \mathbf{R},$$

$$V_s^{\varepsilon}(p,t) = \min_{(x,y)} \; q_s(x) + \tfrac{y}{1-\alpha}$$

$$-\varepsilon \log(y) - \varepsilon \log(t + y - q_s(x)) - \varepsilon \sum \log(x_i)$$

$$\text{s.t.} \quad Wx = h_s - T_s p.$$

Numerical Experiments for 2-Stage Stochastic Probs.

I. Deák's two-stage LPs, with
modified 2nd stage objective functions:

$$V_s(p,t) = \min_{(x,y)} \langle q_s, x\rangle + \gamma\|x\|^2 + \frac{y}{1-\alpha}$$
$$\text{s.t. } Wx = h_s - T_s p,\ \langle q_s, x\rangle + \gamma\|x\|^2 - y \le t,$$
$$x \ge 0,\ y \ge 0.$$

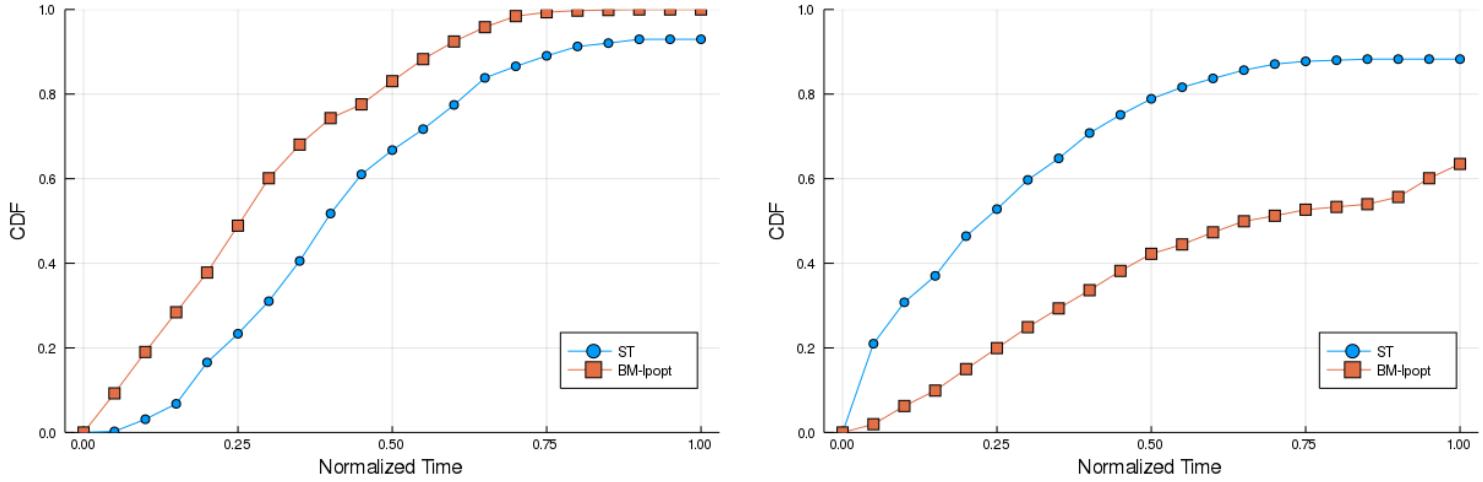
Thus, in the 2nd stage the Bundle Method solves

- LPs if $\gamma = 0$ (easy)
- QCQPs if $\gamma > 0$ (more difficult)

Smoothing Method solves linearly-constrained NLPs.

Numerical Experiments for 2-Stage Stochastic Probs.

$S \in \{1, \ldots, 20\},\ \varepsilon \in \{0.01, 0.1\},\ r \in \{0, 0.1, 1\},\ \gamma \in \{0, 0.01, 0.1, 1\}$

