Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 11: Constrained convex minimization I

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2018)

lions@epfl

License Information for Mathematics of Data Slides

Outline

- Today
  1. Primal-Dual methods
- Next week
  1. Frank-Wolfe method
  2. Primal-dual Frank Wolfe methods

Recommended readings

- Quoc Tran-Dinh, Olivier Fercoq and Volkan Cevher, *A Smooth Primal-Dual Optimization Framework for Nonsmooth Composite Convex Minimization.* to appear in SIOPT, 2017.
- Y. Nesterov, *Smooth Minimization of Non-smooth Functions.* Math. Program., Ser. A, 103:127-152, 2005.

Swiss army knife of convex formulations

A **primal problem** prototype

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K},\ \mathbf{x} \in \mathcal{X} \Big\}, \tag{1}$$

- f is a proper, closed and convex function
- $\mathcal{X}$ and $\mathcal{K}$ are nonempty, closed convex sets
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{b} \in \mathbb{R}^n$ are known
- An optimal solution $\mathbf{x}^\star$ to (1) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$ and $\mathbf{x}^\star \in \mathcal{X}$

The role of convexity

An example from sparseland $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \|\mathbf{w}\|_2, \|\mathbf{x}\|_\infty \leq 1 \right\}. \qquad \text{(SOCP)}$$

Theorem (A model recovery guarantee [20])

Let $\mathbf{A} \in \mathbb{R}^{n\times p}$ be a matrix of i.i.d. Gaussian random variables with zero mean and variances $1/n$. For any $t > 0$ with probability at least $1 - 6\exp\left(-t^2/26\right)$, we have

$$\left\|\mathbf{x}^\star - \mathbf{x}^\natural\right\|_2 \leq \left[\frac{2\sqrt{2s\log(\frac{p}{s}) + \frac{5}{4}s}}{\sqrt{n} - \sqrt{2s\log(\frac{p}{s}) + \frac{5}{4}s} - t} \right] \|\mathbf{w}\|_2 := \varepsilon, \quad \textit{when } \|\mathbf{x}^\natural\|_0 \leq s.$$

Observations:

- perfect recovery (i.e., $\varepsilon = 0$) with $n \geq 2s\log(\frac{p}{s}) + \frac{5}{4}s$ whp when $\mathbf{w} = 0$.
- ϵ-accurate solution in $k = \mathcal{O}\left(\sqrt{2p+1}\log(\frac{1}{\epsilon})\right)$ iterations via IPM[1]
  with each iteration requiring the solution of a structured $n \times 2p$ linear system.[2]
- robust to noise.

[1] There is a subtle yet important caveat here that I am sweeping under the carpet!

[2] When $\mathbf{w} = 0$, the IPM complexity (# of iterations $\times$ cost per iteration) amounts to $\mathcal{O}(n^2 p^{1.5}\log(\frac{1}{\epsilon}))$.

Swiss army knife of convex formulations

A **primal problem** prototype

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K},\ \mathbf{x} \in \mathcal{X} \right\}, \tag{2}$$

- f is a proper, closed and convex function
- $\mathcal{X}$ and $\mathcal{K}$ are nonempty, closed convex sets
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{b} \in \mathbb{R}^n$ are known
- An optimal solution $\mathbf{x}^\star$ to (2) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$ and $\mathbf{x}^\star \in \mathcal{X}$

An example from the sparseland

$$\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \kappa, \|\mathbf{x}\|_\infty \leq c \right\} \tag{SOCP}$$

Broad context for (2):

- Standard convex optimization formulations: *linear programming, convex quadratic programming, second order cone programming, semidefinite programming and geometric programming*.
- Reformulations of existing unconstrained problems via **convex splitting**: *composite convex minimization, consensus optimization, ...*

Swiss army knife of convex formulations

A **primal problem** prototype

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K},\ \mathbf{x} \in \mathcal{X} \right\}, \quad (2)$$

- f is a proper, closed and convex function
- $\mathcal{X}$ and $\mathcal{K}$ are nonempty, closed convex sets
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{b} \in \mathbb{R}^n$ are known
- An optimal solution $\mathbf{x}^\star$ to (2) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$ and $\mathbf{x}^\star \in \mathcal{X}$

A key advantage of the unified formulation (2): **Primal-dual methods**

- decentralized collection & storage of data
- cheap per-iteration costs & distributed computation

Broad context for (2):

- Standard convex optimization formulations: *linear programming, convex quadratic programming, second order cone programming, semidefinite programming and geometric programming*.
- Reformulations of existing unconstrained problems via **convex splitting**: *composite convex minimization, consensus optimization, . . .*

Performance of optimization algorithms

Exact vs. approximate solutions

- Computing an **exact solution** $\mathbf{x}^\star$ to (1) is **impracticable**
- Algorithms seek $\mathbf{x}^\star_\epsilon$ that approximates $\mathbf{x}^\star$ up to ϵ in some sense

A performance metric: Time-to-reach ϵ

`time-to-reach` ϵ `= number of iterations to reach` ϵ $\times$ `per iteration time`

Performance of optimization algorithms

Exact vs. approximate solutions

- Computing an **exact solution** $\mathbf{x}^\star$ to (1) is **impracticable**
- Algorithms seek $\mathbf{x}^\star_\epsilon$ that approximates $\mathbf{x}^\star$ up to ϵ in some sense

A performance metric: Time-to-reach ϵ

`time-to-reach` ϵ `= number of iterations to reach` $\epsilon \times$ `per iteration time`

Per-iteration time:

first-order methods: Multiplication with $\mathbf{A}$, $\mathbf{A}^T$, and appropriate "prox-operators"

Performance of optimization algorithms

Exact vs. approximate solutions

- Computing an **exact solution** $\mathbf{x}^\star$ to (1) is **impracticable**
- Algorithms seek $\mathbf{x}^\star_\epsilon$ that approximates $\mathbf{x}^\star$ up to ϵ in some sense

A performance metric: Time-to-reach ϵ

`time-to-reach` ϵ `= number of iterations to reach` ϵ $\times$ `per iteration time`

Per-iteration time:

first-order methods: Multiplication with $\mathbf{A}$, $\mathbf{A}^T$, and appropriate "prox-operators"

A key issue: Number of iterations to reach ϵ

The notion of ϵ-accuracy is elusive in constrained optimization!

Numerical ϵ-accuracy

- **Unconstrained case:** All iterates are feasible (no advantage from infeasibility)!

$$f(\mathbf{x}^\star_\epsilon) - f^\star \leq \epsilon$$

$$\boxed{f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})}$$

Numerical ϵ-accuracy

- **Unconstrained case:** All iterates are feasible (no advantage from infeasibility)!

$$f(\mathbf{x}^\star_\epsilon) - f^\star \leq \epsilon$$

- **Constrained case:** We need to also measure the infeasibility of the iterates!

$$f^\star - f(\mathbf{x}^\star_\epsilon) \leq \epsilon \quad !!!$$

$$\boxed{f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K}, \ \mathbf{x} \in \mathcal{X} \right\}}$$

Numerical ϵ-accuracy

- **Unconstrained case:** All iterates are feasible (no advantage from infeasibility)!

$$f(\mathbf{x}^\star_\epsilon) - f^\star \leq \epsilon$$

- **Constrained case:** We need to also measure the infeasibility of the iterates!

$$f^\star - f(\mathbf{x}^\star_\epsilon) \leq \epsilon \quad !!!$$

$$\boxed{f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K}, \ \mathbf{x} \in \mathcal{X} \right\}}$$

Our definition of ϵ-accurate solutions [22]

Given a numerical tolerance $\epsilon \geq 0$, a point $\mathbf{x}^\star_\epsilon \in \mathbb{R}^p$ is called an ϵ-solution of (1) if

$$\begin{cases} f(\mathbf{x}^\star_\epsilon) - f^\star \leq \epsilon & \text{(objective residual)}, \\ \text{dist}\left(\mathbf{A}\mathbf{x}^\star_\epsilon - \mathbf{b}, \mathcal{K}\right) \leq \epsilon & \text{(feasibility gap)}, \\ \mathbf{x}^\star_\epsilon \in \mathcal{X} & \text{(exact feasibility for the simple set)}. \end{cases}$$

- When $\mathbf{x}^\star$ is unique, we can also obtain $\|\mathbf{x}^\star_\epsilon - \mathbf{x}^\star\| \leq \epsilon$ (iterate residual).

Numerical ϵ-accuracy

- **Unconstrained case:** All iterates are feasible (no advantage from infeasibility)!

$$f(\mathbf{x}^\star_\epsilon) - f^\star \leq \epsilon$$

- **Constrained case:** We need to also measure the infeasibility of the iterates!

$$f^\star - f(\mathbf{x}^\star_\epsilon) \leq \epsilon \quad !!!$$

$$\boxed{f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K}, \ \mathbf{x} \in \mathcal{X} \right\}}$$

Our definition of ϵ-accurate solutions [22]

Given a numerical tolerance $\epsilon \geq 0$, a point $\mathbf{x}^\star_\epsilon \in \mathbb{R}^p$ is called an ϵ-solution of (1) if

$$\begin{cases} f(\mathbf{x}^\star_\epsilon) - f^\star \leq \epsilon & \text{(objective residual)},\\ \operatorname{dist}(\mathbf{A}\mathbf{x}^\star_\epsilon - \mathbf{b}, \mathcal{K}) \leq \epsilon & \text{(feasibility gap)},\\ \mathbf{x}^\star_\epsilon \in \mathcal{X} & \text{(exact feasibility for the simple set)}. \end{cases}$$

- When $\mathbf{x}^\star$ is unique, we can also obtain $\|\mathbf{x}^\star_\epsilon - \mathbf{x}^\star\| \leq \epsilon$ (iterate residual).

- ϵ can be different for the objective, feasibility gap, or the iterate residual.

Primal-dual methods for (1):

Plenty . . .

- Variants of the **Arrow-Hurwitz's method**:
  - Chambolle-Pock's algorithm [3], and its variants, e.g., He-Yuan's variant [14].
  - Primal-dual Hybrid Gradient (PDHG) method and its variants [10, 12].
  - Proximal-based decomposition (Chen-Teboulle's algorithm) [4].
- Splitting techniques from monotone inclusions:
  - Primal-dual splitting algorithms [1, 5, 26, 6, 7].
  - Three-operator splitting [8].
- Dual splitting techniques:
  - Alternating minimization algorithms (AMA) [11, 26].
  - Alternating direction methods of multipliers (ADMM) [9, 16].
  - Accelerated variants of AMA and ADMM [7, 13].
  - Preconditioned ADMM, Linearized ADMM and inexact Uzawa algorithms [3, 19].
- **Second-order decomposition methods:**
  - Dual (quasi) Newton methods [27].
  - Smoothing decomposition methods via barriers functions [17, 23, 29].

Performance of optimization algorithms

A performance metric: Time-to-reach ϵ

`time-to-reach` ϵ `= number of iterations to reach` ϵ $\times$ `per iteration time`

Finding the fastest algorithm within the zoo is tricky!

- heuristics & tuning parameters
- non-optimal rates & strict assumptions
- lack of precise characterizations

Outline

Methods	
Primal methods	Primal-dual methods
Quadratic penalty method (QP)*	Augmented lagrangian method(ALM)*
$\rightarrow$ Inexact	$\rightarrow$ Inexact
$\rightarrow$ Linearized (and accelerated)	$\rightarrow$ Linearized (and accelerated)
	Dual subgradient method*
	Chambolle Pock's method**
	Primal-dual hybrid gradient method
	ADMM**
	AMA

* Covered in this lecture. ** Covered in the appendix of the lecture.

