A single-loop proximal-conditional-gradient penalty method

Ting Kei Pong
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong

DOML 2025
May 2025
(Joint work with Hao Zhang and Liaoyuan Zeng)

Motivating examples

- Compressed sensing with heavy-tailed noise:

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{Minimize}} & \|x\|_1 \\ \text{Subject to} & \|Ax - b\|_p \le \sigma, \end{array}$$

  where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $m \ll n$, $p \in [1, 2)$ and $\sigma > 0$.

- System realization / low-rank Hankel matrix recovery:

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{Minimize}} & \|x - \bar{x}\| \\ \text{Subject to} & \|\mathcal{H}(x)\|_* \le s, \end{array}$$

  where $\bar{x} \in \mathbb{R}^n$ is a given proxy, $\mathcal{H}$ maps a vector linearly to a Hankel matrix of suitable size, $s > 0$.

Motivating examples

- Compressed sensing with heavy-tailed noise:

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{Minimize}} & \|x\|_1 \\ \text{Subject to} & \|Ax-b\|_p \le \sigma, \end{array}$$

  where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $m \ll n$, $p \in [1,2)$ and $\sigma > 0$.

- System realization / low-rank Hankel matrix recovery:

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{Minimize}} & \|x-\bar{x}\| \\ \text{Subject to} & \|\mathcal{H}(x)\|_* \le s, \end{array}$$

  where $\bar{x} \in \mathbb{R}^n$ is a given proxy, $\mathcal{H}$ maps a vector linearly to a Hankel matrix of suitable size, $s > 0$.

⋆ Projection-based algorithms?

Motivating examples

- Compressed sensing with heavy-tailed noise:

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{Minimize}} & \|x\|_1 \\ \text{Subject to} & \|Ax - b\|_p \le \sigma, \end{array}$$

  where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $m \ll n$, $p \in [1, 2)$ and $\sigma > 0$.

- System realization / low-rank Hankel matrix recovery:

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{Minimize}} & \|x - \bar{x}\| \\ \text{Subject to} & \|\mathcal{H}(x)\|_* \le s, \end{array}$$

  where $\bar{x} \in \mathbb{R}^n$ is a given proxy, $\mathcal{H}$ maps a vector linearly to a Hankel matrix of suitable size, $s > 0$.

$\star$ Projection-based algorithms? ☹

Motivating examples

- Compressed sensing with heavy-tailed noise:

$$\begin{array}{rl} \underset{x\in\mathbb{R}^n}{\text{Minimize}} & \|x\|_1 \\ \text{Subject to} & \|Ax-b\|_p \le \sigma, \end{array}$$

  where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $m \ll n$, $p \in [1,2)$ and $\sigma > 0$.

- System realization / low-rank Hankel matrix recovery:

$$\begin{array}{rl} \underset{x\in\mathbb{R}^n}{\text{Minimize}} & \|x-\bar{x}\| \\ \text{Subject to} & \|\mathcal{H}(x)\|_* \le s, \end{array}$$

  where $\bar{x} \in \mathbb{R}^n$ is a given proxy, $\mathcal{H}$ maps a vector linearly to a Hankel matrix of suitable size, $s > 0$.

⋆ Projection-based algorithms? ☹

⋆ Try **reformulation**: Let $y = Ax - b$ (resp., $y = \mathcal{H}(x)$)...

Problem setting

$$\begin{array}{rl} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x)+g(y) \\ \text{Subject to} & Ax+By=c, \end{array}$$

where

- $\mathcal{E}_1$, $\mathcal{E}_1$, $\mathcal{E}$ are finite dimensional Hilbert spaces, $c \in \mathcal{E}$, A, B are linear maps and f and g are proper, closed and convex.
- The solution set is nonempty.

