Faster Lagrangian Methods for Convex Optimization

Marc Teboulle

School of Mathematical Sciences
Tel Aviv University

Joint work with Shoham Sabach (Technion)

Workshop on Optimization and Operator Theory
Dedicated to Lev Bregman's 80th Birthday
Technion, November 15–17, 2021 Haifa

HAPPY BIRTHDAY PROF. BREGMAN!

Goal/Outline

Main goal. **To unify, simplify, and improve the convergence rate analysis of Lagrangian-based methods for convex optimization.**

Goal/Outline

Main goal. To unify, simplify, and improve the convergence rate analysis of Lagrangian-based methods for convex optimization.

- **A central tool for Lagrangian methods: Nice Primal Algorithmic Map**
- **A framework of Faster LAGrangian (FLAG) methods**
- **New non-ergodic rate of convergence results in terms of function values and feasibility violation**

The Problem

The Linearly Constrained Convex Optimization Model

We focus on the **linearly constrained** convex optimization problem defined by

$$(\mathrm{P}) \qquad \min_{x \in \mathbb{R}^n} \left\{ \Psi(x) : \mathcal{A}x = b \right\},$$

where

- $\Psi : \mathbb{R}^n \to (-\infty, +\infty]$ is proper, lsc and σ-strongly convex with $\sigma \geq 0$.
- $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear mapping, and $b \in \mathbb{R}^m$.
- The feasible set of problem (P) is denoted by $\mathcal{F} = \{x \in \mathbb{R}^n : \mathcal{A}x = b\} \neq \emptyset$.

Despite its apparent simplicity, this model is very rich and encompasses most convex optimization models.

- **Linear composite model**

$$\min_{u\in\mathbb{R}^p}\{f(u)+g(Au)\}=\min_{u\in\mathbb{R}^p,v\in\mathbb{R}^q}\{f(u)+g(v):\ Au=v\},$$

where $f:\mathbb{R}^p\to(-\infty,+\infty]$ and $g:\mathbb{R}^q\to(-\infty,+\infty]$ are proper, lower semi-continuous and convex functions, and $A:\mathbb{R}^p\to\mathbb{R}^q$ is a linear mapping. **It fits into model (P)**, with $x=\left(u^T,v^T\right)^T$, $\Psi(x):=f(u)+g(v)$ and $\mathcal{A}x=Au-v$.

Important Particular Instances of Model (P)

- **Linear composite model**

$$\min_{u\in\mathbb{R}^p}\left\{f\left(u\right)+g\left(Au\right)\right\}=\min_{u\in\mathbb{R}^p,v\in\mathbb{R}^q}\left\{f\left(u\right)+g\left(v\right):\ Au=v\right\},$$

  where $f:\mathbb{R}^p\rightarrow(-\infty,+\infty]$ and $g:\mathbb{R}^q\rightarrow(-\infty,+\infty]$ are proper, lower semi-continuous and convex functions, and $A:\mathbb{R}^p\rightarrow\mathbb{R}^q$ is a linear mapping. **It fits into model (P)**, with $x=\left(u^T,v^T\right)^T$, $\Psi\left(x\right):=f\left(u\right)+g\left(v\right)$ and $\mathcal{A}x=Au-v$.

- **Block linear constrained model**

$$\min_{u\in\mathbb{R}^p,v\in\mathbb{R}^q}\left\{f\left(u\right)+g\left(v\right):\ Au+Bv=b\right\},$$

  where $f:\mathbb{R}^p\rightarrow(-\infty,+\infty]$ and $g:\mathbb{R}^q\rightarrow(-\infty,+\infty]$ are proper, lower semi-continuous and convex functions, $A:\mathbb{R}^p\rightarrow\mathbb{R}^m$ and $B:\mathbb{R}^q\rightarrow\mathbb{R}^m$ are linear mappings. **It fits into model (P)**, with $x=\left(u^T,v^T\right)^T$, $\Psi\left(x\right):=f\left(u\right)+g\left(v\right)$ and $\mathcal{A}x=Au+Bv$.

Important Particular Instances of Model (P)

- **Linear composite model**

$$\min_{u\in\mathbb{R}^p} \{f(u) + g(Au)\} = \min_{u\in\mathbb{R}^p, v\in\mathbb{R}^q} \{f(u) + g(v) : Au = v\},$$

  where $f : \mathbb{R}^p \to (-\infty, +\infty]$ and $g : \mathbb{R}^q \to (-\infty, +\infty]$ are proper, lower semi-continuous and convex functions, and $A : \mathbb{R}^p \to \mathbb{R}^q$ is a linear mapping. **It fits into model (P)**, with $x = \left(u^T, v^T\right)^T$, $\Psi(x) := f(u) + g(v)$ and $\mathcal{A}x = Au - v$.