Outline

Penalty and linearization concepts for constrained optimization

Lagrange duality and dual based algorithms

Warm-up: Quadratic penalty approach

Constrained and penalized formulations:

- Simplified problem (1), with $\mathcal{X} = \mathbb{R}^p$:

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b} \Big\}.$$

- Penalized function with penalty parameter $\mu_k > 0$:

$$F_{\mu_k}(\mathbf{x}) := \Big\{ f(\mathbf{x}) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \Big\}.$$

Main intuition: "mimic" the constrained problem

As $\mu_k \to \infty$, $F_{\mu_k}(x)$ enforces more and more the feasibility.

A formal justification of the intuition

Theorem

Suppose $\{\mathbf{x}_k\}$ are the solutions of $\min_{\mathbf{x}} F_{\mu_k}(\mathbf{x})$ and $\mu_k \to \infty$.
Then, every limit point $\bar{\mathbf{x}}$ of the sequence $\{\mathbf{x}_k\}$ is a solution of the constrained problem.

Proof

Suppose $\mathbf{x}^\star$ is the solution of the constrained problem, then,

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}), \forall \mathbf{x} \text{ with } \mathbf{A}\mathbf{x} = \mathbf{b}. \quad (3)$$

Since $\mathbf{x}_k$ minimizes $F_{\mu_k}(\mathbf{x})$ and $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$,

$$f(\mathbf{x}_k) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\|^2 \leq f(\mathbf{x}^\star) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x}^\star - \mathbf{b}\|^2 = f(\mathbf{x}^\star). \quad (4)$$

Rearranging, we get

$$\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\|^2 \leq \frac{2}{\mu_k}\left(f(\mathbf{x}^\star) - f(\mathbf{x}^k)\right). \quad (5)$$

$\bar{x}$ satisfies $\lim_{k\in\mathcal{K}} \mathbf{x}_k = \bar{\mathbf{x}}$, for an infinite subsequence $\mathcal{K}$.
Taking the limits of (4) and (5), we obtain that $\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\| = 0$ and $f(\bar{\mathbf{x}}) \leq f(\mathbf{x}^\star)$, by using (3) and the assumption that $\mu_k \to \infty$.

Quadratic penalty method

Algorithmic idea

At iteration k:

- Solve

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) + \frac{\mu_k}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}$$

- Set $\mu_{k+1} > \mu_k$.

Quadratic penalty method (QP):
1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$.
2. For $k = 0, 1, \cdots$, perform:
2.a. $\mathbf{x}_{k+1} := \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ f(\mathbf{x}) + \frac{\mu_k}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}$.
2.b. Update $\mu_{k+1} \geq \mu_k$.

Limitations of the quadratic penalty approach

Quadratic penalty method (QP):
1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$.
2. For $k = 0, 1, \cdots,$ perform:
2.a. $\mathbf{x}_{k+1} := \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ f(\mathbf{x}) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}$.
2.b. Update $\mu_{k+1} \geq \mu_k$.

Limitations

- Solving the subproblems exactly in each iteration (ill-conditioning as $\mu_k \to \infty$):

$$\mathbf{x}_{k+1} := \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ f(\mathbf{x}) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}.$$

Common strategies:

- Solve the subproblem inexactly, *i.e.*, up to ϵ accuracy.
- Linearization to simplify subproblems.

Introducing linearization

Bottleneck

How to avoid computing at each iteration:

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}.$$

Linearization idea

- Fact: $\frac{\mu_k}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ has $\mu_k\|\mathbf{A}\|^2$ Lipschitz gradient.
- Estimate around $\mathbf{x}_k$:

$$\begin{aligned} F_{\mu_k}(\mathbf{x}) &= f(\mathbf{x}) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \\ &\le f(\mathbf{x}) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\|^2 + \mu_k\langle \mathbf{A}^\top(\mathbf{A}\mathbf{x}_k - \mathbf{b}), \mathbf{x} - \mathbf{x}_k\rangle + \frac{\mu_k\|\mathbf{A}\|^2}{2}\|\mathbf{x} - \mathbf{x}_k\|^2 \\ &=: F_{\mu_k}^{\mathbf{x}_k}(\mathbf{x}). \end{aligned}$$

- Minimize the upper bound $F_{\mu_k}^{\mathbf{x}_k}(\mathbf{x})$ instead of $F_{\mu_k}(\mathbf{x})$.

Introducing linearization

Bottleneck

How to avoid computing at each iteration:

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}) + \frac{\mu_k}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}. \tag{6}$$

Linearization idea

- We have $F_{\mu_k}(\mathbf{x}) \leq F_{\mu_k}^{\mathbf{x}_k}(\mathbf{x})$.
- At each iteration:

$$\begin{aligned} \mathbf{x}_{k+1} &= \arg\min_{\mathbf{x}} F_{\mu_k}^{\mathbf{x}_k}(\mathbf{x}) \\ &= \mathrm{prox}_{\frac{1}{\mu_k \|\mathbf{A}\|^2} f} \left(\mathbf{x}_k - \frac{1}{\|\mathbf{A}\|^2} \mathbf{A}^\top (\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\|) \right). \end{aligned}$$

- One proximal operator instead of a (potentially) difficult subproblem (6)!

Per-iteration time: The key role of the prox-operator

$$\mathbf{x}_{k+1} = \mathrm{prox}_{\frac{1}{\mu_k \|A\|^2} f}\left(\mathbf{x}_k - \frac{1}{\|A\|^2} A^\top (\|A\mathbf{x}_k - b\|)\right)$$

Recall: Prox-operator

$$\mathrm{prox}_f(\mathbf{x}) := \arg\min_{\mathbf{z}\in\mathbb{R}^p} \left\{ f(\mathbf{z}) + (1/2)\|\mathbf{z}-\mathbf{x}\|^2 \right\}.$$

Key properties:

- single valued & non-expansive since f is a proper convex function.
- distributes when the primal problem has decomposable structure:

$$f(\mathbf{x}) := \sum_{i=1}^{m} f_i(\mathbf{x}_i), \quad \text{and} \quad \mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_m.$$

  where $m \geq 1$ is the number of components.
- often efficient & has closed form expression. For instance, if $f(\mathbf{z}) = \|\mathbf{z}\|_1$, then the prox-operator performs coordinate-wise soft-thresholding by 1.

Linearized quadratic penalty methods

Linearized quadratic penalty method (LQP):

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$.
2. For $k = 0, 1, \cdots,$ perform:
   - **2.a.** $\mathbf{x}_{k+1} := \mathrm{prox}_{\frac{1}{\mu_k \|\mathbf{A}\|^2} f}\left(\mathbf{x}_k - \frac{1}{\|\mathbf{A}\|^2}\mathbf{A}^\top(\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\|)\right).$
   - **2.b.** Update σ_k such that $\frac{(1-\sigma_k)^2}{\sigma_k} = \frac{1}{\sigma_{k-1}}$.
   - **2.c.** Update $\mu_{k+1} = \sqrt{\sigma_k}$.

Accelerated linearized quadratic penalty method (ALQP):

1. Choose $\mathbf{x}^0, \mathbf{y}^0 \in \mathbb{R}^p$.
2. For $k = 0, 1, \cdots,$ perform:
   - **2.a.** $\mathbf{x}_{k+1} := \mathrm{prox}_{\frac{1}{\mu_k \|\mathbf{A}\|^2} f}\left(\mathbf{x}_k - \frac{1}{\|\mathbf{A}\|^2}\mathbf{A}^\top(\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\|)\right).$
   - **2.b.** $\mathbf{y}_{k+1} := \mathbf{x}_{k+1} + \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}(\mathbf{x}_{k+1} - \mathbf{x}_k).$
   - **2.c.** Update $\mu_{k+1} = \mu_k(1 + \tau_{k+1})$.
   - **2.d.** Update $\tau_{k+1} \in (0,1)$ the unique positive root of $\tau^3 + \tau^2 + \tau_k^2\tau - \tau_k^2 = 0$.

Convergence of LQP and FLQP

Theorem (Convergence [25])

• *Let us denote as* $\{\lambda^\star\}$ *the optimal Lagrange multiplier (more on this later this lecture!):*

• *LQP:*
$$\begin{cases} f(\mathbf{x}_k) - f(x^\star) \leq \frac{\|\mathbf{A}\|^2}{2\beta_0\sqrt{k}}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \|\lambda^\star\|^2\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\| + \frac{1}{\sqrt{k}}\|\lambda^\star\|^2 \\ f(\mathbf{x}) - f(x^\star) \geq -\|\lambda^\star\|\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\| \\ \|\mathbf{A}\mathbf{x}_k - \mathbf{b}\| \leq \frac{1}{\sqrt{k+1}}\left[\|\lambda^\star\| + \left(\|\lambda^\star\|^2 + \frac{1}{\beta_0^2}\|\mathbf{A}\|^2\|\mathbf{x}_0 - \mathbf{x}^\star\|^2\right)^{1/2}\right] \end{cases}$$

• *ALQP:*
$$\begin{cases} f(\mathbf{x}_k) - f(x^\star) \leq \frac{\|\mathbf{A}\|^2}{2\beta_0 k}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \|\lambda^\star\|^2\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\| + \frac{2\beta_0}{k}\|\lambda^\star\|^2 \\ f(\mathbf{x}) - f(x^\star) \geq -\|\lambda^\star\|\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\| \\ \|\mathbf{A}\mathbf{x}_k - \mathbf{b}\| \leq \frac{\beta_0}{k+1}\left[\|\lambda^\star\| + \left(\|\lambda^\star\|^2 + \frac{1}{\beta_0^2}\|\mathbf{A}\|^2\|\mathbf{x}_0 - \mathbf{x}^\star\|^2\right)^{1/2}\right] \end{cases}$$

• These methods almost never work better than the worst case.

• Duality concept is needed for the convergence rate analysis.

Outline

Penalty and linearization concepts for constrained optimization

Lagrange duality and dual based algorithms

Lagrange duality and the optimal solution set

Lagrangian function

$$\mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b}).$$

Here, $\lambda \in \mathbb{R}^n$ is the vector of Lagrange multipliers (or dual variables) w.r.t. $\mathbf{A}\mathbf{x} = \mathbf{b}$.

- **Primal problem:**

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b} \Big\},$$

- **Dual function:**

$$d(\lambda) := \min_{\mathbf{x}\in\mathcal{X}} \Big\{ \mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \Big\}. \quad (7)$$

→ Let $\mathbf{x}^\star(\lambda)$ be a solution of (7) then $d(\lambda)$ is finite if $x^\star(\lambda)$ exists.

- **Dual problem**: The following dual problem is concave

$$\boxed{d^\star := \max_{\lambda\in\mathbb{R}^n} d(\lambda)}$$

Min-max formulation and dual problem

- **Primal problem:**

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b} \Big\},$$

- **Dual function:**

$$d(\lambda) := \min_{\mathbf{x}\in\mathcal{X}} \{\mathcal{L}(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b})\}. \tag{8}$$

→ Let $\mathbf{x}^\star(\lambda)$ be a solution of (8) then $d(\lambda)$ is finite if $x^\star(\lambda)$ exists.

- **Dual problem**: The following dual problem is concave

$$\boxed{d^\star := \max_{\lambda\in\mathbb{R}^n} d(\lambda)}$$

Min-max formulation

$$\begin{aligned} d^\star &= \max_{\lambda\in\mathbb{R}^n} d(\lambda) = \max_{\lambda\in\mathbb{R}^n} \min_{\mathbf{x}\in\mathcal{X}} \{ f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x}-\mathbf{b})\} \\ &\leq \min_{\mathbf{x}\in\mathcal{X}} \max_{\lambda\in\mathbb{R}^n} \{ f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x}-\mathbf{b})\} = \begin{cases} \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) & \text{if } \mathbf{A}\mathbf{x}=\mathbf{b}, \\ +\infty & \text{otherwise} \end{cases} \end{aligned}$$

Here, the inequality is due to **the max-min theorem** [21].