Problem setting

$$\begin{array}{ll} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x) + g(y) \\ \text{Subject to} & Ax + By = c, \end{array}$$

where

- $\mathcal{E}_1$, $\mathcal{E}_1$, $\mathcal{E}$ are finite dimensional Hilbert spaces, $c \in \mathcal{E}$, A, B are linear maps and f and g are proper, closed and convex.
- The solution set is nonempty.
- $f = f_1 + f_2$: here, f_1 has Hölderian gradient, and γf_2 admits easy proximal mappings for every $\gamma > 0$. i.e., $\forall u \in \mathcal{E}_1, \gamma f_2(\cdot) + 0.5\|\cdot - u\|^2$ is easy to minimize.
- $g = g_1 + g_2$: here, g_1 has Hölderian gradient, and, $\forall v \in \mathcal{E}_2$, a minimizer of $\langle v, \cdot\rangle + g_2(\cdot)$ exists and can be computed efficiently.
- $c \in A\,\mathrm{ri}(\mathrm{dom}\, f) + B\,\mathrm{ri}(\mathrm{dom}\, g)$, and $\mathrm{dom}\, f$ and $\mathrm{dom}\, g$ are bounded.

Problem setting

$$\begin{array}{rl} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x)+g(y) \\ \text{Subject to} & Ax+By=c, \end{array}$$

where

- $\mathcal{E}_1$, $\mathcal{E}_1$, $\mathcal{E}$ are finite dimensional Hilbert spaces, $c \in \mathcal{E}$, A, B are linear maps and f and g are proper, closed and convex.
- The solution set is nonempty.
- $f = f_1 + f_2$: here, f_1 has Hölderian gradient, and γf_2 admits easy proximal mappings for every $\gamma > 0$. i.e., $\forall u \in \mathcal{E}_1$, $\gamma f_2(\cdot)+0.5\|\cdot - u\|^2$ is easy to minimize.
- $g = g_1 + g_2$: here, g_1 has Hölderian gradient, and, $\forall v \in \mathcal{E}_2$, a minimizer of $\langle v, \cdot\rangle + g_2(\cdot)$ exists and can be computed efficiently.
- $c \in A\,\mathrm{ri}(\mathrm{dom}\, f) + B\,\mathrm{ri}(\mathrm{dom}\, g)$, and dom f and dom g are bounded.
- In our motivating examples, $B = -I$ and A is the CS matrix or the Hankel map, $f(x) = f_2(x) = \|x\|_1 + \delta_{\mathcal{B}}(x)$ or $\|x - \bar{x}\| + \delta_{\mathcal{B}}(x)$, $g = g_2$ is the indicator function of the p-norm or nuc. norm ball.

Problem setting

$$\begin{array}{ll} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x)+g(y) \\ \text{Subject to} & Ax+By=c, \end{array}$$

where

- $\mathcal{E}_1$, $\mathcal{E}_1$, $\mathcal{E}$ are finite dimensional Hilbert spaces, $c \in \mathcal{E}$, A, B are linear maps and f and g are proper, closed and convex.
- The solution set is nonempty.
- $f = f_1 + f_2$: here, f_1 has Hölderian gradient, and γf_2 admits easy proximal mappings for every $\gamma > 0$. i.e., $\forall u \in \mathcal{E}_1$, $\gamma f_2(\cdot)+0.5\|\cdot - u\|^2$ is easy to minimize.
- $g = g_1 + g_2$: here, g_1 has Hölderian gradient, and, $\forall v \in \mathcal{E}_2$, a minimizer of $\langle v, \cdot\rangle + g_2(\cdot)$ exists and can be computed efficiently.
- $c \in A\,\mathrm{ri}(\mathrm{dom}\, f) + B\,\mathrm{ri}(\mathrm{dom}\, g)$, and $\mathrm{dom}\, f$ and $\mathrm{dom}\, g$ are bounded.
- In our motivating examples, $B = -I$ and A is the CS matrix or the Hankel map, $f(x) = f_2(x) = \|x\|_1 + \delta_{\mathcal{B}}(x)$ or $\|x - \bar{x}\| + \delta_{\mathcal{B}}(x)$, $g = g_2$ is the indicator function of the p-norm or nuc. norm ball.
- Apply/adapt ADMM?

