Notes for ORIE 7391: Topics in Mathematical Programming (Kurdyka-Lojasiewicz Inequality)

Qingxuan Jiang

Fall 2020

General Information

This is a set of notes taken for the course ORIE 7391: Topics in Mathematical Programming, taught by Prof. Adrian Lewis at Cornell University in Fall 2020. This graduate topics course mainly covers the motivations, theory and applications of the Kurdyka-Lojasiewicz inequality as a unifying framework for analyzing optimization algorithms.

The course is based on lecture notes from Prof. Lewis, with no official textbook. The scribe personally finds the following references to be useful:

- Course material for ORIE 6328/ORIE 7390 on Convex Analysis, which is a prequel for this course. A scribe of lecture notes for an earlier version of this class is available at this link.
- Rockafellar, R. Tyrrell, and Roger J-B. Wets. Variational analysis. Vol. 317. Springer Science and Business Media, 2009.

Note that the presentation of the material here is informal and does not necessarily follow the lectures - the scribe have significantly rearranged the material based on own ideas, omitted certain technical details, and also added new thoughts that may or may not be accurate. If you find any errors – which are the fault of the scribe, not the lecturer – feel free to let the scribe know.

From time to time, I may make small changes to the notes. The current version is updated in April 2021.

Contents

1 Introduction **3**

2 Motivation: Continuous-Time Steepest Descent Trajectories **5**
2.1 Preliminaries from Convex Analysis . . . 5
2.2 Slope, Steepest Descent, and the Chain Rule . . . 7
2.3 Steepest Descent Trajectory, Brezis's Theorem, and Relation to Proximal Point Method . . . 10

3 KL Convergence Analysis on Convex Problems **15**
3.1 Slope Descent Sequences . . . 15
3.1.1 Motivation and Definition . . . 15
3.1.2 Examples of Slope Descent Sequences . . . 15
3.2 Convergence Theorem from KL Property . . . 19
3.2.1 KL Property and Convergence Theorem . . . 19
3.2.2 Convergence Comparison to Proximal Point Iteration . . . 20
3.2.3 Deriving Concrete Convergence Rates from KL Desingularizer . . . 22
3.3 Examples of KL Analysis for Convex Problems . . . 24
3.3.1 Alternating Projection Method on Convex Feasibility Problem . . . 25
3.3.2 ISTA Method on the LASSO Problem . . . 26

4 KL Convergence Analysis on Non-Convex Problems: Proximal Alternating Linearized Minimization (PALM) **28**
4.1 Proximal Alternating Linearized Minimization (PALM) . . . 28
4.2 Preliminaries from Variational Analysis . . . 29
4.3 KL Analysis of PALM Algorithm . . . 31
4.3.1 Convergence Theorem and Overview of Proof Steps . . . 31
4.3.2 Detailed Convergence Proof . . . 32
4.4 Examples of PALM Algorithm . . . 37
4.4.1 Alternating Projection . . . 37
4.4.2 Nonnegative Matrix Factorization . . . 38

5 Modified KL Convergence Analysis: The Majorization-Minimization Framework **40**
5.1 Motivation: Composite Optimization . . . 40
5.2 Majorization-Minimization Framework and KL Analysis . . . 41
5.2.1 Majorize-Minimize Framework and Overview of Convergence Proof . . . 41
5.2.2 Detailed Convergence Proof . . . 43
5.3 Examples of Algorithms under Majorization-Minimization Framework . . . 46
5.3.1 Proximal Point Method . . . 46
5.3.2 Gradient Descent . . . 47
5.3.3 Composite Optimization . . . 47
5.4 Extended Majorization-Minimization: Sequential Quadratic Programming . . . 48
5.4.1 Motivation: Moving Balls Method . . . 48
5.4.2 Extended Majorization-Minimization Framework . . . 49
5.4.3 Application to Moving Balls Method . . . 50

6 References **54**

1 Introduction

This section introduces basic concepts about continuous-time steepest descent trajectories and discrete-time optimization models, and outlines the relationship between them that will be discussed in detail in later parts of the course.