Saddle point

A point $(\mathbf{x}^\star, \lambda^\star) \in \mathcal{X} \times \mathbb{R}^n$ is called a saddle point of the Lagrange function $\mathcal{L}$ if

$$\mathcal{L}(\mathbf{x}^\star, \lambda) \leq \mathcal{L}(\mathbf{x}^\star, \lambda^\star) \leq \mathcal{L}(\mathbf{x}, \lambda^\star), \ \forall \mathbf{x} \in \mathcal{X}, \ \lambda \in \mathbb{R}^n.$$

Recall the minimax form:

$$\max_{\lambda} \min_{\mathbf{x} \in \mathcal{X}} \left\{ \mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \lambda^T (\mathbf{A}\mathbf{x} - \mathbf{b}) \right\}.$$

Saddle point

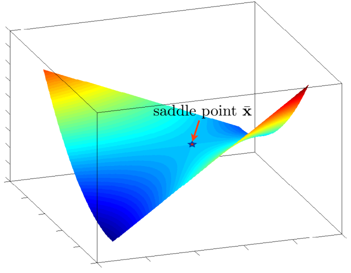
A point $(\mathbf{x}^\star, \lambda^\star) \in \mathcal{X} \times \mathbb{R}^n$ is called a saddle point of the Lagrange function $\mathcal{L}$ if

$$\mathcal{L}(\mathbf{x}^\star, \lambda) \le \mathcal{L}(\mathbf{x}^\star, \lambda^\star) \le \mathcal{L}(\mathbf{x}, \lambda^\star),\ \forall \mathbf{x} \in \mathcal{X},\ \lambda \in \mathbb{R}^n.$$

Recall the minimax form:

$$\max_{\lambda} \min_{\mathbf{x}\in\mathcal{X}} \left\{ \mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \right\}.$$

Illustration of saddle point: $\mathcal{L}(x, \lambda) := (1/2)x^2 + \lambda(x - 1)$ in $\mathbb{R}^2$



Necessary and sufficient condition

Recall the minimax form:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\lambda\in\mathbb{R}^n} \left\{\mathcal{L}(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x}-\mathbf{b})\right\}$$

Theorem (Necessary and sufficient optimality condition)

Under Slater's condition: $\mathrm{relint}(\mathcal{X}) \cap \{\mathbf{x} \,:\, \mathbf{A}\mathbf{x}=\mathbf{b}\} \neq \emptyset$*, the* ***KKT condition***

$$\begin{cases} 0 & \in \partial_{\mathbf{x}}\mathcal{L}(\mathbf{x}^\star,\lambda^\star) \; = \mathbf{A}^T\lambda^\star + \partial f(\mathbf{x}^\star), \\ 0 & = \nabla_\lambda \mathcal{L}(\mathbf{x}^\star,\lambda^\star) = \mathbf{A}\mathbf{x}^\star - \mathbf{b}. \end{cases}$$

is necessary and sufficient for a point $(\mathbf{x}^\star,\lambda^\star)\in\mathcal{X}\times\mathbb{R}^n$ *being an optimal solution for the primal problem and dual problem:*

$$f^\star := \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^p} & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{A}\mathbf{x}=\mathbf{b}, \end{cases} \qquad \textit{and} \quad d^\star := \max_{\mathbf{x}\in\mathbb{R}^n} d(\lambda).$$

- By definition of $f^\star$ and $d^\star$, we always have $\boxed{d^\star \leq f^\star}$ (**weak duality**).
- Under Slater's condition and $\mathcal{X}^\star \neq \emptyset$, we have $\boxed{d^\star = f^\star}$ (**strong duality**).
- Any solution $(\mathbf{x}^\star,\lambda^\star)$ of the KKT condition is also a saddle point.

*Slater's qualification condition

Recall $\mathrm{relint}(\mathcal{X})$ the relative interior of the **feasible set** $\mathcal{X}$. The Slater condition requires

$$\boxed{\mathrm{relint}(\mathcal{X}) \cap \{\mathbf{x} \ : \ \mathbf{A}\mathbf{x} = \mathbf{b}\} \neq \emptyset.} \tag{9}$$

Special cases

- If $\mathcal{X}$ is absent, then (9) $\Leftrightarrow$ $\boxed{\exists \bar{\mathbf{x}} \ : \ \mathbf{A}\bar{\mathbf{x}} = \mathbf{b}}$.
- If $\mathbf{A}\mathbf{x} = \mathbf{b}$ is absent, then (9) $\Leftrightarrow$ $\boxed{\mathrm{relint}(\mathcal{X}) \neq \emptyset}$.
- If $\mathbf{A}\mathbf{x} = \mathbf{b}$ is absent and $\mathcal{X} := \{\mathbf{x} : h(\mathbf{x}) \leq 0\}$, where h is $\mathbb{R}^p \to R^q$ is convex, then

$$(9) \Leftrightarrow \boxed{\exists \bar{\mathbf{x}} \ : \ h(\bar{\mathbf{x}}) < 0.}$$

*Example: Slater's condition

Example

Let us consider the feasible set $\mathcal{D}_\alpha := \mathcal{X} \cap \mathcal{A}_\alpha$ as

$$\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^2 \ : \ x_1^2 + x_2^2 \leq 1\} \ \mathcal{A}_\alpha := \{\mathbf{x} \in \mathbb{R}^2 \ : \ x_1 + x_2 = \alpha\},$$

where $\alpha \in \mathbb{R}$.

⋆Example: Slater's condition

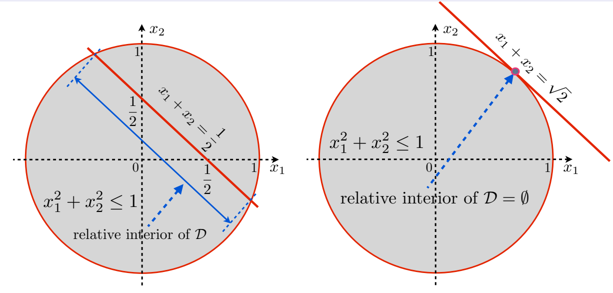
Example

Let us consider the feasible set $\mathcal{D}_\alpha := \mathcal{X} \cap \mathcal{A}_\alpha$ as

$$\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^2 \ : \ x_1^2 + x_2^2 \leq 1\} \ \mathcal{A}_\alpha := \{\mathbf{x} \in \mathbb{R}^2 \ : \ x_1 + x_2 = \alpha\},$$

where $\alpha \in \mathbb{R}$.

Slater's condition holds and does not hold



$\mathcal{D}_{1/2}$ satisfies Slater's condition – $\mathcal{D}_{\sqrt{2}}$-does not satisfy Slater's condition

Dual subgradient method

Recall the dual problem:

$$d^\star := \max_{\lambda \in \mathbb{R}^n} d(\lambda) \tag{10}$$

Subgradient ascent method can be applied to solve it.

Dual subgradient method

Recall the dual problem:

$$d^\star := \max_{\lambda \in \mathbb{R}^n} d(\lambda) \tag{10}$$

Subgradient ascent method can be applied to solve it.

A plausible algorithmic strategy for $\min_{\mathbf{x}\in\mathcal{X}}\{f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}\}$:

A natural minimax formulation:

$$(\mathbf{x}^\star, \lambda^\star) \in \arg\max_{\lambda}\min_{\mathbf{x}\in\mathcal{X}}\{\mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle\}.$$

Lagrangian subproblem: $\mathbf{x}^\star(\lambda) \in \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}(\mathbf{x}, \lambda)$
Dual problem: $\lambda^\star \in \arg\max_{\lambda}\{d(\lambda) := \mathcal{L}(\mathbf{x}^\star(\lambda), \lambda)\}$

- λ is called the Lagrange multiplier.
- The function $d(\lambda)$ is called the dual function, and it is concave!
- The optimal dual objective value is $d^\star = d(\lambda^\star)$.

A basic strategy $\Rightarrow$ Find $\lambda^\star$ and then solve for $\mathbf{x}^\star = \mathbf{x}^\star(\lambda^\star)$

Dual subgradient method

Properties of dual function

- d is **concave**, but **not necessarily differentiable**.
- **Subgradient:** $\mathbf{A}\mathbf{x}^\star(\lambda) - \mathbf{b} \in \partial d(\lambda)$, where $\mathbf{x}^\star(\lambda)$ is such that

$$\mathbf{x}^\star(\lambda) := \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\mathcal{L}(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b})\right\}.$$

Dual subgradient method (DSGM):

1. Choose $\lambda^0 \in \mathbb{R}^p$.
2. For $k = 0, 1, \cdots$, perform:
   - **2.a.** $\mathbf{x}^\star(\lambda_k) := \arg\min_{\mathbf{x}\in\mathcal{X}}\{\mathcal{L}(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b})\}$.
   - **2.b.** Compute the subgradient $\nabla d(\lambda^k) := \mathbf{A}\mathbf{x}^\star(\lambda^k) - \mathbf{b}$.
   - **2.c.** Update $\boxed{\lambda^{k+1} := \lambda^k + \frac{R}{\sqrt{k+1}}\nabla d(\lambda^k)}$, where R is a given constant.

Example: Nonsmoothness of the dual function

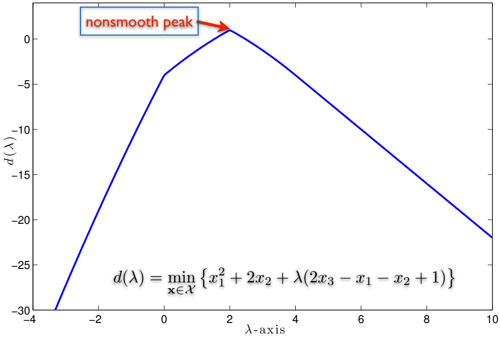
Consider a constrained convex problem:

$$\begin{aligned}\min_{\mathbf{x}\in\mathbb{R}^3} \quad & \left\{ f(\mathbf{x}) := x_1^2 + 2x_2 \right\}, \\ \text{s.t.} \quad & 2x_3 - x_1 - x_2 = 1, \\ & \mathbf{x} \in \mathcal{X} := [-2,2]\times[-2,2]\times[0,2].\end{aligned}$$

The **dual function** is defined as

$$d(\lambda) := \min_{\mathbf{x}\in\mathcal{X}} \left\{ x_1^2 + 2x_2 + \lambda(2x_3 - x_1 - x_2 - 1) \right\}$$

is concave and nonsmooth as illustrated in the figure below.



Convergence of DSGM

Well-definedness

- Problem below may not have solution $\mathbf{x}^\star(\lambda)$ for any λ. Then DSGM is not well-defined except if $\mathcal{X}$ is bounded.

$$\mathbf{x}^\star(\lambda) := \arg\min_{\mathbf{x}\in\mathcal{X}} \{\mathcal{L}(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b})\}.$$

- Impractical to evaluate $R_\star := \|\lambda^0 - \lambda^\star\|_2$, use an upper bound R of $R_\star$.