Problem setting

$$\begin{array}{ll} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x)+g(y) \\ \text{Subject to} & Ax+By=c, \end{array}$$

where

- $\mathcal{E}_1$, $\mathcal{E}_1$, $\mathcal{E}$ are finite dimensional Hilbert spaces, $c \in \mathcal{E}$, A, B are linear maps and f and g are proper, closed and convex.
- The solution set is nonempty.
- $f = f_1 + f_2$: here, f_1 has Hölderian gradient, and γf_2 admits easy proximal mappings for every $\gamma > 0$. i.e., $\forall u \in \mathcal{E}_1$, $\gamma f_2(\cdot)+0.5\|\cdot - u\|^2$ is easy to minimize.
- $g = g_1 + g_2$: here, g_1 has Hölderian gradient, and, $\forall v \in \mathcal{E}_2$, a minimizer of $\langle v, \cdot\rangle + g_2(\cdot)$ exists and can be computed efficiently.
- $c \in A\,\mathrm{ri}(\mathrm{dom}\,f) + B\,\mathrm{ri}(\mathrm{dom}\,g)$, and $\mathrm{dom}\,f$ and $\mathrm{dom}\,g$ are bounded.
- In our motivating examples, $B = -I$ and A is the CS matrix or the Hankel map, $f(x) = f_2(x) = \|x\|_1 + \delta_{\mathcal{B}}(x)$ or $\|x - \bar{x}\| + \delta_{\mathcal{B}}(x)$, $g = g_2$ is the indicator function of the p-norm or nuc. norm ball.
- Apply/adapt ADMM? ☹

Existing work

In Argyriou, Signoretto, Suykens '14, they considered

$$\begin{array}{ll} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x)+g(y) \\ \text{Subject to} & By - x = 0, \end{array}$$

where

- γf admits easy proximal mappings for every $\gamma > 0$;
- $g = g_1 + g_2$: here, g_1 has Hölderian gradient with exponent $\nu \in (0, 1]$, dom g is bounded, and, $\forall v \in \mathcal{E}_2$, a minimizer of $\langle v, \cdot\rangle + g_2(\cdot)$ exists and can be computed efficiently.

Existing work

In Argyriou, Signoretto, Suykens '14, they considered

$$\begin{array}{ll} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x) + g(y) \\ \text{Subject to} & By - x = 0, \end{array}$$

where

- γf admits easy proximal mappings for every $\gamma > 0$;
- $g = g_1 + g_2$: here, g_1 has Hölderian gradient with exponent $\nu \in (0, 1]$, $\operatorname{dom} g$ is bounded, and, $\forall v \in \mathcal{E}_2$, a minimizer of $\langle v, \cdot\rangle + g_2(\cdot)$ exists and can be computed efficiently.

Idea: Alternating (approx.) minimization of F_β with an immediate update of β, where $F_\beta(x, y) := f(x) + g(y) + \frac{\beta}{2}\|By - x\|^2$.

Existing work

In Argyriou, Signoretto, Suykens '14, they considered

$$\begin{array}{ll} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x) + g(y) \\ \text{Subject to} & By - x = 0, \end{array}$$

where

- γf admits easy proximal mappings for every $\gamma > 0$;
- $g = g_1 + g_2$: here, g_1 has Hölderian gradient with exponent $\nu \in (0, 1]$, $\operatorname{dom} g$ is bounded, and, $\forall v \in \mathcal{E}_2$, a minimizer of $\langle v, \cdot\rangle + g_2(\cdot)$ exists and can be computed efficiently.

Idea: Alternating (approx.) minimization of F_β with an immediate update of β, where $F_\beta(x, y) := f(x) + g(y) + \frac{\beta}{2}\|By - x\|^2$. NO INNER LOOPS

Specifically, for $\beta_0 > 0$ and $t = 1, \ldots$,

$$\begin{aligned} x^{t+1} &= \underset{x\in\mathcal{E}_1}{\arg\min}\ f(x) + \frac{\beta_0\sqrt{t}}{2}\|By^t - x\|^2, \\ u^t &\in \underset{y\in\mathcal{E}_2}{\operatorname{Arg\,min}}\ \langle \beta_0\sqrt{t}B^*(By^t - x^{t+1}) + \nabla g_1(y^t), y\rangle + g_2(y), \\ y^{t+1} &= y^t + \frac{2}{t+1}(u^t - y^t). \end{aligned}$$

Existing work cont.

