Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 10: Constrained convex minimization I

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

EE-556 (Fall 2019)

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

- Jorge Nocedal, Stephen Wright, *Numerical Optimization, Chapter 17.* Springer, 2016.
- Yangyang Xu, *Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming.* SIAM J. Optim. 27(3):1459-1484, 2017.

Swiss army knife of convex formulations

A **primal problem** prototype

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

- 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 $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star - \mathbf{b} \in \mathcal{K}$ and $\mathbf{x}^\star \in \mathcal{X}$

An example from the sparseland

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

Broad context for (1):

- 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*, ...

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 [24])

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 × cost per iteration) amounts to $\mathcal{O}(n^2 p^{1.5}\log(\frac{1}{\epsilon}))$.

An alternative formulation

- For a lighter notation, we focus on the following problem.

A simplified template

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \Big\}, \qquad (2)$$

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

- This is equivalent with:

A **primal problem** prototype

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K}, \ \mathbf{x} \in \mathcal{X} \Big\}, \qquad (3)$$

- 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 (3) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star - \mathbf{b} \in \mathcal{K}$ and $\mathbf{x}^\star \in \mathcal{X}$

*How do we reformulate?

A primal problem template

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

- Let $\mathbf{r}_1 = \mathbf{A}\mathbf{x}-\mathbf{b}\in\mathbb{R}^n$ and $\mathbf{r}_2 = \mathbf{x}\in\mathbb{R}^p$.

First step

$$\min_{\mathbf{x},\mathbf{r}_1,\mathbf{r}_2}\left\{f(\mathbf{x}) : \mathbf{r}_1\in\mathcal{K}, \mathbf{r}_2\in\mathcal{X}, \mathbf{A}\mathbf{x}-\mathbf{b}=\mathbf{r}_1, \mathbf{x}=\mathbf{r}_2\right\}.$$

- Define $\mathbf{z} = \begin{bmatrix}\mathbf{x}\\ \mathbf{r}_1\\ \mathbf{r}_2\end{bmatrix}\in\mathbb{R}^{2p+n}$, $\bar{\mathbf{A}} = \begin{bmatrix}\mathbf{A} & -\mathbf{I}_{n\times n} & \mathbf{0}_{n\times p}\\ \mathbf{I}_{p\times p} & \mathbf{0}_{p\times n} & -\mathbf{I}_{p\times p}\end{bmatrix}$, $\bar{\mathbf{b}} = \begin{bmatrix}\mathbf{b}\\ \mathbf{0}\end{bmatrix}$,

$\bar{f}(\mathbf{z}) = f(\mathbf{x}) + \delta_{\mathcal{K}}(\mathbf{r}_1) + \delta_{\mathcal{X}}(\mathbf{r}_2)$ where $\delta_{\mathcal{X}}(\mathbf{x}) = 1$, if $\mathbf{x}\in\mathcal{X}$, and $\delta_{\mathcal{X}}(\mathbf{x}) = +\infty$, o/w.

The simplified template

$$\min_{\mathbf{z}\in\mathbb{R}^{2p+n}}\left\{\bar{f}(\mathbf{z}) : \bar{\mathbf{A}}\mathbf{z} = \bar{\mathbf{b}}\right\}.$$

Performance of optimization algorithms

Exact vs. approximate solutions

- Computing an **exact solution** $\mathbf{x}^\star$ to (2) 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`

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)!

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

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

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

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

$$\boxed{f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b} \right\}} \tag{4}$$

Our definition of ϵ-accurate solutions [26]

Given a numerical tolerance $\epsilon \geq 0$, a point $\mathbf{x}^\star_\epsilon \in \mathbb{R}^p$ is called an ϵ-solution of (4) if

$$\begin{cases} f(\mathbf{x}^\star_\epsilon) - f^\star & \leq \epsilon \text{ (objective residual)}, \\ \|\mathbf{A}\mathbf{x}^\star_\epsilon - \mathbf{b}\| & \leq \epsilon \text{ (feasibility gap)}, \end{cases}$$

- 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 ...

- Penalty and augmented Lagrangian methods:
  - Quadratic penalty method [4].
  - Exact penalty method [3].
  - Augmented Lagrangian method [18, 25].
- Variants of the **Arrow-Hurwitz's method**:
  - Chambolle-Pock's algorithm [5], and its variants, e.g., He-Yuan's variant [16].
  - Primal-dual Hybrid Gradient (PDHG) method and its variants [12, 14].
  - Proximal-based decomposition (Chen-Teboulle's algorithm) [6].
- Splitting techniques from monotone inclusions:
  - Primal-dual splitting algorithms [2, 7, 30, 8, 9].
  - Three-operator splitting [10].
- Dual splitting techniques:
  - Alternating minimization algorithms (AMA) [13, 30].
  - Alternating direction methods of multipliers (ADMM) [11, 17].
  - Accelerated variants of AMA and ADMM [9, 15].
  - Preconditioned ADMM, Linearized ADMM and inexact Uzawa algorithms [5, 22].
- **Second-order decomposition methods:**
  - Dual (quasi) Newton methods [31].
  - Smoothing decomposition methods via barriers functions [19, 27, 34].

Performance of optimization algorithms

Finding the fastest algorithm within the zoo is tricky!