Convergence of DSGM

Well-definedness

- Problem below may not have solution $\mathbf{x}^\star(\lambda)$ for any λ. Then DSGM is not well-defined except if $\mathcal{X}$ is bounded.

$$\mathbf{x}^\star(\lambda) := \arg\min_{\mathbf{x}\in\mathcal{X}} \{\mathcal{L}(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b})\}.$$

- Impractical to evaluate $R_\star := \|\lambda^0 - \lambda^\star\|_2$, use an upper bound R of $R_\star$.

Theorem (Convergence)

Assume that $\|\mathbf{A}\mathbf{x}^\star(\lambda^k) - \mathbf{b}\| \leq M_d$ *for all* $k \geq 0$. *Then* $\{\lambda^k\}$ *generated by DSGM satisfies*

$$\boxed{d^\star - d(\lambda^k) \leq \frac{M_d R_\star}{\sqrt{k+1}}, \forall k \geq 0,}$$

where $R_\star := \min_{\lambda^\star} \|\lambda^0 - \lambda^\star\|_2$. *Convergence rate of DSGM is* $\mathcal{O}(1/\sqrt{k})$.

Convergence of DSGM

Well-definedness

- Problem below may not have solution $\mathbf{x}^\star(\lambda)$ for any λ. Then DSGM is not well-defined except if $\mathcal{X}$ is bounded.

$$\mathbf{x}^\star(\lambda) := \arg\min_{\mathbf{x}\in\mathcal{X}} \{\mathcal{L}(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x}-\mathbf{b})\}.$$

- Impractical to evaluate $R_\star := \|\lambda^0 - \lambda^\star\|_2$, use an upper bound R of $R_\star$.

Theorem (Convergence)

Assume that $\|\mathbf{A}\mathbf{x}^\star(\lambda^k) - \mathbf{b}\| \leq M_d$ *for all* $k \geq 0$. *Then* $\{\lambda^k\}$ *generated by DSGM satisfies*

$$\boxed{d^\star - d(\lambda^k) \leq \frac{M_d R_\star}{\sqrt{k+1}}, \forall k \geq 0,}$$

where $R_\star := \min_{\lambda^\star} \|\lambda^0 - \lambda^\star\|_2$. *Convergence rate of DSGM is* $\mathcal{O}(1/\sqrt{k})$.

Special cases

1. If f is strongly convex, then d is smooth and its gradient is Lipschitz continuous, $d \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$. **Gradient** and **fast gradient methods** can be used to solve the dual problem.

Efficiency considerations for the dual problem

Subgradient method
1. Choose $\lambda^0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \cdots$, perform: $$\lambda^{k+1} = \lambda^k + \alpha_k \mathbf{v}^k,$$
where $\mathbf{v}^k \in \partial d(\lambda^k)$ and α_k is the step-size.

Subgradient method for the dual

Assume that the following conditions

1. $\|\mathbf{v}\|_2 \leq G$ for all $\mathbf{v} \in \partial d(\lambda)$, $\lambda \in \mathbb{R}^n$.
2. $\|\lambda^0 - \lambda^\star\|_2 \leq R$

Let the step-size be chosen as $\alpha_k = \frac{R}{G\sqrt{k}}$. Then, the subgradient method satisfies

$$\min_{0 \leq i \leq k} d^\star - d(\lambda^i) \leq \frac{RG}{\sqrt{k}}$$

Efficiency considerations for the dual problem

Subgradient method
1. Choose $\lambda^0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \cdots$, perform: $\lambda^{k+1} = \lambda^k + \alpha_k \mathbf{v}^k,$
where $\mathbf{v}^k \in \partial d(\lambda^k)$ and α_k is the step-size.

Subgradient method for the dual

Assume that the following conditions

1. $\|\mathbf{v}\|_2 \leq G$ for all $\mathbf{v} \in \partial d(\lambda)$, $\lambda \in \mathbb{R}^n$.
2. $\|\lambda^0 - \lambda^\star\|_2 \leq R$

Let the step-size be chosen as $\alpha_k = \frac{R}{G\sqrt{k}}$. Then, the subgradient method satisfies

$$\min_{0 \leq i \leq k} d^\star - d(\lambda^i) \leq \frac{RG}{\sqrt{k}} \leq \bar{\epsilon}$$

Efficiency considerations for the dual problem

Gradient method
1. Choose $\lambda^0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \cdots$, perform: $\lambda^{k+1} = \lambda^k + \frac{1}{L}\nabla d(\lambda^k)$,
where L is the Lipschitz constant.

SGM: $\mathcal{O}\left(\frac{1}{\bar{\epsilon}^2}\right) \times$ subgradient calculation

GM: $\mathcal{O}\left(\frac{1}{\bar{\epsilon}}\right) \times$ gradient calculation

Impact of smoothness

(Lipschitz gradient) $d(\lambda)$ has Lipschitz continuous gradient iff

$$\|\nabla d(\lambda) - \nabla d(\eta)\|_2 \leq L\|\lambda - \eta\|_2$$

for all $\lambda, \eta \in \mathrm{dom}(d)$ and we indicate this structure as $d(\lambda) \in \mathcal{F}_L$.

For all $d(\lambda) \in \mathcal{F}_L$, the gradient method with step-size $1/L$ obeys

$$d^\star - d(\lambda^k) \leq \frac{2LR^2}{k+4} \leq \bar{\epsilon}.$$

Efficiency considerations for the dual problem

Gradient method

1. Choose $\lambda^0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \cdots$, perform:
$$\lambda^{k+1} = \lambda^k + \frac{1}{L}\nabla d(\lambda^k),$$
where L is the Lipschitz constant.

SGM: $\mathcal{O}\left(\frac{1}{\bar{\epsilon}^2}\right) \times$ subgradient calculation

GM: $\mathcal{O}\left(\frac{1}{\bar{\epsilon}}\right) \times$ gradient calculation

Impact of smoothness

(Lipschitz gradient) $d(\lambda)$ has Lipschitz continuous gradient iff

$$\|\nabla d(\lambda) - \nabla d(\eta)\|_2 \leq L\|\lambda - \eta\|_2$$

for all $\lambda, \eta \in \mathrm{dom}(d)$ and we indicate this structure as $d(\lambda) \in \mathcal{F}_L$.

For all $d(\lambda) \in \mathcal{F}_L$, the gradient method with step-size $1/L$ obeys

$$d^\star - d(\lambda^k) \leq \frac{2LR^2}{k+4} \leq \bar{\epsilon}.$$

This is NOT the best we can do.

There exists a complexity lower-bound

$$d^\star - d(\lambda^k) \geq \frac{3LR^2}{32(k+1)^2}, \forall d(\lambda) \in \mathcal{F}_L,$$

for any iterative method based only on function and gradient evaluations.

Efficiency considerations for the dual problem

Accelerated gradient method

1. Choose $\mathbf{u}^0 = \lambda^0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \cdots$, perform:

$$\lambda^k = \mathbf{u}^k + \frac{1}{L}\nabla d(\mathbf{u}^k),$$
$$\mathbf{u}^{k+1} = \lambda^k + \rho_k(\lambda^k - \lambda^{k-1}),$$

where L is the Lipschitz constant, and ρ_k is a momentum parameter.

SGM: $\mathcal{O}\left(\frac{1}{\epsilon^2}\right) \times$ subgradient calculation
GM: $\mathcal{O}\left(\frac{1}{\epsilon}\right) \times$ gradient calculation
AGM: $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right) \times$ gradient calculation

Impact of smoothness

(Lipschitz gradient) $d(\lambda)$ has Lipschitz continuous gradient iff

$$\|\nabla d(\lambda) - \nabla d(\eta)\|_2 \leq L\|\lambda - \eta\|_2$$

for all $\lambda, \eta \in \mathrm{dom}(d)$ and we indicate this structure as $d(\lambda) \in \mathcal{F}_L$.

For all $d(\lambda) \in \mathcal{F}_L$, the accelerated gradient method with momentum $\rho_k = \frac{k+1}{k+3}$ obeys

$$d^\star - d(\lambda^k) \leq \frac{2LR^2}{(k+2)^2} \leq \bar{\epsilon}$$

This is NEARLY the best we can do.

There exists a complexity lower-bound

$$d^\star - d(\lambda^k) \geq \frac{3LR^2}{32(k+1)^2}, \forall d(\lambda) \in \mathcal{F}_L,$$

for any iterative method based only on function and gradient evaluations.

When is the dual function smooth?

Smoothness of dual function

- When $f(\mathbf{x})$ is γ-strongly convex, the dual function $d(\lambda)$ is $\frac{\|\mathbf{A}\|^2}{\gamma}$-Lipschitz gradient.

(Strong convexity) $f(\mathbf{x})$ is γ-strongly convex iff $f(\mathbf{x}) - \frac{\gamma}{2}\|\mathbf{x}\|_2^2$ is convex.

- However, in general, dual problem is convex but nonsmooth.

Augmented Lagrangian

Augmented Lagrangian: $\mathcal{L}_\mu(\mathbf{x}, \lambda) := \mathcal{L}(\mathbf{x}, \lambda) + (\mu/2)\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, where $\mu > 0$ is a penalty parameter.

Augmented dual function:

$$d_\mu(\lambda) := \min_{\mathbf{x}\in\mathcal{X}} \left\{ \mathcal{L}_\mu(\mathbf{x}, \lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b}) + (\mu/2)\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \right\}.$$

- d_μ is smooth and Lipschitz gradient

Different perspectives

- We will motivate Augmented Lagrangian Method (ALM) from dual perspective.
- ALM can also be motivated by penalty approach, see [2, 18].

Augmented Lagrangian method

Augmented dual function:

$$d_\mu(\lambda) := \min_{\mathbf{x}\in\mathcal{X}} \left\{ \mathcal{L}_\mu(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x}-\mathbf{b}) + (\mu/2)\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 \right\}. \qquad (11)$$

Key properties of d_μ

- d_μ is concave and smooth and

$$\nabla d_\mu(\lambda) = \mathbf{A}\mathbf{x}^\star_\mu(\lambda) - \mathbf{b},$$

where $\mathbf{x}^\star_\mu(\lambda)$ is the solution of (11).

- ∇d_μ is Lipschitz continuous with a Lipschitz constant $L_d := \mu^{-1}$, i.e.:

$$\|\nabla d_\mu(\lambda) - \nabla d_\mu(\hat{\lambda})\| \leq \mu^{-1}\|\lambda - \hat{\lambda}\|, \ \forall \lambda, \hat{\lambda} \in \mathbb{R}^n.$$

Example: Behavior of the augmented Lagrangian dual function

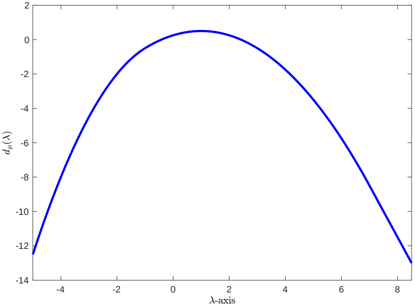
Consider a constrained convex problem:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^3} \quad & \left\{ f(\mathbf{x}) := x_1^2 + x_2^2 \right\}, \\ \text{s.t.} \quad & 2x_3 - x_1 - x_2 = 1, \\ & \mathbf{x} \in \mathcal{X} := [-2,2]\times[-2,2]\times[0,2]. \end{aligned}$$

The **augmented Lagrangian dual function** is defined as

$$d_\mu(\lambda) := \min_{\mathbf{x}\in\mathcal{X}} \left\{ x_1^2 + x_2^2 + \lambda(2x_3 - x_1 - x_2 - 1) + (\mu/2)\|2x_3 - x_1 - x_2 - 1\|_2^2 \right\}$$

is concave and smooth as illustrated in the figure below.



