Double Smoothing technique for infinite-dimensional optimization problems

O. Devolder (F.R.S.-FNRS Research Fellow),
F. Glineur and Y. Nesterov

Center for Operations Research and Econometrics (CORE),
Université catholique de Louvain (UCL)

EUROPT 2010, Aveiro, July 9, 2010

Outline

1. Studied problem class.
2. Dual Approach.
3. Double Regularization.
4. Solving the dual problem.
5. Reconstruction of a primal solution.
6. Conclusion and Further Research.

A class of infinite-dimensional optimization problems

$$P^* = \inf_{u \in L^2([0,T],\mathbb{R}^m)} \int_0^T F(t, u(t))dt$$

$$\int_0^T A(t)u(t)dt \in C$$

$$u(t) \in P(t) \quad \text{a.e. in } [0, T]$$

where

- $C \subset \mathbb{R}^n$ is convex, closed and bounded
- $P(t) \subset \mathbb{R}^m$ is convex, closed such that $P = \cup_{t\in[0,T]} P(t)$ is bounded
- $F : [0, T] \times P \to \mathbb{R}$ is convex in u, bounded and continuously differentiable in (t, u).

The problem is easy without the coupling constraint

$$P^* = \inf_{u \in L^2([0,T],\mathbb{R}^m)} \int_0^T F(t, u(t))dt$$

$$u(t) \in P(t) \quad \text{a.e. in } [0, T].$$

We can minimize in a pointwise way i.e. solve for each $t \in [0, T]$:

$$min_{u \in P(t)} F(t, u).$$

Basic idea: dualize the difficult coupling constraint

$$\int_0^T A(t)u(t)dt \in C$$

Outline

Primal Problem

Let

$$\mathcal{A} : L^2([0,T], \mathbb{R}^m) \to \mathbb{R}^n, \quad u \to \int_0^T A(t)u(t)dt$$

we have:

$$\mathcal{A}u \in C \Leftrightarrow \langle \mathcal{A}u, z\rangle \leq \sigma_C(z) \quad \forall z \in \mathbb{R}^n$$

where $\sigma_C(z) = \sup_{x\in C}\langle x, z\rangle$.
If we dualize the coupling constraint, we obtain the equivalent formulation:

$$P^* = \inf_{u\in U}\left[\int_0^T F(t,u(t))dt + \sup_{z\in\mathbb{R}^n}\left(\langle \mathcal{A}u, z\rangle - \sigma_C(z)\right)\right]$$

where $U = \{u \in L^2([0,T], \mathbb{R}^m) : u(t) \in P(t) \text{ a.e. in } [0,T]\}$.

Dual Problem

$$
\begin{aligned}
D^* &= \sup_{z\in\mathbb{R}^n}\left[-\sigma_C(z) + \inf_{u\in U}\left(\int_0^T F(t,u(t))dt + \langle \mathcal{A}u, z\rangle\right)\right] \\
&= \sup_{z\in\mathbb{R}^n} -\sigma_C(z) - \phi(z)
\end{aligned}
$$

where

$$
\phi(z) = \sup_{u\in U}\left(\int_0^T -F(t,u(t)) - \langle u(t), A(t)^T z\rangle dt\right).
$$

Why a Dual Approach ?

Advantages of the dual problem:

- The dual is an unconstrained optimization problem in finite-dimension ($z \in \mathbb{R}^n$).
- The infinite-dimensional problem defining $\phi(z) = \sup_{u \in U} \left(\int_0^T -F(t, u(t)) - \langle u(t), A(t)^T z \rangle dt \right)$. can be solved pointwisely.
  The function u^* defined at each $t \in [0, T]$ by:

  $$u^*(t) = \arg \max_{v \in P(t)} \left\{ -F(t, v) - \langle v, A(t)^T z \rangle \right\}$$

  is an optimal solution of this problem.
  We assume that these subproblems can be solved in closed-form.
- Strong Duality holds i.e : $P^* = D^*$ when the dual is solvable.

But the dual function can be non-differentiable...

