Convex–Composite Optimization

SIAM Conference on Optimization, Seattle, 2023

Jim Burke

Collaborators

Aleksandr Aravkin, Dmitriy Drusvyatskiy, Abraham Engle,

Michael Ferris, Michael Friedlander, Tim Hoheisel, Quang Nguyen,

René Poliquin, Aleksei Sholokhov, Peng Zheng

Funding: Bill and Melinda Gates Foundation

U.S. NSF grant DMS-1514559

Convex-Composite Optimization

$$\mathbf{P} \qquad \min_{x \in \mathbb{R}^n} f(x) := h(c(x)) + \ g(x)$$

$h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex
$c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^2$-smooth
$g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex

Convex-Composite Optimization

$$\mathbf{P} \qquad \min_{x \in \mathbb{R}^n} f(x) := h(c(x)) + \ g(x)$$

$h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex   The Model
$c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^2$-smooth
$g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex

Convex-Composite Optimization

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} f(x) := h(c(x)) + \; g(x)$$

$h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex — The Model
$c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^2$-smooth — Model input (Data)
$g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex

Convex-Composite Optimization

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} f(x) := h(c(x)) + \ g(x)$$

$h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex — The Model
$c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^2$-smooth — Model input (Data)
$g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex — Regularizer

Convex-Composite Optimization

$$\mathbf{P} \qquad \min_{x \in \mathbb{R}^n} f(x) := h(c(x)) + \ g(x)$$

$h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex
$c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^2$-smooth
$g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex

In general, these problems are **neither convex nor smooth**.

Convex-Composite Optimization

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} f(x) := h(c(x)) + \; g(x)$$

$h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex
$c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^2$-smooth
$g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex

Note that g can be absorbed into h.

Set

$$\tilde{h}(y,x) := h(y) + g(x) \quad \text{and} \quad \tilde{c}(x) \quad := (c(x), x),$$

then $f = \tilde{h} \circ \tilde{c}$ is convex-composite.

Convex-Composite Optimization

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} f(x) := h(c(x)) + \ g(x)$$

$h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex
$c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^2$-smooth
$g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex

Note that g can be absorbed into h.

Set

$$\tilde{h}(y, x) := h(y) + g(x) \quad \text{and} \quad \tilde{c}(x) \quad := (c(x), x),$$

then $f = \tilde{h} \circ \tilde{c}$ is convex-composite.

For simplicity, we usually take $g \equiv 0$.

But in the context of algorithmic implementations, it is often essential to treat g explicitly.

Convex-Composite Optimization

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} f(x) := h(c(x)) + \ g(x)$$

$h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex
$c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^2$-smooth
$g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex

1805 The Gauss-Newton method : $\min_x \frac{1}{2}\|c(x)\|_2^2$
Legendre 1805, Gauss 1809 (1795?)

Gauss, in 1809 at the age of 24, used the method to track the newly discovered asteroid Ceres. He also advanced Legendre's work by establishing connections to probability and statistics using the normal distribution.

Gauss also claimed to have been using the method for celestial computations since 1795 at the age of 10.

Convex-Composite Optimization

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} f(x) := h(c(x)) + \ g(x)$$

$h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex
$c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^2$-smooth
$g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex

1805 The Gauss-Newton method :
Legendre 1805, Gauss 1809 (1795?)

70's
Anderson, Osborne, Watson: *Algorithms for nonlinear approximation*

80-90's
B., Conn, Ferris, **Fletcher**, Kawasaki, Masden, Poliquin, **Powell**, Osborne, Rockafellar, Womersley, Wright, Yuan

Recent (15-)
Aravkin, Bell, B., Chang, Cui, Duchi, Davis, Drusvyatskiy, Engle, Hoheisel, Hong, Lewis, Ioffe, Mohammadi, Mordukhovich, Pang, Paquette, Royset, Ruan, Sarabi, Zheng ...

Examples:

Non-linear least-squares: $f(x) = \|c(x)\|_2^2$

Examples:

Non-linear least-squares: $f(x) = \|c(x)\|_2^2$

Feasibility Problems: $c(x) \in C : \quad f(x) = \operatorname{dist}(c(x) \,|\, C)$,
where $C \subset \mathbb{R}^m$ is closed, convex (e.g., $C = \{0\}^p \times \mathbb{R}^q_-$), and $\operatorname{dist}(y \,|\, C) := \inf \{\|y - z\| \mid z \in C\}$.

Examples:

Non-linear least-squares: $f(x) = \|c(x)\|_2^2$

Feasibility Problems: $c(x) \in C: \quad f(x) = \operatorname{dist}(c(x) \,|\, C),$
where $C \subset \mathbb{R}^m$ is closed, convex (e.g., $C = \{0\}^p \times \mathbb{R}^q_-$), and
$\operatorname{dist}(y \,|\, C) := \inf\{\|y - z\| \mid z \in C\}$.

Non-linear programming (NLP): $\min \varphi(x) + \delta_C(\hat{c}(x))$.
Here $c(x) := (\varphi(x), \hat{c}(x))$ and $h(\mu, y) := \mu + \delta_C(y)$, where

$$\delta_C(y) = \begin{cases} 0, & y \in C, \\ +\infty, & \text{else.} \end{cases}$$

Examples:

Non-linear least-squares: $f(x) = \|c(x)\|_2^2$

Feasibility Problems: $c(x) \in C: \quad f(x) = \operatorname{dist}(c(x) \,|\, C)$,
where $C \subset \mathbb{R}^m$ is closed, convex (e.g., $C = \{0\}^p \times \mathbb{R}^q_-$), and
$\operatorname{dist}(y \,|\, C) := \inf \{\|y - z\| \mid z \in C\}$.

Non-linear programming (NLP): $\min \varphi(x) + \delta_C(\hat{c}(x))$.
Here $c(x) := (\varphi(x), \hat{c}(x))$ and $h(\mu, y) := \mu + \delta_C(y)$, where

$$\delta_C(y) = \begin{cases} 0, & y \in C, \\ +\infty, & \text{else}. \end{cases}$$

δ_C is called the convex indicator function for C, typically

$$C = \{0\}^s \times \mathbb{R}^{m-s}_-.$$

Examples:

Non-linear least-squares: $f(x) = \|c(x)\|_2^2$

Feasibility Problems: $c(x) \in C : \quad f(x) = \mathrm{dist}\,(c(x)\,|C\,)$,
where $C \subset \mathbb{R}^m$ is closed, convex (e.g., $C = \{0\}^p \times \mathbb{R}^q_-$), and
$\mathrm{dist}\,(y\,|C\,) := \inf\{\|y - z\| \mid z \in C\}$.

Non-linear programming (NLP): $\min \varphi(x) + \delta_C(\hat{c}(x))$.

Here $c(x) := (\varphi(x), \hat{c}(x))$ and $h(\mu, y) := \mu + \delta_C(y)$, where

$$\delta_C(y) = \begin{cases} 0, & y \in C, \\ +\infty, & \text{else.} \end{cases}$$

δ_C is called the convex indicator function for C, typically

$$C = \{0\}^s \times \mathbb{R}^{m-s}_-.$$

Exact Penalization: $f(x) = \varphi(x) + \alpha \mathrm{dist}\,(\hat{c}(x)\,|C\,)$

Here $c(x) := (\varphi(x), \hat{c}(x))$ and $h(\mu, y) := \mu + \alpha \mathrm{dist}\,(y\,|C\,)$

Examples:

Non-linear least-squares: $f(x) = \|c(x)\|_2^2$

Feasibility Problems: $c(x) \in C: \quad f(x) = \mathrm{dist}\,(c(x)\,|C\,)$,
where $C \subset \mathbb{R}^m$ is closed, convex (e.g., $C = \{0\}^p \times \mathbb{R}^q_-$), and
$\mathrm{dist}\,(y\,|C\,) := \inf\{\|y - z\| \mid z \in C\}$.

Non-linear programming (NLP): $\min \varphi(x) + \delta_C(\hat{c}(x))$.
Here $c(x) := (\varphi(x), \hat{c}(x))$ and $h(\mu, y) := \mu + \delta_C(y)$, where

$$\delta_C(y) = \begin{cases} 0, & y \in C, \\ +\infty, & \text{else.} \end{cases}$$

δ_C is called the convex indicator function for C, typically

$$C = \{0\}^s \times \mathbb{R}^{m-s}_-.$$

Exact Penalization: $f(x) = \varphi(x) + \alpha \mathrm{dist}\,(\hat{c}(x)\,|C\,)$
Here $c(x) := (\varphi(x), \hat{c}(x))$ and $h(\mu, y) := \mu + \alpha \mathrm{dist}\,(y\,|C\,)$

Additive composite problems: $f(x) = \psi(x) + g(x)$ with $\psi \in \mathcal{C}^1$

Examples

Robust Phase Retrieval:

$$\min_x \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x \rangle^2 - b_i^2|$$

Examples

Robust Phase Retrieval:

$$\min_{x} \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x \rangle^2 - b_i^2|$$

Sparse Dictionary Learning:

$$\min_{D \in \mathbb{R}^{d \times n}, r_i \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \|x_i - D r_i\|_2 + \lambda \|r_i\|_1 \qquad \text{subject to} \quad \|D_i\| \leq 1$$

Examples

Robust Phase Retrieval:

$$\min_x \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x\rangle^2 - b_i^2|$$

Sparse Dictionary Learning:

$$\min_{D \in \mathbb{R}^{d\times n}, r_i \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \|x_i - Dr_i\|_2 + \lambda \|r_i\|_1 \qquad \text{subject to} \quad \|D_i\| \leq 1$$

Robust PCA:

$$\min_{U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r}} \left\| UV^T - M \right\|_1$$

Examples

Robust Phase Retrieval:

$$\min_x \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x \rangle^2 - b_i^2|$$

Sparse Dictionary Learning:

$$\min_{D \in \mathbb{R}^{d \times n}, r_i \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \|x_i - Dr_i\|_2 + \lambda \|r_i\|_1 \qquad \text{subject to} \quad \|D_i\| \le 1$$

Robust PCA:

$$\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}} \left\| UV^T - M \right\|_1$$

Sparse/Robust Estimation and Kalman Smoothing:

$$\min_x V(k(x, z)) + W(q(x)),$$

where V and W are convex piecewise linear-quadratic penalties:

$$\rho(y) = \sup_{u \in U} \left\{ \langle u, b + By \rangle - \frac{1}{2} y^T M y \right\}.$$

ℓ_1, least-squares, elastic net, Vapnik Huber, ...

Rockafellar '88

Outline

1. First-Order Properties: directional derivatives and subgradient
2. The Convex-Composite Lagrangian
3. Second-Order Properties
4. Exact Penalization
5. Convexity of Convex-Composite Functions
6. Algorithms
   i. Sharpness
   ii. Newton's Method
   iii. Globalization
   iv. Complexity
   v. Stochastic Prox-Linear
7. Feature Selection for Mixed Effects Models

First-Order Properties

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} f(x) := h(c(x))$$

Standard first-order necessary conditions for optimality in **P** are

$$f'(x;d) \geq 0 \quad \forall\, d \in \mathbb{R}^n,$$

where $f'(x;d)$ is the directional derivative of f at x given by

$$f'(x;d) := \lim_{t\downarrow 0} \frac{f(x+td)-f(x)}{t}.$$

First-Order Properties

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} f(x) := h(c(x))$$

Standard first-order necessary conditions for optimality in **P** are

$$f'(x;d) \geq 0 \quad \forall\, d \in \mathbb{R}^n,$$

where $f'(x;d)$ is the directional derivative of f at x given by

$$f'(x;d) := \lim_{t\downarrow 0} \frac{f(x+td) - f(x)}{t}.$$

Does the directional derivative exists?