Augmented dual problem

Dual problem:

$$d^\star := \max_{\lambda \in \mathbb{R}^n} \left\{ d(\lambda) = \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \right\}. \tag{12}$$

Augmented dual problem:

$$d_\mu^\star := \max_{\lambda \in \mathbb{R}^n} \left\{ d_\mu(\lambda) = \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \frac{\mu}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2, \quad \mu > 0 \right\}. \tag{13}$$

Augmented dual problem

Dual problem:

$$d^\star := \max_{\lambda \in \mathbb{R}^n} \left\{ d(\lambda) = \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \right\}. \tag{12}$$

Augmented dual problem:

$$d^\star_\mu := \max_{\lambda \in \mathbb{R}^n} \left\{ d_\mu(\lambda) = \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \frac{\mu}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2, \quad \mu > 0 \right\}. \tag{13}$$

Relation between augmented dual problem and dual problem

Under Slater's condition and $\mathcal{X}^\star \neq \emptyset$, we have

- The dual solution set of (13) coincides with the one of the dual problem (12).
- $f^\star = d^\star = d^\star_\mu$ for any $\mu > 0$.

The augmented dual problem (13) is smooth and convex ⇒ **Gradient and Fast gradient methods** can be applied to solve it.

Augmented Lagrangian method

Augmented Lagrangian method (ALM):
1. Choose $\lambda^0 \in \mathbb{R}^p$ and $\mu > 0$.
2. For $k = 0, 1, \cdots,$ perform:
2.a. Solve (11) to compute $\nabla d_\mu(\lambda^k) := \mathbf{A}\mathbf{x}^\star_\mu(\lambda^k) - \mathbf{b}$.
2.b. Update $\boxed{\lambda^{k+1} := \lambda^k + \mu\nabla d_\mu(\lambda^k).}$

Augmented Lagrangian method

Augmented Lagrangian method (ALM):
1. Choose $\lambda^0 \in \mathbb{R}^p$ and $\mu > 0$.
2. For $k = 0, 1, \cdots,$ perform:
2.a. Solve (11) to compute $\nabla d_\mu(\lambda^k) := \mathbf{A}\mathbf{x}^\star_\mu(\lambda^k) - \mathbf{b}$.
2.b. Update $\boxed{\lambda^{k+1} := \lambda^k + \mu \nabla d_\mu(\lambda^k).}$

ALM can be accelerated by Nesterov's optimal method.

Fast augmented Lagrangian method (FALM)

1. Choose $\lambda^0 \in \mathbb{R}^p$ and $\mu > 0$. Set $\tilde{\lambda}^0 := \lambda^0$ and $t_0 := 1$
2. For $k = 0, 1, \cdots,$ perform:
   2.a. Solve (11) to compute $\nabla d_\mu(\tilde{\lambda}^k) := \mathbf{A}\mathbf{x}^\star_\mu(\tilde{\lambda}^k) - \mathbf{b}$.
   2.b. Update

$$\begin{cases} \lambda^{k+1} & := \tilde{\lambda}^k + \mu \nabla d_\mu(\tilde{\lambda}^k), \\ \tilde{\lambda}^{k+1} & := \lambda^{k+1} + ((t_k - 1)/t_{k+1})(\lambda^{k+1} - \lambda^k), \\ t_{k+1} & := (1 + \sqrt{1 + 4t_k^2})/2. \end{cases}$$

Convergence of ALM and FALM

Theorem (Convergence [15])

- *Let $\{\lambda^k\}$ be the sequence generated by ALM. Then*

$$\boxed{d^\star - d_\mu(\lambda^k) \leq \frac{\|\lambda^0 - \lambda^\star\|_2^2}{2\mu(k+1)},\ k \geq 0.}$$

- *Let $\{\lambda^k\}$ be the sequence generated by FALM. Then*

$$\boxed{d^\star - d_\mu(\lambda^k) \leq \frac{2\|\lambda^0 - \lambda^\star\|_2^2}{\mu(k+2)^2},\ k \geq 0.}$$

- Important observation: The right-hand side of both estimates depends on μ. When μ gets large, the right-hand side decreases.
- Guarantees are given for the dual problem and not for the primal!
- We can show guarantees for the primal iterate and averaged primal iterate, see [22].

Drawbacks and enhancements

At each step, ALM solves

$$\mathbf{x}^\star_\mu(\lambda) := \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \mathcal{L}_\mu(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x}-\mathbf{b}) + (\mu/2)\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 \right\}. \quad (14)$$

Drawbacks

1. Drawback 1: The quadratic term $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$ in (14) destroys the separability as well as the tractable proximity of f.
2. Drawback 2: Solving (14) exactly is impractical.
3. Drawback 3: No theoretical guarantee for choosing appropriate values of κ.

Drawbacks and enhancements

At each step, ALM solves

$$\mathbf{x}^\star_\mu(\lambda) := \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \mathcal{L}_\mu(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x}-\mathbf{b}) + (\mu/2)\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 \right\}. \quad (14)$$

Drawbacks

1. Drawback 1: The quadratic term $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$ in (14) destroys the separability as well as the tractable proximity of f.
2. Drawback 2: Solving (14) exactly is impractical.
3. Drawback 3: No theoretical guarantee for choosing appropriate values of κ.

Enhancements

1. Allow inexactness of solving (14), while guaranteeing the same convergence rate.
2. Update the penalty parameter κ
   - Increasing ρ: Lead to the increase of ill-condition in (14).
   - Adaptively update κ: Often heuristic
3. Process the quadratic term $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$ by linearization, alternating, etc.

Going back to primal: Linearized Augmented Lagrangian method

Bottleneck

$$\mathbf{x}_{k+1} := \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \mathcal{L}_\mu(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) + (\mu/2)\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 \right\}.$$

- Recall: Linearization idea

$$\begin{aligned}
\mathbf{x}_{k+1} &:= \arg\min_{\mathbf{x}\in\mathcal{X}} L_\mu^{\mathbf{x}_k}(\mathbf{x},\lambda_k) \\
&:= \mathrm{prox}_{\frac{1}{\mu\|\mathbf{A}\|^2} f}\left(\mathbf{x}_k - \frac{1}{\|\mathbf{A}\|^2}\mathbf{A}^\top\left(\frac{1}{\mu}\lambda_k + (\mathbf{A}\mathbf{x}_k - \mathbf{b})\right)\right)
\end{aligned}$$

Linearized augmented Lagrangian method (LALM)

1. Choose $\mathbf{x}_n \in \mathbb{R}^n$, $\lambda^0 \in \mathbb{R}^p$ and $\kappa, \mu > 0$.
2. For $k = 0, 1, \cdots$, perform:
   2.a. Update

$$\begin{cases}
\mathbf{x}_{k+1} := \mathrm{prox}_{\frac{1}{\mu\|\mathbf{A}\|^2} f}\left(\mathbf{x}_k - \frac{1}{\|\mathbf{A}\|^2}\mathbf{A}^\top\left(\frac{1}{\mu}\lambda_k + (\mathbf{A}\mathbf{x}_k - \mathbf{b})\right)\right), \\
\lambda_{k+1} := \tilde{\lambda}_k + \mu(\mathbf{A}\mathbf{x}_k - \mathbf{b}).
\end{cases}$$

Convergence of Linearized ALM

Theorem (Convergence [28])

Let $\mu > 0$ *and define* $\bar{\mathbf{x}}_{k+1} = \frac{1}{k}\sum_{i=1}^{k} \mathbf{x}_{i+1}$. *Then, the iterates of LALM satisfy:*

$$\left\| A\bar{\mathbf{x}}^k - \mathbf{b} \right\| \leq \frac{1}{k\|} \left(\frac{1}{2}\|x_1 - x^\star\|^2 + \frac{\max\left\{(1+\|\lambda^\star\|)^2, 4\|\lambda^\star\|^2\right\}}{\mu} \right)$$

$$|f(\bar{\mathbf{x}}^k) - f(\mathbf{x}^\star)| \leq \frac{1}{k} \left(\frac{1}{2}\|x_1 - x^\star\|^2 + \frac{\max\left\{(1+\|\lambda^\star\|)^2, 4\|\lambda^\star\|^2\right\}}{\mu} \right)$$

- Guarantees are for the primal.
- No need to solve difficult subproblems at each iteration.
- Guarantees are of the same order as ALM, but slower than FALM at the expense of easy subproblems.
- Guarantees are for $\bar{\mathbf{x}}_k$, and not $\mathbf{x}_k$.

Example: Last iterate vs average iterate of LALM

Problem: Basis pursuit

Given $\mathbf{A} \in \mathbb{R}^{n \times p}$ and $\mathbf{b} \in \mathbb{R}^n$, solve

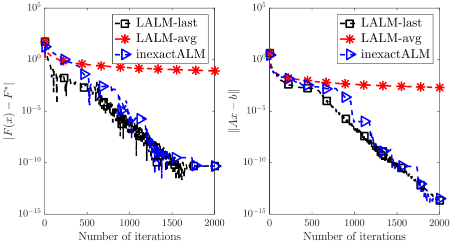
$$F^\star := \min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \mathbf{A}\mathbf{x} = \mathbf{b} \right\}.$$

Data generation

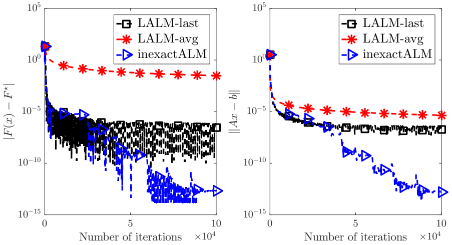
- $\mathbf{A}$ is a row-normalized standard Gaussian matrix.
- $x^\star$ is a k-sparse vector generated randomly.
- Noiseless case: $\mathbf{b} := \mathbf{A}\mathbf{x}^\star$.
- Noisy case: $\mathbf{b} := \mathbf{A}\mathbf{x}^\star + \mathcal{N}(0, 10^{-3})$.

Example: Last iterate vs average iterate of LALM

- Noiseless case.



- Noisy case.