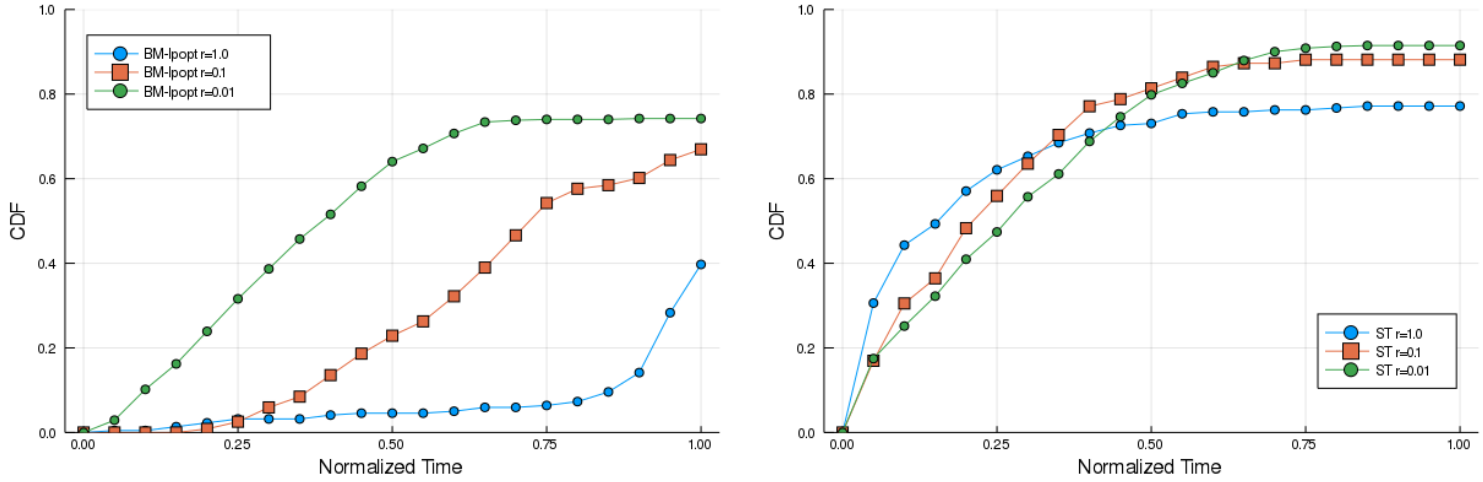


Performance for linear (left, $\gamma = 0$) and quadratic (right, $\gamma > 0$) instances with risk.

Time normalized by max time; given time budget, probability of delivering the best iterate.

Numerical Experiments for 2-Stage Stochastic Probs.

Affect of nonlinearity (of the value of γ)



How the "size" of the quadratic term affects the Bundle Method (left) and the Smoothing Method (right).

Single-Leader Multi-Follower Games

For a given parameter p, agents $a \in A$ determine their decisions independently:

$$x_a(p) = \operatorname{argmin}_x \ f_a(x,p)$$
$$\text{s.t. } B_a(p)x = b_a(p),\ g_a(x,p) \leq 0.$$

Then, some criterion F is optimized, coupling all the agents' decisions: $X_A(p) = (x_a(p), a \in A)$,

$$\min_p \ F(X_A(p)) \quad \text{s.t.} \quad p \in \Pi.$$

The leader's problem involves solution mappings of the followers (not value-functions).

The Smoothing Approach

Recall that $x_a(\cdot)$ are, in general, nonsmooth. Define

$$x_a^{\varepsilon,r}(p) = \quad \operatorname{argmin}_x \; f_a(x,p) - \varepsilon \sum \log(-g_a(x,p)) + \varepsilon \frac{r(\varepsilon)}{2} \|x\|^2$$
$$\text{s.t.} \;\; B_a(p)x = b_a(p),$$

$\varepsilon > 0$, $r(\varepsilon) \geq 0$, and approximate the leader's problem

$$\min_p \; F(X_A(p)) \quad \text{s.t.} \quad p \in \Pi,$$

by

$$\min_p \; F(X_A^{\varepsilon,r}(p)) \quad \text{s.t.} \quad p \in \Pi,$$

where

$$X_A^{\varepsilon,r}(p) = (x_a^{\varepsilon,r}(p), a \in A).$$

Smoothing Properties

Under the previous standing assumptions,

$$X_A^{\varepsilon,r}(\cdot) \text{ and } F(X_A^{\varepsilon,r}(\cdot)) \text{ are smooth,}$$

$$\lim_{\varepsilon \searrow 0, p' \in \Pi, p' \to p} x_a^{\varepsilon,r}(p') = x_a(p) \quad \forall\, p \in \Pi \text{ and } a \in A.$$

With proper management of ε and $r(\varepsilon)$,

$$F(X_A^{\varepsilon,r}(p)) \text{ converges epigraphically to } F(X_A(p)).$$

The latter implies (among other things):

$$\limsup_{\varepsilon \searrow 0} \{\varepsilon - \operatorname*{argmin}_{p \in \Pi} F(X_A^{\varepsilon,r}(p)\} \subset \operatorname*{argmin}_{p \in \Pi} F(X_A(p)),\ \varepsilon \to 0.$$

Decomposition

Agent-wise decomposition is readily available:

1. Given p^k, compute p^{k+1}, an approx. solution of

$$\min_p \; F(X_A^{\varepsilon_k, r_k}(p)) \quad \text{s.t.} \quad p \in \Pi,$$

taking $p^{k,0} = p^k$ as the starting point.