In Argyriou, Signoretto, Suykens '14:

$$\begin{array}{ll} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x) + g(y) \\ \text{Subject to} & By - x = 0, \end{array}$$

Then it holds that:

- if f is Lipschitz, then $|f(By^t) + g(y^t) - \text{opt\_val}| = O(t^{-\min\{\nu,1/2\}})$;
- (Yurtsever, Fercoq, Locatello, Cevher '18) if $f = \delta_{\mathcal{C}}$ for some closed convex set $\mathcal{C}$, ∇g_1 is Lipschitz and $g_2 = \delta_{\mathcal{D}}$ for some compact convex set $\mathcal{D}$, and $0 \in B\,\text{ri}\,\mathcal{D} - \text{ri}\,\mathcal{C}$, then
$$\max\{|g(y^t) - \text{opt\_val}|, \text{dist}(By^t, \mathcal{C})\} = O(t^{-\frac{1}{2}}).$$

Existing work cont.

In Argyriou, Signoretto, Suykens '14:

$$\begin{array}{ll} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x)+g(y) \\ \text{Subject to} & By-x=0, \end{array}$$

Then it holds that:

- if f is Lipschitz, then $|f(By^t)+g(y^t)-\text{opt\_val}| = O(t^{-\min\{\nu,1/2\}})$;
- (Yurtsever, Fercoq, Locatello, Cevher '18) if $f=\delta_{\mathcal{C}}$ for some closed convex set $\mathcal{C}$, ∇g_1 is Lipschitz and $g_2=\delta_{\mathcal{D}}$ for some compact convex set $\mathcal{D}$, and $0\in B\,\text{ri}\,\mathcal{D}-\text{ri}\,\mathcal{C}$, then
$$\max\{|g(y^t)-\text{opt\_val}|, \text{dist}(By^t,\mathcal{C})\} = O(t^{-\frac{1}{2}}).$$

Another related work (Silveti-Falls, Molinari, Fadili '20) is based on penalty and augmented Lagrangian functions. Their A has to be an injective negative partial identity map. ☹

Existing work cont.

In Argyriou, Signoretto, Suykens '14:

$$\begin{array}{ll} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x)+g(y) \\ \text{Subject to} & By - x = 0, \end{array}$$

Then it holds that:

- if f is Lipschitz, then $|f(By^t)+g(y^t)-\text{opt\_val}| = O(t^{-\min\{\nu,1/2\}})$;
- (Yurtsever, Fercoq, Locatello, Cevher '18) if $f=\delta_{\mathcal{C}}$ for some closed convex set $\mathcal{C}$, ∇g_1 is Lipschitz and $g_2=\delta_{\mathcal{D}}$ for some compact convex set $\mathcal{D}$, and $0\in B\,\text{ri}\,\mathcal{D}-\text{ri}\,\mathcal{C}$, then

$$\max\{|g(y^t)-\text{opt\_val}|, \text{dist}(By^t,\mathcal{C})\} = O(t^{-\frac{1}{2}}).$$

Another related work (Silveti-Falls, Molinari, Fadili '20) is based on penalty and augmented Lagrangian functions. Their A has to be an injective negative partial identity map. ☹

Aim: Develop a single-loop algorithm that allows a general A and our generally structured f and g, and analyze its iteration complexity / convergence.

Our algorithm

Idea: Alternating (approx.) minimization of F_β with an immediate update of β, where $F_\beta(x, y) := f(x) + g(y) + \frac{\beta}{2}\|Ax + By - c\|^2$.

Our algorithm

Idea: Alternating (approx.) minimization of F_β with an immediate update of β, where $F_\beta(x,y) := f(x) + g(y) + \frac{\beta}{2}\|Ax + By - c\|^2$.