*A composite reformulation

- Focus the following template in the sequel:

$$\boxed{\min_{\mathbf{x}} \{f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathcal{X}\}}$$

- Fundamentally the same as the composite form: $\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})$

Lasso	$\mathcal{X} = \mathbb{R}^p$	$f(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$	$g(\mathbf{z}) = \frac{1}{n}\|\mathbf{z}-\mathbf{b}\|_2^2$
Square-root Lasso	$\mathcal{X} = \mathbb{R}^p$	$f(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$	$g(\mathbf{z}) = \frac{1}{\sqrt{n}}\|\mathbf{z}-\mathbf{b}\|_2$
SDP	$\mathcal{X} = \{\mathbf{x} \succeq 0, \mathbf{x}' = \mathbf{x}\}$	$f(\mathbf{x}) = \text{tr}(\mathbf{b}\mathbf{x})$	$g(\mathbf{z}) = \begin{cases} 0 & \text{if } \mathbf{z} = \mathbf{b} \\ +\infty & \text{otherwise} \end{cases}$

*Lasso is essentially "easy"

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})$$

- Revelation: Lasso can be solved as if the problem is fully smooth!
  - **not with subgradient descent!**

- Structures in the composite form
  - g has Lipschitz gradient in ℓ_2-norm (i.e., $\|\nabla g(\mathbf{u}) - \nabla g(\mathbf{v})\|_2 \leq L\|\mathbf{u}-\mathbf{v}\|_2$)

    Lasso: $g(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2 \Rightarrow L = 1.$

  - $f : \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$ has a "tractable" proximal operator

$$\mathrm{prox}_f(\mathbf{x}) := \arg\min_{\mathbf{u}\in\mathcal{X}} f(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2$$

Lasso: $f(\mathbf{x}) = \|\mathbf{x}\|_1, \mathcal{X} = \mathbb{R}^p \Rightarrow \mathrm{prox}_f$ is soft thresholding.

$$[\mathrm{prox}_f(\mathbf{x})]_i = \begin{cases} 0, & \text{if } |\mathbf{x}_i| \leq \lambda \\ \mathbf{x}_i - \lambda \mathrm{sign}(\mathbf{x}_i), & \text{if } |\mathbf{x}_i| > \lambda \end{cases}$$

*Famous Algorithms I

$$\boxed{\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})}$$

- FISTA (aka. accelerated proximal gradient method, aka. Nesterov acceleration):

  At iteration k:

$$\mathbf{x}^{k+1} = \mathrm{prox}_{f/L\|\mathbf{A}\|^2}\left(\mathbf{y}^k - \frac{1}{L\|\mathbf{A}\|^2}\mathbf{A}^\top\nabla g(\mathbf{A}\mathbf{y}^k)\right)$$

$$\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \frac{k+1}{k+3}\left(\mathbf{x}^{k+1} - \mathbf{x}^k\right)$$

- Convergence: We have

$$f(\mathbf{x}^k) + g(\mathbf{A}\mathbf{x}^k) - f(\mathbf{x}^\star) - g(\mathbf{A}\mathbf{x}^\star) \leq \frac{4L\|\mathbf{A}\|^2\|\mathbf{x}^\star - \mathbf{x}^0\|_2^2}{(k+1)^2}$$

*Famous Algorithms I

$$\boxed{\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})}$$

- FISTA (aka. accelerated proximal gradient method, aka. Nesterov acceleration):

  At iteration k:

$$\mathbf{x}^{k+1} = \mathrm{prox}_{f/L\|\mathbf{A}\|^2}\left(\mathbf{y}^k - \frac{1}{L\|\mathbf{A}\|^2}\mathbf{A}^\top\nabla g(\mathbf{A}\mathbf{y}^k)\right)$$
$$\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \frac{k+1}{k+3}\left(\mathbf{x}^{k+1} - \mathbf{x}^k\right)$$

- Convergence: We have

$$f(\mathbf{x}^k) + g(\mathbf{A}\mathbf{x}^k) - f(\mathbf{x}^\star) - g(\mathbf{A}\mathbf{x}^\star) \leq \frac{4L\|\mathbf{A}\|^2\|\mathbf{x}^\star - \mathbf{x}^0\|_2^2}{(k+1)^2}$$

- Problem: Strong convexity, otherwise optimal!

*Famous Algorithms I

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})$$

- FISTA (aka. accelerated proximal gradient method, aka. Nesterov acceleration):

  At iteration k:

$$\mathbf{x}^{k+1} = \mathrm{prox}_{f/L\|\mathbf{A}\|^2}\left(\mathbf{y}^k - \frac{1}{L\,\|\mathbf{A}\|^2}\mathbf{A}^\top \nabla g(\mathbf{A}\mathbf{y}^k)\right)$$

$$\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \frac{k+1}{k+3}\left(\mathbf{x}^{k+1} - \mathbf{x}^k\right)$$

- Convergence: We have

$$f(\mathbf{x}^k) + g(\mathbf{A}\mathbf{x}^k) - f(\mathbf{x}^\star) - g(\mathbf{A}\mathbf{x}^\star) \leq \frac{4L\,\|\mathbf{A}\|^2\,\|\mathbf{x}^\star - \mathbf{x}^0\|_2^2}{(k+1)^2}$$

- Problem: Strong convexity, otherwise optimal!

- Solution: Use a corrected momentum term or periodically **restart** the momentum.

*A useful minimax reformulation for the general case

$$\boxed{\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})}$$

- If $0 \in \text{ri}(\text{dom}g - A\text{dom}f)$ then the optimization problem is equivalent to

$$\max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + \langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle - g^*(\mathbf{y})$$

  where g^* is the Fenchel conjugate of g: $g^*(\mathbf{y}) := \max_{\mathbf{x}} \langle \mathbf{x}, \mathbf{y}\rangle - g(\mathbf{x})$.

  - Constrained case: $g(\mathbf{z}) = \begin{cases} 0 & \text{if } \mathbf{z} = \mathbf{b} \\ +\infty & \text{otherwise} \end{cases}$, and hence, $g^*(\mathbf{y}) = \langle \mathbf{b}, \mathbf{y}\rangle$

⋆Duality gap

- The duality gap:

$$G(\mathbf{x},\mathbf{y}) = f(\mathbf{x}) + g(\mathbf{A}\mathbf{x}) + g^*(\mathbf{y}) + f^*(-\mathbf{A}^\top\mathbf{y})$$

$$= \max_{\bar{\mathbf{y}}\in\mathcal{Y}} \Big(f(\mathbf{x}) + \langle \bar{\mathbf{y}}, \mathbf{A}\mathbf{x}\rangle - g^*(\bar{\mathbf{y}}) \Big) - \min_{\bar{\mathbf{x}}\in\mathcal{X}} \Big(- g^*(\mathbf{y}) + \langle \bar{\mathbf{x}}, \mathbf{A}^\top\mathbf{y}\rangle + f(\bar{\mathbf{x}}) \Big)$$

  - Note the symmetric roles between $(f, g, \mathbf{A})$ and $(-g^*, -f^*, \mathbf{A}^\top)$

- Useful properties:
  - Convex as a function of $(\mathbf{x},\mathbf{y})$
  - $G(\mathbf{x},\mathbf{y}) = 0$ iff $(\mathbf{x},\mathbf{y}) = (\mathbf{x}^\star,\mathbf{y}^\star)$

*Famous algorithms II

- Chambolle-Pock method (dual perspective):

  At iteration k:

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} f(x) + \langle y^k, Ax - c \rangle + \frac{\beta}{2} \left\| x - x^k \right\|_{\mathcal{X}}^2$$

$$y^{k+1} = y^k + \frac{\beta - \epsilon}{\|\mathbf{A}\|^2} \left(\mathbf{A}(2x^{k+1} - x^k) - c \right)$$

- Convergence: We have

$$G(\mathbf{x}^k, \mathbf{y}^k) \leq \frac{1}{k} \left(\frac{\beta}{2} D_{\mathcal{X}}^2 + \frac{\|\mathbf{A}\|^2}{2(\beta - \epsilon)} D_{\mathcal{Y}}^2 \right)$$

  where $D_{\mathcal{X}}$ is the diameter of dom$f$ and $D_{\mathcal{Y}}$ is the diameter of domg^*.

*Famous algorithms II

- Chambolle-Pock method (dual perspective):

  At iteration k:

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} f(x) + \langle y^k, Ax - c \rangle + \frac{\beta}{2} \left\| x - x^k \right\|_{\mathcal{X}}^2$$
$$y^{k+1} = y^k + \frac{\beta - \epsilon}{\|\mathbf{A}\|^2} \left(\mathbf{A}(2x^{k+1} - x^k) - c \right)$$

- Convergence: We have

$$G(\mathbf{x}^k, \mathbf{y}^k) \leq \frac{1}{k} \left(\frac{\beta}{2} D_{\mathcal{X}}^2 + \frac{\|\mathbf{A}\|^2}{2(\beta - \epsilon)} D_{\mathcal{Y}}^2 \right)$$

  where $D_{\mathcal{X}}$ is the diameter of dom$f$ and $D_{\mathcal{Y}}$ is the diameter of domg^*.

- Problem: We have $D_{\mathcal{Y}} = +\infty$.

*Smoothing the indicator function (primal perspective)

$$\boxed{\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})}$$

- Constrained case: $g(\mathbf{z}) = \begin{cases} 0 & \text{if } \mathbf{z} = \mathbf{b} \\ +\infty & \text{otherwise} \end{cases}$, and hence, $g^*(\mathbf{y}) = \langle \mathbf{b}, \mathbf{y} \rangle$

• A smoothed estimate of g by Nesterov around a center point λ:

$$g_\beta(\mathbf{z};\lambda) = \max_{\mathbf{y}\in\mathcal{Y}} \left(\langle \mathbf{z}, \mathbf{y} \rangle - g^*(\mathbf{y}) - \frac{\beta}{2} \|\mathbf{y} - \lambda\|^2 \right)$$

- $g_\beta(\mathbf{z};\lambda)$ is differentiable wrt $\mathbf{z}$ and $\nabla_\mathbf{z} g_\beta(\mathbf{z};\lambda)$ is $\frac{1}{\beta}$-Lipschitz
- $g_\beta(\mathbf{A}\mathbf{x}^k, \lambda) = \langle \lambda, \mathbf{A}\mathbf{x}^k - \mathbf{b} \rangle + \frac{1}{2\beta} \left\| \mathbf{A}\mathbf{x}^k - \mathbf{b} \right\|^2$

*A first attempt

- Nesterov's smooth minimization of non-smooth functions approach:

  Choose $\beta > 0$ and λ.

  Run FISTA on $\mathbf{x} \mapsto f(\mathbf{x}) + g_\beta(\mathbf{A}\mathbf{x}, \lambda)$ as a proxy for $f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})$.

- Convergence:

$$f(\mathbf{x}^k) + g_\beta(\mathbf{A}\mathbf{x}^k, \lambda) - f(\mathbf{x}^\star) - g(\mathbf{A}\mathbf{x}^\star) \leq \frac{4\|\mathbf{A}\|^2 \left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2}{\beta(k+1)^2}$$

$$f(\mathbf{x}^k) + g(\mathbf{A}\mathbf{x}^k) - f(\mathbf{x}^\star) - g(\mathbf{A}\mathbf{x}^\star) \leq \frac{4\|\mathbf{A}\|^2 \left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2}{\beta(k+1)^2} + \beta D_{\mathcal{Y}}$$

- Problem: The optimal choice for β is $\beta = \frac{\epsilon}{2D_{\mathcal{Y}}}$ where $D_{\mathcal{Y}} = +\infty$.

*Our fundamental theorem

- Denote the (primal) smoothed gap function at $y^\star$ as

$$S_\beta(x, \dot{y}) := f(x) + g_\beta(Ax; \dot{y}) - f(x^\star)$$

Theorem

If β and $S_\beta(x, \dot{y})$ are small, we have an approximate solution:

$$\|A\mathbf{x} - \mathbf{b}\| \leq \beta \left[\|\lambda^\star - \lambda\| + \left(\|\lambda^\star - \lambda\|^2 + 2\beta^{-1} S_\beta(\mathbf{x}; \lambda) \right)^{1/2} \right]$$

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \geq -\|\lambda^\star\| \|A\mathbf{x} - \mathbf{b}\|$$

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \leq S_\beta(\mathbf{x}, \lambda) + \|\lambda^\star\| \|A\mathbf{x} - \mathbf{b}\| + \frac{\beta}{2} \|\lambda^\star - \lambda\|^2$$

Algorithmic idea:

- Minimize the smoothed problem (*i.e.* augmented Lagrangian),

$$\min_{\mathbf{x}} f(\mathbf{x}) + \lambda^\top (A\mathbf{x} - \mathbf{b}) + \frac{1}{2\beta} \|A\mathbf{x} - \mathbf{b}\|^2, \tag{15}$$

with any method and obtain $S_\beta(\mathbf{x}, \lambda)$.

- Make sure $\beta \to 0$.

- Use the previous theorem to obtain guarantees for primal objective and feasibility, instead of dual problem!

