A Unified Primal-Dual Algorithm Framework for Inequality Constrained Problems

Zaiwen Wen

Beijing International Center For Mathematical Research
Peking University

Joint work with Zhenyuan Zhu, Fan Chen, Junyu Zhang
https://arxiv.org/abs/2208.14196

Introduction

Consider the convex composite optimization problem:

$$\begin{aligned} \min_x \ & \Phi(x) = f(x) + h(x), \\ \text{s.t.} \ & Ax - b \in \mathcal{K}. \end{aligned} \tag{P}$$

- $f(x) : \mathbb{R}^n \mapsto \mathbb{R}$ is a differentiable convex function whose gradient $\nabla f(\cdot)$ is L_f-Lipschitz continuous
- $h(x) : \mathbb{R}^n \mapsto \mathbb{R}$ is a simple convex function whose proximal operator can be efficiently evaluated:

$$\mathbf{prox}_h(x) := \arg\min_u \{h(u) + \frac{1}{2}\|x - u\|^2\}$$

- $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $\mathcal{K} \subset \mathbb{R}^m$ is either $\{0\}$ or a proper cone

Applications

- Compressive sensing
  - basis pursuit
  - LASSO
- Image processing
  - TV denoising
  - TVL1 denoising
  - globally convex segmentation
- Statistics and machine learning
  - latent variable Gaussian graphical model selection
  - robust principal component analysis
  - support vector machine

Outline

Saddle point formulation

- Saddle point problem based on the Lagrangian function:

$$\min_{x} \max_{y \in \mathcal{K}^*} \; \mathcal{L}(x, y) := f(x) + h(x) - y^{\mathrm{T}}(Ax - b), \qquad \text{(SP-L)}$$

  where $\mathcal{K}^*$ is the dual cone of $\mathcal{K}$, i.e., $\mathcal{K}^* = \{x | \langle x, y \rangle \geq 0, \forall y \in \mathcal{K}\}$.

- Denote $\Psi(x, y) = f(x) - y^{\mathrm{T}}(Ax - b)$, $s(y) = \mathbb{1}_{\mathcal{K}^*}(y)$.
- (SP-L) can be rewritten as

$$\min_{x} \max_{y} \; h(x) + \Psi(x, y) - s(y).$$

PDHG and CP

- Let $g_x^k = \nabla_x \Psi(x^k, y^k), g_y^k = \nabla_y \Psi(x^k, y^k) = -Ax^k + b$.
- One primal gradient-type step + one dual gradient-type step
- **Primal-dual hybrid gradient (PDHG):** "Gauss-Seidel iteration"

$$x^{k+1} = \mathbf{prox}_{\tau h}\left[x^k - \tau g_x^k\right],$$
$$y^{k+1} = \mathbf{prox}_{\sigma s}\left[y^k + \sigma g_y^{k+1}\right].$$

- **The Chambolle-Pock method (CP):** "Gauss-Seidel iteration" with an extrapolation step on the dual variable

$$x^{k+1} = \mathbf{prox}_{\tau h}\left[x^k - \tau g_x^k\right],$$
$$y^{k+1} = \mathbf{prox}_{\sigma s}\left[y^k + \sigma\left(2g_y^{k+1} - g_y^k\right)\right].$$

GDA and OGDA

- **Gradient descent ascent (GDA):** "Jacobian iteration"

$$x^{k+1} = \mathbf{prox}_{\tau h}\left[x^k - \tau g_x^k\right],$$
$$y^{k+1} = \mathbf{prox}_{\sigma s}\left[y^k + \sigma g_y^k\right].$$

- **Optimistic gradient descent ascent (OGDA):** "Jacobian iteration" with extrapolation steps on both primal and dual variables

$$x^{k+1} = \mathbf{prox}_{\tau h}\left[x^k - \tau\left(2g_x^k - g_x^{k-1}\right)\right],$$
$$y^{k+1} = \mathbf{prox}_{\sigma s}\left[y^k + \sigma\left(2g_y^k - g_y^{k-1}\right)\right].$$

Motivations

- Primal-dual algorithms: PDHG, CP, OGDA ...
  - Existing ergodic convergence results are almost established on the duality gap of the Lagrangian function: for **bounded** $\mathcal{X}$ and $\mathcal{Y}$,

$$\mathrm{DualGap}(\bar{x}_N, \bar{y}_N) := \max_{y \in \mathcal{Y}} \mathcal{L}(\bar{x}_N, y) - \min_{x \in \mathcal{X}} \mathcal{L}(x, \bar{y}_N) \sim \mathcal{O}(1/N),$$

    where $\bar{x}_N = \sum_{k=1}^N x_k/N$, $\bar{y}_N = \sum_{k=1}^N y_k/N$.
  - If $\mathcal{Y}$ is not bounded and $A\bar{x}_N - b \notin \mathcal{K}$: $\mathrm{DualGap}(\bar{x}_N, \bar{y}_N) = +\infty$.
- Dual ascent class algorithms: ALM, ADMM ...
  - need to solve subproblems
  - multi-block ADMM may not necessarily converge

We aim to

1. design a class of easy-to-implement algorithms with good convergence properties
2. give an error bound w.r.t constraint violation and function value gap without the boundedness assumption of $\mathcal{Y}$