First-Order Properties

$$\mathbf{P} \qquad \min_{x \in \mathbb{R}^n} f(x) := h(c(x))$$

Standard first-order necessary conditions for optimality in **P** are

$$f'(x; d) \geq 0 \quad \forall\, d \in \mathbb{R}^n,$$

where $f'(x; d)$ is the directional derivative of f at x given by

$$f'(x; d) := \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}.$$

Does the directional derivative exists?

We begin by assuming that h is finite valued.
Convexity implies that h is locally Lipschitz continuous, i.e.

$$\forall\, \bar{u} \quad \exists\, L > 0 \ : \ |h(u) - h(v)| \leq L\|u - v\| \quad \forall\, u, v \text{ near } \bar{u}.$$

The Directional Derivative $f'(x;d)$

$$|h(c(x)) - h(c(\overline{x}) + c'(\overline{x})(x - \overline{x}))| \le L|c(x) - [c(\overline{x}) + c'(\overline{x})(x - \overline{x})]| = o(\|x - \overline{x}\|)$$

The Directional Derivative $f'(x;d)$

$$|h(c(x))-h(c(\overline{x})+c'(\overline{x})(x-\overline{x}))| \le L|c(x)-[c(\overline{x})+c'(\overline{x})(x-\overline{x})]| = o(\|x-\overline{x}\|)$$

Consequently,

$$h(c(x)) = h(c(\overline{x})+c'(\overline{x})(x-\overline{x}))+o(\|x-\overline{x}\|).$$

The Directional Derivative $f'(x;d)$

$$|h(c(x)) - h(c(\overline{x}) + c'(\overline{x})(x - \overline{x}))| \leq L|c(x) - [c(\overline{x}) + c'(\overline{x})(x - \overline{x})]| = o(\|x - \overline{x}\|)$$

Consequently,

$$h(c(x)) = h(c(\overline{x}) + c'(\overline{x})(x - \overline{x})) + o(\|x - \overline{x}\|).$$

$$\begin{aligned} f'(x;d) = (h \circ c)'(x;d) &= \lim_{t \downarrow 0} \frac{h(c(x + td)) - h(c(x))}{t} \\ &= \lim_{t \downarrow 0} \frac{h(c(x) + tc'(x)d) - h(c(x))}{t} \\ &= h'(c(x); c'(x)d). \end{aligned}$$

The Subdifferential $\partial f(x)$

Recall that for a convex function φ, we have

$$\varphi'(y;v) = \sup\{\langle z, v\rangle \mid z \in \partial\varphi(y)\},$$

whenever $\partial\varphi(\overline{y}) \neq \emptyset$, where

$$\partial\varphi(\overline{y}) := \{z \mid \varphi(\overline{y}) + \langle z, y - \overline{y}\rangle \leq \varphi(y)\ \forall\, y \in \mathbb{R}^m\},$$

is the convex subdifferential of $\varphi(\overline{y})$ at $\overline{y}$.

The Subdifferential $\partial f(x)$

Recall that for a convex function φ, we have

$$\varphi'(y;v) = \sup\{\langle z, v\rangle \mid z \in \partial\varphi(y)\},$$

whenever $\partial\varphi(\overline{y}) \neq \emptyset$, where

$$\partial\varphi(\overline{y}) := \{z \mid \varphi(\overline{y}) + \langle z, y - \overline{y}\rangle \leq \varphi(y) \ \forall\, y \in \mathbb{R}^m\},$$

is the convex subdifferential of $\varphi(\overline{y})$ at $\overline{y}$.

Consequently,

$$f'(x;d) = h'(c(x);c'(x)d) = \sup\left\{\langle z, c'(x)d\rangle \mid z \in \partial h(c(x))\right\}$$

The Subdifferential $\partial f(x)$

Recall that for a convex function φ, we have

$$\varphi'(y;v) = \sup\{\langle z, v\rangle \mid z \in \partial\varphi(y)\},$$

whenever $\partial\varphi(\overline{y}) \neq \emptyset$, where

$$\partial\varphi(\overline{y}) := \{z \mid \varphi(\overline{y}) + \langle z, y - \overline{y}\rangle \leq \varphi(y)\ \forall\, y \in \mathbb{R}^m\},$$

is the convex subdifferential of $\varphi(\overline{y})$ at $\overline{y}$.

Consequently,

$$\begin{aligned} f'(x;d) = h'(c(x);c'(x)d) &= \sup\left\{\langle z, c'(x)d\rangle \mid z \in \partial h(c(x))\right\} \\ &= \sup\left\{\langle c'(x)^T z, d\rangle \mid z \in \partial h(c(x))\right\} \end{aligned}$$

The Subdifferential $\partial f(x)$

Recall that for a convex function φ, we have

$$\varphi'(y;v) = \sup\{\langle z, v\rangle \mid z \in \partial\varphi(y)\},$$

whenever $\partial\varphi(\overline{y}) \neq \emptyset$, where

$$\partial\varphi(\overline{y}) := \{z \mid \varphi(\overline{y}) + \langle z, y - \overline{y}\rangle \leq \varphi(y)\ \forall\, y \in \mathbb{R}^m\},$$

is the convex subdifferential of $\varphi(\overline{y})$ at $\overline{y}$.

Consequently,

$$\begin{aligned} f'(x;d) = h'(c(x);c'(x)d) &= \sup\{\langle z, c'(x)d\rangle \mid z \in \partial h(c(x))\} \\ &= \sup\left\{\langle c'(x)^T z, d\rangle \mid z \in \partial h(c(x))\right\} \\ &= \sup\left\{\langle w, d\rangle \,\middle|\, w \in c'(x)^T \partial h(c(x))\right\}. \end{aligned}$$

The Subdifferential $\partial f(x)$

Recall that for a convex function φ, we have

$$\varphi'(y;v) = \sup\{\langle z, v\rangle \mid z \in \partial\varphi(y)\},$$

whenever $\partial\varphi(\overline{y}) \neq \emptyset$, where

$$\partial\varphi(\overline{y}) := \{z \mid \varphi(\overline{y}) + \langle z, y - \overline{y}\rangle \leq \varphi(y)\ \forall\, y \in \mathbb{R}^m\},$$

is the convex subdifferential of $\varphi(\overline{y})$ at $\overline{y}$.

Consequently,

$$\begin{aligned} f'(x;d) = h'(c(x); c'(x)d) &= \sup\{\langle z, c'(x)d\rangle \mid z \in \partial h(c(x))\} \\ &= \sup\left\{\langle c'(x)^T z, d\rangle \mid z \in \partial h(c(x))\right\} \\ &= \sup\left\{\langle w, d\rangle \;\middle|\; w \in c'(x)^T \partial h(c(x))\right\}. \end{aligned}$$

Define $\quad \partial f(x) := c'(x)^T \partial h(c(x)).$

The Subdifferential $\partial f(x)$

Recall that for a convex function φ, we have

$$\varphi'(y; v) = \sup\{\langle z, v\rangle \mid z \in \partial\varphi(y)\},$$

whenever $\partial\varphi(\overline{y}) \neq \emptyset$, where

$$\partial\varphi(\overline{y}) := \{z \mid \varphi(\overline{y}) + \langle z, y - \overline{y}\rangle \leq \varphi(y) \ \forall\, y \in \mathbb{R}^m\},$$

is the convex subdifferential of $\varphi(\overline{y})$ at $\overline{y}$.

Consequently,

$$\begin{aligned} f'(x; d) = h'(c(x); c'(x)d) &= \sup\{\langle z, c'(x)d\rangle \mid z \in \partial h(c(x))\} \\ &= \sup\left\{\langle c'(x)^T z, d\rangle \mid z \in \partial h(c(x))\right\} \\ &= \sup\left\{\langle w, d\rangle \;\middle|\; w \in c'(x)^T \partial h(c(x))\right\}. \end{aligned}$$

Define $\partial f(x) := c'(x)^T \partial h(c(x)).$

f is subdifferentially regular.

The Basic Constraint Qualification

What happens when h is not finite-valued?

The Basic Constraint Qualification

What happens when h is not finite-valued?

In this case, $f'(x;d)$ does not adequately describe the variational behavior of f on the *relative boundary* of $\operatorname{dom}(h)$.

The Basic Constraint Qualification

What happens when h is not finite-valued?

In this case, $f'(x;d)$ does not adequately describe the variational behavior of f on the *relative boundary* of $\mathrm{dom}\,(h)$.

Consequently, we employ the more general *subderivative*:

$$df(x)(d) := \liminf_{t\downarrow 0, d'\to d} \frac{f(x+td') - f(x)}{t} \ .$$

The Basic Constraint Qualification

What happens when h is not finite-valued?

In this case, $f'(x;d)$ does not adequately describe the variational behavior of f on the *relative boundary* of $\mathrm{dom}\,(h)$.

Consequently, we employ the more general *subderivative*:

$$df(x)(d) := \liminf_{t \downarrow 0, d' \to d} \frac{f(x+td') - f(x)}{t} \ .$$

In addition, c may be "deficient" at $\overline{x} \in \mathrm{rbdry}\,(\mathrm{dom}\,(f))$ in the sense that

$$\nexists\, \tilde{x} \quad s.t. \quad c(\overline{x}) + c'(\overline{x})(\tilde{x} - \overline{x}) \in \mathrm{ri}\,(\mathrm{dom}\,(h))\,.$$

That is, $c(\overline{x}) + c'(\overline{x})(x - \overline{x})$ does not enter $\mathrm{ri}\,(\mathrm{dom}\,(h))$ from $c(\overline{x})$.

The Basic Constraint Qualification

What happens when h is not finite-valued?

In this case, $f'(x;d)$ does not adequately describe the variational behavior of f on the *relative boundary* of $\mathrm{dom}\,(h)$.

Consequently, we employ the more general *subderivative*:

$$df(x)(d) := \liminf_{t \downarrow 0, d' \to d} \frac{f(x + td') - f(x)}{t} \ .$$

In addition, c may be "deficient" at $\overline{x} \in \mathrm{rbdry}\,(\mathrm{dom}\,(f))$ in the sense that

$$\nexists\, \tilde{x} \quad s.t. \quad c(\overline{x}) + c'(\overline{x})(\tilde{x} - \overline{x}) \in \mathrm{ri}\,(\mathrm{dom}\,(h))\,.$$

That is, $c(\overline{x}) + c'(\overline{x})(x - \overline{x})$ does not enter $\mathrm{ri}\,(\mathrm{dom}\,(h))$ from $c(\overline{x})$.

A constraint qualification is employed to address this deficiency.