- **Block linear constrained model**

$$\min_{u\in\mathbb{R}^p, v\in\mathbb{R}^q} \{f(u) + g(v) : Au + Bv = b\},$$

  where $f : \mathbb{R}^p \to (-\infty, +\infty]$ and $g : \mathbb{R}^q \to (-\infty, +\infty]$ are proper, lower semi-continuous and convex functions, $A : \mathbb{R}^p \to \mathbb{R}^m$ and $B : \mathbb{R}^q \to \mathbb{R}^m$ are linear mappings. **It fits into model (P)**, with $x = \left(u^T, v^T\right)^T$, $\Psi(x) := f(u) + g(v)$ and $\mathcal{A}x = Au + Bv$.

- **Additive smooth/non-smooth composite objective**

$$\min_{x\in\mathbb{R}^n} \{f(x) + h(x) : \mathcal{A}x = b\},$$

  with $\Psi(x) := f(x) + h(x)$ where $h : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function with a Lipschitz continuous gradient.

Lagrangians for the Convex Model (P)

We recall problem (P)

$$(\mathrm{P}) \qquad \min_{x \in \mathbb{R}^n} \left\{ \Psi(x) : \mathcal{A}x = b \right\},$$

The corresponding Lagrangian and augmented Lagrangian, are respectively given by:

$$\mathcal{L}(x, y) = \Psi(x) + \langle y, \mathcal{A}x - b \rangle, \quad y \in \mathbb{R}^m,$$

and, for any $\rho > 0$,

$$\mathcal{L}_\rho(x, y) = \mathcal{L}(x, y) + \frac{\rho}{2} \|\mathcal{A}x - b\|^2.$$

Assumption
The Lagrangian $\mathcal{L}$ has a saddle point, that is, there exists (x^*, y^*) such that

$$\mathcal{L}(x^*, y) \leq \mathcal{L}(x^*, y^*) \leq \mathcal{L}(x, y^*), \quad \forall\, x \in \mathbb{R}^n, \quad \forall\, y \in \mathbb{R}^m.$$

It can be warranted, for instance, **under standard CQ on the problem's data**.

Lagrangian-based Methods for Model (P)

Starting point: **all Lagrangian-based methods** update a couple (x, y) via

$$x^+ \in \mathcal{P}(x, y),$$
$$y^+ = y + \mu\rho\left(\mathcal{A}x^+ - b\right),$$

where $\mathcal{P}(\cdot, \cdot)$ is a primal algorithmic map and $\mu > 0$ is a scaling parameter.

Lagrangian-based Methods for Model (P)

Starting point: **all Lagrangian-based methods** update a couple (x, y) via

$$x^+ \in \mathcal{P}(x, y),$$
$$y^+ = y + \mu\rho\left(\mathcal{A}x^+ - b\right),$$

where $\mathcal{P}(\cdot, \cdot)$ is a primal algorithmic map and $\mu > 0$ is a scaling parameter.

- The main difference between Lagrangian-based methods is encapsulated through the choice of a primal algorithmic map that updates the primal variable.

- This primal map can be seen as the step of any optimization method that is applied on the augmented Lagrangian itself or a variation of it.

Examples – Some Classical Lagrangian Based Schemes

$$\mathcal{L}_\rho(x,y) = \Psi(x) + \langle y, \mathcal{A}x - b\rangle + \frac{\rho}{2}\|\mathcal{A}x - b\|^2$$

Augmented Lagrangian (Hestenes (69), Powell (69))

Main step: Given (x,y), update the new point $\left(x^+, y^+\right)$ via:

$$x^+ \in \operatorname{argmin}\left\{\mathcal{L}_\rho(\xi, y) : \xi \in \mathbb{R}^n\right\},$$
$$y^+ = y + \mu\rho\left(\mathcal{A}x^+ - b\right).$$

In this case, $\mathcal{P}$ is an **exact minimization** applied on the augmented Lagrangian.

Examples – Some Classical Lagrangian Based Schemes

$$\mathcal{L}_\rho(x,y) = \Psi(x) + \langle y, \mathcal{A}x - b\rangle + \frac{\rho}{2}\|\mathcal{A}x - b\|^2$$

Augmented Lagrangian (Hestenes (69), Powell (69))

Main step: Given (x, y), update the new point $\left(x^+, y^+\right)$ via:

$$x^+ \in \operatorname{argmin}\left\{\mathcal{L}_\rho(\xi, y) : \xi \in \mathbb{R}^n\right\},$$
$$y^+ = y + \mu\rho\left(\mathcal{A}x^+ - b\right).$$

In this case, $\mathcal{P}$ is an **exact minimization** applied on the augmented Lagrangian.

Proximal Linearized Augmented Lagrangian Main step: Given (x, y), update the new point $\left(x^+, y^+\right)$ via:

$$x^+ \in \operatorname{argmin}\left\{\Psi(\xi) + \left\langle \xi, \mathcal{A}^T\left(y + \rho\left(\mathcal{A}x - b\right)\right)\right\rangle + \frac{1}{2}\|\xi - x\|_M^2 : \xi \in \mathbb{R}^n\right\}, (M \succ 0)$$
$$y^+ = y + \mu\rho\left(\mathcal{A}x^+ - b\right).$$

In this case, $\mathcal{P}$ is a **proximal gradient** applied on the augmented Lagrangian.