Algorithm $\text{proxCG}^{\text{pen}}_{1\ell}$ (Zhang, Zeng, P. '24)

Step 0. Choose $x^0 \in \operatorname{dom} f$, $y^0 \in \operatorname{dom} g$, $\beta_0 > 0$, $H_0 > 0$.

Step 1. For $t = 0, 1, \ldots$, let $\alpha_t = \frac{2}{t+2}$. Compute $R^t = Ax^t + By^t - c$,

$$x^{t+1} = \operatorname*{arg\,min}_{x \in \mathcal{E}_1} \langle \nabla f_1(x^t) + \beta_t A^* R^t, x\rangle + \frac{H_t + \beta_t \lambda_{\max}(A^*A)}{2}\|x - x^t\|^2 + f_2(x),$$

$$u^t \in \operatorname*{Arg\,min}_{y \in \mathcal{E}_2} \langle \nabla g_1(y^t) + \beta_t B^*(Ax^{t+1} + By^t - c), y\rangle + g_2(y),$$

$$y^{t+1} = y^t + \alpha_t(u^t - y^t),$$

$$H_{t+1} = \max\left\{H_0, \frac{2M_f}{\mu+1}\right\}(t+1)^{1-\mu}, \quad \beta_{t+1} = \beta_0(t+2)^{1-\min\{0.5,\mu,\nu\}}.$$

Our algorithm

Idea: Alternating (approx.) minimization of F_β with an immediate update of β, where $F_\beta(x,y) := f(x) + g(y) + \frac{\beta}{2}\|Ax + By - c\|^2$.

Algorithm $\text{proxCG}^{\text{pen}}_{\mathbb{1}\ell}$ (Zhang, Zeng, P. '24)

Step 0. Choose $x^0 \in \operatorname{dom} f$, $y^0 \in \operatorname{dom} g$, $\beta_0 > 0$, $H_0 > 0$.

Step 1. For $t = 0, 1, \ldots$, let $\alpha_t = \frac{2}{t+2}$. Compute $R^t = Ax^t + By^t - c$,

$$x^{t+1} = \operatorname*{arg\,min}_{x \in \mathcal{E}_1} \langle \nabla f_1(x^t) + \beta_t A^* R^t, x\rangle + \frac{H_t + \beta_t \lambda_{\max}(A^*A)}{2}\|x - x^t\|^2 + f_2(x),$$

$$u^t \in \operatorname*{Arg\,min}_{y \in \mathcal{E}_2} \langle \nabla g_1(y^t) + \beta_t B^*(Ax^{t+1} + By^t - c), y\rangle + g_2(y),$$

$$y^{t+1} = y^t + \alpha_t(u^t - y^t),$$

$$H_{t+1} = \max\left\{H_0, \tfrac{2M_f}{\mu+1}\right\}(t+1)^{1-\mu}, \quad \beta_{t+1} = \beta_0(t+2)^{1-\min\{0.5,\mu,\nu\}}.$$

Remark: M_f is the Hölderian modulus of ∇f_1, and μ and ν are the Hölderian exponents of ∇f_1 and ∇g_1 respectively.

Complexity result

Recall that we consider

$$\begin{array}{rl} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x)+g(y) \\ \text{Subject to} & Ax+By=c, \end{array}$$

under the blanket assumptions.

Theorem 1. (Zhang, Zeng, P. '24)
Let $\{(x^t, y^t)\}$ be generated by $\text{proxCG}^{\text{pen}}_{\mathbb{1}\ell}$ and let (x^*, y^*) solve the above problem. Then it holds that

$$\|Ax^t + By^t - c\| = O(t^{-\frac{1}{2}}),$$
$$|f(x^t)+g(y^t)-f(x^*)-g(y^*)| = O(t^{-\min\{0.5,\mu,\nu\}}).$$

Complexity result

Recall that we consider

$$\begin{array}{rl}\underset{x\in\mathcal{E}_1,y\in\mathcal{E}_2}{\text{Minimize}} & f(x)+g(y)\\ \text{Subject to} & Ax+By=c,\end{array}$$

under the blanket assumptions.