The Basic Constraint Qualification

Basic Constraint Qualification (BCQ) (Rockafellar '85):

$$\ker\left(c'(x)^T\right) \cap N\left(c(x)\,|\,\mathrm{dom}\,(h)\right) = \{0\}$$

where

$$N\left(\overline{y}\,|\,C\right) := \partial\delta_C(\overline{y}) = \{z \mid \langle z, y - \overline{y}\rangle \leq 0 \;\forall\, y \in C\}$$

is the normal cone to the convex set C at $\overline{y} \in C$.

The Basic Constraint Qualification

Basic Constraint Qualification (BCQ) (Rockafellar '85):

$$\ker\left(c'(x)^T\right) \cap N\left(c(x) \,|\, \mathrm{dom}\,(h)\right) = \{0\}$$

where

$$N\left(\overline{y} \,|\, C\right) := \partial \delta_C(\overline{y}) = \{z \mid \langle z, y - \overline{y}\rangle \leq 0 \ \forall\, y \in C\}$$

is the normal cone to the convex set C at $\overline{y} \in C$.

• If $f = h \circ c$ satisfies the BCQ at $x \in \mathrm{dom}\,(f)$, then f is *subdifferentially regular* at x with

$$\partial f(x) = c'(x)^T \partial h(c(x)) \quad \text{and}$$
$$df(x)(d) = \sup\{\langle z, d\rangle \mid z \in \partial f(x)\}.$$

The Basic Constraint Qualification

Basic Constraint Qualification (BCQ) (Rockafellar '85):

$$\ker\left(c'(x)^T\right) \cap N\left(c(x) \,|\, \mathrm{dom}\,(h)\right) = \{0\}$$

where

$$N\left(\overline{y} \,|\, C\right) := \partial\delta_C(\overline{y}) = \{z \mid \langle z, y - \overline{y}\rangle \le 0 \;\forall\, y \in C\}$$

is the normal cone to the convex set C at $\overline{y} \in C$.

- $f = h \circ c$ satisfies the BCQ at $x \in \mathrm{dom}\,(f)$ if and only if

$$\left\{ y \in \partial h(c(x)) \;\middle|\; v = c'(x)^T y \right\} \text{ is compact } \forall\, v \in \partial f(x).$$

The Basic Constraint Qualification

Basic Constraint Qualification (BCQ) (Rockafellar '85):

$$\ker\left(c'(x)^T\right) \cap N\left(c(x) \,|\, \mathrm{dom}\,(h)\right) = \{0\}$$

where

$$N\left(\overline{y} \,|\, C\right) := \partial\delta_C(\overline{y}) = \{z \mid \langle z, y - \overline{y}\rangle \leq 0 \ \forall\, y \in C\}$$

is the normal cone to the convex set C at $\overline{y} \in C$.

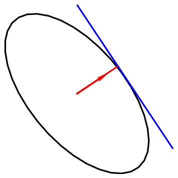
- $f = h \circ c$ satisfies the BCQ at $x \in \mathrm{dom}\,(f)$ if and only if

$$\left\{ y \in \partial h(c(x)) \,\middle|\, v = c'(x)^T y \right\} \text{ is compact } \forall\, v \in \partial f(x).$$

In the case of NLP, the BCQ is precisely the Mangasarian-Fromovitz constraint qualification (MFCQ).

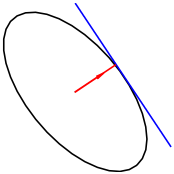
Interlude: Support Functions and Conjugacy

$\sigma_S(z) := \sup_{x \in S} \langle z, x \rangle$



Interlude: Support Functions and Conjugacy

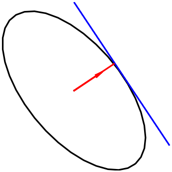
$\sigma_S(z) := \sup_{x \in S} \langle z, x \rangle$



$\sigma_S = \sigma_{\overline{\text{conv}}(S)}$

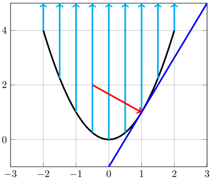
Interlude: Support Functions and Conjugacy

$$\sigma_S(z) := \sup_{x \in S} \langle z, x \rangle$$



$$\sigma_S = \sigma_{\overline{\text{conv}}(S)}$$

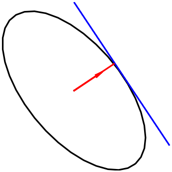
$$\text{epi}\,(\varphi) := \{(x, \mu) \mid \varphi(x) \le \mu\}$$



$$\varphi^*(z) := \sigma_{\text{epi}(\varphi)}(z, -1)$$

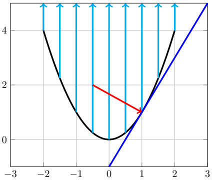
Interlude: Support Functions and Conjugacy

$\sigma_S(z) := \sup_{x \in S} \langle z, x \rangle$



$\sigma_S = \sigma_{\overline{\text{conv}}(S)}$

$\text{epi}\,(\varphi) := \{(x, \mu) \mid \varphi(x) \leq \mu\}$

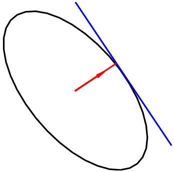


$\varphi^*(z) := \sigma_{\text{epi}(\varphi)}(z, -1)$

$$\varphi^*(z) := \sigma_{\text{epi}(\varphi)}(z, -1) = \sup_x \{\langle z, x \rangle - \varphi(x)\}$$

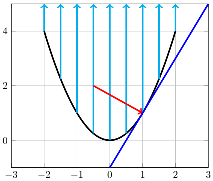
Interlude: Support Functions and Conjugacy

$$\sigma_S(z) := \sup_{x\in S}\langle z, x\rangle$$



$$\sigma_S = \sigma_{\overline{\text{conv}}(S)}$$

$$\text{epi}\,(\varphi) := \{(x,\mu)\mid \varphi(x) \le \mu\}$$



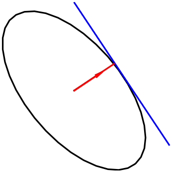
$$\varphi^*(z) := \sigma_{\text{epi}(\varphi)}(z,-1)$$

$$\varphi^*(z) := \sigma_{\text{epi}(\varphi)}(z,-1) = \sup_x\{\langle z, x\rangle - \varphi(x)\}$$

$$\varphi(x) + \varphi^*(z) \ge \langle z, x\rangle\ \forall\, x, z \quad \overset{\text{equality}}{\Longrightarrow} \quad \varphi(x) = (\varphi^*)^*(x).$$

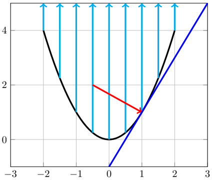
Interlude: Support Functions and Conjugacy

$\sigma_S(z) := \sup_{x \in S} \langle z, x \rangle$



$\sigma_S = \sigma_{\overline{\mathrm{conv}}(S)}$

$\mathrm{epi}\,(\varphi) := \{(x, \mu) \mid \varphi(x) \leq \mu\}$



$\varphi^*(z) := \sigma_{\mathrm{epi}(\varphi)}(z, -1)$

$$\varphi^*(z) := \sigma_{\mathrm{epi}(\varphi)}(z, -1) = \sup_x \{\langle z, x \rangle - \varphi(x)\}$$

$$\varphi(x) + \varphi^*(z) \geq \langle z, x \rangle \ \forall\, x, z \quad \overset{\text{equality}}{\Longrightarrow} \quad \varphi(x) = (\varphi^*)^*(x).$$

Bi-conjugacy: If there exists x such that $-\infty < \varphi(x) < +\infty$, then

$$\mathrm{epi}\left(\varphi^{**}\right) = \overline{\mathrm{conv}}\,(\mathrm{epi}\,(\varphi)) \quad \text{so} \quad \varphi(x) \geq \varphi^{**}(x)\ \forall\, x.$$

If, in addition, $\mathrm{epi}\,(\varphi)$ is closed and convex, then $\varphi(x) = \varphi^{**}(x)$.

The Convex-Composite Lagrangian

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} h(c(x))$$

- The Lagrangian for $\mathbf{P}$:

$$L(x, y \quad) := \langle y, c(x)\rangle - h^*(y)$$

The Convex-Composite Lagrangian

$$\mathbf{P} \qquad \min_{x \in \mathbb{R}^n} h(c(x)) \quad + g(x)$$

- The Lagrangian for $\mathbf{P}$:

$$L(x, y, v) := \langle y, c(x) \rangle - h^*(y) \quad + \langle v, x \rangle - g^*(v)$$

The Convex-Composite Lagrangian

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} h(c(x)) \quad + g(x)$$

- The Lagrangian for $\mathbf{P}$:

$$L(x,y) := \langle y, c(x)\rangle - h^*(y) \quad + g(x)$$

The Convex-Composite Lagrangian

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} h(c(x))$$

- The Lagrangian for $\mathbf{P}$: $L(x,y) := \langle y, c(x)\rangle - h^*(y)$

$$\min_x (h\circ c)(x) \;=\; \min_x \;\sup_y [\langle y, c(x)\rangle - h^*(y)] \;=\; \min_x \;\sup_y \; L(x,y)$$

The Convex-Composite Lagrangian

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} h(c(x))$$

- The Lagrangian for $\mathbf{P}$: $L(x,y) := \langle y, c(x)\rangle - h^*(y)$

$$\min_x (h\circ c)(x) \;=\; \min_x \;\sup_y [\langle y, c(x)\rangle - h^*(y)] \;=\; \min_x \;\sup_y \; L(x,y)$$

- First-Order Optimality Conditions:

$$\overline{x} \in \operatorname{argmin}_x f \;\implies\; 0 \in \partial f(\overline{x}) \iff \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial_x L(\overline{x},\overline{y}) \\ \partial_y(-L)(\overline{x},\overline{y}) \end{pmatrix}$$

In the case of NLP, the Lagrangian optimality conditions are precisely the KKT conditions.

The Convex-Composite Lagrangian

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} h(c(x))$$

- The Lagrangian for $\mathbf{P}$: $L(x,y) := \langle y, c(x)\rangle - h^*(y)$

$$\min_x (h\circ c)(x) \;=\; \min_x \;\sup_y [\langle y, c(x)\rangle - h^*(y)] \;=\; \min_x \;\sup_y \; L(x,y)$$

- First-Order Optimality Conditions:

$$\overline{x} \in \operatorname{argmin}_x f \;\implies\; 0 \in \partial f(\overline{x}) \iff \begin{pmatrix}0\\0\end{pmatrix} \in \begin{pmatrix}\partial_x L(\overline{x},\overline{y})\\ \partial_y(-L)(\overline{x},\overline{y})\end{pmatrix}$$

Rockafellar ('23) has recently introduced a notion of augmented Lagrangians for convex-composite functions and proposed an associated AL method.