Examples: Some Classical Schemes for Block Models

As discussed above, Model (P) covers the following block model

$$\min_{(u,v)\in\mathbb{R}^n} \left\{ f(u) + g(v) \, : \, Au + Bv = b \right\}.$$

$$\mathcal{L}_\rho(u, v, y) = f(u) + g(v) + \langle y, Au + Bv - b \rangle + \frac{\rho}{2} \|Au + Bv - b\|^2 .$$

However, the block structure can be exploited in designing Lagrangian-based methods.

As discussed above, Model (P) covers the following block model

$$\min_{(u,v)\in\mathbb{R}^n} \{f(u) + g(v) : Au + Bv = b\}.$$

$$\mathcal{L}_\rho(u,v,y) = f(u) + g(v) + \langle y, Au + Bv - b\rangle + \frac{\rho}{2}\|Au + Bv - b\|^2.$$

However, the block structure can be exploited in designing Lagrangian-based methods.

Alternating Direction Method of Multipliers (ADMM)

(Glowinski and Marroco (75), Gabay and Mercier (76))

Main step: Given (u, v, y), update the new point $\left(u^+, v^+, y^+\right)$ via:

$$\begin{aligned} u^+ &= \operatorname{argmin}\{\mathcal{L}_\rho(\xi, v, y) : \xi \in \mathbb{R}^n\}, \\ v^+ &= \operatorname{argmin}\left\{\mathcal{L}_\rho\left(u^+, \eta, y\right) : \eta \in \mathbb{R}^m\right\}, \\ y^+ &= y + \mu\rho\left(Au^+ + Bv^+ - b\right). \end{aligned}$$

In this case, $\mathcal{P}$ is an **alternating minimization** applied on the augmented Lagrangian.

$$\mathcal{L}_{\rho}\left(u, v, y\right) = f\left(u\right) + g\left(v\right) + \left\langle y, Au + Bv - b\right\rangle + \frac{\rho}{2}\left\|Au + Bv - b\right\|^{2}.$$

Previous steps can be difficult to implement. Instead, we can *approximate them*:

$$\mathcal{L}_\rho\left(u, v, y\right) = f\left(u\right) + g\left(v\right) + \langle y, Au + Bv - b\rangle + \frac{\rho}{2}\left\|Au + Bv - b\right\|^2 .$$

Previous steps can be difficult to implement. Instead, we can *approximate them*:

Proximal Linearized ADMM

Main step: Given (u, v, y), update the new point $\left(u^+, v^+, y^+\right)$ via:

$$u^+ = \underset{\xi}{\operatorname{argmin}}\left\{f\left(\xi\right) + \left\langle A^T\left(y + \rho\left(Au + Bv - b\right)\right), \xi - u\right\rangle + \frac{1}{2}\left\|\xi - u\right\|_{M_1}^2\right\},$$

$$v^+ = \underset{\eta}{\operatorname{argmin}}\left\{g\left(\eta\right) + \left\langle B^T\left(y + \rho\left(Au^+ + Bv - b\right)\right), \eta - v\right\rangle + \frac{1}{2}\left\|\eta - v\right\|_{M_2}^2\right\},$$

$$y^+ = y + \mu\rho\left(Au^+ + Bv^+ - b\right).$$

(Here $M_1, M_2 \succ 0$).

In this case, $\mathcal{P}$ is a **alternating proximal gradient** applied on the augmented Lagrangian.

Unified Framework

Nice Primal Algorithmic Map

It captures the essential ingredients and plays a central role in unifying the analysis of all Lagrangian-based methods into a single and simple framework.

Nice Primal Algorithmic Map

It captures the essential ingredients and plays a central role in unifying the analysis of all Lagrangian-based methods into a single and simple framework.

Definition [Nice primal algorithmic map] Given the parameters $\rho, t > 0$, let

$$(\rho_t, \tau_t) := \begin{cases} (\rho, t^{-1}) & \text{if } \sigma = 0 \\ (\rho t, \quad t) & \text{if } \sigma > 0. \end{cases}$$

A primal algorithmic map $\mathbb{S}_t : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ which is applied on the augmented Lagrangian $\mathcal{L}_{\rho_t}(z, \lambda)$ that generates $z^+ \in \mathbb{S}_t(z, \lambda)$, is called **nice**, **if there exist** $\delta \in (0, 1]$ **and** $P, Q \succeq 0$, such that for any $\xi \in \mathcal{F}$ we have

$$\mathcal{L}_{\rho_t}\left(z^+, \lambda\right) - \mathcal{L}_{\rho_t}(\xi, \lambda) \leq \tau_t \Delta_P\left(\xi, z, z^+\right) - \frac{\tau_t}{2}\left\|z^+ - z\right\|_Q^2 - \frac{\sigma}{2}\left\|\xi - z^+\right\|^2 - \frac{\delta \rho_t}{2}\left\|\mathcal{A}z^+ - b\right\|^2$$

- For any matrix $P \succeq 0$ and any three vectors $u, v, w \in \mathbb{R}^n$:
$$\Delta_P(u, v, w) := \frac{1}{2}\|u - v\|_P^2 - \frac{1}{2}\|u - w\|_P^2 .$$
- When $P \equiv I_n$, the identity matrix, we simply write $\Delta_P(u, v, w) \equiv \Delta(u, v, w)$.