⋆A Linearized Accelerated Quadratic Penalty Method

• Apply accelerated proximal gradient method (or FISTA) to minimize the augmented Lagrangian, with $\lambda := 0$, *i.e.* quadratic penalty function.

Accelerated Smoothed Gap Reduction (ASGARD)
1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $\beta > 0$. Set $\bar{\mathbf{x}}^0 := \hat{\mathbf{x}}^0 := \mathbf{x}^0$ and $\tau_0 := 1$
2. For $k = 0, 1, \cdots$, perform: $$\begin{cases} \bar{\mathbf{x}}^{k+1} & := \mathrm{prox}_{\beta\|A\|^{-2} f}\left(\bar{\mathbf{x}}^k - \|A\|^{-2} A^\top (A\hat{\mathbf{x}}^k - \mathbf{b})\right), \\ \hat{\mathbf{x}}^{k+1} & := \bar{\mathbf{x}}^{k+1} + \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}(\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k), \\ \tau_{k+1} & \in (0,1) \text{ root of } \tau^3 + \tau^2 + \tau_k^2 \tau - \tau_k^2 = 0, \\ \beta_{k+2} & = \frac{\beta_{k+1}}{1+\tau_{k+1}}. \end{cases}$$

• Recall: ASGARD corresponds to linearized, accelerated quadratic penalty method!

⋆Convergence theorem

Theorem

The iterates of ASGARD drive the smoothed gap to zero: $S_{\beta_k}(\bar{\mathbf{x}}^k, \lambda) = \mathcal{O}(1/k)$, and also provides a $\mathcal{O}(1/k)$ convergence guarantee in function value as well as feasibility:

$$\left\|A\bar{\mathbf{x}}^k - \mathbf{b}\right\| \leq \frac{\beta_1}{k+1}\left[\,\left\|\lambda^\star\right\| + \sqrt{\|\lambda^\star\|^2 + \frac{\|A\|^2}{\beta_1^2}\,\|\bar{\mathbf{x}}^0 - \mathbf{x}^\star\|^2}\,\right]$$

$$f(\bar{\mathbf{x}}^k) - f(\mathbf{x}^\star) \geq -\|\lambda^\star\|\|A\bar{\mathbf{x}}^k - \mathbf{b}\|$$

$$f(\bar{\mathbf{x}}^k) - f(\mathbf{x}^\star) \leq \frac{1}{k}\frac{\|A\|^2}{2\beta_1}\left\|\bar{\mathbf{x}}^0 - \mathbf{x}^\star\right\|^2 + \left\|\lambda^\star\right\|\left\|A\bar{\mathbf{x}}^k - \mathbf{b}\right\| + \frac{\beta_1}{k+1}\left\|\lambda^\star\right\|^2$$

*Effect of restarting

- Periodically restarting the algorithm with nonzero λ helps tremendously in practice. How to formalize?

*Restarted algorithm

• ASGARD changes β_k each iteration. How about decreasing $S_{\beta_k}(\mathbf{x}^k, \lambda^k)$ in a sequential manner?

• A double loop procedure:

- Apply accelerated proximal gradient method (or FISTA) to $\min_{\mathbf{x}} f(\mathbf{x}) + g_{\beta_k}(A\mathbf{x}; \lambda)$ for some number of iterations m_k
- Restart λ and decrease β_k
- Repeat for $k = k + 1$.

*Double Loop ASGARD

Double Loop ASGARD

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $\beta > 0$. Set $\bar{\mathbf{x}}^0 := \hat{\mathbf{x}}^0 := \mathbf{x}^0$ and $\tau_0 := 1$
2. For $k = 0, 1, \cdots$, perform:
   - **2.a** For $i = 0, 1, \cdots, m_k - 1$, perform:

$$
\begin{cases}
\hat{\mathbf{x}}_i^k & = (1 - \tau_k)\bar{\mathbf{x}}_i^k + \tau_k \tilde{\mathbf{x}}_i^k, \\
\tilde{\mathbf{x}}_{i+1}^k & = \mathrm{prox}_{\beta\|A\|^{-2} f}\left(\tilde{\mathbf{x}}_i^k - \|A\|^{-2} A^\top (\beta\lambda_k + A\hat{\mathbf{x}}_i^k - \mathbf{b})\right), \\
\bar{\mathbf{x}}_{i+1}^k & = \hat{\mathbf{x}}_i^k + \tau_k(\tilde{\mathbf{x}}_{i+1}^k - \tilde{\mathbf{x}}_i^k), \\
\tau_{k+1} & = \frac{2}{k+2},
\end{cases}
$$

   - **2.b** Restart primal and dual variable updates

$$
\begin{cases}
\bar{\mathbf{x}}_0^{k+1} & = \tilde{\mathbf{x}}_{m_k}^{k+1} \\
\lambda_{k+1} & = \lambda_k + \frac{1}{\beta_k}(A\bar{\mathbf{x}}_i^{k+1} - \mathbf{b}) \\
\tau_0 & = 1 \\
\beta_{k+1} & = \frac{\beta_k}{\omega} \\
m_{k+1} & = m_k \omega
\end{cases}
$$

- Corresponds to inexact augmented Lagrangian method with explicit inner termination rule.
- We can prove guarantees of the same order as ASGARD for the last iterate $\bar{\mathbf{x}}^k$, see [24].

*ADMM[3]

Primal problem with a specific decomposition structure

$$f^\star := \min_{\mathbf{x}:=(\mathbf{u},\mathbf{v})} \left\{ f(\mathbf{x}) := g(\mathbf{u}) + h(\mathbf{v}) : \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{v} = \mathbf{b},\ \mathbf{u} \in \mathcal{U},\ \mathbf{v} \in \mathcal{V} \right\}$$

- $\mathcal{X} := \mathcal{U} \times \mathcal{V}$ - nonempty, closed, convex and bounded.
- $\mathbf{A} := [\mathbf{B}, \mathbf{C}]$.

The Fenchel dual problem

$$d^\star := \max_{\lambda \in \mathbb{R}^n} \left\{ d(\lambda) := -g^*_{\mathcal{U}}(-\mathbf{B}^T\lambda) - h^*_{\mathcal{V}}(-\mathbf{C}^T\lambda) + \langle \mathbf{b}, \lambda \rangle \right\}$$

- $g^*_{\mathcal{U}}$ and $h^*_{\mathcal{U}}$ are the Fenchel conjugate of $g_{\mathcal{U}} := g + \delta_{\mathcal{U}}$ and $h_{\mathcal{V}} := h + \delta_{\mathcal{V}}$, resp.

The dual function

$$d(\lambda) := \underbrace{\min_{\mathbf{u} \in \mathcal{U}} \left\{ g(\mathbf{u}) + \langle \mathbf{B}^T\lambda, \mathbf{u} \rangle \right\}}_{d^1(\lambda)} + \underbrace{\min_{\mathbf{v} \in \mathcal{V}} \left\{ h(\mathbf{v}) + \langle \mathbf{C}^T\lambda, \mathbf{v} \rangle \right\}}_{d^2(\lambda)} - \langle \mathbf{b}, \lambda \rangle.$$

---

[3] Q. Tran-Dinh and V. Cevher, *Splitting the Smoothed Primal-dual Gap: Optimal Alternating Direction Methods* Tech. Report, 2015, (http://arxiv.org/pdf/1507.03734.pdf) / (http://lions.epfl.ch/publications)

*Standard ADMM as the dual Douglas-Rachford method

We can derive ADMM via the Douglas-Rachford splitting on the dual:

$$0 \in \mathbf{B}\partial g^*_{\mathcal{U}}(-\mathbf{B}^T\lambda) + \mathbf{C}\partial h^*{}_{\mathcal{V}}(-\mathbf{C}^T\lambda) + \boldsymbol{c},$$

which is the optimality condition of the **dual problem**.

Douglas-Rachford splitting method

$$\begin{cases} \mathbf{z}_g^k & := \mathrm{prox}_{\eta_k^{-1} g^*_{\mathcal{U}}(-\mathbf{B}^T \cdot)}(\lambda^k) \\ \mathbf{z}_h^k & := \mathrm{prox}_{\eta_k^{-1} h^*_{\mathcal{V}}(-\mathbf{C}^T \cdot)}(2\mathbf{z}_g^k - \lambda^k) \\ \lambda^{k+1} & := \lambda^k + (\mathbf{z}_g^k - \mathbf{z}_h^k). \end{cases}$$

Standard ADMM

$$\begin{cases} \mathbf{u}^{k+1} & := \arg\min_{\mathbf{u}\in\mathcal{U}} \left\{ g(\mathbf{u}) + \langle \lambda^k, \mathbf{B}\mathbf{u}\rangle + \frac{\eta_k}{2}\|\mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{v}^k - \mathbf{b}\|^2 \right\} \\ \mathbf{v}^{k+1} & := \arg\min_{\mathbf{v}\in\mathcal{V}} \left\{ h(\mathbf{v}) + \langle \lambda^k, \mathbf{C}\mathbf{v}\rangle + \frac{\eta_k}{2}\|\mathbf{B}\mathbf{u}^{k+1} + \mathbf{C}\mathbf{v} - \mathbf{b}\|^2 \right\} \\ \lambda^{k+1} & := \lambda^k + \eta_k\left(\mathbf{B}\mathbf{u}^{k+1} + \mathbf{C}\mathbf{v}^{k+1} - \mathbf{b}\right). \end{cases}$$

Here, $\eta_k > 0$ is a given penalty parameter.

*Splitting the smoothed gap

Smoothing the gap

- The **dual components** d^1 and d^2 are nonsmooth. We smooth one, e.g., d^1, using:

$$d^1_\gamma(\lambda) := \min_{\mathbf{u}\in\mathcal{U}} \left\{ g(\mathbf{u}) + \frac{\gamma}{2}\|\mathbf{B}(\mathbf{u}-\mathbf{u}_c)\|^2 + \langle \lambda, \mathbf{B}\mathbf{u}\rangle \right\}$$

- Recall: We also approximate f by f_β as:

$$f_\beta(\mathbf{x}) := f(\mathbf{x}) + \frac{1}{2\beta}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 \to f(\mathbf{x}) \quad \text{as } \mathbf{x} \text{ becomes feasible}$$

Three key properties of d^1_γ

- d^1_γ is concave and smooth.
- ∇d^1_γ is Lipschitz continuous with $L := \gamma^{-1}$.
- d^1_γ approximates d^1 as:

$$d^1_\gamma(\lambda) - \gamma D_{\mathcal{U}} \le d^1(\lambda) \le d^1_\gamma(\lambda),$$

where $D_{\mathcal{U}} := \max\left\{(1/2)\|\mathbf{B}(\mathbf{u}-\mathbf{u}_c)\|^2 : \mathbf{u}\in\mathcal{U}\right\}$.

*Our ADMM scheme: D-R on the smoothed gap

- Our new ADMM scheme consists of three steps: ADMM step, acceleration step, and primal averaging.

Step 1: The main ADMM steps

$$\begin{cases} \hat{\mathbf{u}}^{k+1} & := \arg\min_{\mathbf{u}\in\mathcal{U}} \left\{ g_{\gamma_{k+1}}(\mathbf{u}) + \langle \hat{\lambda}^k, \mathbf{B}\mathbf{u}\rangle + \frac{\rho_k}{2}\|\mathbf{B}\mathbf{u} + \mathbf{C}\hat{\mathbf{v}}^k - \mathbf{b}\|^2 \right\} \\ \hat{\mathbf{v}}^{k+1} & := \arg\min_{\mathbf{v}\in\mathcal{V}} \left\{ h(\mathbf{v}) + \langle \hat{\lambda}^k, \mathbf{C}\mathbf{v}\rangle + \frac{\eta_k}{2}\|\mathbf{B}\hat{\mathbf{u}}^{k+1} + \mathbf{C}\mathbf{v} - \mathbf{b}\|^2 \right\} \\ \lambda^{k+1} & := \hat{\lambda}^k + \eta_k\left(\mathbf{B}\hat{\mathbf{u}}^{k+1} + \mathbf{C}\hat{\mathbf{v}}^{k+1} - \mathbf{b}\right). \end{cases}$$

where $g_\gamma(\cdot) := g(\cdot) + \frac{\gamma}{2}\|\mathbf{B}(\cdot - \mathbf{u}_c)\|^2$.

