Kurdyka-Łojasiewicz exponent for a class of Hadamard-difference-parameterized models

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

The 6th Conference on Discrete Optimization & Machine Learning
National Graduate Institute for Policy Studies, Tokyo
July 2024

(Joint work with Yuncheng Liu, Wenqing Ouyang & Hao Wang)

Over-parameterized models

Example: Deep neural network training:

$$\underset{\substack{W_\ell:\mathbb{R}^{p_{\ell-1}}\to\mathbb{R}^{p_\ell},\text{affine},\\ \ell=1,\dots,L}}{\text{Minimize}} \quad \sum_{i=1}^{m}(\varrho_L(W_L(\varrho_{L-1}(W_{L-1}(\cdots\varrho_1(W_1(x_i))\cdots)))) - y_i)^2$$

where $\varrho_\ell : \mathbb{R} \to \mathbb{R}$ is the activation function for the ℓ-th layer (acting entrywise on vectors), $\ell = 1, \dots, L$, $p_0, \cdots, p_{L-1}$ are positive integers, $p_L = 1$, $(x_i, y_i) \in \mathbb{R}^{p_0} \times \mathbb{R}$ for $i = 1, \dots, m$ are data points.

Over-parameterized models

Example: Deep neural network training:

$$\underset{\substack{W_\ell:\mathbb{R}^{p_{\ell-1}}\to\mathbb{R}^{p_\ell},\textit{affine},\\ \ell=1,\dots,L}}{\text{Minimize}} \quad \sum_{i=1}^{m}(\varrho_L(W_L(\varrho_{L-1}(W_{L-1}(\cdots\varrho_1(W_1(x_i))\cdots))))-y_i)^2$$

where $\varrho_\ell : \mathbb{R} \to \mathbb{R}$ is the activation function for the ℓ-th layer (acting entrywise on vectors), $\ell = 1, \dots, L$, $p_0, \cdots, p_{L-1}$ are positive integers, $p_L = 1$, $(x_i, y_i) \in \mathbb{R}^{p_0} \times \mathbb{R}$ for $i = 1, \dots, m$ are data points.

Over-parametrized models can have desirable properties.

- Better generalization. (Allen-Zhu, Li, Liang '19, Pandey, Kumar '23, Subramanian, Arya, Sahal '22, ...)
- Implicit bias / regularization. (Belkin, Hsu, Ma, Mandal '19, Dai, Karzand, Srebro '21, Gunasekar, Lee, Soudry, Srebro '18, Li, Nguyen, Hegde, Wong, '21, ...)
- ...

Hadamard parametrized model

For $\mu > 0$ and $h \in C^2(\mathbb{R}^n)$, consider

$$f(x) := h(x) + \mu \|x\|_1 \quad \text{and} \quad G(u, v) := h(u \circ v) + \frac{\mu}{2}(\|u\|^2 + \|v\|^2).$$

Hadamard parametrized model

For $\mu > 0$ and $h \in C^2(\mathbb{R}^n)$, consider

$$f(x) := h(x) + \mu\|x\|_1 \text{ and } G(u, v) := h(u \circ v) + \frac{\mu}{2}(\|u\|^2 + \|v\|^2).$$

- It holds that $G(u, v) \geq f(u \circ v)$ and $\inf f = \inf G$, thanks to the AM-GM inequality.

Hadamard parametrized model

For $\mu > 0$ and $h \in C^2(\mathbb{R}^n)$, consider

$$f(x) := h(x) + \mu\|x\|_1 \quad \text{and} \quad G(u, v) := h(u \circ v) + \frac{\mu}{2}(\|u\|^2 + \|v\|^2).$$

- It holds that $G(u, v) \geq f(u \circ v)$ and $\inf f = \inf G$, thanks to the AM-GM inequality.
- f commonly arises in compressed sensing / variable selections, with popular choices of h being $h(x) = \sum_{i=1}^{m} \ln(1 + \exp(\langle y_i, x \rangle))$ or $\frac{1}{2}\|Ax - z\|^2$ for some $A \in \mathbb{R}^{m \times n}$, $z \in \mathbb{R}^m$ and $y_i \in \mathbb{R}^n$ for $i = 1, \ldots, m$.