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

Outline

Primal approach: Penalization

Dual approach: Lagrangian-based method

Primal-dual approach: Augmented Lagrangian

A primal approach: Penalty methods

$$\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b} \right\}$$

- Rule of thumb: Convert constrained problem (difficult) to unconstrained (easy)

Penalization

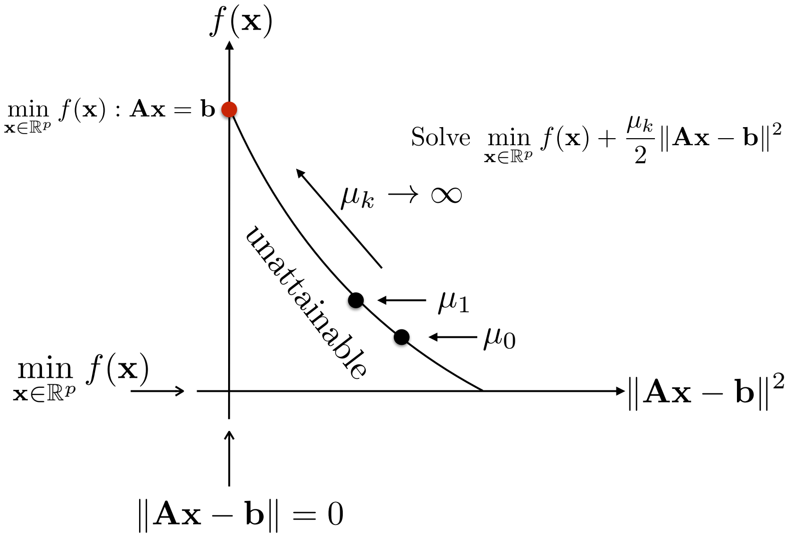
- Penalized function with penalty parameter $\mu > 0$:

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

Observations:

- Minimize a weighted combination of $f(\mathbf{x})$ and $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ at the same time.
- μ determines the weight of $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$.
- As $\mu \to \infty$, we enforce $\mathbf{A}\mathbf{x} = \mathbf{b}$.
- Other functions than the quadratic $\frac{1}{2}\|\cdot\|^2$ are also possible. For example, exact nonsmooth penalty functions:
  - $\mu\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ or $\mu\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1$
  - They work with finite μ, but they are difficult to solve [21, Section 17.2], [3]

Quadratic penalty: Intuition



Quadratic penalty: Conceptual algorithm

Quadratic penalty method (QP):
1. Choose $\mathbf{x}_0 \in \mathbb{R}^p$ and $\mu_0 > 0$.
2. For $k = 0, 1, \cdots$, perform:
2.a. $\mathbf{x}_k := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}$.
2.b. Update $\mu_{k+1} > \mu_k$.

Theorem [21, Theorem 17.1]

Assume that f is smooth 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,

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) \colon \mathbf{A}\mathbf{x} = \mathbf{b} \right\}.$$

*Quadratic penalty: Proof of convergence

Theorem [21, Theorem 17.1]

Assume that f is smooth 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,

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}), \forall \mathbf{x} \text{ with } \mathbf{A}\mathbf{x} = \mathbf{b}. \tag{5}$$

Since $\mathbf{x}_k \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} F_{\mu_k}(\mathbf{x})$ and $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$,

$$f(\mathbf{x}_k) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\|^2 \leq f(\mathbf{x}^\star) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x}^\star - \mathbf{b}\|^2 = f(\mathbf{x}^\star). \tag{6}$$

Rearranging, we get

$$\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\|^2 \leq \frac{2}{\mu_k}\left(f(\mathbf{x}^\star) - f(\mathbf{x}_k)\right). \tag{7}$$

$\bar{x}$ is a limit point: $\lim_{k\in\mathcal{K}} \mathbf{x}_k = \bar{\mathbf{x}}$, for a subsequence $\mathcal{K}$.

- Taking the limit of (7) and using $\mu_k \to \infty$ gives $\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\| = 0$.
- Taking the limit of (6) and using that $\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\| = 0$ gives $f(\bar{\mathbf{x}}) \leq f(\mathbf{x}^\star)$.
- Since $f(x^\star)$ is the minimum value and $\bar{x}$ is feasible, we conclude that $\bar{x}$ is a solution.

Quadratic penalty: Limitations

Quadratic penalty method (QP):
1. Choose $\mathbf{x}_0 \in \mathbb{R}^p$ and $\mu_0 > 0$.
2. For $k = 0, 1, \cdots$, perform:
2.a. $\mathbf{x}_k := \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ f(\mathbf{x}) + \frac{\mu_k}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}$.
2.b. Update $\mu_{k+1} > \mu_k$.

- Ill-conditioned subproblems as $\mu_k \to \infty$:

$$\mathbf{x}_k := \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ f(\mathbf{x}) + \frac{\mu_k}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}.$$

Common improvements:

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

We cover this idea in the sequel.

Quadratic penalty: Linearization

Generalized quadratic penalty method:
1. Choose $\mathbf{x}_0 \in \mathbb{R}^p$, $\mu_0 > 0$ and positive semidefinite matrix $\mathbf{Q}_k$.
2. For $k = 0, 1, \cdots,$ perform:
2.a. $\mathbf{x}_k := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) + \frac{\mu_k}{2}\lVert \mathbf{A}\mathbf{x} - \mathbf{b}\rVert^2 + \frac{1}{2}\lVert \mathbf{x} - \mathbf{x}_{k-1}\rVert^2_{\mathbf{Q}_k} \right\}$.
2.b. Update $\mu_{k+1} > \mu_k$.