FLAG – Faster LAGrangian based method

1. **Input:** **Problem data** $[\Psi, \mathcal{A}, b, \sigma]$, and a **nice primal algorithmic map** $\mathbb{S}_t(\cdot)$.
2. **Initialization:** Set $t_0 = 1$, $\mu \in (0, \delta]$ and $\rho > 0$. Start with any (x^0, z^0, y^0).

FLAG – Faster LAGrangian based method

1. **Input:** **Problem data** $[\Psi, \mathcal{A}, b, \sigma]$, and a **nice primal algorithmic map** $\mathbb{S}_t(\cdot)$.
2. **Initialization:** Set $t_0 = 1$, $\mu \in (0, \delta]$ and $\rho > 0$. Start with any $\left(x^0, z^0, y^0\right)$.
3. **Iterations:** Generate $\left\{\left(x^k, z^k, y^k\right)\right\}_{k \in \mathbb{N}}$ and $\{t_k\}_{k \in \mathbb{N}}$ via
   3.1. Compute
   $$\lambda^k = y^k + \rho_k (t_k - 1)\left(\mathcal{A}x^k - b\right), \quad \text{with} \quad \rho_k = \begin{cases} \rho, & \text{if } \sigma = 0 \\ \rho t_k & \text{if } \sigma > 0. \end{cases}$$
   3.2. Update the sequence $\left\{\left(x^k, z^k, y^k\right)\right\}_{k \in \mathbb{N}}$ by
   $$\begin{aligned} z^{k+1} &\in \mathbb{S}_{t_k}\left(z^k, \lambda^k\right), \\ y^{k+1} &= y^k + \mu\rho_k\left(\mathcal{A}z^{k+1} - b\right), \\ x^{k+1} &= \left(1 - t_k^{-1}\right)x^k + t_k^{-1}z^{k+1}. \end{aligned}$$
   3.3. Update the sequence $\{t_k\}_{k \in \mathbb{N}}$ by solving the equation $t_{k+1}^p - t_k^p = t_{k+1}^{p-1}$, *i.e.*,
   $$t_{k+1} = \begin{cases} t_k + 1, & p = 1 \text{ (convex case)}, \\ \left(1 + \sqrt{1 + 4t_k^2}\right)/2, & p = 2 \text{ (strongly convex case)}. \end{cases}$$

FLAG - A Few Comments

- Setting $t_k \equiv 1$ in FLAG, implies $\rho_k \equiv \rho$, $\lambda^k \equiv y^k$, and $x^k \equiv z^k$, thus **recovering the classical basic Lagrangian-based methods**.

FLAG - A Few Comments

- Setting $t_k \equiv 1$ in FLAG, implies $\rho_k \equiv \rho$, $\lambda^k \equiv y^k$, and $x^k \equiv z^k$, thus **recovering the classical basic Lagrangian-based methods**.

- Borrows ideas from acceleration of FOM (Nesterov (83), Auslender and T. (06)).
- The choice of t_k plays a key role in accelerating the nice primal map $\mathbb{S}_t$. Both the augmented parameter ρ_k and the prox parameter τ_k are determined and chosen through the recursion which defines the sequence t_k.

FLAG - A Few Comments

- Setting $t_k \equiv 1$ in FLAG, implies $\rho_k \equiv \rho$, $\lambda^k \equiv y^k$, and $x^k \equiv z^k$, thus **recovering the classical basic Lagrangian-based methods**.

- Borrows ideas from acceleration of FOM (Nesterov (83), Auslender and T. (06)).
- The choice of t_k plays a key role in accelerating the nice primal map $\mathbb{S}_t$. Both the augmented parameter ρ_k and the prox parameter τ_k are determined and chosen through the recursion which defines the sequence t_k.
- A main new feature of FLAG is the **auxiliary variable** λ^k defined by:

$$\lambda^k = y^k + \rho_k (t_k - 1) \left(\mathcal{A} x^k - b \right),$$

which enable us to derive the new faster **non-ergodic** rate of convergence results!

FLAG - A Few Comments

- Setting $t_k \equiv 1$ in FLAG, implies $\rho_k \equiv \rho$, $\lambda^k \equiv y^k$, and $x^k \equiv z^k$, thus **recovering the classical basic Lagrangian-based methods**.

- Borrows ideas from acceleration of FOM (Nesterov (83), Auslender and T. (06)).
- The choice of t_k plays a key role in accelerating the nice primal map $\mathbb{S}_t$. Both the augmented parameter ρ_k and the prox parameter τ_k are determined and chosen through the recursion which defines the sequence t_k.
- A main new feature of FLAG is the **auxiliary variable** λ^k defined by:

$$\lambda^k = y^k + \rho_k (t_k - 1) \left(\mathcal{A}x^k - b\right),$$

  which enable us to derive the new faster **non-ergodic** rate of convergence results!
- $\mathbb{S}_t$ is assumed to be **nice primal algorithmic map** and this is **all we need** to guarantee rate of convergence results (classical and fast)!