Hadamard parametrized model

For $\mu > 0$ and $h \in C^2(\mathbb{R}^n)$, consider

$$f(x) := h(x) + \mu\|x\|_1 \quad \text{and} \quad G(u, v) := h(u \circ v) + \frac{\mu}{2}(\|u\|^2 + \|v\|^2).$$

- It holds that $G(u, v) \geq f(u \circ v)$ and $\inf f = \inf G$, thanks to the AM-GM inequality.
- f commonly arises in compressed sensing / variable selections, with popular choices of h being $h(x) = \sum_{i=1}^{m} \ln(1 + \exp(\langle y_i, x\rangle))$ or $\frac{1}{2}\|Ax - z\|^2$ for some $A \in \mathbb{R}^{m\times n}$, $z \in \mathbb{R}^m$ and $y_i \in \mathbb{R}^n$ for $i = 1, \ldots, m$.
- G is called the Hadamard parametrization of f. (Hoff '17)
- The smoothness of G has been recently exploited for algorithmic design. (Hoff '17, Kolb, Müller, Bischl, Rügamer '23, Poon, Peyré '21, '23)

Hadamard difference parameterization

For $\mu > 0$ and $h \in C^2(\mathbb{R}^n)$, consider

$$f(x) := h(x) + \mu\|x\|_1 \quad \text{and} \quad G(u, v) := h(u \circ v) + \frac{\mu}{2}(\|u\|^2 + \|v\|^2).$$

In view of the subsitution

$$a = (u + v)/2 \quad \text{and} \quad b = (u - v)/2,$$

it follows that minimizing G is equivalent to minimizing F defined as

$$F(a, b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

Hadamard difference parameterization

For $\mu > 0$ and $h \in C^2(\mathbb{R}^n)$, consider

$$f(x) := h(x) + \mu\|x\|_1 \quad \text{and} \quad G(u, v) := h(u \circ v) + \frac{\mu}{2}(\|u\|^2 + \|v\|^2).$$

In view of the substitution

$$a = (u + v)/2 \quad \text{and} \quad b = (u - v)/2,$$

it follows that minimizing G is equivalent to minimizing F defined as

$$F(a, b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

Remark:

- F is called the Hadamard difference parameterization (HDP) of f. (Vaškevičius, Kanade, Rebeschini '19) We focus on F from now on.

Hadamard difference parameterization

For $\mu > 0$ and $h \in C^2(\mathbb{R}^n)$, consider

$$f(x) := h(x) + \mu\|x\|_1 \quad \text{and} \quad G(u, v) := h(u \circ v) + \frac{\mu}{2}(\|u\|^2 + \|v\|^2).$$

In view of the subsitution

$$a = (u + v)/2 \quad \text{and} \quad b = (u - v)/2,$$

it follows that minimizing G is equivalent to minimizing F defined as

$$F(a, b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

Remark:

- F is called the Hadamard difference parameterization (HDP) of f. (Vaškevičius, Kanade, Rebeschini '19) We focus on F from now on.

Questions:

- How do the stationary points of F correspond to those of f?

Hadamard difference parameterization

For $\mu > 0$ and $h \in C^2(\mathbb{R}^n)$, consider

$$f(x) := h(x) + \mu\|x\|_1 \text{ and } G(u,v) := h(u \circ v) + \frac{\mu}{2}(\|u\|^2 + \|v\|^2).$$

In view of the subsitution

$$a = (u+v)/2 \quad \text{and} \quad b = (u-v)/2,$$

it follows that minimizing G is equivalent to minimizing F defined as

$$F(a,b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

Remark:

- F is called the Hadamard difference parameterization (HDP) of f. (Vaškevičius, Kanade, Rebeschini '19) We focus on F from now on.

Questions:

- How do the stationary points of F correspond to those of f?
- (Roughly) If a stationary point of f can be found efficiently, how about F?

2nd-order stationary points of F

Recall that for $\mu > 0$,

$$f(x) := h(x) + \mu\|x\|_1 \text{ and } F(a,b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

Theorem 1. (Ouyang, Liu, P., Wang '24)
For all $(a,b) \in \mathbb{R}^n \times \mathbb{R}^n$, the following statements are equivalent:

(i) The point (a,b) is a 2nd-order stationary point of F.