Rewrite the dual problem as a minimization problem:

$$-D^* = \theta^* = \min_{z \in \mathbb{R}^n} \theta(z) = \min_{z \in \mathbb{R}^n} \sigma_C(z) + \phi(z)$$

σ_C and ϕ can be **non-differentiable**:

- $\partial \sigma_C(z) = \{\tilde{x} \in C : \langle \tilde{x}, z \rangle = \sigma_C(z)\}$
- $\partial \phi(z) =$
  $\{-\mathcal{A}\tilde{u}$ for any optimal solution $\tilde{u}$ of the problem defining $\phi(z)\}$.

Conclusion:

We have to solve a non-smooth convex optimization problem.

How to solve a non-smooth convex problem ?

- **The classical approach: subgradient-type scheme.**
  Advantage : Can be applied directly on the dual objective function without any regularization
  Disadvantage: Slow Convergence
  $$\theta(z_k) \to \theta^* \text{ in } O\left(\frac{1}{\epsilon^2}\right).$$
- **The smoothing approach.**
  We modify the dual objective function in order to be able to apply more efficient scheme of smooth convex optimization.
  Advantage : Faster convergence, we will obtain a scheme such that
  $$\theta(z_k) \to \theta^* \text{ in } O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right).$$
  Disadvantage : We have to modify the dual objective function with some regularizations.

Outline

1. Studied problem class.
2. Dual Approach.
3. Double Regularization.
4. Solving the dual problem.
5. Reconstruction of a primal solution.
6. Conclusion and Further Research.

Double Regularization of the dual objective function

In order to

- be able to solve efficiently the dual problem
- to be able to obtain a nearly optimal and feasible primal solution from a nearly optimal dual solution

we will modify the dual objective function with two regularizations:

1. A first regularization that makes the dual objective function **gradient Lipschitz-continuous**
2. A second regularization that makes the dual objective function **strongly convex.**

First regularization

Why?

In order to obtain a smooth objective function for the dual problem with gradient Lipschitz-continuous i.e.:

$$\|\nabla g(z) - \nabla g(\bar{z})\| \leq L \|z - \bar{z}\| \quad \forall z, \bar{z} \text{ with } L < +\infty$$

and therefore be able to apply efficient schemes of smooth convex optimization.

First regularization: How ?

How?

Modify the dual objective function:

$$\begin{aligned}\theta(z) &= \sigma_C(z) + \phi(z) \\ &= \sup_{x\in C}\langle x, z\rangle + \sup_{u\in U}\int_0^T \Big(-F(t,u(t)) - \langle u(t), A(t)^T z\rangle\Big)\,dt\end{aligned}$$

in

$$\begin{aligned}\theta_{\rho,\mu}(z) &= \sigma_{\rho,C}(z) + \phi_\mu(z) \\ &= \sup_{x\in C}\{\langle x, z\rangle - \frac{\rho}{2}\|x\|_2^2\} \\ &\quad + \sup_{u\in U}\int_0^T \Big(-F(t,u(t)) - \langle u(t), A(t)^T z\rangle - \frac{\mu}{2}\|u(t)\|_2^2\Big)\,dt\end{aligned}$$

with $\rho, \mu > 0$.

First regularization: $\theta_{\rho,\mu}$ is differentiable

With the addition of the strongly concave functions $-\frac{\rho}{2}\|x\|_2^2$ and $-\frac{\mu}{2}\int_0^T \|u(t)\|_2^2\,dt$, the optimization problems defining $\sigma_{\rho,C}(z)$ and $\phi_\mu(y)$ have both only one optimal solution:

$$x_{\rho,z} = \arg\min_{x\in C}\{\langle x, z\rangle - \frac{\rho}{2}\|x\|_2^2\}$$

$$u_{\mu,z} = \arg\min_{u\in U}\left\{\int_0^T \left(-F(t,u(t)) - \langle u(t), A(t)^T z\rangle - \frac{\mu}{2}\|u(t)\|_2^2\right)dt\right\}.$$

The function $\theta_{\rho,\mu}$ is therefore differentiable with gradient:

$$\nabla\theta_{\rho,\mu}(z) = x_{\rho,z} - \mathcal{A}u_{\mu,z}.$$

First regularization: Further properties of $\theta_{\rho,\mu}$

- $\theta_{\rho,\mu}$ is gradient Lipschitz-continuous with constant

$$L_{\rho,\mu} = \frac{1}{\rho} + \frac{\|\mathcal{A}\|_2^2}{\mu}$$

- $\theta_{\rho,\mu}$ is a good approximation of θ with absolute accuracy bound depending on ρ and μ:

$$\theta_{\rho,\mu}(z) \leq \theta(z) \leq \theta_{\rho,\mu}(z) + \rho D_1 + \mu D_2 \quad \forall z \in \mathbb{R}^n$$

where

$$\begin{aligned} D_1 &= \max_{x \in C} \frac{1}{2} \|x\|_2^2 \\ D_2 &= \max_{u \in U} \frac{1}{2} \|u\|_2^2 . \end{aligned}$$

Second regularization

Why?