Second-Order Optimality Conditions

Theorem: (B.-Poliquin '92) (Necessity) *If* $\overline{x}$ *is a local solution to* $\min_x f(x)$ *at which the BCQ is satisfied, then*

$$h''(c(\overline{x});\, c'(\overline{x})d) + \max_{y \in M(\overline{x})} d^T \nabla^2_{xx} L(\overline{x}, y) d \ \geq \ 0$$

for all $d \in \mathbb{R}^n$ *such that* $df(\overline{x})(d) \leq 0$ *where*

$$h''(c(\overline{x});\, c'(\overline{x})d) := \liminf_{u \to d,\, t \downarrow 0} \frac{h(c(\overline{x}) + tc'(\overline{x})u) - f(\overline{x}) - tdf(\overline{x})(d)}{\frac{1}{2}t^2}$$

$$M(\overline{x}) := \left\{ y \in \partial h(c(\overline{x})) \;\middle|\; c'(\overline{x})^T y = 0 \right\}.$$

Second-Order Optimality Conditions

Theorem: (B.-Poliquin '92) (Necessity) *If* $\overline{x}$ *is a local solution to* $\min_x f(x)$ *at which the BCQ is satisfied, then*

$$h''(c(\overline{x});\, c'(\overline{x})d) + \max_{y\in M(\overline{x})} d^T \nabla^2_{xx} L(\overline{x}, y)d \;\geq\; 0$$

for all $d \in \mathbb{R}^n$ *such that* $df(\overline{x})(d) \leq 0$ *where*

$$h''(c(\overline{x});\, c'(\overline{x})d) := \liminf_{u\to d,\, t\downarrow 0} \frac{h(c(\overline{x}) + tc'(\overline{x})u) - f(\overline{x}) - tdf(\overline{x})(d)}{\frac{1}{2}t^2}$$

$$M(\overline{x}) := \left\{ y \in \partial h(c(\overline{x})) \;\middle|\; c'(\overline{x})^T y = 0 \right\}.$$

Example:

$$h \in \mathcal{C}^2 \implies \nabla^2 f(x) = c'(x)^T \nabla^2 h(c(x)) c'(x) \;+\; \sum_{i=1}^{m} y_i \nabla^2 c_i(x),$$

where $y = \nabla h(c(x))$.

Second-Order Optimality Conditions

Theorem: (B.-Poliquin '92) (Necessity) *If $\overline{x}$ is a local solution to $\min_x f(x)$ at which the BCQ is satisfied, then*

$$h''(c(\overline{x});\, c'(\overline{x})d) + \max_{y\in M(\overline{x})} d^T\nabla^2_{xx}L(\overline{x},y)d \ \geq\ 0$$

for all $d\in\mathbb{R}^n$ such that $df(\overline{x})(d)\leq 0$ where

$$h''(c(\overline{x});\, c'(\overline{x})d) := \liminf_{u\to d,\, t\downarrow 0} \frac{h(c(\overline{x})+tc'(\overline{x})u) - f(\overline{x}) - tdf(\overline{x})(d)}{\frac{1}{2}t^2}$$

$$M(\overline{x}) := \left\{ y\in\partial h(c(\overline{x})) \;\middle|\; c'(\overline{x})^T y = 0 \right\}.$$

Theorem: (Rockafellar '89) (Sufficiency) *Suppose that h is a piecewise linear-quadratic function. If $\overline{x}$ is such that $0\in\partial f(\overline{x})$ and*

$$h''(c(\overline{x});\, c'(\overline{x})d) + \max_{y\in M(\overline{x})} d^T\nabla^2_{xx}L(\overline{x},y)d \ >\ 0$$

for all $d\neq 0$ such that $df(\overline{x})(d)\leq 0$, then there is an $\alpha>0$ such that $f(x)\geq f(\overline{x})+\alpha\|x-\overline{x}\|_2^2$ for all x near $\overline{x}$.

Second-Order Optimality Conditions

Theorem: (B.-Poliquin '92) (Necessity) *If $\overline{x}$ is a local solution to $\min_x f(x)$ at which the BCQ is satisfied, then*

$$h''(c(\overline{x});\, c'(\overline{x})d) + \max_{y \in M(\overline{x})} d^T \nabla^2_{xx} L(\overline{x}, y) d \ \geq \ 0$$

for all $d \in \mathbb{R}^n$ such that $df(\overline{x})(d) \leq 0$ where

$$h''(c(\overline{x});\, c'(\overline{x})d) := \liminf_{u \to d,\, t \downarrow 0} \frac{h(c(\overline{x}) + tc'(\overline{x})u) - f(\overline{x}) - tdf(\overline{x})(d)}{\frac{1}{2}t^2}$$

$$M(\overline{x}) := \left\{ y \in \partial h(c(\overline{x})) \;\middle|\; c'(\overline{x})^T y = 0 \right\}.$$

Theorem: (Rockafellar '89) (Sufficiency) *Suppose that h is a piecewise linear-quadratic function. If $\overline{x}$ is such that $0 \in \partial f(\overline{x})$ and*

$$h''(c(\overline{x});\, c'(\overline{x})d) + \max_{y \in M(\overline{x})} d^T \nabla^2_{xx} L(\overline{x}, y) d \ > \ 0$$

for all $d \neq 0$ such that $df(\overline{x})(d) \leq 0$, then there is an $\alpha > 0$ such that $f(x) \geq f(\overline{x}) + \alpha \|x - \overline{x}\|_2^2$ for all x near $\overline{x}$.

Mohammadi and Sarabi '20 use Rockafellar's notion of *parabolic regularity* '85 and metric subregularity to give a new approach to the necessity theorem and extend the sufficiency theorem.

"Exactness" and the Pasch-Hausdorff Envelope

Pasch-Hausdorff Envelope:

$$h_\alpha(y) := \inf_w [h(w) + \alpha \|y - w\|]$$

h_α is finite-valued and globally α-Lipschitz.

“Exactness” and the Pasch-Hausdorff Envelope

Pasch-Hausdorff Envelope:

$$h_\alpha(y) := \inf_w [h(w) + \alpha\|y - w\|]$$

h_α is finite-valued and globally α-Lipschitz.

Example:

$h(y) := \delta_\Omega(y) \implies h_\alpha(y) := \alpha \inf_{w\in\Omega} \|y - w\| = \alpha \mathrm{dist}\,(y\,|\Omega\,)$.

“Exactness” and the Pasch-Hausdorff Envelope

Pasch-Hausdorff Envelope:

$$h_\alpha(y) := \inf_w [h(w) + \alpha \|y - w\|]$$

h_α is finite-valued and globally α-Lipschitz.

Define $f_\alpha(x) := h_\alpha(c(x))$.

"Exactness" and the Pasch-Hausdorff Envelope

Pasch-Hausdorff Envelope:

$$h_\alpha(y) := \inf_w [h(w) + \alpha\|y - w\|]$$

h_α is finite-valued and globally α-Lipschitz.

Define $f_\alpha(x) := h_\alpha(c(x))$.

Exactness: Does $\operatorname{argmin} f \ = \ \operatorname{argmin} f_\alpha$?

"Exactness" and the Pasch-Hausdorff Envelope

Pasch-Hausdorff Envelope:

$$h_\alpha(y) := \inf_w [h(w) + \alpha\|y - w\|]$$

h_α is finite-valued and globally α-Lipschitz.

Define $f_\alpha(x) := h_\alpha(c(x))$.

Theorem:(B.-Poliquin '92)
If $\overline{x}$ is a local solution to $\min_x f(x)$ at which c is locally Lipschitz and the BCQ is satisfied, then there is an $\bar{\alpha} > 0$ such that $\overline{x}$ is a local solution to $\min_x f_\alpha(x)$ with $f(\overline{x}) = f_\alpha(\overline{x})$ for all $\alpha > \bar{\alpha}$.

“Exactness” and the Pasch-Hausdorff Envelope

Pasch-Hausdorff Envelope:

$$h_\alpha(y) := \inf_w [h(w) + \alpha\|y - w\|]$$

h_α is finite-valued and globally α-Lipschitz.

Define $f_\alpha(x) := h_\alpha(c(x))$.

Theorem:(B.-Poliquin ’92)
If $\overline{x}$ is a local solution to $\min_x f(x)$ at which c is locally Lipschitz and the BCQ is satisfied, then there is an $\bar{\alpha} > 0$ such that $\overline{x}$ is a local solution to $\min_x f_\alpha(x)$ with $f(\overline{x}) = f_\alpha(\overline{x})$ for all $\alpha > \bar{\alpha}$.

NLP exact penalization as well as other exact penalization results for this class follow from this theorem since $(\delta_\Omega)_\alpha(x) = \alpha \mathrm{dist}\,(y \,|\Omega\,)$.

When is a convex-composite function convex?

Observe that

$$f(x) = h(c(x)) = h^{**}(c(x)) = \sup_y [\langle y, c(x) \rangle - h^*(y)].$$

,

When is a convex-composite function convex?

Observe that

$$f(x) = h(c(x)) = h^{**}(c(x)) = \sup_y [\langle y, c(x)\rangle - h^*(y)].$$

So f is convex if $\langle y, c\rangle(\cdot)$ is convex for all $y \in \operatorname{dom}\left(h^*\right)$,

,

When is a convex-composite function convex?

Observe that

$$f(x) = h(c(x)) = h^{**}(c(x)) = \sup_y [\langle y, c(x)\rangle - h^*(y)].$$

So f is convex if $\langle y, c\rangle(\cdot)$ is convex for all $y \in \mathrm{dom}\left(h^*\right)$, i.e.,

$\forall\, y \in K := \mathbb{R}_+ \mathrm{dom}\left(h^*\right),\ u, v \in \mathbb{R}^n$, $\lambda \in [0,1]$

$$\langle y, c\rangle((1-\lambda)u + \lambda v) \leq (1-\lambda)\langle y, c\rangle(u) + \lambda\langle y, c\rangle(v)$$

,

When is a convex-composite function convex?

Observe that

$$f(x) = h(c(x)) = h^{**}(c(x)) = \sup_y [\langle y, c(x)\rangle - h^*(y)].$$

So f is convex if $\langle y, c\rangle(\cdot)$ is convex for all $y \in \operatorname{dom}\left(h^*\right)$, i.e.,

$\forall\, y \in K := \mathbb{R}_+ \operatorname{dom}\left(h^*\right),\ u, v \in \mathbb{R}^n,\ \lambda \in [0,1]$

$$\langle y, c\rangle((1-\lambda)u + \lambda v) \leq (1-\lambda)\langle y, c\rangle(u) + \lambda \langle y, c\rangle(v)$$

$$\Longleftrightarrow$$

$$\langle y, c((1-\lambda)u + \lambda v)) - [(1-\lambda)c(u) + \lambda c(v)]\rangle \leq 0$$

,

When is a convex-composite function convex?

Observe that

$$f(x) = h(c(x)) = h^{**}(c(x)) = \sup_y [\langle y, c(x)\rangle - h^*(y)].$$

So f is convex if $\langle y, c\rangle(\cdot)$ is convex for all $y \in \mathrm{dom}\left(h^*\right)$, i.e.,

$\forall\, y \in K := \mathbb{R}_+ \mathrm{dom}\left(h^*\right),\ u, v \in \mathbb{R}^n,\ \lambda \in [0,1]$

$$\langle y, c\rangle((1-\lambda)u + \lambda v) \leq (1-\lambda)\langle y, c\rangle(u) + \lambda\langle y, c\rangle(v)$$

$$\Longleftrightarrow$$

$$\langle y, c((1-\lambda)u + \lambda v)) - [(1-\lambda)c(u) + \lambda c(v)]\rangle \leq 0$$

$$\Longleftrightarrow$$

$$c((1-\lambda)u + \lambda v) - [(1-\lambda)c(u) + \lambda c(v)] \in K^\circ$$

,

When is a convex-composite function convex?