(ii) The point $s := a^2 - b^2$ is a stationary point of f, $\min\{a^2, b^2\} = 0$, and

$$w^T \nabla^2 h(s) w \geq 0 \quad \forall w \in \{v : \ v_i = 0 \text{ when } s_i = 0\}.$$

Strict saddle property

Recall that for $\mu > 0$,

$$f(x) := h(x) + \mu\|x\|_1 \text{ and } F(a,b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

Theorem 2. (Ouyang, Liu, P., Wang '24)
Suppose that h is convex. Then there exists a $\delta > 0$ such that for all $(a,b) \in \mathbb{R}^n \times \mathbb{R}^n$, the following statements are equivalent:

(i) The point (a,b) is a stationary point of F and it holds that $\lambda_{\min}(\nabla^2 F(a,b)) > -\delta$.

(ii) The point $a^2 - b^2$ minimizes f, and $\min\{a^2, b^2\} = 0$.

(iii) The point (a,b) minimizes F.

(iv) The point (a,b) is a 2nd-order stationary point of F.

Remark: The above result was established in (Poon, Peyré '21) when h is a convex quadratic function.

KL property & exponent

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

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

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

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

KL property & exponent

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

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

KL property & exponent

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

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

Prototypical local convergence results

Fact 1. (Attouch, Bolte '09)
For proximal gradient algorithm and its variants:
Let $\{x^k\}$ be a bounded sequence generated. If g satisfies the KL property with exponent $\alpha \in [0, 1)$ at every cluster point of $\{x^k\}$, then:

- if $\alpha = 0$, then $\{x^k\}$ converges finitely;
- if $\alpha \in (0, \frac{1}{2}]$, then $\{x^k\}$ converges locally linearly;
- if $\alpha \in (\frac{1}{2}, 1)$, then $\{x^k\}$ converges locally sublinearly.

Prototypical local convergence results

Fact 1. (Attouch, Bolte '09)
For proximal gradient algorithm and its variants:
Let $\{x^k\}$ be a bounded sequence generated. If g satisfies the KL property with exponent $\alpha \in [0, 1)$ at every cluster point of $\{x^k\}$, then:

- if $\alpha = 0$, then $\{x^k\}$ converges finitely;
- if $\alpha \in (0, \frac{1}{2}]$, then $\{x^k\}$ converges locally linearly;
- if $\alpha \in (\frac{1}{2}, 1)$, then $\{x^k\}$ converges locally sublinearly.

KL exponent calculus?

- The KL exponent of $f := h + \mu\|\cdot\|_1$ is known for many loss functions h, such as least squares loss and logistic loss.
- Can we deduce the KL exponent of the corresponding HDP model F?

KL exponent under strict complementarity

Recall that for $\mu > 0$,

$$f(x) := h(x) + \mu\|x\|_1 \text{ and } F(a,b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

KL exponent under strict complementarity

Recall that for $\mu > 0$,

$$f(x) := h(x) + \mu\|x\|_1 \text{ and } F(a,b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

Theorem 3. (Ouyang, Liu, P., Wang '24)
Let (a^*, b^*) be a 2nd-order stationary point of F and set $s^* = (a^*)^2 - (b^*)^2$. Suppose that f satisfies the KL property with exponent $\alpha \in (0,1)$ at s^*. If $0 \in \mathrm{ri}\,\partial f(s^*)$, then F satisfies the KL property at (a^*, b^*) with exponent $\max\{\alpha, \frac{1}{2}\}$.

KL exponent under strict complementarity

Recall that for $\mu > 0$,

$$f(x) := h(x) + \mu\|x\|_1 \text{ and } F(a,b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

Theorem 3. (Ouyang, Liu, P., Wang '24)
Let (a^*, b^*) be a 2nd-order stationary point of F and set $s^* = (a^*)^2 - (b^*)^2$. Suppose that f satisfies the KL property with exponent $\alpha \in (0,1)$ at s^*. If $0 \in \operatorname{ri} \partial f(s^*)$, then F satisfies the KL property at (a^*, b^*) with exponent $\max\{\alpha, \frac{1}{2}\}$.