We want not only

- to solve the dual problem

but also

- to reconstruct from the obtained nearly optimal dual solution, a nearly optimal and feasible primal solution.

Reconstruction of a primal solution

For a given dual iterate z_k, if we consider the function

$$u_k(t) = u_{\mu,z_k}(t),$$

the unique optimal solution of the problem defining $\phi_\mu(z_k)$, we have:

- $$\left| \int_0^T F(t, u_k(t))dt - P^* \right| \leq Cste \left|\theta(z_k) - \theta^*\right| + Cste \left\|\nabla\theta_{\rho,\mu}(z_k)\right\|_2$$

- $$\left\|\mathcal{A}u_k - x_{\rho,z_k}\right\|_2 = \left\|\nabla\theta_{\rho,\mu}(z_k)\right\|_2$$

  where $x_{\rho,z_k} \in C$.

The quality of this primal solution depends not only on the convergence rate of $\theta(z_k)$ to θ^* but also on the convergence rate of $\|\nabla\theta_{\rho,\mu}(z_k)\|$ to 0.

Convexity and Gradient Lipschitz-continuity are not enough

If the dual objective function is convex, gradient Lipschitz-continuous and if we apply the optimal scheme for $F_L^{1,1}(\mathbb{R}^n)$:

$$g(z_k) \to g^* \text{ in } O\left(\frac{1}{\sqrt{\epsilon}}\right)$$

but the convergence of the gradient is slower:

$$\|\nabla g(z_k)\|_2 \to 0 \text{ in } O\left(\frac{1}{\epsilon}\right).$$

Convexity and Gradient Lipschitz-continuity are not enough

In our case, if we apply this scheme to our function $\theta_{\rho,\mu} \in F^{1,1}_{L_{\rho,\mu}}(\mathbb{R}^n)$ with a good choice for ρ and μ, we have

$$\theta(z_k) - \theta^* \to 0 \text{ in } O\left(\frac{1}{\epsilon}\right)$$

but

$$\|\nabla\theta_{\rho,\mu}(z_k)\| \to 0 \text{ in } O\left(\frac{1}{\epsilon^2}\right).$$

Therefore if the dual objective function is only convex and gradient Lipschitz-continuous, we have a convergence rate in $O\left(\frac{1}{\epsilon^2}\right)$ for the primal sequence.

This is not better than with the subgradient scheme!

Is the smoothing approach useless ?

No!!!

If the dual objective function is also strongly convex, we can apply the optimal scheme for $S^{1,1}_{\kappa,L}(\mathbb{R}^n)$ for wich we have the same rate of convergence for $g(z_k) - g^*$ and $\|\nabla g(z_k)\|_2$ in

$$O\left(\ln\left(\frac{1}{\epsilon}\right)\right).$$

In our case, if we apply this scheme to the function $\theta_{\rho,\mu,\kappa} = \theta_{\rho,\mu}(z) + \frac{\kappa}{2}\|z\|_2^2 \in S^{1,1}_{\kappa,L_{\rho,\mu}+\kappa}(\mathbb{R}^n)$ with a good choice for ρ, μ and κ, we have

$$\theta(z_k) - \theta^* \to 0 \text{ in } O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)$$

and

$$\|\nabla\theta_{\rho,\mu,\kappa}(z_k)\| \to 0 \text{ in } O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right).$$

Modified Dual Objective Function

$$\theta_{\rho,\mu,\kappa}(z) = \sigma_{\rho,C}(z) + \phi_\mu(z) + \frac{\kappa}{2}\|z\|_2^2$$

where

- $\sigma_{\rho,C}(z) = \sup_{x \in C}\{\langle x, z\rangle - \frac{\rho}{2}\|x\|_2^2\}$
- $\phi_\mu(z) =$
  $\sup_{u \in U} \int_0^T \left(-F(t, u(t)) - \langle u(t), A(t)^T z\rangle - \frac{\mu}{2}\|u(t)\|_2^2\right) dt$
- $\rho, \mu, \kappa > 0$.