Observe that

$$f(x) = h(c(x)) = h^{**}(c(x)) = \sup_y [\langle y, c(x)\rangle - h^*(y)].$$

So f is convex if $\langle y, c\rangle(\cdot)$ is convex for all $y \in \operatorname{dom}\left(h^*\right)$, i.e.,

$\forall\, y \in K := \mathbb{R}_+ \operatorname{dom}\left(h^*\right),\ u, v \in \mathbb{R}^n,\ \lambda \in [0, 1]$

$$\langle y, c\rangle((1-\lambda)u + \lambda v) \leq (1-\lambda)\langle y, c\rangle(u) + \lambda\langle y, c\rangle(v)$$

$$\Longleftrightarrow$$

$$\langle y, c((1-\lambda)u + \lambda v)) - [(1-\lambda)c(u) + \lambda c(v)]\rangle \leq 0$$

$$\Longleftrightarrow$$

$$c((1-\lambda)u + \lambda v) - [(1-\lambda)c(u) + \lambda c(v)] \in K^\circ$$

$$\Longleftrightarrow$$

c is concave wrt K° ,

When is a convex-composite function convex?

Observe that

$$f(x) = h(c(x)) = h^{**}(c(x)) = \sup_y [\langle y, c(x)\rangle - h^*(y)].$$

So f is convex if $\langle y, c\rangle(\cdot)$ is convex for all $y \in \mathrm{dom}\,(h^*)$, i.e.,

$\forall\, y \in K := \mathbb{R}_+ \mathrm{dom}\,(h^*)\,,\ u, v \in \mathbb{R}^n,\ \lambda \in [0,1]$

$$\langle y, c\rangle((1-\lambda)u + \lambda v) \leq (1-\lambda)\langle y, c\rangle(u) + \lambda\langle y, c\rangle(v)$$

$$\Longleftrightarrow$$

$$\langle y, c((1-\lambda)u + \lambda v)) - [(1-\lambda)c(u) + \lambda c(v)]\rangle \leq 0$$

$$\Longleftrightarrow$$

$$c((1-\lambda)u + \lambda v) - [(1-\lambda)c(u) + \lambda c(v)] \in K^\circ$$

$$\Longleftrightarrow$$

c is concave wrt $K^\circ = \mathrm{hzn}\,(h)\,,$

where $\mathrm{hzn}\,(h) := \{ z \mid h(x + \lambda z) \leq h(x)\ \forall x \in \mathrm{dom}\,(h),\ \lambda > 0 \}$.

Convex convex-composite functions

Theorem:(B.-Hoheisel-Nguyen '21)
If $c : \Omega \to \mathbb{R}^m$ *is convex wrt* $(-\mathrm{hzn}\,(h))$, *then* $f = h \circ c$ *is convex.*

If, in addition,

$$c(\mathrm{ri}\,(\Omega) \cap \mathrm{ri}\,(\mathrm{dom}\,(h)) \neq \emptyset,$$

then

$$(h \circ c)^*(p) = \min_{v \in \mathbb{R}^m} h^*(v) + \langle v, c(\cdot)\rangle^*(p)$$

and

$$\partial(h \circ c)(\bar{x}) = \bigcup_{v \in \partial h(c(\bar{x}))} \partial\langle v, c(\cdot)\rangle(\bar{x}) \quad (\bar{x} \in \mathrm{dom}\,(h \circ c)).$$

Convex convex-composite functions

Theorem:(B.-Hoheisel-Nguyen '21)
If $c : \Omega \to \mathbb{R}^m$ *is convex wrt* $(-\mathrm{hzn}\,(h))$, *then* $f = h \circ c$ *is convex.*

If, in addition,

$$c(\mathrm{ri}\,(\Omega) \cap \mathrm{ri}\,(\mathrm{dom}\,(h)) \neq \emptyset,$$

then

$$(h \circ c)^*(p) = \min_{v \in \mathbb{R}^m} h^*(v) + \langle v, c(\cdot) \rangle^*(p)$$

and

$$\partial(h \circ c)(\bar{x}) = \bigcup_{v \in \partial h(c(\bar{x}))} \partial \langle v, c(\cdot) \rangle(\bar{x}) \quad (\bar{x} \in \mathrm{dom}\,(h \circ c)).$$

Borwein '74, Bot-Wanka-Grad-Hodrea '06-'10,
Combari-Lagdhir-Thibault '94, Pennanen '99

Convex convex-composite functions

Theorem:(B.-Hoheisel-Nguyen '21)
If $c : \Omega \to \mathbb{R}^m$ *is convex wrt* $(-\mathrm{hzn}\,(h))$, *then* $f = h \circ c$ *is convex.*

If, in addition,

$$c(\mathrm{ri}\,(\Omega) \cap \mathrm{ri}\,(\mathrm{dom}\,(h)) \neq \emptyset,$$

then

$$(h \circ c)^*(p) = \min_{v \in \mathbb{R}^m} h^*(v) + \langle v, c(\cdot)\rangle^*(p)$$

and

$$\partial(h \circ c)(\bar{x}) = \bigcup_{v \in \partial h(c(\bar{x}))} \partial\langle v, c(\cdot)\rangle(\bar{x}) \quad (\bar{x} \in \mathrm{dom}\,(h \circ c)).$$

Applications: conic programming, Kiefer-Gaffe-Krafft inequalities, matrix-fractional functions, variational Gram functions, spectral functions, generalized Farkas theorems, ...

Algorithms

$$\mathbf{P}_k \qquad \min_{\left\|x-x^k\right\|\leq\eta_k} h\left(c(x^k)+\nabla c(x^k)[x-x^k]\right)+\frac{1}{2}(x-x^k)^\top H_k(x-x^k),$$

Algorithms

$$\mathbf{P}_k \qquad \min_{\left\|x-x^k\right\|\leq\eta_k} h\left(c(x^k)+\nabla c(x^k)[x-x^k]\right)+\frac{1}{2}(x-x^k)^\top H_k(x-x^k),$$

- Newton-like method: $H_k \approx \nabla^2_{xx} L(x^k, y^k)$

Algorithms

$$\mathbf{P}_k \qquad \min_{\left\|x-x^k\right\| \leq \eta_k} h\left(c(x^k)+\nabla c(x^k)[x-x^k]\right)+\frac{1}{2}(x-x^k)^\top H_k(x-x^k),$$

- Newton-like method: $H_k \approx \nabla^2_{xx} L(x^k, y^k)$
- Prox-linear method: $H_k = \alpha_k I$

Algorithms

$$\mathbf{P}_k \qquad \min_{\left\| x - x^k \right\| \le \eta_k} h\left(c(x^k) + \nabla c(x^k)[x - x^k] \right) + \frac{1}{2}(x - x^k)^\top H_k (x - x^k),$$

- Newton-like method: $H_k \approx \nabla^2_{xx} L(x^k, y^k)$
- Prox-linear method: $H_k = \alpha_k I$
- $\mathbf{P}_k$ may or may not be convex depending on whether $H_k \succeq 0$.

Algorithm for NLP

NLP minimize $\phi(x)$

subject to $f_i(x) = 0,\ i = 1, \ldots, s,\ f_i(x) \leq 0,\ i = s+1, \ldots, m.$

Algorithm for NLP

NLP minimize $\phi(x)$

subject to $f_i(x) = 0, \; i = 1, \ldots, s, \; f_i(x) \leq 0, \; i = s+1, \ldots, m.$

- Convex-Composite Framework

$$h(\mu, y) = \mu + \delta_K(y), \qquad K := \{0\}^s \times \mathbb{R}_-^{m-s}$$

$$c(x) = (\phi(x), \hat{c}(x))$$

$$L(x, y) = \phi(x) + \sum_{k=1}^{m} y_i \hat{c}_i(x) \; - \delta_{K^\circ}(y), \quad K^\circ = \mathbb{R}^s \times \mathbb{R}_+^{m-s}$$

Algorithm for NLP

NLP minimize $\phi(x)$

subject to $f_i(x)=0,\ i=1,\ldots,s,\ f_i(x)\leq 0,\ i=s+1,\ldots,m.$

- Convex-Composite Framework

$$h(\mu,y)=\mu+\delta_K(y), \qquad K:=\{0\}^s\times\mathbb{R}_-^{m-s}$$

$$c(x)=(\phi(x),\hat{c}(x))$$

$$L(x,y)=\phi(x)+\sum_{k=1}^{m}y_i\hat{c}_i(x)\ -\delta_{K^\circ}(y), \quad K^\circ=\mathbb{R}^s\times\mathbb{R}_+^{m-s}$$

- Subproblems:

$\mathbf{P_k}$ minimize $\phi(x^k)+\nabla\phi(x^k)^T(x-x^k)+\frac{1}{2}[x-x^k]^\top H_k[x-x^k]$

subject to $\hat{c}_i(x^k)+\nabla\hat{c}_i(x^k)^T(x-x^k)=0,\ i=1,\ldots,s$

$\hat{c}_i(x^k)+\nabla\hat{c}_i(x^k)^T(x-x^k)\leq 0,\ i=s+1,\ldots,m.$

Algorithm for NLP

NLP minimize $\phi(x)$

subject to $f_i(x)=0,\ i=1,\ldots,s,\ f_i(x)\leq 0,\ i=s+1,\ldots,m.$

- Convex-Composite Framework

$$h(\mu,y)=\mu+\delta_K(y), \qquad K:=\{0\}^s\times\mathbb{R}_-^{m-s}$$

$$c(x)=(\phi(x),\hat{c}(x))$$

$$L(x,y)=\phi(x)+\sum_{k=1}^{m}y_i\hat{c}_i(x)\ -\delta_{K^\circ}(y), \quad K^\circ=\mathbb{R}^s\times\mathbb{R}_+^{m-s}$$

- Subproblems: Sequential quadratic programming (SQP)

$\mathbf{P_k}$ minimize $\phi(x^k)+\nabla\phi(x^k)^T(x-x^k)+\frac{1}{2}[x-x^k]^\top H_k[x-x^k]$

subject to $\hat{c}_i(x^k)+\nabla\hat{c}_i(x^k)^T(x-x^k)=0,\ i=1,\ldots,s$

$\hat{c}_i(x^k)+\nabla\hat{c}_i(x^k)^T(x-x^k)\leq 0,\ i=s+1,\ldots,m.$

The Sharp Case

The set $C := \operatorname{argmin} h$ is said to be a set of *sharp minima* for h if

$$\exists\, \alpha > 0 \quad s.t. \quad h(c) \geq h_{\min} + \alpha \operatorname{dist}(c \,|\, C) \;\; \forall\, c \in \mathbb{R}^m.$$

The Sharp Case

The set $C := \operatorname{argmin} h$ is said to be a set of *sharp minima* for h if

$$\exists\, \alpha > 0 \quad s.t. \quad h(c) \geq h_{\min} + \alpha \operatorname{dist}\left(c \,|\, C\right) \;\; \forall\, c \in \mathbb{R}^m.$$