Remark:

- The condition $0 \in \operatorname{ri} \partial f(s^*)$ is typically referred to as the strict complementarity condition.

KL exponent without strict complementarity

Recall that for $\mu > 0$,

$$f(x) := h(x) + \mu\|x\|_1 \text{ and } F(a, b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

Theorem 4. (Ouyang, Liu, P., Wang '24)
Let (a^*, b^*) be a 2nd-order stationary point of F and set $s^* = (a^*)^2 - (b^*)^2$. Suppose that h is convex and $\Omega := \operatorname{Arg\,min} f$ is polyhedral. If f satisfies the KL property with exponent $\alpha \in (0, 1)$ at s^*, then F satisfies the KL property at (a^*, b^*) with exponent $(1 + \alpha)/2$.

KL exponent without strict complementarity

Recall that for $\mu > 0$,

$$f(x) := h(x) + \mu\|x\|_1 \text{ and } F(a,b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

Theorem 4. (Ouyang, Liu, P., Wang '24)
Let (a^*, b^*) be a 2nd-order stationary point of F and set $s^* = (a^*)^2 - (b^*)^2$. Suppose that h is convex and $\Omega := \operatorname{Arg\,min} f$ is polyhedral. If f satisfies the KL property with exponent $\alpha \in (0,1)$ at s^*, then F satisfies the KL property at (a^*, b^*) with exponent $(1+\alpha)/2$.

Remark:

- Ω is polyhedral when $h(x) = \ell(Ax)$ for some strictly convex function $\ell : \mathbb{R}^m \to \mathbb{R}$ and $A \in \mathbb{R}^{m\times n}$. (Zhou, So '17)

Example: tightness of exponent

Example: Let $\alpha \in [\frac{1}{2}, 1)$ and define $h : \mathbb{R} \to \mathbb{R}$ as $h(x) = (1-\alpha)|x|^{\frac{1}{1-\alpha}} - x$. Consider

$$f(x) := h(x) + |x| \quad \text{and} \quad F(a,b) := h(a^2 - b^2) + (a^2 + b^2).$$

Then $h \in C^2(\mathbb{R})$ is convex, $\operatorname{Arg\,min} f = \{0\}$ and $(0,0) \in \operatorname{Arg\,min} F$.

Example: tightness of exponent

Example: Let $\alpha \in [\frac{1}{2}, 1)$ and define $h : \mathbb{R} \to \mathbb{R}$ as $h(x) = (1-\alpha)|x|^{\frac{1}{1-\alpha}} - x$. Consider

$$f(x) := h(x) + |x| \quad \text{and} \quad F(a,b) := h(a^2 - b^2) + (a^2 + b^2).$$

Then $h \in C^2(\mathbb{R})$ is convex, $\operatorname{Arg\,min} f = \{0\}$ and $(0,0) \in \operatorname{Arg\,min} F$. Moreover,

$$f(x) = \begin{cases} (1-\alpha)|x|^{\frac{1}{1-\alpha}} & \text{if } x \geq 0, \\ (1-\alpha)|x|^{\frac{1}{1-\alpha}} - 2x & \text{if } x < 0. \end{cases}$$

$$f'(x) = \begin{cases} |x|^{\frac{\alpha}{1-\alpha}} & \text{if } x > 0, \\ -|x|^{\frac{\alpha}{1-\alpha}} - 2 & \text{if } x < 0. \end{cases}$$

Thus, the KL exponent of f at 0 is α.

Example cont.: tightness of exponent

Example cont.: On the other hand, we have

$$\begin{aligned} F(a,b) &= h(a^2 - b^2) + a^2 + b^2 \\ &= (1-\alpha)|a^2 - b^2|^{\frac{1}{1-\alpha}} - (a^2 - b^2) + a^2 + b^2 \\ &= (1-\alpha)|a^2 - b^2|^{\frac{1}{1-\alpha}} + 2b^2. \end{aligned}$$

Take $t > 0$. Then we have

$$\nabla F(t,0) = \begin{bmatrix} 2t^{\frac{1+\alpha}{1-\alpha}} & 0 \end{bmatrix}^\top \text{ and } F(t,0) = (1-\alpha)t^{\frac{2}{1-\alpha}}.$$

This implies that $\|\nabla F(t,0)\| = 2(\frac{1}{1-\alpha}F(t,0))^{\frac{1+\alpha}{2}}$, which shows that the KL exponent of F at 0 is no less than $\frac{1+\alpha}{2}$.