Outline

Assumptions

1. **The optimal solution of (P) is attainable.** There exists $x^* \in \mathbb{R}^n$ such that $Ax^* - b \in \mathcal{K}$ and $\Phi(x^*)$ equals to the optimal value Φ^*.
2. **Slater's condition:**
   - There exists $x \in \mathrm{relint}\ \mathcal{D}$ such that $Ax - b \in \mathrm{int}\ \mathcal{K}$, where $\mathcal{D} = \mathrm{dom}\ \Phi$, $\mathrm{relint}\ \mathcal{D}$ denotes the relative interior of $\mathcal{D}$ and $\mathrm{int}\ \mathcal{K}$ denotes the interior of $\mathcal{K}$.
   - When $\mathcal{K}$ is a polyhedral cone (including the case of $\mathcal{K} = \{0\}$), the condition can be relaxed to the existence of $x \in \mathrm{relint}\ \mathcal{D}$ such that $Ax - b \in \mathcal{K}$.

Slater's condition guarantees (P) is equivalent to (SP-L).

Augmented Lagrangian duality

- Generalize the formulation to the *augmented* Lagrangian function:

Lemma

Suppose that $\mathcal{K}$ is $\{0\}$ or a proper cone. Given any penalty coefficient $\rho > 0$, we define the augmented Lagrangian function as

$$\mathcal{L}_\rho(x, y) := f(x) + h(x) + \frac{\rho}{2} \left\| \mathcal{P}_{\mathcal{K}^\circ} \left(Ax - b - \frac{y}{\rho} \right) \right\|^2 - \frac{\|y\|^2}{2\rho},$$

where $\mathcal{K}^\circ = -\mathcal{K}^*$. The strong duality holds for $\mathcal{L}_\rho(x, y)$, that is,

$$\min_x \max_y \; \mathcal{L}_\rho(x, y) = \max_y \min_x \; \mathcal{L}_\rho(x, y), \tag{SP-AL}$$

where both sides are equivalent to (P).

Proof: case of $\mathcal{K} = \{0\}$

- The augmented Lagrangian function degenerates into

$$\mathcal{L}_\rho(x, y) = f(x) + g(x) - y^{\mathrm{T}}(Ax - b) + \frac{\rho}{2}\|Ax - b\|^2.$$

- Consider the following equivalent problem:

$$\begin{aligned} \min_x \ & f(x) + g(x) + \frac{\rho}{2}\|Ax - b\|^2, \\ \text{s.t. } & Ax = b, \end{aligned}$$

whose Lagrangian function is $\mathcal{L}_\rho(x, y)$.

- By Slater's condition, the strong duality holds, that is,

$$\min_x \max_y \ \mathcal{L}_\rho(x, y) = \max_y \min_x \ \mathcal{L}_\rho(x, y).$$

Why *augmented* Lagrangian duality?

$$\text{(SP-L): } \min_x \max_{y \in \mathcal{K}^*} \mathcal{L}(x,y) \qquad \text{(SP-AL): } \min_x \max_y \mathcal{L}_\rho(x,y)$$

- **Make the constraint $y \in \mathcal{K}^*$ optional.** Removing the constraint improves the flexibility of algorithm design since it may cause difficulty in solving subproblems.
- **Make the framework more versatile.** The framework based on (SP-AL) can cover a wider range of algorithms, e.g. linearized ALM.
- **Make the objective function have better convexity.** $\mathcal{L}_\rho(x,y)$ is a strongly convex function along at least one direction of x, which can bring the benefits of convergence.

A unified primal-dual algorithm framework

Define

$$s(y) = \begin{cases} 0, & \rho > 0, \\ \mathbb{1}_{\mathcal{K}^*}(y), & \rho = 0, \end{cases}$$

and

$$\Psi(x, y) = \begin{cases} f(x) + \frac{\rho}{2} \left\| \mathcal{P}_{\mathcal{K}^\circ} \left(Ax - b - \frac{y}{\rho} \right) \right\|^2 - \frac{\|y\|^2}{2\rho}, & \rho > 0, \\ f(x) - y^{\mathrm{T}}(Ax - b), & \rho = 0. \end{cases}$$

We can rewrite both problem (SP-L) and (SP-AL) as

$$\min_x \max_y \; \mathcal{L}_\rho(x, y) := h(x) + \Psi(x, y) - s(y). \tag{SP}$$

A unified primal-dual algorithm framework

$$
\begin{aligned}
x^{k+1} &= \mathbf{prox}_{\tau h}\left[x^k - \boldsymbol{\tau}\left((1+\alpha)g_x^k - \alpha g_x^{k-1}\right)\right] \\
y^{k+1} &= \mathbf{prox}_{\sigma s}\Big[y^k + \boldsymbol{\sigma}\mu\left((1+\beta)g_y^k - \beta g_y^{k-1}\right) \\
&\qquad + \boldsymbol{\sigma}(1-\mu)\left((1+\beta)g_y^{k+1} - \beta g_y^k\right)\Big]
\end{aligned}
\tag{PD}
$$

- $g_x^k = \nabla_x \Psi(x^k, y^k), g_y^k = \nabla_y \Psi(x^k, y^k)$
- $\boldsymbol{\tau}, \boldsymbol{\sigma} > 0$: primal and dual step sizes
- $\alpha \in [0,1], \beta \geq 0$: gradient extrapolation coefficients
- $\mu \in [0,1]$: the ratio of "Gauss-Seidel iteration" versus "Jacobian iteration"

A unified primal-dual algorithm framework

- When $\rho = 0$, $g_y^{k+1} = -Ax^{k+1} - b$, the scheme is an explicit update.
- When $\rho > 0$, $g_y^{k+1} = -(Ax^{k+1} - b) + \mathcal{P}_{\mathcal{K}}(w^{k+1})$ where $w^{k+1} = Ax^{k+1} - b - \frac{\boldsymbol{y^{k+1}}}{\rho}$. The y^{k+1} update is an implicit scheme.