Consider the following algorithm with $\Delta > 0$:

$$x^{k+1} \quad \text{solves} \quad \min_{\left\|x - x^k\right\| \leq \Delta} h(c(x^k) + c'(x^k)(x - x^k)).$$

The Sharp Case

The set $C := \operatorname{argmin} h$ is said to be a set of *sharp minima* for h if

$$\exists\, \alpha > 0 \quad s.t. \quad h(c) \geq h_{\min} + \alpha \operatorname{dist}(c \,|\, C\,) \;\; \forall\, c \in \mathbb{R}^m.$$

Consider the following algorithm with $\Delta > 0$:

$$x^{k+1} \quad \text{solves} \quad \min_{\left\|x - x^k\right\| \leq \Delta} h(c(x^k) + c'(x^k)(x - x^k)).$$

Theorem:(B.-Ferris '95) *If* $\{x^k\}$ *is generated by the algorithm above with* x^0 *such that* $c(x^0)$ *is sufficiently close to* C *and*

$$\ker(c'(x^0)^T) \cap \left[\mathbb{R}_+(C - c(x^0))\right]^\circ = \{0\},$$

then there exists $\overline{x}$ *such that* $c(\overline{x}) \in C$ *with* $x^k \to \overline{x}$ *at a quadratic rate.*

The Sharp Case

The set $C := \operatorname{argmin} h$ is said to be a set of *sharp minima* for h if

$$\exists\, \alpha > 0 \quad s.t. \quad h(c) \geq h_{\min} + \alpha \mathrm{dist}\,(c\,|C\,) \;\; \forall\, c \in \mathbb{R}^m.$$

Consider the following algorithm with $\Delta > 0$:

$$x^{k+1} \quad \text{solves} \quad \min_{\left\|x - x^k\right\| \leq \Delta} h(c(x^k) + c'(x^k)(x - x^k)).$$

Theorem:(B.-Ferris '95) *If* $\{x^k\}$ *is generated by the algorithm above with* x^0 *such that* $c(x^0)$ *is sufficiently close to* C *and*

$$\ker(c'(x^0)^T) \cap \left[\mathbb{R}_+(C - c(x^0))\right]^\circ = \{0\},$$

then there exists $\overline{x}$ *such that* $c(\overline{x}) \in C$ *with* $x^k \to \overline{x}$ *at a quadratic rate.*

Li-Wang '02 use the same proof technique but slightly weaken the sharpness hypothsis.

Newton's Method in General: Hypotheses

Assume h is convex piecewise linear-quadratic (PLQ), i.e.,

$\mathrm{dom}\,(h) = \bigcup_{i=1}^{N} C_i$ with each C_i convex polyhedral, and

$h(z) = \frac{1}{2}\langle z, Q_k z\rangle + \langle b_k, z\rangle + \beta_k$ on C_i with $Q_k \in \mathbb{S}^m$.

Newton's Method in General: Hypotheses

Assume h is convex piecewise linear-quadratic (PLQ), i.e., $\mathrm{dom}\,(h) = \bigcup_{i=1}^{N} C_i$ with each C_i convex polyhedral, and $h(z) = \frac{1}{2}\langle z, Q_k z\rangle + \langle b_k, z\rangle + \beta_k$ on C_i with $Q_k \in \mathbb{S}^m$.

$(\overline{x}, \overline{y})$ a primal-dual optimal pair for $\min f = h \circ c$.

Newton's Method in General: Hypotheses

Assume h is convex piecewise linear-quadratic (PLQ), i.e.,

$\operatorname{dom}(h) = \bigcup_{i=1}^{N} C_i$ with each C_i convex polyhedral, and

$h(z) = \frac{1}{2}\langle z, Q_k z\rangle + \langle b_k, z\rangle + \beta_k$ on C_i with $Q_k \in \mathbb{S}^m$.

$(\overline{x}, \overline{y})$ a primal-dual optimal pair for $\min f = h \circ c$.

Assume $c \in \mathcal{C}^3$ and $(\overline{x}, \overline{y})$ satisfy NLP-like conditions:

LICQ,

strict complementarity, and

second-order sufficiency.

Newton's Method in General: Hypotheses

Assume h is convex piecewise linear-quadratic (PLQ), i.e.,
$\operatorname{dom}(h) = \bigcup_{i=1}^{N} C_i$ with each C_i convex polyhedral, and
$h(z) = \frac{1}{2}\langle z, Q_k z\rangle + \langle b_k, z\rangle + \beta_k$ on C_i with $Q_k \in \mathbb{S}^m$.

$(\overline{x}, \overline{y})$ a primal-dual optimal pair for $\min f = h \circ c$.

Assume $c \in \mathcal{C}^3$ and $(\overline{x}, \overline{y})$ satisfy NLP-like conditions:

- LICQ,
- strict complementarity, and
- second-order sufficiency.

In the case of NLP, these assumptions reduce the usual NLP assumptions.

Convergence of Newton's Method

Theorem: (B.-Engle '19) If (x^0, y^0) is sufficiently close to $(\overline{x}, \overline{y})$, then the Newton sequence $\{(x^k, y^k)\}$ satisfies

(i) $c(x^{k-1}) + \nabla c(x^{k-1})(x^k - x^{k-1}) \in$ active manifold (active constr. ID),

(ii) $y^k \in \mathrm{ri}\left(\partial h(c(x^{k-1}) + \nabla c(x^{k-1})(x^k - x^{k-1}))\right)$ (str. compl.),

(iii) $$\begin{array}{rl} y^k & \in \partial h(c(x^k) + c'(x^k)(x^k - x^{k-1}) \\ 0 & = \nabla c(x^{k-1})^\top y^k + \nabla^2_{xx} L(x^k, y^k)(x^k - x^{k-1}) \end{array}$$ (1st-order opt.),

(iv) x^{k+1} is a strong local minimizer of $\mathbf{P_k}$ (2nd order suff.),

(v) $(x^k, y^k) \to (\overline{x}, \overline{y})$ at a quadratic rate.

Convergence of Newton's Method

Theorem: (B.-Engle '19) If (x^0, y^0) is sufficiently close to $(\overline{x}, \overline{y})$, then the Newton sequence $\{(x^k, y^k)\}$ satisfies

(i) $c(x^{k-1}) + \nabla c(x^{k-1})(x^k - x^{k-1}) \in$ active manifold (active constr. ID),

(ii) $y^k \in \mathrm{ri}\left(\partial h(c(x^{k-1}) + \nabla c(x^{k-1})(x^k - x^{k-1}))\right)$ (str. compl.),

(iii) $\begin{array}{cl} y^k & \in \partial h(c(x^k) + c'(x^k)(x^k - x^{k-1}) \\ 0 & = \nabla c(x^{k-1})^\top y^k + \nabla^2_{xx} L(x^k, y^k)(x^k - x^{k-1}) \end{array}$ (1st-order opt.),

(iv) x^{k+1} is a strong local minimizer of $\mathbf{P_k}$ (2nd order suff.),

(v) $(x^k, y^k) \to (\overline{x}, \overline{y})$ at a quadratic rate.

Proof uses Robinson's *generalized equations*, Rockafellar's PLQ 2^{nd}-order theory, metric subregularity, and Lewis' *partial smoothness* techniques.

Globalization and descent

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} f(x) := h(c(x)) + g(x),$$

where $h : \mathbb{R}^m \to \mathbb{R}$ convex, $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ proper, convex, loc. Lipschitz relative to $\mathrm{dom}\,(g)$, and $c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^1$.

$$\mathbf{P}_k \qquad \min_{\|d\|\leq\eta_k} h(c(x^k) + \nabla c(x^k)d) + \frac{1}{2}d^T H_k d + g(x^k + d)$$

Globalization and descent

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} f(x) := h(c(x)) + g(x),$$

where $h : \mathbb{R}^m \to \mathbb{R}$ convex, $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ proper, convex, loc. Lipschitz relative to $\operatorname{dom}(g)$, and $c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^1$.

$$\mathbf{P}_k \qquad \min_{\|d\|\le\eta_k} h(c(x^k) + \nabla c(x^k)d) + \frac{1}{2}d^T H_k d + g(x^k + d)$$

Define

$$\Delta f(x; d) := h(c(x) + \nabla c(x)d) + \frac{1}{2}d^T H_k d + g(x + d) - f(x).$$

Globalization and descent

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} f(x) := h(c(x)) + g(x),$$

where $h : \mathbb{R}^m \to \mathbb{R}$ convex, $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ proper, convex, loc. Lipschitz relative to $\operatorname{dom}(g)$, and $c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^1$.

$$\mathbf{P}_k \qquad \min_{\|d\|\le\eta_k} h(c(x^k) + \nabla c(x^k)d) + \frac{1}{2}d^T H_k d + g(x^k + d)$$

Define

$$\Delta f(x; d) := h(c(x) + \nabla c(x)d) + \frac{1}{2}d^T H_k d + g(x + d) - f(x).$$

Recall that

$$f'(x; d) = \lim_{t\downarrow 0} \tfrac{\Delta f(x;td)}{t} = \inf_{t>0} \tfrac{\Delta f(x;td)}{t}.$$

Backtracking, Weak Wolfe, Trust Regions

(B. –Engle '19)

Assume $f'(x;d) \leq \Delta f(x;d) \leq \tau \min_{\|d\| \leq \eta} \Delta f(x;d) < 0$ for $\tau \in (0,1)$.

Backtracking, Weak Wolfe, Trust Regions

(B. –Engle '19)
Assume $f'(x;d) \leq \Delta f(x;d) \leq \tau \min_{\|d\| \leq \eta} \Delta f(x;d) < 0$ for $\tau \in (0,1)$.

Backtracking: With $\sigma \in (0,1)$ choose $t > 0$ to satisfy

$$f(x+td) > f(x) + \sigma t \Delta f(x;d).$$

Backtracking, Weak Wolfe, Trust Regions

(B. –Engle '19)
Assume $f'(x;d) \le \Delta f(x;d) \le \tau \min_{\|d\|\le\eta} \Delta f(x;d) < 0$ for $\tau \in (0,1)$.

Backtracking: With $\sigma \in (0,1)$ choose $t > 0$ to satisfy

$$f(x+td) > f(x) + \sigma t \Delta f(x;d).$$

Weak Wolfe: With $0<\sigma_1<\sigma_2<1$ choose $t>0$ to satisfy

$$\begin{array}{ll} \mathrm{WW1} & f(x+td) \le f(x) + \sigma_1 t \Delta f(x;d), \text{ and} \\ \mathrm{WW2} & \sigma_2 \Delta f(x;d) \le \Delta f(x+td;d)\ . \end{array}$$

Backtracking, Weak Wolfe, Trust Regions

(B. –Engle '19)
Assume $f'(x;d) \le \Delta f(x;d) \le \tau \min_{\|d\|\le\eta} \Delta f(x;d) < 0$ for $\tau \in (0,1)$.