Analysis

The Two Main Pillars of the Analysis

- The analysis of Lagrangian based methods is usually complicated, and relies on very lengthy and nontrivial proofs.
- Here, it relies on two key lemmas, admitting simple proofs; half-page each!

The Two Main Pillars of the Analysis

- The analysis of Lagrangian based methods is usually complicated, and relies on very lengthy and nontrivial proofs.
- Here, it relies on two key lemmas, admitting simple proofs; half-page each!

Lemma 1 Let $\left\{\left(x^k, z^k, y^k\right)\right\}_{k\in\mathbb{N}}$ generated by FLAG. Then, for any $\xi \in \mathcal{F}$, $\eta \in \mathbb{R}^m$ and $k \geq 0$,

$$\mathcal{L}_{\rho_k}\left(z^{k+1}, \eta\right) - \mathcal{L}_{\rho_k}\left(\xi, \eta\right) \leq \tau_k \Delta_P\left(\xi, z^k, z^{k+1}\right) - \frac{\sigma}{2}\left\|\xi - z^{k+1}\right\|^2 + \frac{1}{\mu\rho_k}\Delta\left(\eta, y^k, y^{k+1}\right)$$
$$- \rho t_{k-1}^p \left\langle \mathcal{A}x^k - b, \mathcal{A}z^{k+1} - b\right\rangle .$$

The Two Main Pillars of the Analysis

- The analysis of Lagrangian based methods is usually complicated, and relies on very lengthy and nontrivial proofs.
- Here, it relies on two key lemmas, admitting simple proofs; half-page each!

Lemma 1 Let $\left\{\left(x^k, z^k, y^k\right)\right\}_{k\in\mathbb{N}}$ generated by FLAG. Then, for any $\xi \in \mathcal{F}$, $\eta \in \mathbb{R}^m$ and $k \geq 0$,

$$\mathcal{L}_{\rho_k}\left(z^{k+1}, \eta\right) - \mathcal{L}_{\rho_k}(\xi, \eta) \leq \tau_k \Delta_P\left(\xi, z^k, z^{k+1}\right) - \frac{\sigma}{2}\left\|\xi - z^{k+1}\right\|^2 + \frac{1}{\mu\rho_k}\Delta\left(\eta, y^k, y^{k+1}\right) - \rho t_{k-1}^p \left\langle \mathcal{A}x^k - b, \mathcal{A}z^{k+1} - b\right\rangle.$$

Lemma 2 Let $\left\{\left(x^k, z^k, y^k\right)\right\}_{k\in\mathbb{N}}$ generated by FLAG. Then, for any $\xi \in \mathcal{F}$, $\eta \in \mathbb{R}^m$ and $k \geq 0$,

$$t_k^p s_{k+1} - t_{k-1}^p s_k \leq \frac{\tau_k \rho_k}{\rho}\Delta_P\left(\xi, z^k, z^{k+1}\right) - \frac{\rho_k \sigma}{2\rho}\left\|\xi - z^{k+1}\right\|^2 + \frac{1}{\mu\rho}\Delta\left(\eta, y^k, y^{k+1}\right),$$

where $s_k = \mathcal{L}_{\rho t_{k-1}^p}\left(x^k, \eta\right) - \mathcal{L}_{\rho t_{k-1}^p}(\xi, \eta)$.

Main results: Rate of Convergence

Types of Rate of Convergence – Many Results

We focus on non-asymptotic rate of convergence (iteration complexity) using the following two classical measures:

(i) **Function values gap** in terms of $\Psi\left(x^k\right) - \Psi\left(x^*\right)$.

(ii) **Feasibility violation** of the constraints of problem (P) in terms of $\left\|\mathcal{A}x^k - b\right\|$.

Other measures in the literature: Lagrangian, $\left\|x^k - x^*\right\|^2$, $\left\|x^{k+1} - x^k\right\|^2$, etc.

When discussing these measures, we also distinguish between rates expressed in terms of the **original produced sequence** or of the **ergodic sequence**.

Types of Rate of Convergence – Many Results

We focus on non-asymptotic rate of convergence (iteration complexity) using the following two classical measures:

(i) **Function values gap** in terms of $\Psi\left(x^k\right) - \Psi\left(x^*\right)$.

(ii) **Feasibility violation** of the constraints of problem (P) in terms of $\left\|\mathcal{A}x^k - b\right\|$.

Other measures in the literature: Lagrangian, $\left\|x^k - x^*\right\|^2$, $\left\|x^{k+1} - x^k\right\|^2$, etc.

When discussing these measures, we also distinguish between rates expressed in terms of the **original produced sequence** or of the **ergodic sequence**.

- Many rate of convergence results in the literature for variants of Lagrangian-based methods. **Mostly ergodic!** (Chambolle and Pock (11),He and Yuan (12), Monteiro-Svaiter (13),...)

Types of Rate of Convergence – Many Results

We focus on non-asymptotic rate of convergence (iteration complexity) using the following two classical measures:

(i) **Function values gap** in terms of $\Psi\left(x^k\right) - \Psi\left(x^*\right)$.