When the solver asks function and gradient of $F(X_A^{\varepsilon_k, r_k}(\cdot))$ at its inner iterate $p^{k,i}$, $i = 0, 1, \ldots$

- For each $a \in A$ solve independently

$$\begin{aligned} &\min_x \quad f_a(x, p^{k,i}) - \varepsilon \sum \log(-g_a(x, p^{k,i})) + \varepsilon_k \tfrac{r_k}{2} \|x\|^2 \\ &\text{s.t.} \quad B_a(p^{k,i})x = b_a(p^{k,i}). \end{aligned}$$

2. Update ε and r. Go to Step 1.

Numerical Experiments for Walrasian Equilibrium Problems

2 agents exchanging $n \in \{2, 10, 20, 30\}$ goods.
(from J. Deride, A. Jofré and R.J-B. Wets, 2017)

Deterministic WEP with symmetric agents

n	SM-Time	SM-Clearing	PATH-Time	PATH-Clearing
2	0.03 / 0.01	10^{-6} / 10^{-6}	0.01 / 0.00	10^{-33} / 10^{-33}
10	0.13 / 0.01	10^{-5} / 10^{-5}	0.57 / 1.09	10^{-33} / 10^{-33}
20	0.16 / 0.02	10^{-4} / 10^{-4}	0.07 / 0.00	10^{-32} / 10^{-32}
30	0.21 / 0.02	10^{-4} / 10^{-4}	0.13 / 0.00	10^{-32} / 10^{-32}

Except for the experiments with $n = 10$, where PATH seems to have struggled for one run,
Newtonian iterations in PATH make the output more precise and faster for this set of problems, as expected.

Numerical Experiments for Walrasian Equilibrium Problems

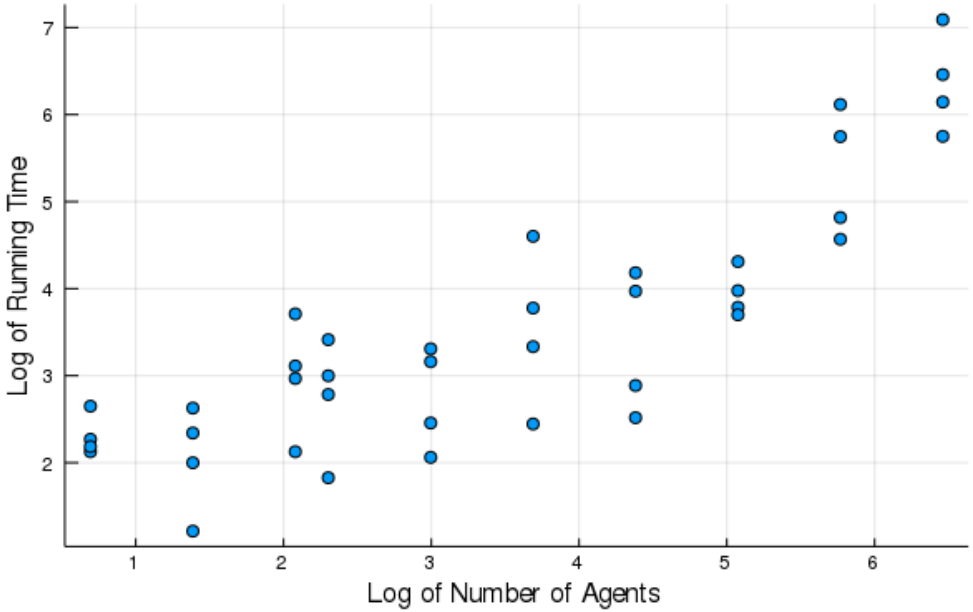
5 agents exchanging 10 goods
(from H.E. Skarf, 1973; also
J. Deride, A. Jofré and R.J-B. Wets, 2017)

	SM-Time	SM-Clearing	PATH-Time	PATH-Clearing
Avg.	2.30	10^{-5}	10.24	10^{-10}
Std.	1.76	10^{-2}	5.45	10^{-8}

- While PATH is still more precise,
  the smoothing approach is now faster
  (to termination).
- Increasing the size, as the PATH formulation
  does not allow decomposition, PATH starts to fail.

Scaling Capabilities of Decomposition Across Agents

We extend the previous example to an economy with 80 goods, and the number of agents ranging from 2 to 640.



In log scale, running times grow linearly with respect to the number of agents (log scale because the number of agents in the experiment grows exponentially).

Decomposition

When the agents' problems are two-stage stochastic programs, our construction allows

- decomposition across agents
- decomposition across scenarios
- decomposition across both agents and scenarios

Conclusions

- Optimization problems that invlove solutions of other optimization problems
  (i.e., solution mappings or value-functions)
- Fully parameterized convex problems
  - Solutions and value-functions are nonsmooth
  - Solutions and value-functions are implicit
  - Smoothing via Tikhonov-regularized log-barrier
  - The approximation "converges"
  - Derivatives are computable
    (the approach is computonally tractable)
- Applications to, and numerical experiments for:
  - Two-stage stochastic programming
  - Walrasian equilibrium problems

Details:

- Mathematical Programming, 2020,
  DOI 10.1007/s10107-020-01582-2
- Computational Optimization and Applications, 2021,
  Vol. 78, pp. 675–704.

or

`http://www.impa.br/~optim/solodov.html`

Thanks!