Lemma (Explicit form of dual update)

Let $\rho > 0$ and $s(y) = 0$, the dual update rule of (PD) *can be written as*

$$y^{k+1} = \omega + \frac{\kappa}{\kappa + 1}\mathcal{P}_{\mathcal{K}}\left(\nu - \omega\right),$$

where $\omega =$ $y^k + \sigma\mu\left((1+\beta)g_y^k - \beta g_y^{k-1}\right) - \sigma(1-\mu)\left((1+\beta)(Ax^{k+1} - b) + \beta g_y^k\right)$, $\kappa = \sigma(1-\mu)(1+\beta)/\rho \geq 0$, *and* $\nu = \rho(Ax^{k+1} - b)$.

Consequences of the unified framework

- Well-known algorithms:
  - PDHG: $\mu = 0, \alpha = 0, \beta = 0, \rho \geq 0$
  - CP: $\mu = 0, \alpha = 0, \beta = 1, \rho \geq 0$
  - GDA: $\mu = 1, \alpha = 0, \beta = 0, \rho \geq 0$
  - OGDA: $\mu = 1, \alpha = 1, \beta = 1, \rho \geq 0$
  - Linearized ALM: $\mu = 0, \alpha = 0, \beta = 0, \rho > 0$
- New algorithms: e.g. SOGDA: $\mu = 1, \alpha = 0, \beta = 1, \rho \geq 0$

$$
\begin{aligned}
x^{k+1} &= \mathbf{prox}_{\tau h}\left[x^k - \tau g_x^k\right], \\
y^{k+1} &= \mathbf{prox}_{\sigma s}\left[y^k + \sigma\left(2g_y^k - g_y^{k-1}\right)\right].
\end{aligned}
$$

Interpertation: Jacobian-type of CP based on AL
or OGDA with only dual variables extrapolated

Outline

Ergodic convergence: affine equality constrained problem

We first consider the case of $\mathcal{K} = \{0\}$.

- In this case, we define the weight c as

$$c = C^{\text{affine}}_{\alpha,\beta,\mu}(\tau,\sigma,\rho) := \alpha\tau L_{f_\rho} + \max\{|\mu\beta|\sqrt{\sigma\tau}\,\|A\|\,, \alpha\sqrt{\sigma\tau}\,\|A\|\},$$

  where $f_\rho(x) := f(x) + \frac{\rho}{2}\|Ax - b\|^2$ and $L_{f_\rho} := L_f + \rho\|A\|^2$ is the Lipschitz constant of $\nabla f_\rho(x)$.

- Given any coefficient c, we define the matrix P_c as

$$P_c := \begin{bmatrix} \rho I_m & 0_{m\times n} & \frac{1-\alpha-\beta+\mu}{2} I_m \\ 0_{n\times m} & \left(\frac{1-2c}{2\tau} - \frac{(1-\alpha)L_{f_\rho}}{2}\right) I_n & \frac{\beta-\mu}{2} A^{\mathrm{T}} \\ \frac{1-\alpha-\beta+\mu}{2} I_m & \frac{\beta-\mu}{2} A & \frac{1-2c}{2\sigma} I_m \end{bmatrix}.$$

Ergodic convergence: affine equality constrained problem

Theorem

If the parameters $\tau, \sigma, \rho, \alpha, \beta$ *and* μ *are properly chosen so that* $\boldsymbol{P_c} \succeq \mathbf{0}$, *then for* $\forall N \geq 1$ *and* $\forall \gamma \geq 0$, *we have*

$$\Phi(\bar{x}_N) - \Phi(x^*) + \gamma \|A\bar{x}_N - b\| \leq \frac{1}{N} \left(\frac{\left\|x^0 - x^*\right\|^2}{\tau} + \frac{\gamma^2}{\sigma} \right),$$

where $\bar{x}_N = \frac{1}{N} \sum_{k=1}^{N} x^k$. *Moreover, it holds that*

$$|\Phi(\bar{x}_N) - \Phi(x^*)| \leq \frac{4}{N} \left(\frac{\left\|x^0 - x^*\right\|^2}{\tau} + \frac{\|y^*\|^2}{\sigma} \right),$$

$$\|A\bar{x}_N - b\| \leq \frac{3}{N} \left(\frac{\left\|x^0 - x^*\right\|}{\sqrt{\tau\sigma}} + \frac{\|y^*\|}{\sigma} \right).$$