Theorem 1. (Zhang, Zeng, P. '24)
Let $\{(x^t,y^t)\}$ be generated by $\text{proxCG}_{\mathbb{1}\ell}^{\text{pen}}$ and let (x^*,y^*) solve the above problem. Then it holds that

$$\|Ax^t+By^t-c\| = O(t^{-\frac{1}{2}}),$$
$$|f(x^t)+g(y^t)-f(x^*)-g(y^*)| = O(t^{-\min\{0.5,\mu,\nu\}}).$$

Note: There are explicit estimates for the constants in the big O.

KL property & exponent

Definition: (Attouch, Bolte, Redont, Soubeyran '10)
Let h be proper closed and $\alpha \in [0, 1)$.

- h is said to satisfy the Kurdyka-Łojasiewicz (KL) property with exponent α at $\bar{x} \in \operatorname{dom} \partial h$ if there exist c, ν, $\epsilon > 0$ so that

$$c[h(x) - h(\bar{x})]^\alpha \leq \operatorname{dist}(0, \partial h(x))$$

whenever $x \in \operatorname{dom} \partial h$, $\|x - \bar{x}\| \leq \epsilon$ and $h(\bar{x}) < h(x) < h(\bar{x}) + \nu$.

KL property & exponent

Definition: (Attouch, Bolte, Redont, Soubeyran '10)
Let h be proper closed and $\alpha \in [0, 1)$.

- h is said to satisfy the Kurdyka-Łojasiewicz (KL) property with exponent α at $\bar{x} \in \operatorname{dom} \partial h$ if there exist c, ν, $\epsilon > 0$ so that
$$c[h(x) - h(\bar{x})]^{\alpha} \leq \operatorname{dist}(0, \partial h(x))$$
whenever $x \in \operatorname{dom} \partial h$, $\|x - \bar{x}\| \leq \epsilon$ and $h(\bar{x}) < h(x) < h(\bar{x}) + \nu$.
- If h satisfies the KL property at any $\bar{x} \in \operatorname{dom} \partial h$ with the same α, then h is said to be a KL function with exponent α.

KL property & exponent

Definition: (Attouch, Bolte, Redont, Soubeyran '10)
Let h be proper closed and $\alpha \in [0,1)$.

- h is said to satisfy the Kurdyka-Łojasiewicz (KL) property with exponent α at $\bar{x} \in \operatorname{dom} \partial h$ if there exist c, ν, $\epsilon > 0$ so that

$$c[h(x) - h(\bar{x})]^{\alpha} \leq \operatorname{dist}(0, \partial h(x))$$

  whenever $x \in \operatorname{dom} \partial h$, $\|x - \bar{x}\| \leq \epsilon$ and $h(\bar{x}) < h(x) < h(\bar{x}) + \nu$.
- If h satisfies the KL property at any $\bar{x} \in \operatorname{dom} \partial h$ with the same α, then h is said to be a KL function with exponent α.

Examples:

- Proper closed semialgebraic functions are KL functions with exponent $\alpha \in [0,1)$. (Bolte, Daniilidis, Lewis, Shiota '07)
- If h is the maximum of m polynomials of degree at most d, then the KL exponent is $1 - \frac{1}{\max\{1,(d+1)(3d)^{n+m-2}\}}$. (Li, Mordukovich, Pham '15)

Convergence rate

Theorem 2. (Zhang, Zeng, P. '24)
Let $h : \mathcal{E} \to \mathbb{R}$ be a real-valued convex function, $\Theta \subset \mathcal{E}$ be a compact convex set, G be a linear map and $b \in G\,\mathrm{ri}\,\Theta$. Let

$$H(x) = h(x) + \delta_\Theta(x) + \delta_{\{b\}}(Gx).$$

If H is a KL function with exponent $\alpha \in [0, 1)$, then there exist $\epsilon > 0$, $c_0 > 0$ and $\eta > 0$ such that

$$\mathrm{dist}\,(x, \operatorname{Arg\,min} H) \leq c_0 \Big| h(x) + \eta \|Gx - b\| - \inf H \Big|^{1-\alpha}$$

whenever $x \in \Theta$ and $\mathrm{dist}(x, \operatorname{Arg\,min} H) \leq \epsilon$.