(ii) **Feasibility violation** of the constraints of problem (P) in terms of $\left\|\mathcal{A}x^k - b\right\|$.

Other measures in the literature: Lagrangian, $\left\|x^k - x^*\right\|^2$, $\left\|x^{k+1} - x^k\right\|^2$, etc.

When discussing these measures, we also distinguish between rates expressed in terms of the **original produced sequence** or of the **ergodic sequence**.

- Many rate of convergence results in the literature for variants of Lagrangian-based methods. **Mostly ergodic!** (Chambolle and Pock (11),He and Yuan (12), Monteiro-Svaiter (13),...)
- Non-ergodic $O(1/N)$ result for $\left\|x^{k+1} - x^k\right\|^2$ (He and Yuan (15)).
- Non-ergodic $O(1/N^2)$ result for $\left\|x^k - x^*\right\|^2$, in the strongly convex setting (Chambolle and Pock (11)).

Types of Rate of Convergence – Many Results

We focus on non-asymptotic rate of convergence (iteration complexity) using the following two classical measures:

(i) **Function values gap** in terms of $\Psi\left(x^k\right) - \Psi\left(x^*\right)$.

(ii) **Feasibility violation** of the constraints of problem (P) in terms of $\left\|\mathcal{A}x^k - b\right\|$.

Other measures in the literature: Lagrangian, $\left\|x^k - x^*\right\|^2$, $\left\|x^{k+1} - x^k\right\|^2$, etc.

When discussing these measures, we also distinguish between rates expressed in terms of the **original produced sequence** or of the **ergodic sequence**.

- Many rate of convergence results in the literature for variants of Lagrangian-based methods. **Mostly ergodic!** (Chambolle and Pock (11),He and Yuan (12), Monteiro-Svaiter (13),...)
- Non-ergodic $O(1/N)$ result for $\left\|x^{k+1} - x^k\right\|^2$ (He and Yuan (15)).
- Non-ergodic $O(1/N^2)$ result for $\left\|x^k - x^*\right\|^2$, in the strongly convex setting (Chambolle and Pock (11)).
- Non-ergodic $O(1/N)$ rate of convergence result in terms of function values and feasibility violation for the **specific Linearized ADMM** (Li and Lin (19)).

The strongly convex case $\sigma > 0$.

Theorem 1. (A fast non-ergodic function values and feasibility violation rates) Let $\left\{\left(x^k, z^k, y^k\right)\right\}_{k\in\mathbb{N}}$ be a sequence generated by FLAG. **Suppose that $\sigma > 0$ and $0 \preceq P \preceq (\sigma/2)\, I_n$.** Let $c > 0$ with $c \geq 2\|y^*\|$, where y^* is an optimal solution of the dual problem. Then, for any optimal solution x^* of problem (P) and $N \geq 1$,

$$\Psi\left(x^N\right) - \Psi\left(x^*\right) \leq \frac{B_{\rho,c}\left(x^*\right)}{2N^2} \qquad \text{and} \qquad \left\|\mathcal{A}x^N - b\right\| \leq \frac{B_{\rho,c}\left(x^*\right)}{cN^2},$$

where $B_{\rho,c}\left(x^*\right) := 4\left(\left\|x^* - z^0\right\|_P^2 + \frac{1}{\mu\rho}\left(\left\|y^0\right\| + c\right)^2\right)$.

The convex case $\sigma = 0$.

Theorem 2. (A non-ergodic function values and feasibility violation rates) Let $\left\{\left(x^k, z^k, y^k\right)\right\}_{k\in\mathbb{N}}$ be a sequence generated by FLAG and **suppose that $\sigma = 0$**. Let $c > 0$ with $c \geq 2\|y^*\|$, where y^* is an optimal solution of the dual problem. Then for any optimal solution x^* of problem (P) and $N \geq 1$,

$$\Psi\left(x^N\right) - \Psi\left(x^*\right) \leq \frac{B_{\rho,c}\left(x^*\right)}{2N} \qquad \text{and} \qquad \left\|\mathcal{A}x^N - b\right\| \leq \frac{B_{\rho,c}\left(x^*\right)}{cN},$$

where $B_{\rho,c}\left(x^*\right) := 2\left(\left\|x^* - z^0\right\|_P^2 + \frac{1}{\mu\rho}\left(\left\|y^0\right\| + c\right)^2\right)$.

Ergodic type results

FLAG is versatile

- Deriving weaker results of ergodic type was not our primary goal.
- Nevertheless, our main framework FLAG easily adapt to that task.
  (The sequences $\left\{x^k\right\}_{k\in\mathbb{N}}$ and $\left\{\lambda^k\right\}_{k\in\mathbb{N}}$ are not used for that scenario!)

See details in the paper.