Proof sketch

- Denote $z = [x; y]$, define $\Lambda = \mathrm{diag}(\tau I_n, \sigma I_m)$, $\Xi = \mathrm{diag}(I_n, \mu I_m)$, $\Theta = \mathrm{diag}(\alpha I_n, \beta I_m)$, $F(z) = [\nabla_x \Psi(x,y); -\nabla_y \Psi(x,y)]$.
- Define the discrete Lyapunov function:

$$\Delta_k(z) := \frac{1}{2}\|z^k - z\|^2_{\Lambda^{-1}} + \frac{c}{2}\|z^k - z^{k-1}\|^2_{\Lambda^{-1}} + \langle F(z^k) - F(z^{k-1}), z - z^k \rangle_{\Xi\Theta} + (\mu - \beta)\left\langle \nabla_y \Psi(z^k), y^k - y \right\rangle,$$

- Fix $x = x^*$ and denote $\tilde{z} = [x^*, y]$ for arbitrary y. Due to $P_c \succeq 0$, we obtain

(one-step descent): $\Delta_k(\tilde{z}) - \Delta_{k+1}(\tilde{z}) \geq \Phi(x^{k+1}) - \Phi(x^*) - \langle Ax^{k+1} - b, y \rangle$,

(upper and lower bound): $0 \leq \Delta_k(\tilde{z}) \leq \left\|z^k - \tilde{z}\right\|^2_{\Lambda^{-1}} + c\left\|z^k - z^{k-1}\right\|^2_{\Lambda^{-1}}$.

- Let $\hat{y} = \gamma(A\bar{x}_N - b)/\|A\bar{x}_N - b\|$ and $\hat{z} = [x^*, \hat{y}]$, combining the convexity yields

$$\begin{aligned}\Phi(\bar{x}_N) - \Phi(x^*) + \gamma\|A\bar{x}_N - b\| &= \Phi(\bar{x}_N) - \Phi(x^*) - \langle A\bar{x}_N - b, \hat{y} \rangle \\ &\leq \frac{1}{N}\sum_{k=0}^{N-1}\left(\Phi(x^{k+1}) - \Phi(x^*) - \langle A\bar{x}^{k+1} - b, \hat{y} \rangle\right) \\ &\leq \frac{\Delta_0(\hat{z}) - \Delta_N(\hat{z})}{N} \leq \frac{1}{N}\left(\frac{\|x^0 - x^*\|^2}{\tau} + \frac{\gamma^2}{\sigma}\right).\end{aligned}$$

Step size conditions

- **SOGDA** Set $\mu = 1, \alpha = 0, \beta = 1$. For any $\rho > 0$, $P_c \succeq 0$ is guaranteed if

$$2\sqrt{\sigma\tau}\,\|A\| + \max\left(\frac{\sigma}{2\rho}, \tau L_{f_\rho}\right) \leq 1.$$

  When $\rho = 0$, SOGDA has no convergence guarantee, which is also observed numerically.

- **PDHG** Set $\mu = 0, \alpha = 0, \beta = 0$. For any $\rho > 0$, $P_c \succeq 0$ is guaranteed if

$$\sigma \leq 2\rho, \quad \frac{1}{\tau} \geq L_f + \rho\,\|A\|^2.$$

  When $\rho = 0$, PDHG is potentially non-convergent. When $\rho > 0$, the algorithm becomes the linearized ALM, which is proved to be convergent.

Step size conditions

- **CP** Set $\mu = 0, \alpha = 0, \beta = 1$. For any $\rho \geq 0$, $P_c \succeq 0$ is guaranteed if
$$\frac{1}{\tau} \geq L_f + (\rho + \sigma) \|A\|^2 .$$
- **GDA** Set $\mu = 1, \alpha = 0, \beta = 0$. For any $\rho > 0$, $P_c \succeq 0$ is guaranteed if
$$\sigma < \frac{\rho}{2}, \quad \frac{1}{\tau} \geq L_f + \rho \|A\|^2 \frac{\rho - \sigma}{\rho - 2\sigma}.$$
- **OGDA** Set $\mu = 1, \alpha = 1, \beta = 1$. For any $\rho \geq 0$, $P_c \succeq 0$ is guaranteed if
$$\tau L_{f_\rho} + \sqrt{\sigma\tau} \|A\| \leq \frac{1}{2}.$$

Remark: The above analysis is also feasible to the case of general cone with $\rho = 0$. We only need to set $s(y) = \mathbb{1}_{\mathcal{K}^*}(y)$.

Next, we only need to consider general problems with $\rho > 0$.

- In this case, we define

$$c = C^{\mathrm{conic}}_{\alpha,\beta,\mu}(\tau,\sigma,\rho) := \max\left\{\alpha\tau L_{f_\rho}, |\mu\beta|\,\frac{\sigma}{\rho}\right\} + \max\left\{\alpha, |\mu\beta|\right\}\|A\|\sqrt{\sigma\tau}.$$

- Define $\gamma_y = (\mu-\beta)^2 + (1+\alpha)|\mu-\beta| + 4(\mu-\beta)$, $\gamma_w = t\big(2-2\alpha, (1-\alpha)^2 + (1+\alpha)|\mu-\beta|\big)$ where the function $t(\cdot,\cdot)$ is given by

$$t(a,b) := \begin{cases} b + \frac{(a-b)^2}{2a-b}, & a > b, \\ b, & a \le b. \end{cases}$$