Convergence rate cont.

Recall that we consider

$$\begin{array}{ll} \underset{x\in\mathcal{E}_1, y\in\mathcal{E}_2}{\text{Minimize}} & f(x)+g(y) \\ \text{Subject to} & Ax+By=c, \end{array}$$

under the blanket assumptions.

Corollary. (Zhang, Zeng, P. '24)
Suppose that $F(x,y) := f(x)+g(y)+\delta_{\{c\}}(Ax+By)$ is a KL function with exponent $\alpha\in[0,1)$. Suppose we further assume that $f=f_0+\delta_\Xi$ and $g=g_0+\delta_\Delta$ for some real-valued convex functions f_0 and g_0 and compact convex sets Ξ and Δ.

If $\{(x^t,y^t)\}$ is generated by $\text{proxCG}^{\text{pen}}_{1\ell}$, then

$$\text{dist}((x^t,y^t),\operatorname{Arg\,min} F) = O(t^{-\min\{0,5,\mu,\nu\}(1-\alpha)})$$

Example: Explicit KL exponent

Consider the compressed sensing problem with heavy-tailed noise:

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n, y\in\mathbb{R}^m}{\text{Minimize}} & \|x\|_1 \\ \text{Subject to} & \|y\|_p \le \sigma,\ Ax - y = b, \end{array}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $m \ll n$, $p \in (1,2)$ and $\sigma > 0$. We discuss how to derive the KL exponent of the following associated function:

$$F(x,y) := \|x\|_1 + \delta_{\|\cdot\|_p \le \sigma}(y) + \delta_{\{b\}}(Ax - y).$$

Example: Explicit KL exponent

Consider the compressed sensing problem with heavy-tailed noise:

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n, y\in\mathbb{R}^m}{\text{Minimize}} & \|x\|_1 \\ \text{Subject to} & \|y\|_p \le \sigma,\ Ax - y = b, \end{array}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $m \ll n$, $p \in (1, 2)$ and $\sigma > 0$. We discuss how to derive the KL exponent of the following associated function:

$$F(x, y) := \|x\|_1 + \delta_{\|\cdot\|_p \le \sigma}(y) + \delta_{\{b\}}(Ax - y).$$

Step 1: (Conic lifting) Note that $F(x, y) = \inf_{w,s} \widehat{F}(x, w, y, s)$, where

$$\widehat{F}(x, w, y, s) = w + \delta_{\mathcal{F}}(x, w, y, s),$$
$$\mathcal{F} := \{(x, w, y, s) : s = \sigma,\ Ax - y = b,\ (y, s) \in \mathcal{K}_p^{m+1},\ (x, w) \in \mathcal{K}_1^{n+1}\},$$

with $\mathcal{K}_p^{m+1}$ being the p-cone in $\mathbb{R}^{m+1}$, i.e., $\{(y, s) \in \mathbb{R}^m \times \mathbb{R} : \|y\|_p \le s\}$. It is known that the KL exponent of $\widehat{F}$ gives that of F (Yu, Li, P. '22).

Example cont.: Explicit KL exponent

Observation: Let $\theta = \inf \widehat{F}$. Then letting $z := (x, w, y, s)$, we have

$$\operatorname{Arg\,min} \widehat{F} = \underbrace{\{z : \; w = \theta, \; s = \sigma, \; Ax - y = b\}}_{\mathcal{S}_1} \cap \underbrace{(\mathcal{K}_1^{n+1} \times \mathcal{K}_p^{m+1})}_{\mathcal{S}_2}$$

Example cont.: Explicit KL exponent

Observation: Let $\theta = \inf \widehat{F}$. Then letting $z := (x, w, y, s)$, we have

$$\operatorname{Arg\,min} \widehat{F} = \underbrace{\{z : \ w = \theta,\ s = \sigma,\ Ax - y = b\}}_{\mathcal{S}_1} \cap \underbrace{(\mathcal{K}_1^{n+1} \times \mathcal{K}_p^{m+1})}_{\mathcal{S}_2}$$