Backtracking: With $\sigma \in (0,1)$ choose $t > 0$ to satisfy

$$f(x+td) > f(x) + \sigma t \Delta f(x;d).$$

Weak Wolfe: With $0<\sigma_1<\sigma_2<1$ choose $t>0$ to satisfy

$$\text{WW1} \qquad f(x+td) \le f(x) + \sigma_1 t \Delta f(x;d), \text{ and}$$

$$\text{WW2} \qquad \sigma_2 \Delta f(x;d) \le \Delta f(x+td;d) \;.$$

Trust Region: With $\|d\| \le \delta$ and
$0 < \gamma_1 \le \gamma_2 < 1 \le \gamma_3, 0 < \beta_1 \le \beta_2 < \beta_3 < 1$ update δ as follows:

$$r = [f(x+d) - f(x)]/[\Delta f(x;d)]$$

$$\delta \in \begin{cases} [\delta,\ \gamma_3\delta] & \text{, if } r > \beta_3, \\ \{\delta\} & \text{, if } \beta_2 \le r \le \beta_3, \\ [\gamma_1\delta,\ \gamma_2\delta] & \text{, if } r < \beta_2. \end{cases}$$

Global Convergence: $x^{k+1} := x^k + \tau_k d^k$

- **Backtracking**: $\sum_{k=0}^{\infty} \frac{\Delta f(x^k; d^k)^2}{\left\|d^k\right\|_2^2} < \infty$, in particular, $\Delta f(x^k; d^k) \to 0$.

Global Convergence: $x^{k+1} := x^k + \tau_k d^k$

- **Backtracking**: $\sum_{k=0}^{\infty} \frac{\Delta f(x^k; d^k)^2}{\left\|d^k\right\|_2^2} < \infty$, in particular,

  $\Delta f(x^k; d^k) \to 0$.

- **Weak Wolfe**: $\sum_{k=0}^{\infty} \frac{\Delta f(x^k; d^k)^2}{\left\|d^k\right\| + \left\|d^k\right\|^2} < \infty$, in particular,

  $\Delta f(x^k; d^k) \to 0$.

Global Convergence: $x^{k+1} := x^k + \tau_k d^k$

- **Backtracking**: $\sum_{k=0}^{\infty} \frac{\Delta f(x^k; d^k)^2}{\left\|d^k\right\|_2^2} < \infty$, in particular, $\Delta f(x^k; d^k) \to 0$.

- **Weak Wolfe**: $\sum_{k=0}^{\infty} \frac{\Delta f(x^k; d^k)^2}{\left\|d^k\right\| + \left\|d^k\right\|^2} < \infty$, in particular, $\Delta f(x^k; d^k) \to 0$.

- **Trust Region**: $\Delta f(x^k; d^k) \to 0$.

Global Convergence: $x^{k+1} := x^k + \tau_k d^k$

- **Backtracking**: $\displaystyle\sum_{k=0}^{\infty} \frac{\Delta f(x^k; d^k)^2}{\left\|d^k\right\|_2^2} < \infty$, in particular, $\Delta f(x^k; d^k) \to 0$.

- **Weak Wolfe**: $\displaystyle\sum_{k=0}^{\infty} \frac{\Delta f(x^k; d^k)^2}{\left\|d^k\right\| + \left\|d^k\right\|^2} < \infty$, in particular, $\Delta f(x^k; d^k) \to 0$.

- **Trust Region**: $\Delta f(x^k; d^k) \to 0$.

In all cases, cluster points $\overline{x}$ satisfy $0 \in \partial f(\overline{x})$.

Complexity: Drusvyatskiy-Paquette '18

Inexact Prox-Linear Algorithms:

- Additional Assumptions:

(i) h is L-Lipschitz: $\|h(u) - h(v)\| \leq L\|u - v\| \quad \forall\, u, v \in \mathbb{R}^m$.

(ii) c is β-Lipschitz. $\|c(x) - h(z)\| \leq \beta\|x - z\| \quad \forall\, x, z \in \mathbb{R}^n$.

Complexity: Drusvyatskiy-Paquette '18

Inexact Prox-Linear Algorithms:

- Additional Assumptions:
  - **(i)** h is L-Lipschitz:
  - **(ii)** c is β-Lipschitz.
- Prox-Linear ingredients:

$$S_t(x) := \operatorname*{argmin}_z f_t(z;x) := h(c(x) + \nabla c(x)(z-x)) + g(z) + \frac{1}{2t}\|z-x\|_2^2$$

$$\mathcal{G}_t(x) := t^{-1}(x - S_t(x))$$

optimality $\implies \mathcal{G}_t(\overline{x}) = 0 \quad \forall\, t > 0$

Complexity: Drusvyatskiy-Paquette '18

Inexact Prox-Linear Algorithms:

- Additional Assumptions:

(i) h is L-Lipschitz:

(ii) c is β-Lipschitz.

- Prox-Linear ingredients:

$$S_t(x) := \operatorname*{argmin}_z f_t(z;x) := h(c(x) + \nabla c(x)(z-x)) + g(z) + \frac{1}{2t}\|z-x\|_2^2$$

$$\mathcal{G}_t(x) := t^{-1}(x - S_t(x))$$

optimality $\implies \mathcal{G}_t(\overline{x}) = 0 \quad \forall\, t > 0$

- Algorithm: $x^{k+1} \approx S_t(x^k)$ (or an ϵ_k-approx. min of $f_t(z;x^k)$)

Complexity: Drusvyatskiy-Paquette '18

Inexact Prox-Linear Algorithms:

- Additional Assumptions:
  - **(i)** h is L-Lipschitz:
  - **(ii)** c is β-Lipschitz.
- Prox-Linear ingredients:

$$S_t(x) := \operatorname*{argmin}_z f_t(z;x) := h(c(x) + \nabla c(x)(z-x)) + g(z) + \frac{1}{2t}\|z-x\|_2^2$$

$$\mathcal{G}_t(x) := t^{-1}(x - S_t(x))$$

optimality $\implies \mathcal{G}_t(\overline{x}) = 0 \quad \forall\, t > 0$

- Algorithm: $x^{k+1} \approx S_t(x^k)$ (or an ϵ_k-approx. min of $f_t(z;x^k)$)
- Convergence: If $t < (L\beta)^{-1}$, then

$$\min_{j=1,\ldots,N} \left\|\mathcal{G}_t(x^j)\right\|_2^2 \leq \frac{2(f(x^0) - \hat{f} + \sum_{j=1}^N \epsilon_j)}{tN}$$

where $\hat{f} := \liminf_k f(x^k)$.

Stochastic Prox Linear

Duchi-Ruan '17, Davis-Drusvyatskiy '19

$$f(x) = \mathbb{E}_{\xi \sim P}[h(c(x, \xi), \xi)] + g(x),$$

Stochastic Prox Linear

Duchi-Ruan '17, Davis-Drusvyatskiy '19

$$f(x) = \mathbb{E}_{\xi \sim P}[h(c(x,\xi),\xi)] + g(x),$$

Input: $x^0 \in \mathbb{R}^n$, $\bar{\rho} > \rho$ where $h \circ c + g$ is ρ-weakly convex, $\gamma > 0$, an iteration count T.

Step: $t = 1, 2, \ldots, T$

$$\left\{ \begin{array}{l} \text{Sample } \xi_t \sim P \\ \\ \beta_t = \bar{\rho} + \gamma^{-1}\sqrt{T+1} \\ \\ \text{Set} \\ x^{t+1} = \operatorname{argmin}_x \left\{ r(x) + h(c(x^t,\xi_t) + c'(x^t,\xi_t)(x - x^t), \xi_t) + \frac{\beta_t}{2}\left\|x - x^t\right\|_2^2 \right\} \end{array} \right\}$$

Sample: $t^* \in \{0, 1, \ldots, T\}$ according to $\mathbb{P}(t^* = t) \propto \frac{\bar{\rho} - \rho}{\beta_t - \rho}$.

Return: x^{t^*}

Convergence

$$\mathrm{E}\left[\left\|\nabla f_{1/\bar{\rho}}(x^{t^*})\right\|_2^2\right] \leq \frac{2(\bar{\rho}(f_{1/\bar{\rho}}(x^0) - \min_x f) + 2\bar{\rho}^2 L^2 \gamma^2}{\bar{\rho} - \rho} \cdot \left(\frac{\bar{\rho} - \rho}{T+1} + \frac{1}{\gamma\sqrt{T+1}}\right) ,$$

where

$$f_{1/\bar{\rho}}(x) := \min_z [f(z) + \frac{\rho}{2}\|z - x\|_2^2]$$

$$L = \sqrt{\mathbb{E}_\xi[\ell(\xi)]^2]}\sqrt{\mathbb{E}_\xi[M(\xi)]^2]}.$$

Convergence

$$\mathrm{E}\left[\left\|\nabla f_{1/\bar\rho}(x^{t^*})\right\|_2^2\right] \le \frac{2(\bar\rho(f_{1/\bar\rho}(x^0)-\min_x f)+2\bar\rho^2L^2\gamma^2}{\bar\rho-\rho}\cdot\left(\frac{\bar\rho-\rho}{T+1}+\frac{1}{\gamma\sqrt{T+1}}\right),$$

where

$$f_{1/\bar\rho}(x) := \min_z [f(z) + \frac{\rho}{2}\|z-x\|_2^2]$$

$$L = \sqrt{\mathbb{E}_\xi[\ell(\xi)]^2}\sqrt{\mathbb{E}_\xi[M(\xi)]^2}.$$

SIAM Prize Session: 2023 SIAG/OPT Best Paper Prize Lecture:
Stochastic Model-Based Minimization of Weakly Convex Functions

Damek Davis, Cornell University, U.S.

Dmitriy Drusvyatskiy, University of Washington, U.S.

Friday, June 2, 9:15 AM - 10:45 AM Room: Grand Ballroom B/C/D, 2nd floor

Feature Selection in Mixed Effects Models

Linear mixed-effects (LME) models are often used for analyzing nested or combined data across a range of groups or clusters.

Covariates are used to separate the total population variability (the fixed effects) from the group variability (the random effects).

Due to strength across groups, LMEs can estimate key statistics when the within group data is limited or highly variable.

Feature selection in mixed effects models finds a sparse set of covariates that explain
(i) the mean behavior across groups, and
(ii) the variability between groups.

Linear Mixed-Effects (LME) Model

$$\mathbf{y}_i = X_i\beta + Z_i u_i + \varepsilon_i, \quad i = 1 \dots m$$
$$u_i \sim N(0, \Gamma), \quad \Gamma \in \mathbb{S}^q_+$$
$$\varepsilon_i \sim N(0, \Lambda_i), \quad \Lambda_i \in \mathbb{S}^{n_i}_{++}$$

where

- y_i are known observations,
- $\beta \in \mathbb{R}^p$ is an unknown vector of fixed (mean) covariates,
- $u_i \in \mathbb{R}^q$ are unobserved random effects distributed $N(0, \Gamma)$,
- Λ_i known observation error covariance matrices,
- $\Gamma := \operatorname{Diag}\gamma,\ \gamma \in \mathbb{R}^s_+$ unknown random effects covariance matrix,
- $\Omega_i(\Gamma) := Z_i \Gamma Z_i^T + \Lambda_i$ the marginalized covariance.