- Then we define the matrix P'_c for any given $c > 0$ as

$$P'_c = \begin{bmatrix} \left(\frac{1-2c}{2\tau} - \frac{(1-\alpha)L_f}{2}\right) I_n - \frac{\rho\gamma_w}{4} A^{\mathrm{T}} A & \frac{\gamma_w}{4} A^{\mathrm{T}} \\ \frac{\gamma_w}{4} A & \left(\frac{1-2c}{2\sigma} - \frac{\gamma_w+\gamma_y}{4\rho}\right) I_m \end{bmatrix}.$$

Ergodic convergence: conic inequality constrained problem

Theorem

If the parameters $\tau, \sigma, \rho, \alpha, \beta$ *and* μ *are properly chosen such that* $\boldsymbol{P_c'} \succeq \mathbf{0}$ *and* $\boldsymbol{P_c'} + \boldsymbol{c\Lambda}^{-1} \succ \mathbf{0}$. *Then for* $\bar{x}_N = \frac{1}{N}\sum_{k=1}^{N} x^k$, *it holds that*

$$|\Phi(\bar{x}_N) - \Phi(x^*)| \leq \mathcal{O}\left(\frac{1}{N}\right), \qquad \|\mathcal{P}_{\mathcal{K}^\circ}(A\bar{x}_N - b)\| \leq \mathcal{O}\left(\frac{1}{N}\right),$$

where $\mathcal{O}(\cdot)$ *hides constants that depend on* $\left\|x^* - x^0\right\|, \|y^*\|$ *and the parameters.*

The proof is similar to the case of affine equality constrained problem.

Step size conditions

- **SOGDA** Set $\mu = 1, \alpha = 0, \beta = 1$. For any $\rho > 0$, $P_c' \succeq 0$ can be guaranteed if
$$\sqrt{\sigma\tau}\,\|A\| + \frac{\sigma}{\rho} \le \frac{3}{8}, \quad \frac{1}{\tau} \ge 4L_f + \rho\,\|A\|^2\,\frac{\rho}{\frac{3}{8}\rho - \sigma}.$$
- **PDHG** Set $\mu = 0, \alpha = 0, \beta = 0$. For any $\rho > 0$, $P_c' \succeq 0$ can be guaranteed if
$$\sigma < \frac{3}{2}\rho, \quad \frac{1}{\tau} \ge L_f + \rho\,\|A\|^2\,\frac{\rho}{\frac{3}{2}\rho - \sigma}.$$
- **CP** Set $\mu = 0, \alpha = 0, \beta = 1$. For any $\rho > 0$, $P_c' \succeq 0$ can be guaranteed if
$$\frac{1}{\tau} \ge L_f + (\rho + \sigma)\,\|A\|^2\,.$$
- **GDA** Set $\mu = 1, \alpha = 0, \beta = 0$. For any $\rho > 0$, $P_c' \succeq 0$ can be guaranteed if
$$\sigma < \frac{\rho}{4}, \quad \frac{1}{\tau} \ge L_f + \rho\,\|A\|^2\,\frac{\rho - 3\sigma}{\rho - 4\sigma}.$$
- **OGDA** Set $\mu = 1, \alpha = 1, \beta = 1$. For any $\rho > 0$, $P_c' \succeq 0$ can be guaranteed if
$$\max\left\{\tau L_{f_\rho}, \frac{\sigma}{\rho}\right\} + \sqrt{\sigma\tau}\,\|A\| \le \frac{1}{2}.$$

Some observations

The penalty term brings the benefits of convergence.

- For example, PDHG, OGDA and SOGDA based on (SP-L) have no convergence guarantee generally, while these methods based on (SP-AL) are guaranteed to converge.
- Possible interpretation: the penalty term makes the convex objective function into a strongly convex function along at least one direction.

Non-ergodic convergence

- We only consider the affine equality constrained problem:

$$\min_x \ \Phi(x) = f(x) + h(x), \quad \text{s.t.}\, Ax = b.$$

- Denote $z = [x; y] \in \mathbb{R}^{n+m}$ and define a set-valued operator T as

$$T : z = [x; y]^\top \mapsto [\partial\Phi(x) - A^\top y; Ax - b]^\top,$$

- **Optimality condition:** Let $\mathcal{Z}^*$ be the set of all KKT pairs of (P), then for any $z^* \in \mathcal{Z}^*$, we have $0 \in T(z^*)$.

Definition (Local error bound condition)

The operator T satisfies (LEB) if for every $z^* \in \mathcal{Z}^*$, there exists $\epsilon > 0$, $M > 0$ such that

$$\text{dist}\,(z, \mathcal{Z}^*) \leq M\text{dist}\,(T(z), 0)\,, \qquad \forall z \ \text{ s.t.}\, \text{dist}\,(z, z^*) \leq \epsilon.$$

Some examples of (LEB)

Example (affinely constrained strongly convex problem)

$$\min_x f(x), \quad \text{s.t. } Ax = b. \tag{1}$$

(LEB) is satisfied if f is L_f-smooth and μ_f-strongly convex.