Step 2: (Conic error bound) It holds that (Lindstrom, Lourenço, P. '24) for each $r > 0$, $\exists c_r > 0$ such that for all $z \in \mathcal{B}_r := \{u : \ \|u\| \leq r\}$,

$$\operatorname{dist}(z, \mathcal{S}_1 \cap \mathcal{S}_2) \leq c_r \max\{\operatorname{dist}(z, \mathcal{S}_1)^{\frac{1}{2}}, \operatorname{dist}(z, \mathcal{S}_2)^{\frac{1}{2}}\},$$

which implies $\exists \kappa_r > 0$ such that for all $z \in \mathcal{B}_r \cap \mathcal{F}$,

$$\operatorname{dist}(z, \operatorname{Arg\,min} \widehat{F}) = \operatorname{dist}(z, \mathcal{S}_1 \cap \mathcal{S}_2) \leq c_r \operatorname{dist}(z, \mathcal{S}_1)^{\frac{1}{2}} \leq \kappa_r |w - \theta|^{\frac{1}{2}}.$$

Recall that $\widehat{F}(x, w, y, s) = w + \delta_{\mathcal{F}}(x, w, y, s)$. The above display shows that $\widehat{F}$ has KL exponent $\frac{1}{2}$ (Bolte, Nguyen, Peypouquet, Suter '17).

Numerical results

Consider random instances of

$$\begin{array}{rl} \underset{x \in \mathbb{R}^n}{\text{Minimize}} & \|x\|_1 \\ \text{Subject to} & \|Ax - b\|_{1.5} \leq \sigma. \end{array}$$

Numerical results

Consider random instances of

$$\begin{array}{rl} \underset{x \in \mathbb{R}^n}{\text{Minimize}} & \|x\|_1 \\ \text{Subject to} & \|Ax - b\|_{1.5} \leq \sigma. \end{array}$$

If $\{(x^t, y^t)\}$ is generated by $\text{proxCG}^{\text{pen}}_{\mathbb{1}\ell}$, then

$$\max\{|\|x^t\|_1 - \|x^*\|_1|, \|Ax^t - y^t - b\|\} = O(t^{-1/2}).$$

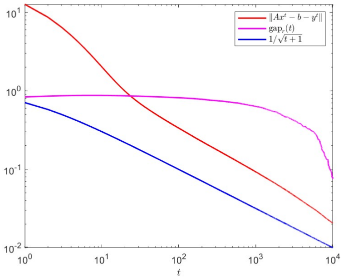
Numerical results

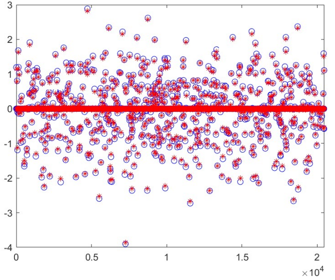
Consider random instances of

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{Minimize}} & \|x\|_1 \\ \text{Subject to} & \|Ax-b\|_{1.5} \le \sigma. \end{array}$$

If $\{(x^t, y^t)\}$ is generated by $\text{proxCG}^{\text{pen}}_{\mathbb{1}\ell}$, then

$$\max\{|\|x^t\|_1 - \|x^*\|_1|, \|Ax^t - y^t - b\|\} = O(t^{-1/2}).$$





Conclusion

Conclusion:

- A single-loop algorithm based on penalty method is developed for linearly constrained convex optimization problems involving prox-friendly and linear-oracle-friendly components.
- Each iteration involves one prox and one linear-oracle call.
- Iteration complexity and (local) convergence rate are derived.

Reference:

- Hao Zhang, Liaoyuan Zeng and Ting Kei Pong. *A single-loop proximal-conditional-gradient penalty method*. Preprint. Available at https://arxiv.org/abs/2409.14957.

Thanks for coming! ☺