Linear Mixed-Effects (LME) Model

$$\begin{aligned} \mathbf{y}_i &= X_i\beta + Z_i u_i + \varepsilon_i, \quad i = 1 \ldots m \\ u_i &\sim N(0, \Gamma), \quad \Gamma \in \mathbb{S}^q_+ \\ \varepsilon_i &\sim N(0, \Lambda_i), \quad \Lambda_i \in \mathbb{S}^{n_i}_{++} \end{aligned}$$

where

- y_i are known observations,
- $\beta \in \mathbb{R}^p$ is an unknown vector of fixed (mean) covariates,
- $u_i \in \mathbb{R}^q$ are unobserved random effects distributed $N(0, \Gamma)$,
- Λ_i known observation error covariance matrices,
- $\Gamma := \operatorname{Diag}\gamma,\ \gamma \in \mathbb{R}^s_+$ unknown random effects covariance matrix,
- $\Omega_i(\Gamma) := Z_i \Gamma Z_i^T + \Lambda_i$ the marginalized covariance.

The marginalized negative log-likelihood function

$$\mathcal{L}(\beta, \gamma) := \sum_{i=1}^{m} \frac{1}{2}(y_i - X_i\beta)^T \Omega_i(\Gamma)^{-1}(y_i - X_i\beta) + \frac{1}{2}\ln\det\Omega_i(\Gamma).$$

Maximum likelihood estimates for β and γ solve

$$\min_{\beta, \gamma \in \mathbb{R}^q_+} \ \mathcal{L}(\beta, \gamma)$$

Convex-Composite Structure

$\frac{1}{2}(y_i - X_i\beta)^T \Omega_i(\Gamma)^{-1}(y_i - X_i\beta)$ is convex-composite.

Matrix Fractional Functions
(B.-Gao-Hoheisel '15,'18)

Given the graph of the mapping $Y \mapsto -\frac{1}{2}YY^T$,

$$\mathcal{G} := \left\{ \left(Y, -\frac{1}{2}YY^T \right) \,\middle|\, Y \in \mathbb{R}^{n\times m} \right\},$$

we have

$$\sigma_{\mathcal{G}}(X,V) = \begin{cases} \frac{1}{2}\mathrm{tr}\left(X^T V^\dagger X\right) & \text{if } \mathrm{rge}\, X \subset \mathrm{rge}\, V,\ V \in \mathbb{S}^n, \\ +\infty & \text{else,} \end{cases}$$

where $V^\dagger$ is the Moore-Penrose pseudo inverse of V.

Feature Selection for Linear Mixed Effects

$$\min_{\beta \in \mathbb{R}^p, \gamma \in \mathbb{R}^q_+} \mathcal{L}(\beta, \gamma) + R(\beta, \gamma)$$

$$\mathcal{L}(\beta, \gamma) := \sum_{i=1}^{m} \frac{1}{2}(y_i - X_i\beta)^T \Omega_i(\Gamma)^{-1}(y_i - X_i\beta) + \frac{1}{2}\ln\det\Omega_i(\Gamma)$$

$\mathcal{L}$ is smooth on its domain.

R is closed, proper, convex with easily computed *prox operator*.

Feature Selection for Linear Mixed Effects

$$\min_{\beta \in \mathbb{R}^p, \gamma \in \mathbb{R}^q_+} \mathcal{L}(\beta, \gamma) + R(\beta, \gamma)$$

$$\mathcal{L}(\beta, \gamma) := \sum_{i=1}^{m} \frac{1}{2}(y_i - X_i\beta)^T \Omega_i(\Gamma)^{-1}(y_i - X_i\beta) + \frac{1}{2}\ln\det\Omega_i(\Gamma)$$

$\mathcal{L}$ is smooth on its domain.
R is closed, proper, convex with easily computed *prox operator*.

$\mathcal{L}$ is weakly convex since

$$\nabla^2\mathcal{L}(\beta, \gamma) = H(\beta, \gamma) - \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{2}(Z_i^T \Omega_i(\gamma)^{-1} Z_i)^{\circ 2} \end{bmatrix},$$

where $H(\beta, \gamma)$ is always positive semi-definite.

Feature Selection for Linear Mixed Effects

$$\min_{\beta \in \mathbb{R}^p, \gamma \in \mathbb{R}^q_+} \mathcal{L}(\beta, \gamma) + R(\beta, \gamma)$$

$$\mathcal{L}(\beta, \gamma) := \sum_{i=1}^{m} \frac{1}{2}(y_i - X_i\beta)^T \Omega_i(\Gamma)^{-1}(y_i - X_i\beta) + \frac{1}{2}\ln\det\Omega_i(\Gamma)$$

$\mathcal{L}$ is smooth on its domain.
R is closed, proper, convex with easily computed *prox operator*.

$\mathcal{L}$ is weakly convex since

$$\nabla^2\mathcal{L}(\beta, \gamma) = H(\beta, \gamma) - \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{2}(Z_i^T \Omega_i(\gamma)^{-1} Z_i)^{\circ 2} \end{bmatrix},$$

where $H(\beta, \gamma)$ is always positive semi-definite.

Apply PGD!

Feature Selection

$$\min_{\beta\in\mathbb{R}^p,\gamma\in\mathbb{R}^q_+} \mathcal{L}(\beta,\gamma) + R(\beta,\gamma)$$

with

$$\mathcal{L}(\beta,\gamma) := \tfrac{1}{2}(y - X\beta)^T \Omega(\Gamma(\gamma))^{-1}(y - X\beta) + \tfrac{1}{2}\ln\det\Omega(\Gamma(\gamma)).$$

Feature Selection

$$\min_{\beta\in\mathbb{R}^p,\gamma\in\mathbb{R}^q_+} \mathcal{L}(\beta,\gamma) + R(\beta,\gamma)$$

with

$$\mathcal{L}(\beta,\gamma) := \tfrac{1}{2}(y - X\beta)^T \Omega(\Gamma(\gamma))^{-1}(y - X\beta) + \tfrac{1}{2}\ln\det\Omega(\Gamma(\gamma)).$$

The relaxed model problem (Decouple and smooth)

$$\min_{(\beta,\gamma),(\tilde{\beta},\tilde{\gamma}),\tilde{\gamma}\geq 0} \mathcal{L}(\beta,\gamma) + \phi_\mu(\gamma) + \frac{\eta}{2}\left\| \begin{array}{c} \beta - \tilde{\beta} \\ \gamma - \tilde{\gamma} \end{array} \right\|_2^2 + R(\tilde{\beta},\tilde{\gamma}),$$

where

$$\varphi(\gamma,\mu) := \begin{cases} -\mu\sum_{i=1}^q \ln(\gamma_i/\mu) & ,\ \mu > 0, \\ \delta_{\mathbb{R}^q_+}(\gamma) & ,\ \mu = 0, \\ +\infty & ,\ \mu < 0. \end{cases}$$

Optimal value function reformulation

$$\min_{(\beta,\gamma),(\tilde{\beta},\tilde{\gamma}),\tilde{\gamma}\geq 0} \mathcal{L}(\beta,\gamma) + \phi_\mu(\gamma) + \frac{\eta}{2}\left\| \begin{array}{c} \beta - \tilde{\beta} \\ \gamma - \tilde{\gamma} \end{array} \right\|_2^2 + R(\tilde{\beta},\tilde{\gamma}),$$

Optimal value function reformulation:

$$\mathcal{P}_{\eta,\mu} \quad \min_{(\tilde{\beta},\tilde{\gamma})} u_{\eta,\mu}(\tilde{\beta},\tilde{\gamma}) + R(\tilde{\beta},\tilde{\gamma}) + \delta_{\mathbb{R}^q_+}(\tilde{\gamma})$$

where

$$u_{\eta,\mu}(\tilde{\beta},\tilde{\gamma}) := \min_{(\beta,\gamma)} \mathcal{L}(\beta,\gamma) + \phi_\mu(\gamma) + \frac{\eta}{2}\left\| \begin{array}{c} \beta - \tilde{\beta} \\ \gamma - \tilde{\gamma} \end{array} \right\|_2^2.$$

Optimal value function reformulation

$$\min_{(\beta,\gamma),(\tilde{\beta},\tilde{\gamma}),\tilde{\gamma}\geq 0} \mathcal{L}(\beta,\gamma) + \phi_\mu(\gamma) + \frac{\eta}{2}\left\| \begin{array}{c} \beta - \tilde{\beta} \\ \gamma - \tilde{\gamma} \end{array} \right\|_2^2 + R(\tilde{\beta},\tilde{\gamma}),$$

Optimal value function reformulation:

$$\mathcal{P}_{\eta,\mu} \quad \min_{(\tilde{\beta},\tilde{\gamma})} u_{\eta,\mu}(\tilde{\beta},\tilde{\gamma}) + R(\tilde{\beta},\tilde{\gamma}) + \delta_{\mathbb{R}^q_+}(\tilde{\gamma})$$

where

$$u_{\eta,\mu}(\tilde{\beta},\tilde{\gamma}) := \min_{(\beta,\gamma)} \mathcal{L}(\beta,\gamma) + \phi_\mu(\gamma) + \frac{\eta}{2}\left\| \begin{array}{c} \beta - \tilde{\beta} \\ \gamma - \tilde{\gamma} \end{array} \right\|_2^2.$$

Apply the PGD algorithm to $\mathcal{P}_{\eta,\mu}$ with

$$\nabla u_{\eta,\mu}(\tilde{\beta},\tilde{\gamma}) = \begin{pmatrix} \tilde{\beta} - \bar{\beta} \\ \tilde{\gamma} - \bar{\gamma} \end{pmatrix}, \quad \text{(locally Lipschitz)}$$

with $\begin{pmatrix} \bar{\beta} \\ \bar{\gamma} \end{pmatrix} = \operatorname{argmin}_{(\beta,\gamma)} \mathcal{L}_{\eta,\mu}((\beta,\gamma),(\tilde{\beta},\tilde{\gamma})).$

Performance

Regilarizer	Model Metric	PGD	MSR3	MSR3-fast
L0	Accuracy	0.89	**0.92**	**0.92**
	Time	41.68	88.54	**0.13**
L1	Accuracy	0.73	**0.88**	**0.88**
	Time	38.39	9.13	**0.13**
ALASSO	Accuracy	0.88	**0.92**	0.91
	Time	34.55	65.19	**0.12**
SCAD	Accuracy	0.71	**0.93**	0.92
	Time	77.62	84.67	**0.17**

The Experiment. The number of fixed effects p and random effects q is 20. $\beta = \gamma = [\frac{1}{2}, \frac{2}{2}, \frac{3}{2}, \dots, \frac{10}{2}, 0, 0, 0, \dots, 0]$

$$\begin{aligned} y_i &= X_i\beta + Z_i u_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, 0.3^2 I) \\ X_i &\sim N(0, I)^p, \quad Z_i = X_i \\ u_i &\sim N(0, \operatorname{Diag}\gamma) \end{aligned}$$

9 groups sizes $[10, 15, 4, 8, 3, 5, 18, 9, 6]$

Each experiment is repeated 100 times.

More Details

MS219: Modeling and Optimization in Global Health II
Aleksei Sholokhov, Friday, June 2, 12:15pm
Room: Medina, 3rd floor

More Details

MS219: Modeling and Optimization in Global Health II
Aleksei Sholokhov, Friday, June 2, 12:15pm
Room: Medina, 3rd floor

Thank You!