Example (two-block affinely constrained convex problem)

$$\min_{x_1,x_2} f(x_1)+h(x_2), \quad \text{s.t. } A_1x_1+A_2x_2 = b.$$

(LEB) holds if the following assumptions are satisfied

- A_1 has full row rank, A_2 has full column rank.
- $f(x_1) = g(Lx_1) + \langle q, x_1\rangle$ with g being smooth and strongly convex.
- h is either a convex piecewise linear-quadratic function, or a $\ell_{1,q}$*-norm regularizer with* $q \in [1, 2]$, or a *sparse-group LASSO regularizer*.

Non-ergodic convergence

We can obtain the linear convergence of the unified algorithm framework:

Theorem

Suppose that (LEB) condition holds. If the parameters are chosen so that $\boldsymbol{P_c} \succ \mathbf{0}$*, then there* $\exists \kappa, R > 0$ *and an integer* K*, s.t. for all* $k \geq K$*, it holds that*

$$\begin{aligned} \operatorname{dist}\left(x^k, \mathcal{X}^*\right) &\leq Re^{-\kappa(k-K)}, \\ \operatorname{dist}\left(\partial\Phi(x^k) - A^{\mathrm{T}}y^k, 0\right) &\leq Re^{-\kappa(k-K)}. \end{aligned}$$

For the affinely constrained strongly convex problem, the optimal solution x^* is unique, and

$$\mathcal{Z}^* = \{x^*\} \times \mathcal{Y}^*, \qquad \mathcal{Y}^* = \left\{ y : A^\top y = \nabla f(x^*) \right\}.$$

Theorem

Let the step sizes be suitably chosen as $\tau = \frac{c_\tau}{L_f}, \sigma = c_\sigma \frac{L_f}{\|A\|^2}, \rho = c_\rho \frac{L_f}{\|A\|^2}$, *where the constants* c_τ, c_σ, c_ρ *are chosen to ensure* $\boldsymbol{P_c \succ 0}$. *Then there exists constant* c_κ *such that for all* $k \geq 0$,

$$\left\| x^k - x^* \right\| \leq \mathcal{O}\left(\exp(-c_\kappa \left(\kappa_f + \kappa_A^2 \right) k) \right),$$

where $\mathcal{O}(\cdot)$ *hides constants that depend on* x^0, y^0 *only.*

SOGDA: $\rho = \sigma = \frac{L_f}{4\|A\|^2}, \tau = \frac{1}{8L_f}$, LALM: $\rho = \sigma = \frac{L_f}{2\|A\|^2}, \tau = \frac{1}{2L_f}$.

A byproduct: proximal OGDA for nonsmooth problems

Since the existing works of OGDA mainly focus on differentiable problems, the proximal OGDA method covered in (PD) is a direct extension of OGDA on the non-differentiable saddle point problems.

Theorem

The sequence $\{z^n\}_{n=0}^{+\infty}$ *is generated by proximal OGDA with the step sizes satisfying* $\tau \le \frac{1}{2L_{xx}}, \sigma \le \frac{1}{2L_{yy}}$ *and* $\left(\frac{1}{\tau} - 2L_{xx}\right)\left(\frac{1}{\sigma} - 2L_{yy}\right) > 4\max\{L_{xy}, L_{yx}\}^2$. *Then the sequence* $\{z^n\}_{n=0}^{+\infty}$ *converges to a saddle point of problem (SP). Furthermore, let* $(\bar{x}_N, \bar{y}_N) = \left(\frac{1}{N}\sum_{k=1}^N x^k, \frac{1}{N}\sum_{k=1}^N y^k\right)$, *then for any* $R_x, R_y > 0$, *it holds*

$$\max_{y \in \mathcal{Y} \cap \mathbb{B}(y_0, R_y)} \mathcal{L}(\bar{x}_N, y) - \min_{x \in \mathcal{X} \cap \mathbb{B}(x_0, R_x)} \mathcal{L}(x, \bar{y}_N) \le \frac{1}{N}\left(\frac{R_x^2}{\tau} + \frac{R_y^2}{\sigma}\right).$$

Outline

Linear programming

- Consider the following linear programming problem:

$$\min_x \ r^{\mathrm{T}} x, \quad \text{s.t. } Ax \le b, \ Cx = d, \ l \le x \le u.$$

- When $\rho_1 = \rho_2 = 0$, the Lagrangian function is

$$\mathcal{L}(x, y, z) = r^{\mathrm{T}} x - y^{\mathrm{T}}(Cx - d) - z^{\mathrm{T}}(Ax - b).$$

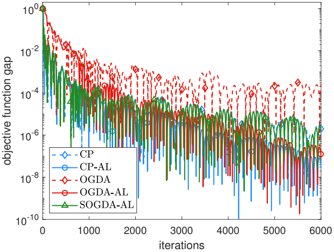
  Let $f(x) = 0, h(x) = \mathbb{1}_{[l,u]}(x), \Psi(x, y, z) = \mathcal{L}(x, y, z)$ and $s(y, z) = \mathbb{1}_{[-\infty,0]}(z)$.