This function is:

- Strongly convex with parameter κ
- Gradient Lipschitz-continuous with constant $L_{\rho,\mu,\kappa} = L_{\rho,\mu} + \kappa$.

Why a double smoothing ? Summary

Method	Dual function	Dual conv.	Primal conv.
Subgradient	Convex but Non-Smooth	$O\left(\frac{1}{\epsilon^2}\right)$	$O\left(\frac{1}{\epsilon^2}\right)$
Simple Smoothing	Convex ∇ Lipschitz-cont	$O\left(\frac{1}{\epsilon}\right)$	$O\left(\frac{1}{\epsilon^2}\right)$
Double Smoothing	Strongly convex ∇ Lipschitz-cont.	$O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)$	$O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)$

Outline

Solving the dual problem using the optimal scheme for $S^{1,1}_{\kappa, L_{\rho,\mu,\kappa}}(\mathbb{R}^n)$

If we apply the optimal scheme for $S^{1,1}_{\kappa, L_{\rho,\mu,\kappa}}(\mathbb{R}^n)$ to our modified dual function, we have:

$$\theta(z_k) - \theta^* \leq \rho D_1 + \mu D_2 + \frac{\kappa}{2} R^2$$

$$+ \frac{25}{8} \left(\theta(0) - \theta(z^{**}) + \rho D_1 + \mu D_2\right) \exp\left(-\frac{k}{2}\sqrt{\frac{\kappa}{L_{\rho,\mu,\kappa}}}\right)$$

where

- z^{**} is an optimal solution of the original dual problem $\min_{z \in \mathbb{R}^n} \theta(z)$
- $\theta^* = -P^*$ is the optimal value of this problem
- R is such that $\|z^{**}\|_2 \leq R$.

If we want an accuracy $\theta(z_k) - \theta^* \leq \epsilon$, we can choose ρ, μ, κ and k such that each of the four terms are $\leq \epsilon/4$.

Choice of ρ, μ and σ

- If we want
$$\rho D_1 \leq \frac{\epsilon}{4}$$
we choose
$$\rho(\epsilon) = \frac{1}{4D_1}\epsilon = C_1\epsilon$$
- If we want
$$\mu D_2 \leq \frac{\epsilon}{4}$$
we choose
$$\mu(\epsilon) = \frac{1}{4D_2}\epsilon = C_2\epsilon$$
- If we want
$$\frac{\kappa}{2} R^2 \leq \frac{\epsilon}{4}$$
we choose
$$\kappa(\epsilon) = \frac{1}{2R^2}\epsilon = C_3\epsilon.$$

Number of iterations needed

If we want

$$\frac{25}{8}\left(\theta(0)-\theta(z^{**})+\rho D_1+\mu D_2\right)\exp\left(-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}\right)\leq\frac{\epsilon}{4}$$

we have to choose

$$k(\epsilon)\geq\sqrt{1+\frac{8}{\epsilon^2}[D_1+D_2\left\|\mathcal{A}\right\|_2^2]R^2}\ \ln\left(\frac{25(\theta(0)-\theta(z^{*}*)+\frac{\epsilon}{2})}{2\epsilon}\right)$$

$$=g_1(\epsilon)=O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right).$$

Convergence of the gradient

In order to reconstruct from the nearly optimal dual solution, a nearly feasible and optimal primal solution with a given accuracy, we need the rate of convergence of $\|\nabla\theta_{\rho,\mu}(z_k)\|$.