*The dual accelerated and primal averaging steps

- **Step 2:** [Dual acceleration] $\hat{\lambda}^k$ is computed as:

$$\hat{\lambda}^k := (1-\tau_k)\lambda_k + \frac{\tau_k}{\beta_k}(\mathbf{B}\mathbf{u}^k + \mathbf{C}\mathbf{v}^k - \mathbf{b}).$$

- **Step 3:** [Averaging] The primal iteration $\mathbf{x}^k := (\mathbf{u}^k, \mathbf{v}^k)$ is updated as:

$$\mathbf{u}^{k+1} := (1-\tau_k)\mathbf{u}^k + \tau_k\hat{\mathbf{u}}^{k+1} \quad \text{and} \quad \mathbf{v}^{k+1} := (1-\tau_k)\mathbf{v}^k + \tau_k\hat{\mathbf{v}}^{k+1}.$$

*How do we update parameters?

Duality gap and smoothed gap functions

- The duality gap: $G(\mathbf{w}) := f(\mathbf{x}) - d(\lambda)$, where $\mathbf{w} := (\mathbf{x}, \lambda)$.
- The smoothed gap: $\boxed{G_{\gamma\beta}(\mathbf{w}) := f_\beta(\mathbf{x}) - d_\gamma(\lambda)}$ with $d_\gamma := d^1_\gamma + d^2$.

Model-based gap reduction

The **gap reduction model** provides conditions to derive parameter update rules:

$$\boxed{G_{\gamma_{k+1}\beta_{k+1}}(\mathbf{w}^{k+1}) \leq (1-\tau_k)G_{\gamma_k\beta_k}(\mathbf{w}^k) + \tau_k(\eta_k + \rho_k)D_{\mathcal{X}}}$$

where $\gamma_{k+1} < \gamma_k$, $\beta_{k+1} < \beta_k$ and $D_{\mathcal{X}} := \max_{\mathbf{x}\in\mathcal{X}} \left\{(1/2)\|\mathbf{Bu} + \mathbf{Cv} - \mathbf{b}\|^2\right\}$.

Update rules

- **The smoothness parameters:** $\gamma_{k+1} := \frac{2\gamma_0}{k+3}$ and $\beta_k := \frac{9(k+3)}{\gamma_0(k+1)(k+7)}$.
- **The penalty parameters:** $\eta_k := \frac{\gamma_0}{k+3}$ and $\rho_k := \frac{3\gamma_0}{(k+3)(k+4)}$.
- **The step-size** $\tau_k := \frac{3}{k+4} \quad \Rightarrow \quad \mathcal{O}\left(\frac{1}{k}\right)$.

*Convergence guarantee & Other cases of interest

Convergence rate guarantee

- Rate on the primal objective residual and constraint feasibility:

$$
\begin{aligned}
f(\mathbf{x}^k) - f^\star &\leq \frac{2\gamma_0 D_{\mathcal{U}}}{k+2} + \frac{3\gamma_0 D_{\mathcal{X}}}{2(k+3)}\left(1 + \frac{6}{k+2}\right) &\Rightarrow \mathcal{O}\left(\frac{1}{k}\right)\\
\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\| &\leq \frac{18D_d^*}{\gamma_0(k+2)} + \frac{6}{k+2}\sqrt{D_{\mathcal{U}} + \frac{3(k+8)}{2(k+3)}D_{\mathcal{X}}} &\Rightarrow \mathcal{O}\left(\frac{1}{k}\right)
\end{aligned}
$$

where D_d^* is the diameter of the **dual solution set** $\Lambda^\star$.

- Lower bound: $-D_d^*\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\| \leq f(\mathbf{x}^k) - f^\star$.
- Rate on the dual objective residual:

$$
d^\star - d(\lambda^k) \leq \frac{18(D_d^*)^2}{\gamma_0(k+2)} + \frac{6D_d^*}{k+2}\sqrt{D_{\mathcal{U}} + \frac{3(k+8)}{2(k+3)}D_{\mathcal{X}}} \Rightarrow \mathcal{O}\left(\frac{1}{k}\right).
$$

Special cases: *cf.*, `http://lions.epfl.ch/publications`

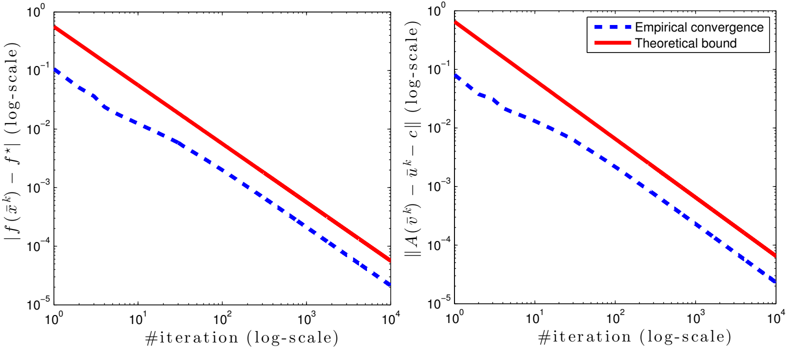
- Full-column rank or orthogonality of $\mathbf{A}$: Using smoothing term $(\gamma/2)\|\mathbf{u} - \mathbf{u}_c\|^2$.
- Strong convexity of g: We do not need to smooth d^1.
- Decomposability of g and $\mathcal{U}$: Using smoothing term

$$
(\gamma/2)\sum_{i=1}^{s}\|\mathbf{B}_i(\mathbf{u}_i - \mathbf{u}_{c,i})\|^2.
$$

*A comparison to the theoretical bounds

A stylized example: Square-root LASSO

$$f^\star := \min_{\mathbf{u}\in\mathcal{U},\mathbf{v}\in\mathcal{V}} \left\{ f(\mathbf{x}) := \|\mathbf{u}\|_2 + \kappa \|\mathbf{v}\|_1 : \mathbf{B}(\mathbf{v}) - \mathbf{u} = \boldsymbol{c} \right\}.$$



- See the preprint for more examples, enhancements, ...

References I

[1] H.H. Bauschke and P. Combettes.
Convex analysis and monotone operators theory in Hilbert spaces.
Springer-Verlag, 2011.

[2] Dimitri P. Bertsekas.
Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series).
Athena Scientific, 1996.

[3] A. Chambolle and T. Pock.
A first-order primal-dual algorithm for convex problems with applications to imaging.
Journal of Mathematical Imaging and Vision, 40(1):120–145, 2011.

[4] G. Chen and M. Teboulle.
A proximal-based decomposition method for convex minimization problems.
Math. Program., 64:81–101, 1994.

[5] P. L. Combettes and V. R. Wajs.
Signal recovery by proximal forward-backward splitting.
Multiscale Model. Simul., 4:1168–1200, 2005.

References II

[6] D. Davis.
Convergence rate analysis of the forward-Douglas-Rachford splitting scheme.
UCLA CAM report 14-73, 2014.

[7] D. Davis and W. Yin.
Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions.
UCLA CAM report 14-58, 2014.

[8] D. Davis and W. Yin.
A three-operator splitting scheme and its optimization applications.
Tech. Report., 2015.

[9] J. Eckstein and D. Bertsekas.
On the Douglas - Rachford splitting method and the proximal point algorithm for maximal monotone operators.
Math. Program., 55:293–318, 1992.

[10] J. E. Esser.
Primal-dual algorithm for convex models and applications to image restoration, registration and nonlocal inpainting.
Phd. thesis, University of California, Los Angeles, Los Angeles, USA, 2010.

References III

[11] D. Gabay and B. Mercier.
A dual algorithm for the solution of nonlinear variational problems via finite element approximation.
Computers & Mathematics with Applications, 2(1):17 – 40, 1976.

[12] T. Goldstein, E. Esser, and R. Baraniuk.
Adaptive Primal-Dual Hybrid Gradient Methods for Saddle Point Problems.
Tech. Report., http://arxiv.org/pdf/1305.0546v1.pdf:1–26, 2013.

[13] T. Goldstein, B. ODonoghue, and S. Setzer.
Fast Alternating Direction Optimization Methods.
SIAM J. Imaging Sci., 7(3):1588–1623, 2012.

[14] B. He and X. Yuan.
Convergence analysis of primal-dual algorithms for saddle-point problem: from contraction perspective.
SIAM J. Imaging Sciences, 5:119–149, 2012.

[15] Bingsheng He and Xiaoming Yuan.
On the acceleration of augmented lagrangian method for linearly constrained optimization.
2010.

References IV

[16] B.S. He and X.M. Yuan.
On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method.
SIAM J. Numer. Anal., 50:700–709, 2012.

[17] I. Necoara and J.A.K. Suykens.
Interior-point lagrangian decomposition method for separable convex optimization.
J. Optim. Theory and Appl., 143(3):567–588, 2009.

[18] J. Nocedal and S.J. Wright.
Numerical Optimization.
Springer Series in Operations Research and Financial Engineering. Springer, 2 edition, 2006.

[19] Y. Ouyang, Y. Chen, G. LanG. Lan., and E. JR. Pasiliao.
An accelerated linearized alternating direction method of multiplier.
Tech, 2014.

[20] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi.
Simple bounds for noisy linear inverse problems with exact side information.
2013.
arXiv:1312.0641v2 [cs.IT].

References V

[21] R. T. Rockafellar.
Convex Analysis, volume 28 of *Princeton Mathematics Series*.
Princeton University Press, 1970.

[22] Q. Tran-Dinh and V. Cevher.
Constrained convex minimization via model-based excessive gap.
In *Proc. the Neural Information Processing Systems Foundation conference (NIPS2014)*, pages 1–9, Montreal, Canada, December 2014.

[23] Q. Tran-Dinh, I. Necoara, C. Savorgnan, and M. Diehl.
An Inexact Perturbed Path-Following Method for Lagrangian Decomposition in Large-Scale Separable Convex Optimization.
SIAM J. Optim., 23(1):95–125, 2013.

[24] Quoc Tran-Dinh, Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher.
An adaptive primal-dual framework for nonsmooth convex minimization.
arXiv preprint arXiv:1808.04648, 2018.

[25] Quoc Tran-Dinh, Olivier Fercoq, and Volkan Cevher.
A smooth primal-dual optimization framework for nonsmooth composite convex minimization.
SIAM Journal on Optimization, 28(1):96–134, 2018.

References VI

[26] P. Tseng.
Applications of splitting algorithm to decomposition in convex programming and variational inequalities.
SIAM J. Control Optim., 29:119–138, 1991.

[27] E. Wei, A. Ozdaglar, and A.Jadbabaie.
A Distributed Newton Method for Network Utility Maximization.
http://web.mit.edu/asuman/www/publications.htm, 2011.

[28] Yangyang Xu.
Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming.
SIAM Journal on Optimization, 27(3):1459–1484, 2017.

[29] G. Zhao.
A Lagrangian dual method with self-concordant barriers for multistage stochastic convex programming.
Math. Progam., 102:1–24, 2005.