- When $\rho_1, \rho_2 > 0$, the augmented Lagrangian function is

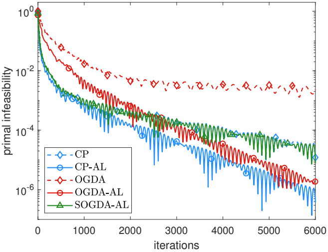
$$\mathcal{L}_\rho(x, y, z) = r^{\mathrm{T}} x - y^{\mathrm{T}}(Cx - d) + \rho_1 \| Cx - d\|^2 + \frac{\rho_2}{2} \left\| \left[Ax - b - \frac{z}{\rho_2} \right]_+ \right\|^2 - \frac{\|z\|^2}{2\rho_2}.$$

  Let $f(x) = 0, h(x) = \mathbb{1}_{[l,u]}(x), \Psi(x, y, z) = \mathcal{L}_\rho(x, y, z)$ and $s(y, z) = 0$.

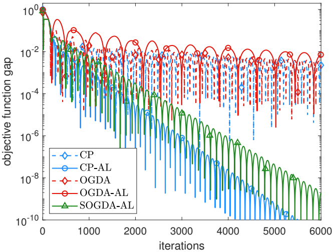
Linear programming



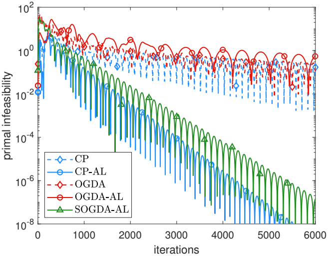
(a) qap8



(b) qap8



(c) sc50a



(d) sc50a

Basis pursuit

- Consider the following problem:

$$\min_{x} \ \|x\|_1, \quad \text{s.t. } Ax = b.$$

- The augmented Lagrangian function can be written as

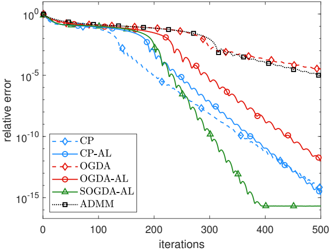
$$\mathcal{L}_\rho(x, y) = \|x\|_1 - y^{\mathrm{T}}(Ax - b) + \frac{\rho}{2}\|Ax - b\|_2^2.$$

  Let $f(x) = 0$, $h(x) = \|x\|_1$, $\Psi(x, y) = -y^{\mathrm{T}}(Ax - b) + \frac{\rho}{2}\|Ax - b\|^2$ and $s(y) = 0$.

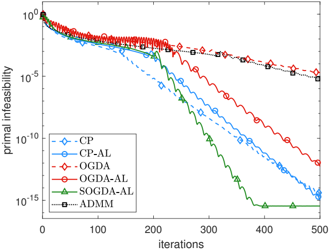
- Relative error and primal infeasibility:

$$\text{RelErr} = \frac{\|x - x^*\|_2}{\max(\|x^*\|_2, 1)}, \qquad \text{Pinf} = \frac{\|Ax - b\|_2}{\|b\|_2}.$$

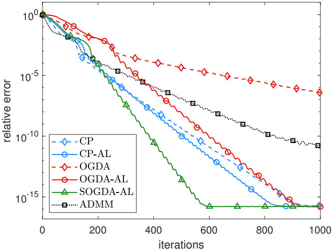
Basis pursuit



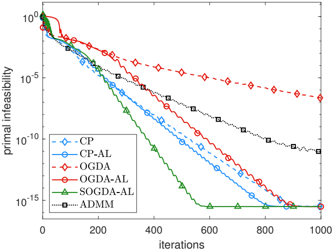
(a) 20dB



(b) 20dB



(c) 40dB



(d) 40dB

L1L1

- Consider the following problem:

$$\min_{x} \ \zeta\|x\|_1 + \|Ax - b\|_1.$$

- Introduce $r := b - Ax$ and the problem becomes

$$\min_{x,r} \ \zeta\|x\|_1 + \|r\|_1, \quad \text{s.t. } Ax - b + r = 0.$$

- The augmented Lagrangian function is

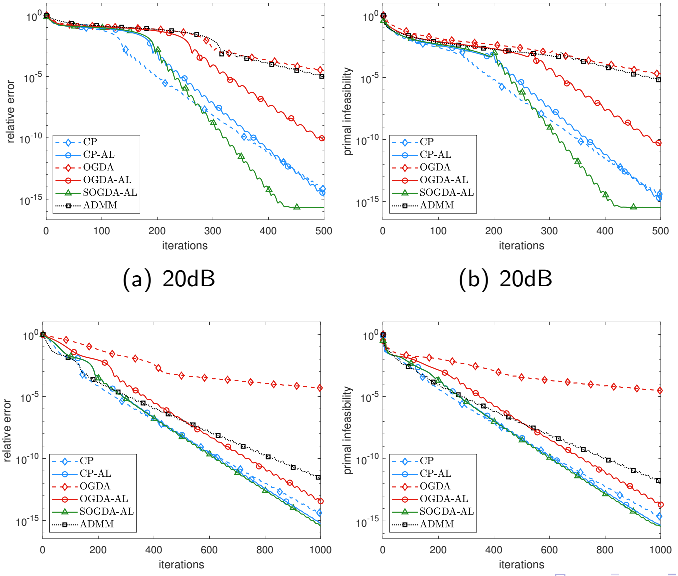
$$\mathcal{L}_\rho(x, r, y) = \zeta\|x\|_1 + \|r\|_1 - y^{\mathrm{T}}(Ax - b + r) + \frac{\rho}{2}\|Ax - b + r\|_2^2.$$

  Let $f(x, r) = 0$, $h(x, r) = \zeta\|x\|_1 + \|r\|_1$,
  $\Psi(x, r, y) = -y^{\mathrm{T}}(Ax - b + r) + \frac{\rho}{2}\|Ax - b + r\|_2^2$ and $s(y) = 0$.