We have:

$$\|\nabla\theta_{\rho,\mu}(z_k)\|_2 \leq$$

$$\sqrt{2L(\rho,\mu,\kappa)(\theta(0)-\theta(z^{**})+\frac{\epsilon}{2})}\exp\left(-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}\right)+2\sqrt{3}\kappa R.$$

Convergence of the gradient: Number of iterations

With the choosen parameters $\rho(\epsilon), \mu(\epsilon), \kappa(\epsilon)$ we have:

$$2\sqrt{3}\kappa R = \frac{\sqrt{3}}{R}\epsilon.$$

Furthermore, if we want

$$\sqrt{2L(\rho,\mu,\kappa)(\theta(0) - \theta(z^{**}) + \frac{\epsilon}{2})}\exp\left(-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}\right) \leq \frac{2-\sqrt{3}}{R}\epsilon$$

we have to take

$$k(\epsilon) \geq g_2(\epsilon) = O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right).$$

Complexity

Let $\epsilon > 0$ and choose $\rho(\epsilon) = C_1\epsilon, \mu(\epsilon) = C_2\epsilon$ and $\kappa(\epsilon) = C_3\epsilon$, after

$$k(\epsilon) = \max\{g_1(\epsilon), g_2(\epsilon)\} = O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)$$

iterations, we have:

- $$\theta(z_{k(\epsilon)}) - \theta^* \leq \epsilon$$
- $$\left\|\nabla\theta_{\rho,\mu}(z_{k(\epsilon)})\right\|_2 \leq \frac{2}{R}\epsilon.$$

Outline

1. Studied problem class.
2. Dual Approach.
3. Double Regularization.
4. Solving the dual problem.
5. Reconstruction of a primal solution.
6. Conclusion and Further Research.

A nearly feasible and optimal primal solution

Using the dual iterate $z_{k(\epsilon)}$, we can obtain a primal solution with the same order of accuracy.
Consider

$$u_{k(\epsilon)}(t) = u_{\mu(\epsilon),z_{k(\epsilon)}}(t)$$

the unique optimal solution of the problem defining $\phi_{\mu(\epsilon)}(z_{k(\epsilon)})$ that we can compute analytically.

A nearly feasible and optimal primal solution (2)

This function $u_{k(\epsilon)}$ is:

- In U by construction i.e. $u_{k(\epsilon)}(t) \in P(t)$ a.e. in $[0, T]$
- Nearly optimal for the primal problem:

$$\left| \int_0^T F(t, u_{k(\epsilon)}(t))dt - P^* \right| \leq 2(1 + 2\sqrt{3})\epsilon$$

- Nearly feasible for the coupling constraints:

$$\text{dist}(\mathcal{A}u_{k(\epsilon)}, C) \leq \|\nabla\theta_{\rho,\mu}(z_k(\epsilon))\| = \left\|\mathcal{A}u_{k(\epsilon)} - x_{\rho,z_k(\epsilon)}\right\| \leq \frac{2}{R}\epsilon.$$

Outline

Double Smoothing Algorithm: Conclusion

After $k(\epsilon) = O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)$ iterations, the double smoothing algorithm provide us with a nearly optimal:

$$\left|\int_0^T F(t, u_{k(\epsilon)}(t))dt - P^*\right| \leq 2(1+2\sqrt{3})\epsilon$$

and nearly feasible:

$$\mathrm{dist}(\mathcal{A}u_{k(\epsilon)}, C) \leq \frac{2}{R}\epsilon$$

$$u_{k(\epsilon)}(t) \in P(t) \quad \forall t \in [0, T]$$

solution of the original infinite-dimensional problem.

Applications

The double smoothing scheme can be applied for solving:

- Optimal Control problems of the form:

$$P^* = \inf_{u \in L^2([0,T],\mathbb{R}^m)} \int_0^T G(t, u(t)) + \langle a(t), x(t) \rangle dt$$

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(0) = x_0$$

$$x(t_i) \in \overline{C}_i \quad i = 1, ..., N$$

$$u(t) \in P(t) \quad \text{a.e. in } [0, T]$$

  (See the Euro 2010 talk for more details)

- Large scale finite-dimensional problems with coupling and pointwise constraints (resulting typically from a discretization)

Further research

- **What is the efficiency of the double smoothing algorithm in practice ?**
  Numerical experimentation and comparison with methods based on preliminary discretization.
- How to compute $\phi_\mu(z)$ and $\nabla\phi_\mu(z)$?
  In order to obtain the exact value of these quantities, we need to compute an infnite number of pointwise minimization wich is impossible in practice.
  What are the consequence on the optimal scheme, if we use inexact first-order informations ?

Thanks for your attention !