- We minimize a majorizer of $F_\mu(\mathbf{x})$, parametrized by $\mathbf{Q}_k$.
- $\mathbf{Q}_k = \mathbf{0}$ recovers the standard QP.
- $\mathbf{Q}_k = \mathbf{I}$ gives strongly convex subproblems.
- $\mathbf{Q}_k = \alpha_k \mathbf{I} - \mu_k \mathbf{A}^\top \mathbf{A}$, with $\alpha_k \geq \mu_k \lVert \mathbf{A}\rVert^2$ gives

$$\begin{aligned}\mathbf{x}_k &= \arg\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) + \frac{\mu_k}{2}\lVert \mathbf{A}\mathbf{x} - \mathbf{b}\rVert^2 + \frac{1}{2}\lVert \mathbf{x} - \mathbf{x}_{k-1}\rVert^2_{\mathbf{Q}_k} \\ &= \mathrm{prox}_{\frac{1}{\alpha_k} f}\left(\mathbf{x}_{k-1} - \frac{\mu_k}{\alpha_k}\mathbf{A}^\top(\mathbf{A}\mathbf{x}_{k-1} - \mathbf{b})\right)\end{aligned}$$

Only one proximal operator!

  ▷ Picking $\alpha_k = \mu_k \lVert \mathbf{A}\rVert^2$ gives

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

⋆Derivation for linearization

- $\mathbf{Q}_k = \alpha_k \mathbf{I} - \mu_k \mathbf{A}^\top \mathbf{A}$, with $\alpha_k \geq \mu_k \|\mathbf{A}\|^2$ gives

$$
\begin{aligned}
\mathbf{x}_k &= \arg\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|^2_{\mathbf{Q}_k} \\
&= \arg\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}_{k-1}\|^2 + \mu_k\langle \mathbf{A}\mathbf{x}_{k-1} - \mathbf{b}, \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}_{k-1}\rangle \\
&+ \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x}_{k-1} - \mathbf{b}\|^2 + \frac{\alpha_k}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|^2 - \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}_{k-1}\|^2 \\
&= \arg\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) + \mu_k\langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x}_{k-1} - \mathbf{b}\rangle + \frac{\alpha_k}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|^2 \\
&= \arg\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) + \frac{\alpha_k}{2}\left\|\mathbf{x} - \left(\mathbf{x}_{k-1} - \frac{\mu_k}{\alpha_k}\mathbf{A}^\top(\mathbf{A}\mathbf{x}_{k-1} - \mathbf{b})\right)\right\|^2 \\
&= \mathrm{prox}_{\frac{1}{\alpha_k}f}\left(\mathbf{x}_{k-1} - \frac{\mu_k}{\alpha_k}\mathbf{A}^\top(\mathbf{A}\mathbf{x}_{k-1} - \mathbf{b})\right)
\end{aligned}
$$

Only one proximal operator!

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

Recall: Prox-operator

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

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.

Quadratic penalty: Linearized methods

Linearized quadratic penalty method (LQP):
1. Choose $\mathbf{x}_0 \in \mathbb{R}^p$, $\sigma_0 = 1$, $\mu_0 > 0$.
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+1} such that $\frac{(1-\sigma_{k+1})^2}{\sigma_{k+1}} = \frac{1}{\sigma_k}$.
2.c. Update $\mu_{k+1} = \sqrt{\sigma_{k+1}}$.

Quadratic penalty: Linearized methods

Linearized quadratic penalty method (LQP):

1. Choose $\mathbf{x}_0 \in \mathbb{R}^p$, $\sigma_0 = 1$, $\mu_0 > 0$.
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+1} such that $\frac{(1-\sigma_{k+1})^2}{\sigma_{k+1}} = \frac{1}{\sigma_k}$.

   2.c. Update $\mu_{k+1} = \sqrt{\sigma_{k+1}}$.

Accelerated linearized quadratic penalty method (ALQP):

1. Choose $\mathbf{x}_0, \mathbf{y}_0 \in \mathbb{R}^p$, $\tau_0 = 1$, $\mu_0 > 0$.
2. For $k = 0, 1, \cdots,$ perform:

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

   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 ALQP

Theorem (Convergence [29])

- *LQP:*

$$\begin{cases} |f(\mathbf{x}_k) - f(x^\star)| & \leq \mathcal{O}\left(\frac{\mu_0}{\sqrt{k}} + \frac{1}{\mu_0 \sqrt{k}}\right) \\ \|\mathbf{A}\mathbf{x}_k - \mathbf{b}\| & \leq \mathcal{O}\left(\frac{1}{\mu_0 \sqrt{k}}\right) \end{cases}$$

- *ALQP:*

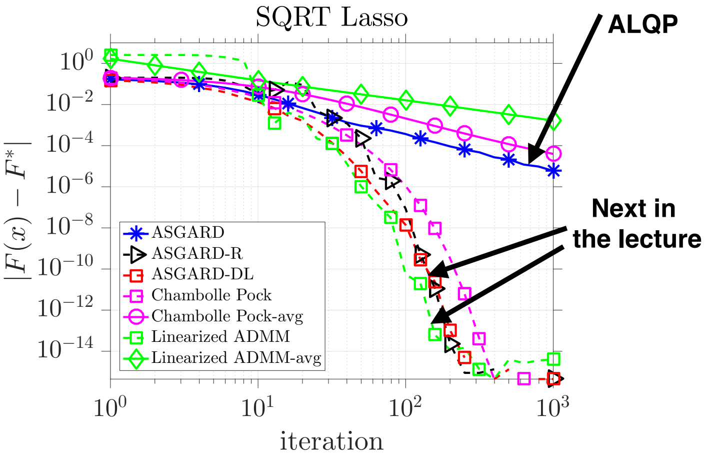
$$\begin{cases} |f(\mathbf{x}_k) - f(x^\star)| & \leq \mathcal{O}\left(\frac{\mu_0}{k} + \frac{1}{\mu_0 k}\right) \\ \|\mathbf{A}\mathbf{x}_k - \mathbf{b}\| & \leq \mathcal{O}\left(\frac{1}{\mu_0 k}\right) \end{cases}$$

- Poor (worst case) performance in practice.