- Relative error and primal infeasibility:

$$\text{RelErr} = \frac{\|x - x^*\|_2}{\max(\|x^*\|_2, 1)}, \qquad \text{Pinf} = \frac{\|Ax - b + r\|_2}{\|b\|_2}.$$

L1L1



(a) 20dB (b) 20dB

(c) 40dB (d) 40dB

Multi-block basis pursuit

- Consider the problem:

$$\min_{x_1,x_2,\cdots,x_N} \sum_{i=1}^N \|x_i\|_1, \quad \text{s.t.} \sum_{i=1}^N A_i x_i = b.$$

- The augmented Lagrangian function is

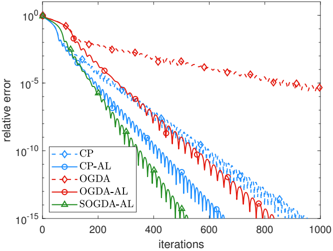
$$\mathcal{L}_\rho(x_1,x_2,\cdots,x_N,y) = \sum_{i=1}^N \|x_i\|_1 - y^{\mathrm{T}}\left(\sum_{i=1}^N A_i x_i - b\right) + \frac{\rho}{2}\left\|\sum_{i=1}^N A_i x_i - b\right\|_2^2.$$

- The derived multi-block algorithm is equivalent to the one-block algorithm.
- For the multi-block ADMM, the subproblem $\min_{x_i} \mathcal{L}_\rho(x_1,\cdots,x_N,y)$ has no explicit solution. We introduce $u_i = x_i$ to get an equivalent form:

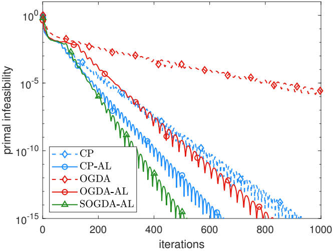
$$\min_{\substack{x_1,x_2,\cdots,x_N \\ u_1,u_2,\cdots,u_N}} \sum_{i=1}^N \|x_i\|_1, \quad \text{s.t.} \sum_{i=1}^N A_i u_i = b, \quad x_i = u_i,\ i=1,\cdots,N.$$

- Then all the subproblems of multi-block ADMM have explicit solutions.

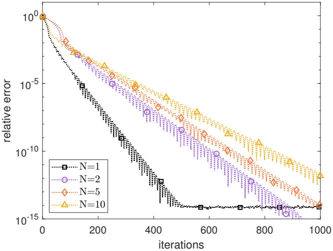
Multi-block basis pursuit



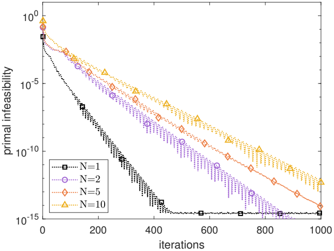
(a) primal-dual methods



(b) primal-dual methods



(c) multi-block ADMM



(d) multi-block ADMM

Non-convergent example for the multi-block ADMM

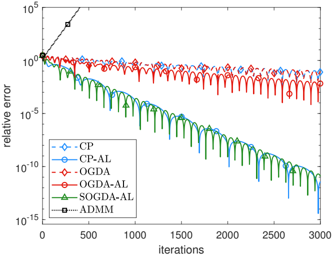
- Example 1:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} x_1 + \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} x_2 + \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} x_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

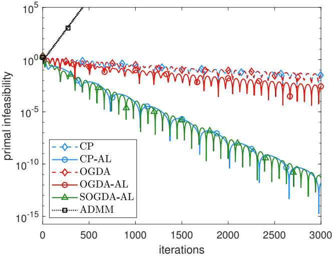
- Example 2:

$$\begin{aligned} \min\ & \frac{1}{2} x_1^2, \\ \text{s.t.}\ & \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} x_1 + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} x_2 + \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} x_3 + \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} x_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \end{aligned}$$

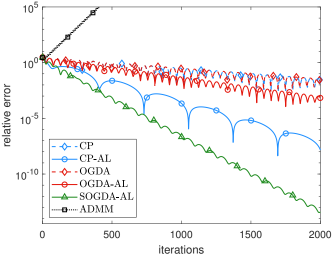
Non-convergent example for the multi-block ADMM



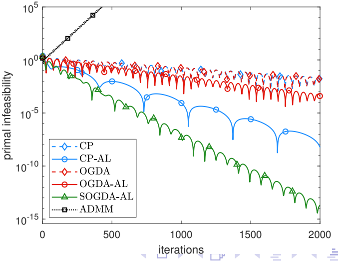
(a) Example1



(b) Example1



(c) Example2



(d) Example2

Many Thanks For Your Attention!

- 教材：刘浩洋, 户将, 李勇锋，文再文，最优化：建模、算法与理论; http://bicmr.pku.edu.cn/~wenzw/optbook.html