Example: new models with explicit KL exponents

Example: Consider

- $h(x) := \frac{1}{2}\|Ax - z\|^2$ for some $A \in \mathbb{R}^{m\times n}$ and $z \in \mathbb{R}^m$; or
- $h(x) := \sum_{i=1}^m \ln(1 + \exp(\langle y_i, x\rangle))$ for $y_i \in \mathbb{R}^n$, $i = 1, \ldots, m$.

Example: new models with explicit KL exponents

Example: Consider

- $h(x) := \frac{1}{2}\|Ax - z\|^2$ for some $A \in \mathbb{R}^{m\times n}$ and $z \in \mathbb{R}^m$; or
- $h(x) := \sum_{i=1}^m \ln(1 + \exp(\langle y_i, x\rangle))$ for $y_i \in \mathbb{R}^n$, $i = 1, \ldots, m$.

For $\mu > 0$, consider

$$f(x) := h(x) + \mu\|x\|_1 \text{ and } F(a, b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

It is known that

- $\operatorname{Arg\,min} f$ is polyhedral. (Zhou, So '17)
- KL exponent of f is $\frac{1}{2}$. (Li, P. '18)

Example: new models with explicit KL exponents

Example: Consider

- $h(x) := \frac{1}{2}\|Ax - z\|^2$ for some $A \in \mathbb{R}^{m\times n}$ and $z \in \mathbb{R}^m$; or
- $h(x) := \sum_{i=1}^m \ln(1 + \exp(\langle y_i, x\rangle))$ for $y_i \in \mathbb{R}^n$, $i = 1, \ldots, m$.

For $\mu > 0$, consider

$$f(x) := h(x) + \mu\|x\|_1 \text{ and } F(a, b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

It is known that

- $\operatorname{Arg\,min} f$ is polyhedral. (Zhou, So '17)
- KL exponent of f is $\frac{1}{2}$. (Li, P. '18)

Consequently, the KL exponent of F at a 2nd-order stationary point (a^*, b^*) is $\frac{1}{2}$ or $\frac{3}{4}$ depending on whether $0 \in \operatorname{ri} \partial f(s^*)$, where $s^* := (a^*)^2 - (b^*)^2$.

Applications

How can we make use of the KL exponents at 2nd-order stationary points of F?

Applications

How can we make use of the KL exponents at 2nd-order stationary points of F?

Recall that for $\mu > 0$,

$$f(x) := h(x) + \mu\|x\|_1 \text{ and } F(a,b) := h(a^2 - b^2) + \mu(\|a\|^2 + \|b\|^2).$$

Theorem 5. (Ouyang, Liu, P., Wang '24)
Suppose that h is subanalytic and lower-bounded.

Consider the steepest descent with backtracking linesearch ($\mathrm{SD}_{\mathrm{ls}}$) with initial stepsize θ_0 and initial point (a^0, b^0) for minimizing F.

Then for almost all $\theta_0 > 0$, there exists a $V \subseteq \mathbb{R}^n \times \mathbb{R}^n$ with full measure such that whenever $(a^0, b^0) \in V$, the sequence $\{(a^k, b^k)\}$ generated by $\mathrm{SD}_{\mathrm{ls}}$ converges to a 2nd-order stationary point of F.

Conclusion

Conclusion:

- 2nd-order stationary points of the HDP model F correspond to some stationary points of f.
- The KL exponent of F at a 2nd-order stationary point can be deduced from the KL exponent at the corresponding stationary point of f, under suitable assumptions.

Reference:

- Wenqing Ouyang, Yuncheng Liu, Ting Kei Pong and Hao Wang. *Kurdyka-Łojasiewicz exponent via Hadamard parametrization*. Preprint. Available at https://arxiv.org/abs/2402.00377.

Thanks for coming! $\ddot{\smile}$