What happens in practice

- A nonsmooth problem: SQRT Lasso

$$\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2 + \lambda\|\mathbf{x}\|_1.$$



Outline

Primal approach: Penalization

Dual approach: Lagrangian-based method

Primal-dual approach: Augmented Lagrangian

An alternative to penalization

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

- Recall the penalization approach:

$$f^\star = f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{A}\mathbf{x}^\star - \mathbf{b}\|^2, \quad \forall \mu > 0.$$

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

▷ Another unconstrained formulation at the solution

$$f^\star = f(\mathbf{x}^\star) + \max_{\lambda\in\mathbb{R}^n} \langle \lambda, \mathbf{A}\mathbf{x}^\star - \mathbf{b}\rangle.$$

▷ We then define

$$F_\lambda(\mathbf{x}) = f(\mathbf{x}) + \max_{\lambda\in\mathbb{R}^n} \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle.$$

$$\max_{\lambda\in\mathbb{R}^n} \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle = \begin{cases} 0, & \text{if } \mathbf{A}\mathbf{x} = \mathbf{b}, \\ +\infty, & \text{if } \mathbf{A}\mathbf{x} \neq \mathbf{b}. \end{cases}$$

$$\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x})\colon \mathbf{A}\mathbf{x} = \mathbf{b} \right\} = \min_{\mathbf{x}\in\mathbb{R}^p} \max_{\lambda\in\mathbb{R}^n} \left\{ f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle \right\}$$

Exchanging max and min

$$\boxed{f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b} \right\} = \min_{\mathbf{x}\in\mathbb{R}^p} \max_{\lambda\in\mathbb{R}^n} \left\{ f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \right\}}$$

$$\boxed{\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle}$$

- Since $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$, it holds for any λ

$$\begin{aligned} \mathcal{L}(\mathbf{x}^\star, \lambda) = f(x^\star) &= f(\mathbf{x}^\star) + \langle \lambda, \mathbf{A}\mathbf{x}^\star - \mathbf{b} \rangle \\ &\geq \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \right\} \\ &= \min_{\mathbf{x}\in\mathbb{R}^p} \mathcal{L}(\mathbf{x}, \lambda). \end{aligned}$$

- Take maximum of both sides in λ:

$$f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} \max_{\lambda\in\mathbb{R}^n} \mathcal{L}(\mathbf{x}, \lambda) \geq \max_{\lambda\in\mathbb{R}^n} \min_{\mathbf{x}\in\mathbb{R}^p} \mathcal{L}(\mathbf{x}, \lambda) =: \max_{\lambda\in\mathbb{R}^n} d(\lambda) = d^\star.$$

- $\max\min$ is the best lower bound to $\min\max$.

Terminology

- We established

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

- **Lagrangian function:**

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

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:**

$$d(\lambda) := \min_{\mathbf{x} \in \mathbb{R}^p} \Big\{ \mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \Big\}. \tag{8}$$

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

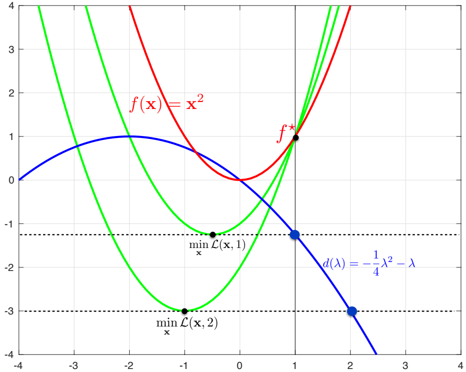
▷ Dual function is concave $\Rightarrow$ (potentially) easier to solve.

▷ $f^\star \geq d^\star$ is called weak duality.

Primal and dual functions

$$f(x^\star) = \min_{\mathbf{x} \in \mathbb{R}} \left\{ \mathbf{x}^2 : \mathbf{x} - 1 = 0 \right\}$$

$$d(\lambda) = \min_{\mathbf{x} \in \mathbb{R}} \mathcal{L}(\mathbf{x}, \lambda) = \min_{\mathbf{x} \in \mathbb{R}} \left\{ \mathbf{x}^2 + \langle \lambda, \mathbf{x} - 1 \rangle \right\}.$$



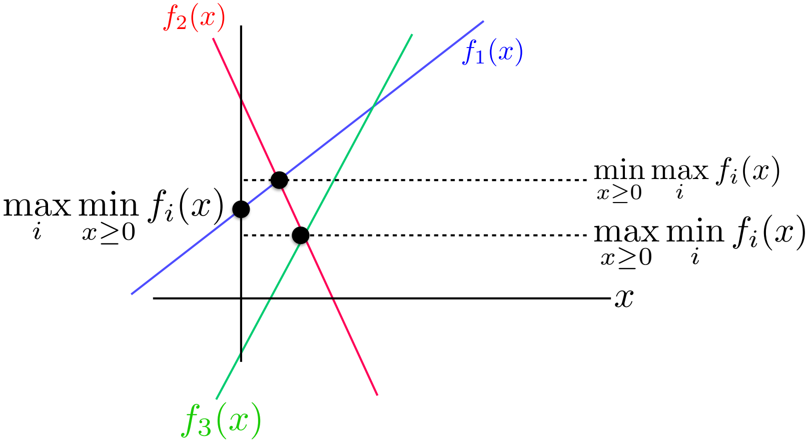
horizontal axis is shared by x and λ

Figure adapted from [1]