Nice Primal Algorithmic Maps and their FLAG

Nice Primal Algorithmic Map for Block Model

The notion of nice algorithmic map is flexible and **easily adapt to the block setting:**

$$\min_{x:=(u,v)\in\mathbb{R}^p\times\mathbb{R}^q=\mathbb{R}^n}\left\{\Psi(x):=f\left(u\right)+g\left(v\right)\;:\;\mathcal{A}x:=Au+Bv=b\right\}.$$

In the block model, we only need to assume that **either f or g is strongly convex**.

Nice Primal Algorithmic Map for Block Model

The notion of nice algorithmic map is flexible and **easily adapt to the block setting:**

$$\min_{x:=(u,v)\in\mathbb{R}^p\times\mathbb{R}^q=\mathbb{R}^n}\left\{\Psi(x):=f(u)+g(v)\,:\,\mathcal{A}x:=Au+Bv=b\right\}.$$

In the block model, we only need to assume that **either f or g is strongly convex**.

Definition [Nice primal algorithmic map - Block version]

Given the parameters $\rho, t > 0$, we let $(\rho_t, \tau_t) = (\rho, t^{-1})$ (when $\sigma = 0$) and $(\rho_t, \tau_t) = (\rho t, t)$ (when $\sigma > 0$). A primal algorithmic map $\mathbb{S}_t : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, which is applied on the augmented Lagrangian $\mathcal{L}_{\rho_t}(z, \lambda)$, that generates $z^+ = \left(u^+, v^+\right)$ via $z^+ \in \mathbb{S}_t(z, \lambda)$, is *nice*, if there exist $\delta \in (0, 1]$ and $P_1, Q_1 \in \mathbb{S}^p_+$ and $P_2, Q_2 \in \mathbb{S}^q_+$ with $P = (P_1, P_2)$ and $Q = (Q_1, Q_2)$, s.t. for any $(\xi_1, \xi_2) \in \mathcal{F}$

$$\begin{aligned}\mathcal{L}_{\rho_t}\left(z^+,\lambda\right)-\mathcal{L}_{\rho_t}(\xi,\lambda)\le{}&\frac{1}{t}\Delta_{P_1}\left(\xi_1,u,u^+\right)-\frac{1}{2t}\left\|u^+-u\right\|^2_{Q_1}+\tau_t\Delta_{P_2}\left(\xi_2,v,v^+\right)\\&-\frac{\tau_t}{2}\left\|v^+-v\right\|^2_{Q_2}-\frac{\sigma}{2}\left\|\xi_2-v^+\right\|^2-\frac{\delta\rho_t}{2}\left\|\mathcal{A}z^+-b\right\|^2.\end{aligned}$$

- **Note:** Here we use g strongly convex. Hence, only the block v uses τ_t (to cover both convex/strongly convex cases). For the other block u (with only convexity. i.e., $\sigma = 0$), we fixed $\tau_t = t^{-1}$.

Iconic Lagrangian-based Methods Admit a Nice Primal Algorithmic Map!

- Augmented Lagrangian Methods (classical, proximal, and prox-linearized)
- Alternating Direction Method of Multipliers ADMM
- Proximal ADMM
- Proximal Linearized ADMM
- Chambolle-Pock Method
- Proximal Jacobi Direction Method of Multipliers
- Predictor Corrector Proximal Multipliers

For each method an explicit parameter δ and matrices P, Q can be found!

(See details in paper.)

Iconic Lagrangian-based Methods Admit a Nice Primal Algorithmic Map!

- Augmented Lagrangian Methods (classical, proximal, and prox-linearized)
- Alternating Direction Method of Multipliers ADMM
- Proximal ADMM
- Proximal Linearized ADMM
- Chambolle-Pock Method
- Proximal Jacobi Direction Method of Multipliers
- Predictor Corrector Proximal Multipliers

For each method an explicit parameter δ and matrices P, Q can be found!

(See details in paper.)

Meaning, they all **admit Nice Primal Algorithmic Map!**

Therefore, our nonergodic convergence rate results can be applied.

In addition, nice primal algorithmic maps, can be also be identified for problems with **a composite objective**, as we illustrate next.

Model (P) with the Sum Composite Objective Function

We consider the sum composite model: nonsmooth + smooth objective

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) + h(x) \, : \, \mathcal{A}x = b \right\},$$

- $f : \mathbb{R}^n \to (-\infty, +\infty]$ is a proper, lsc and σ-strongly convex ($\sigma \geq 0$),
- $h : \mathbb{R}^n \to \mathbb{R}$ is convex C^1 with ***L*-Lipschitz continuous gradient**.

Model (P) with the Sum Composite Objective Function

We consider the sum composite model: nonsmooth + smooth objective

$$\min_{x\in\mathbb{R}^n} \left\{ f\left(x\right) + h\left(x\right) \, : \, \mathcal{A}x = b \right\},$$

- $f : \mathbb{R}^n \to (-\infty, +\infty]$ is a proper, lsc and σ-strongly convex ($\sigma \geq 0$),
- $h : \mathbb{R}^n \to \mathbb{R}$ is convex C^1 with **L-Lipschitz continuous gradient**.