A visual clue

- Recall weak duality: $f^\star \geq \max_{\lambda \in \mathbb{R}^n} \min_{\mathbf{x} \in \mathbb{R}^p} \mathcal{L}(\mathbf{x}, \lambda)$, then it follows

$$\max_{\lambda \in \mathbb{R}^n} \min_{\mathbf{x} \in \mathbb{R}^p} \mathcal{L}(\mathbf{x}, \lambda) \leq \min_{\mathbf{x} \in \mathbb{R}^p} \max_{\lambda \in \mathbb{R}^n} \mathcal{L}(\mathbf{x}, \lambda) = \begin{cases} f^\star, \text{ if } \mathbf{Ax} = \mathbf{b} \\ +\infty, \text{ if } \mathbf{Ax} \neq \mathbf{b} \end{cases}$$



Saddle point

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

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

Recall the minimax form:

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

Saddle point

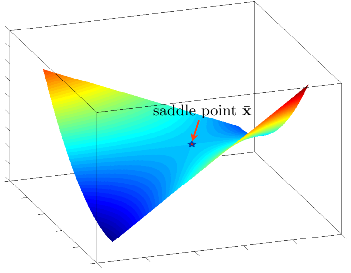
A point $(\mathbf{x}^\star, \lambda^\star) \in \mathbb{R}^p \times \mathbb{R}^n$ is called a saddle point of the Lagrangian function $\mathcal{L}$ if

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

Recall the minimax form:

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

Illustration of saddle point



Necessary and sufficient condition

Minimax form:

$$\min_{\mathbf{x}\in\mathbb{R}^p}\max_{\lambda\in\mathbb{R}^n}\{\mathcal{L}(\mathbf{x},\lambda) := f(\mathbf{x}) + \langle\lambda, \mathbf{A}\mathbf{x}-\mathbf{b}\rangle\}$$

Theorem (Necessary and sufficient optimality condition)

Under the Slater's condition: $\mathrm{relint}(\mathrm{dom}\, f)\cap\{\mathbf{x} \,:\, \mathbf{A}\mathbf{x}=\mathbf{b}\}\neq\emptyset$, *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 $(\mathbf{x}^\star,\lambda^\star)\in\mathbb{R}^p\times\mathbb{R}^n$ *being an optimal solution for the primal problem and dual problem:*

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

- By definition of $f^\star$ and $d^\star$, we always have $\boxed{d^\star \leq f^\star}$ (**weak duality**).
- If a primal solution exists and the Slater's condition holds, 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 $\mathrm{relint}(\mathrm{dom}\, f)$ the relative interior of the domain. The Slater condition requires

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

Special cases

- If $\mathrm{dom}\, f = \mathbb{R}^p$, then (9) $\Leftrightarrow$ $\boxed{\exists \bar{\mathbf{x}} \ : \ \mathbf{A}\bar{\mathbf{x}} = \mathbf{b}}$.
- If $\mathrm{dom}\, f = \mathbb{R}^p$ and instead of $\mathbf{A}\mathbf{x} = \mathbf{b}$, we have the feasible set $\{\mathbf{x} : h(\mathbf{x}) \leq 0\}$, 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

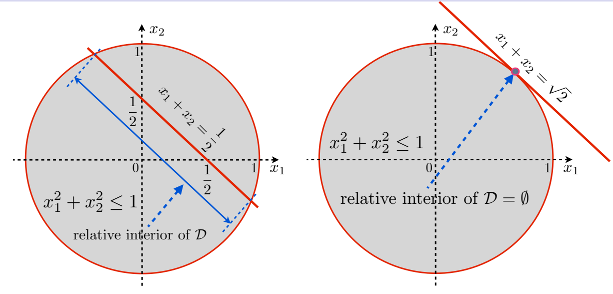
Example

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

Example: Nonsmoothness of the dual function

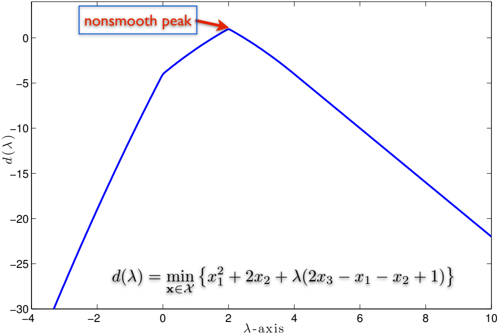
Consider a constrained convex problem:

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

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.



Dual subgradient method

Recall the dual problem:

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

Subgradient ascent method can be applied to solve it.

Dual subgradient method

Recall the dual problem:

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

Subgradient ascent method can be applied to solve it.

A plausible algorithmic strategy for $\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b} \right\}$:

A natural minimax formulation:

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

Lagrangian subproblem: $\mathbf{x}^*(\lambda) \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \mathcal{L}(\mathbf{x}, \lambda)$

Dual problem: $\lambda^\star \in \arg\max_{\lambda\in\mathbb{R}^n} \left\{ d(\lambda) := \mathcal{L}(\mathbf{x}^*(\lambda), \lambda) \right\}$

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

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

- Conceptual, since we do not have exact solution $\lambda^\star$

Dual subgradient method

Properties of dual function

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

$$\mathbf{x}^*(\lambda) := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \right\}.$$

Dual subgradient method (DSGM):

1. Choose $\lambda_0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \cdots$, perform:

   2.a. $\mathbf{x}^*(\lambda_k) := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \mathcal{L}(\mathbf{x}, \lambda_k) := f(\mathbf{x}) + \langle \lambda_k, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \right\}$.

   2.b. Compute the subgradient $\nabla d(\lambda_k) := \mathbf{A}\mathbf{x}^*(\lambda_k) - \mathbf{b}$.

   2.c. Update $\boxed{\lambda_{k+1} := \lambda_k + \frac{R}{\sqrt{k+1}} \nabla d(\lambda_k)}$, where R is a given constant.

Convergence of DSGM

Well-definedness

- Problem below may not have solution $\mathbf{x}^*(\lambda)$ for any λ. Then DSGM is not well-defined except if $f(\mathbf{x}) = f_1(\mathbf{x}) + \delta_{\mathcal{X}}(x)$ and $\mathcal{X}$ is bounded.

$$\mathbf{x}^*(\lambda) := \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \mathcal{L}(\mathbf{x}, \lambda) := f_1(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \right\}.$$

- 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 $\mathbf{x}^*(\lambda)$ for any λ. Then DSGM is not well-defined except if $f(\mathbf{x}) = f_1(\mathbf{x}) + \delta_{\mathcal{X}}(x)$ and $\mathcal{X}$ is bounded.

$$\mathbf{x}^*(\lambda) := \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \mathcal{L}(\mathbf{x},\lambda) := f_1(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle \right\}.$$

- Impractical to evaluate $R_\star := \|\lambda_0 - \lambda^\star\|_2$, use an upper bound R of $R_\star$.

Theorem (Convergence)

Assume that $\|\mathbf{A}\mathbf{x}^*(\lambda_k) - \mathbf{b}\| \leq M_d$ 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 $\mathcal{O}(1/\sqrt{k})$, instead of $\mathcal{O}(1/k)$ of accelerated linearized quadratic penalty.

- Approximate solution for primal via averaging: $\mathbf{x}^\epsilon = \frac{1}{k}\sum_{i=0}^{k-1} \mathbf{x}^*(\lambda_i)$ [33]

Outline

Primal approach: Penalization

Dual approach: Lagrangian-based method

Primal-dual approach: Augmented Lagrangian

Combining Lagrangian and penalty approaches

- Quadratic penalty approach:

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

- Lagrangian approach:

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle.$$

▷ Combine to get *augmented Lagrangian* (AL):

$$\mathcal{L}_\mu(\mathbf{x}, \lambda) = f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \frac{\mu}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2.$$

Properties of augmented Lagrangian

- Corresponding dual function is concave and $\frac{1}{\mu}$-*smooth*:

$$d_\mu(\lambda) = \min_{\mathbf{x} \in \mathbb{R}^p} \left\{ f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \frac{\mu}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}.$$

Can apply gradient or accelerated gradient methods in the dual!

- μ does not need to increase until infinity.

No more ill-conditioned subproblems!

Example: Behavior of the augmented Lagrangian dual function

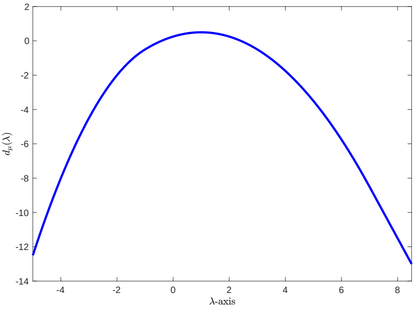
Consider a constrained convex problem:

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

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:

$$d^\star := \max_{\lambda \in \mathbb{R}^n} \left\{ d(\lambda) = \min_{\mathbf{x} \in \mathbb{R}^p} f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \right\}. \tag{10}$$

Augmented dual problem:

$$d^*_\mu := \max_{\lambda \in \mathbb{R}^n} \left\{ d_\mu(\lambda) = \min_{\mathbf{x} \in \mathbb{R}^p} f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \frac{\mu}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}, \quad \mu > 0. \tag{11}$$

Augmented dual problem

Dual problem:

$$d^\star := \max_{\lambda \in \mathbb{R}^n} \left\{ d(\lambda) = \min_{\mathbf{x} \in \mathbb{R}^p} f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \right\}. \tag{10}$$

Augmented dual problem:

$$d^*_\mu := \max_{\lambda \in \mathbb{R}^n} \left\{ d_\mu(\lambda) = \min_{\mathbf{x} \in \mathbb{R}^p} f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \frac{\mu}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}, \quad \mu > 0. \tag{11}$$

Relation between augmented dual problem and dual problem

If a primal solution exists and Slater's condition holds, we have

- The dual solution set of (11) coincides with the one of the dual problem (10).
- $f^\star = d^\star = d^*_\mu$ for any $\mu > 0$.

Recall: The augmented dual problem (11) is smooth and concave $\Rightarrow$ **Gradient and accelerated gradient methods** can be applied to solve it.

Augmented Lagrangian method: Conceptual

$$d_\mu(\lambda) = \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \frac{\mu}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\} \quad (12)$$

$$\mathbf{x}^*_\mu(\lambda) = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \frac{\mu}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}$$

Augmented Lagrangian method (ALM):

1. Choose $\lambda_0 \in \mathbb{R}^n$ and $\mu > 0$.
2. For $k = 0, 1, \cdots$, perform:
   - **2.a**. Solve (12) to compute $\nabla d_\mu(\lambda_k) := \mathbf{A}\mathbf{x}^*_\mu(\lambda_k) - \mathbf{b}$.
   - **2.b**. Update $\lambda_{k+1} := \lambda_k + \mu\nabla d_\mu(\lambda_k)$.