Lemma (Proximal AL is nice)
Let $M \succeq LI_n$, the primal algorithmic map $\mathbb{S}_t\left(\cdot\right)$ defined by

$$z^{+} = \underset{\xi}{argmin} \left\{ f\left(\xi\right) + \left\langle \nabla h\left(z\right), \xi \right\rangle + \left\langle \lambda, \mathcal{A}\xi - b \right\rangle + \frac{\rho_t}{2} \left\| \mathcal{A}\xi - b \right\|^2 + \frac{\tau_t}{2} \left\| \xi - z \right\|_M^2 \right\},$$

is nice with $\delta = 1$ and $P = M$ and $Q = M - LI_n$.

Model (P) with the Sum Composite Objective Function

We consider the sum composite model: nonsmooth + smooth objective

$$\min_{x\in\mathbb{R}^n}\left\{f(x)+h(x)\,:\,\mathcal{A}x=b\right\},$$

- $f:\mathbb{R}^n\to(-\infty,+\infty]$ is a proper, lsc and σ-strongly convex ($\sigma\geq 0$),
- $h:\mathbb{R}^n\to\mathbb{R}$ is convex C^1 with **L-Lipschitz continuous gradient**.

Lemma (Proximal AL is nice)
Let $M\succeq LI_n$, the primal algorithmic map $\mathbb{S}_t(\cdot)$ defined by

$$z^+=\underset{\xi}{argmin}\left\{f(\xi)+\langle\nabla h(z),\xi\rangle+\langle\lambda,\mathcal{A}\xi-b\rangle+\frac{\rho_t}{2}\|\mathcal{A}\xi-b\|^2+\frac{\tau_t}{2}\|\xi-z\|_M^2\right\},$$

is nice with $\delta=1$ and $P=M$ and $Q=M-LI_n$.

Lemma (Proximal Linearized AL is nice)
Let $M\succeq\rho\mathcal{A}^T\mathcal{A}+LI_n$, the primal algorithmic map $\mathbb{S}_t(\cdot)$ defined by

$$z^+=\underset{\xi}{argmin}\left\{f(\xi)+\langle\nabla h(z),\xi\rangle+\langle\lambda,\mathcal{A}\xi-b\rangle+\rho_t\langle\mathcal{A}z-b,\mathcal{A}\xi\rangle+\frac{\tau_t}{2}\|\xi-z\|_M^2\right\},$$

is nice with $\delta=1$ and $P=M-\rho\mathcal{A}^T\mathcal{A}$ and $Q=M-\rho\mathcal{A}^T\mathcal{A}-LI_n$.

Epilogue

Epilogue

A Simple Recipe for Rate of Convergence of Lagrangian-based Methods

(i) Formulate the problem at hand via model (P), *i.e.*, **identify the relevant problem data** $[\Psi, \mathcal{A}, b, \sigma]$. **The value of σ will determine the type of rate that can be achieved (classical or fast)**.

(ii) **Define the desired iterative step(s) of the primal algorithmic map $\mathbb{S}_t(\cdot)$** applied on the augmented Lagrangian $\mathcal{L}_{\rho_t}(\cdot)$ of model (P).

Epilogue

A Simple Recipe for Rate of Convergence of Lagrangian-based Methods

(i) Formulate the problem at hand via model (P), *i.e.*, **identify the relevant problem data** $[\Psi, \mathcal{A}, b, \sigma]$. **The value of σ will determine the type of rate that can be achieved (classical or fast)**.

(ii) **Define the desired iterative step(s) of the primal algorithmic map** $\mathbb{S}_t(\cdot)$ applied on the augmented Lagrangian $\mathcal{L}_{\rho_t}(\cdot)$ of model (P).

(iii) **Show that the defined primal algorithmic map is nice**, *i.e.*, determine the parameter δ and the matrices P and Q.

A Simple Recipe for Rate of Convergence of Lagrangian-based Methods

(i) Formulate the problem at hand via model (P), *i.e.*, **identify the relevant problem data** $[\Psi, \mathcal{A}, b, \sigma]$. **The value of σ will determine the type of rate that can be achieved (classical or fast)**.

(ii) **Define the desired iterative step(s) of the primal algorithmic map** $\mathbb{S}_t(\cdot)$ applied on the augmented Lagrangian $\mathcal{L}_{\rho_t}(\cdot)$ of model (P).

(iii) **Show that the defined primal algorithmic map is nice**, *i.e.*, determine the parameter δ and the matrices P and Q.

(iv) Apply Theorem 1 (if $\sigma > 0$) or Theorem 2 (if $\sigma = 0$) for the corresponding FLAG to **obtain a faster non-ergodic rate of convergence for the designed method**.

Therefore, there is no need any more to enter into the machinery of the proofs!

More details and Results

Sabach, S. and Teboulle, M. Faster Lagrangian-Based Methods in Convex Optimization.
SIAM J. Optimization. To appear.

`http://www.math.tau.ac.il/~teboulle/`

More details and Results

Sabach, S. and Teboulle, M. Faster Lagrangian-Based Methods in Convex Optimization.
***SIAM J. Optimization. To appear*.**

`http://www.math.tau.ac.il/~teboulle/`

Thank you for listening!

Happy Birthday Prof. Bregman !