Augmented Lagrangian method: Conceptual

$$d_\mu(\lambda) = \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle + \frac{\mu}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\} \quad (12)$$

$$\mathbf{x}^*_\mu(\lambda) = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle + \frac{\mu}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}$$

Augmented Lagrangian method (ALM):

1. Choose $\lambda_0 \in \mathbb{R}^n$ and $\mu > 0$.
2. For $k = 0, 1, \cdots,$ perform:
   - **2.a**. Solve (12) to compute $\nabla d_\mu(\lambda_k) := \mathbf{A}\mathbf{x}^*_\mu(\lambda_k) - \mathbf{b}$.
   - **2.b**. Update $\lambda_{k+1} := \lambda_k + \mu\nabla d_\mu(\lambda_k)$.

ALM can be accelerated by Nesterov's optimal method.

Accelerated augmented Lagrangian method (AALM)

1. Choose $\lambda_0 \in \mathbb{R}^n$ and $\mu > 0$. Set $\tilde{\lambda}_0 := \lambda_0$ and $t_0 := 1$
2. For $k = 0, 1, \cdots,$ perform:
   - **2.a**. Solve (12) to compute $\nabla d_\mu(\tilde{\lambda}_k) := \mathbf{A}\mathbf{x}^*_\mu(\tilde{\lambda}_k) - \mathbf{b}$.
   - **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 AALM

Theorem (Convergence [20])

- *Let* $\{\lambda_k\}$ *be the sequence generated by ALM. Then*

$$\boxed{d^\star - d_\mu(\lambda_k) \leq \frac{\|\lambda_0 - \lambda^\star\|_2^2}{2\mu(k+1)}.}$$

- *Let* $\{\lambda_k\}$ *be the sequence generated by AALM. Then*

$$\boxed{d^\star - d_\mu(\lambda_k) \leq \frac{2\|\lambda_0 - \lambda^\star\|_2^2}{\mu(k+1)^2}.}$$

- Guarantees are given for the dual problem and not for the primal!
- Approximate solution for primal via averaging: $\mathbf{x}^\epsilon = \frac{1}{k}\sum_{i=0}^{k-1} \mathbf{x}_\mu^*(\lambda_i)$ [33]

Drawbacks and enhancements

At each step, ALM solves

$$\mathbf{x}^*_\mu(\lambda) := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \mathcal{L}_\mu(\mathbf{x},\lambda) := f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle + \frac{\mu}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 \right\}. \quad (13)$$

Drawbacks

1. Drawback 1: The quadratic term $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2$ in (13) destroys the separability as well as the tractable proximity of f.
2. Drawback 2: Solving (13) exactly is impractical.

Drawbacks and enhancements

At each step, ALM solves

$$\mathbf{x}^*_\mu(\lambda) := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \mathcal{L}_\mu(\mathbf{x},\lambda) := f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle + \frac{\mu}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 \right\}. \tag{13}$$

Drawbacks

1. Drawback 1: The quadratic term $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2$ in (13) destroys the separability as well as the tractable proximity of f.
2. Drawback 2: Solving (13) exactly is impractical.

Enhancements

1. Process the quadratic term $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2$ by linearization.
2. Allow inexactness of solving (13), while guaranteeing the same convergence rate.

Going back to primal: Linearized Augmented Lagrangian method

- Linearization idea from Slide 19: Majorize the augmented Lagrangian

$$\mathbf{x}_{k+1} := \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle + \frac{\mu}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_k\|^2_{\mathbf{Q}_k} \right\}.$$

- When $\mathbf{Q}_k = \alpha_k \mathbf{I} - \mu \mathbf{A}^\top \mathbf{A} \geq 0$ with $\alpha_k \geq \mu\|\mathbf{A}\|^2$ (same calculation as in Slide 19):

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

- We pick $\alpha_k = \mu\|\mathbf{A}\|^2$.

Linearized augmented Lagrangian method (LALM)

1. Choose $\mathbf{x}_0 \in \mathbb{R}^p$, $\lambda_0 \in \mathbb{R}^n$ and $\mu > 0$.
2. For $k = 0, 1, \cdots$, perform:
   2.a. Update

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

Convergence of Linearized ALM

Theorem (Convergence [32])

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

$$\|\mathbf{A}\bar{\mathbf{x}}_k - \mathbf{b}\| \leq \frac{1}{k}\left(\frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \frac{\max\left\{(1+\|\lambda^\star\|)^2, 4\|\lambda^\star\|^2\right\}}{\mu}\right)$$

$$|f(\bar{\mathbf{x}}_k) - f(\mathbf{x}^\star)| \leq \frac{1}{k}\left(\frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \frac{\max\left\{(1+\|\lambda^\star\|)^2, 4\|\lambda^\star\|^2\right\}}{\mu}\right)$$

- Guarantees are for the primal and in fact optimal [23].
- No need to solve difficult subproblems at each iteration.
- Guarantees are for $\bar{\mathbf{x}}_k$, and not $\mathbf{x}_k$.

Alternative approach for subproblems of ALM

- Primal subproblem:

$$\mathbf{x}^*_\mu(\lambda) := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \mathcal{L}_\mu(\mathbf{x},\lambda) := f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x}-\mathbf{b}\rangle + \frac{\mu}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 \right\}. \qquad (14)$$

- This is a composite optimization problem.

Accelerated proximal methods (e.g. FISTA) can be used to solve this up to some accuracy.

Conceptual inexact augmented Lagrangian method:
1. Choose $\lambda_0 \in \mathbb{R}^n$, $\mu > 0$ and a decreasing nonnegative sequence ϵ_k.
2. For $k = 0, 1, \cdots$, perform:
2.a. Solve (14) with FISTA until $\mathcal{L}_\mu(\mathbf{x}^{\epsilon_k}_\mu(\lambda_k), \lambda_k) \leq \mathcal{L}_\mu(\mathbf{x}^*_\mu(\lambda_k), \lambda_k) + \epsilon_k$.
2.b. Update $\lambda_{k+1} := \lambda_k + \mu(\mathbf{A}\mathbf{x}^{\epsilon_k}_\mu(\lambda_k) - \mathbf{b})$.

- Conceptual since $\mathbf{x}^*_\mu(\lambda_k)$ is unknown.

▷ Solve (14) for increasing (explicit) number of iterations $m_k > 0$.

*An explicit inexact ALM

Inexact ALM (Double Loop ASGARD [28])

1. $\mathbf{x}_0 = \hat{x}_{0,0} = \bar{x}_{0,0} = \tilde{x}_{0,0} \in \mathbb{R}^p$, $\lambda_0 \in \mathbb{R}^n$. Set $\mu_k > 0$, $\tau_0 = 1$, $m_0 > 2$.
2. For $k = 0, 1, \cdots$, perform:

   2.a For $i = 0, 1, \cdots, m_k - 1$, perform (accelerated proximal method):

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

   2.b Restart primal and dual variable updates

$$
\begin{cases}
\bar{\mathbf{x}}_{k+1,0} & = \tilde{\mathbf{x}}_{k,m_k} \\
\lambda_{k+1} & = \lambda_k + \mu_k (A\bar{\mathbf{x}}_{k+1,0} - \mathbf{b}), \\
\tau_0 & = 1 \\
\mu_{k+1} & = \mu_k \omega, \\
m_{k+1} & = m_k \omega,
\end{cases}
$$

dual variable update

μ_k needs to increase now

number of inner iterations increase

- Corresponds to inexact ALM with explicit inner termination rule.
- We can prove optimal $\mathcal{O}(1/k)$ on the last iterate.

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

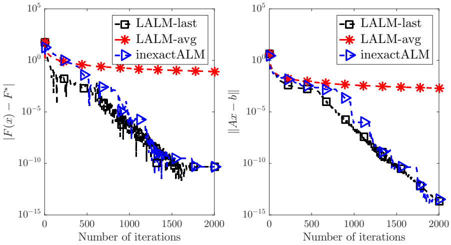
$$F^\star := \min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \mathbf{A}\mathbf{x} = \mathbf{b} \right\}.$$

Data generation

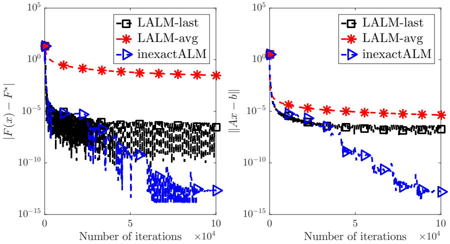
- $\mathbf{A}$ is a row-normalized standard Gaussian matrix.
- $\mathbf{x}^\star$ is a k-sparse vector generated randomly.
- Noiseless case: $\mathbf{b} := \mathbf{A}\mathbf{x}^\star$.
- Noisy case: $\mathbf{b} := \mathbf{A}\mathbf{x}^\star + \mathcal{N}(0, 10^{-3})$.

Example: Last iterate vs average iterate of LALM

- Noiseless case.



- Noisy case.



References I

[1] Intuition behind lagrange duality.
`https://masszhou.github.io/2016/09/10/Lagrange-Duality/`.
Accessed: 2019-11-24.

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

[3] Dimitri P Bertsekas.
Necessary and sufficient conditions for a penalty method to be exact.
Mathematical programming, 9(1):87–99, 1975.

[4] Dimitri P Bertsekas.
On penalty and multiplier methods for constrained minimization.
SIAM Journal on Control and Optimization, 14(2):216–235, 1976.

[5] 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.

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

References II

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

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

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

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

[11] 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.

References III

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

References IV

[17] 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.

[18] Magnus R Hestenes.
Multiplier and gradient methods.
Journal of optimization theory and applications, 4(5):303–320, 1969.

[19] 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.

[20] V. Nedelcu, I. Necoara, and Q. Tran-Dinh.
Computational Complexity of Inexact Gradient Augmented Lagrangian Methods: Application to Constrained MPC.
SIAM J. Optim. Control, 52(5):3109–3134, 2014.

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

References V

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

[23] Yuyuan Ouyang and Yangyang Xu.
Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems.
arXiv preprint arXiv:1808.02901, 2018.

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

[25] Michael JD Powell.
A method for nonlinear constraints in minimization problems.
Optimization, pages 283–298, 1969.

[26] 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.

References VI

[27] 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.

[28] 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.

[29] 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.

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

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

References VII

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

[33] A. Yurtsever, Q. Tran-Dinh, and V. Cevher.
Universal primal-dual proximal-gradient methods.
Tech. Report. (LIONS, EPFL), Available at: http://arxiv.org/pdf/1502.03123.pdf., 2015.

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