Continuous-Time Optimization Models via Subgradient Descent Trajectories

- **Subgradient Descent Trajectory**:
  - **Idea**: We would like to consider a continuous-time version of subgradient descent.
  - **Subgradient**: For convex function $f : \mathbb{R}^n \to \mathbb{R}$, we say $y \in \partial f(x)$ if x minimizes $f(x) - \langle x, y\rangle$.
  - **Shortest subgradient**:
    - **Definition**: Since the subgradient set is convex, we know that it has a unique element closest to 0. We will call that the **shortest subgradient** and denote by $\partial^0 f(x)$.
    - **Property**: $\partial^0 f(x) = -$steepest descent direction
  - **Subgradient descent trajectory**: The subgradient descent trajectory is a path that always follows the steepest descent direction. More concretely, $\chi : \mathbb{R} \to \mathbb{R}^n$ is the **subgradient descent trajectory** starting at $x_0 \in \mathbb{R}^n$ if it satisfies the initial value condition $\chi(0) = x_0$ and $\dot{\chi}(t) = -\partial^0 f(\chi(t))$ a.e.
  - **Theorem (Existence, Uniqueness, and Convergence) (Brezis, 1973)**: The subgradient descent trajectory above is well-defined, unique, and converges to a minimizer.
- **Complexity of Continuous-Time Subgradient Descent**
  - **Idea**: We would like to measure the "time complexity" of continuous-time subgradient descent. Instead of counting the number of steps (as in discrete case), the continuous analog would be to measure the length of the trajectory.
  - **Question**: We know that when the trajectory is unbounded (like when minimum is at infinity), the path will be of infinite length. Is it possible for the trajectory to be of infinite length if we have a bounded trajectory?
  - **Answer**: Yes. We will consider the Mexican Hat function (illustration at this link):

$$f(r,\theta) = \begin{cases} e^{-\frac{1}{1-r^2}}\left[1 - \frac{4r^4}{4r^4+(1-r^2)^4}\sin(\theta - \frac{1}{1-r^2})\right] \text{ , if } r < 1, \\ 0 \text{ , if } r \geq 1. \end{cases}$$

    In this case, the trajectory would spiral around the "hat" for infinite length. Moreover, the function here is actually in C^∞.
  - **Takeaway**:
    - The above example illustrates that to have finite length subgradient descent trajectory, we need a condition on f that is stronger than C^∞. In fact, as is shown in the following theorem, what we need is analyticity of f.
    - To see that analyticity is a stronger property than C^∞, a simple example of a C^∞ function that is not analytic is

$$f(x) = \begin{cases} e^{\frac{1}{x^2}} \text{ , if } x \neq 0 \\ 0 \text{ , if } x = 0 \end{cases}$$

  - **Theorem (Analyticity ⇒ Finite trajectory) (Lojasiewicz 1959, Proof 1984)**: For analytic convex function f, any bounded gradient descent trajectory has finite length.
  - **Key step in proof (Lojasiewicz gradient inequality)**: For analytic function $f : \mathbb{R}^n \to \mathbb{R}$ and a stationary point $x^* \in \mathbb{R}^n$ with $\nabla f(x^*) = 0$, there exists constants $C > 0$ and $0 \leq \mu < 1$ such that for any x in a neighborhood of x^*, we have

$$\|\nabla f(x)\| \geq c|f(x) - f(x^*)|^\mu$$

- **Extending Lojasiewicz Gradient Inequality to Non-Analytic Functions**:

- **Slope**: For general functions $f : \mathbb{R}^n \to \mathbb{R}$, we define the slope at point $x \in \mathbb{R}^n$ to be

$$|\nabla f|(x) = \limsup_{z \to x} \frac{f(x) - f(z)}{\|x - z\|}$$

  Note that this is consistent with the usual definition of slope $\|\nabla f(x)\|$ for smooth functions.

- **Further properties of subgradient descent trajectory**
  - **Speed = Slope**: $\|\dot{\chi}(t)\| = |\nabla f|(\chi(t))$.
  - **Rate of decrease = Speed^2**: $-\frac{d}{dt} f(\chi(t)) = \|\dot{\chi}(t)\|^2$
- **Extension to sharp functions**: For "sharp functions" with subgradient bounded uniformly away from 0 (i.e. for each local minimizer x^*, we have $|\nabla f|(x) \geq \epsilon > 0$ in some neighborhood of x^*), we can apply Lojasiewicz inequality above to show that $\chi(t)$ attains minimal value in finite time.
- **Theorem (Kurdyka-Lojasiewicz inequality)**: For certain categories of smooth functions $f : \mathbb{R}^n \to \mathbb{R}$, there exists a reparametrization $\phi : \mathbb{R} \to \mathbb{R}$, so that the resulting function is "sharp", in the sense that it satisfies $|\nabla(\phi \circ f)| \geq 1$ around local minimizers.
- **Simple example for illustration**: Consider $f(x) = x^2$, then we can pick the parametrization $\phi(y) = \sqrt{y}$, and this makes $(\phi \circ f)(x) = |x|$ a sharp function with norm of gradient at least 1 almost everywhere.
- **Categories for which KL inequality holds**: A large category of f is the semi-algebraic function (whose graphs are defined by polynomial inequalities).

Application to Discrete-Time Optimization Models via Slope Descent Sequences

- **Idea**: We would like to design algorithms that satisfy conditions similar to that of subgradient descent trajectory. This would allow us to apply a time complexity analysis similar to that of KL inequality.
- **Slope Descent Conditions**:
  - **Speed = Slope**: $\beta|x_k - x_{k+1}| \geq |\nabla f|(x_{k+1})$.
  - **Rate of decrease = Speed^2**: $f(x_k) - f(x_{k+1}) \geq \alpha\|x_k - x_{k+1}\|^2$
- Under the above pair of conditions, we will be able to get convergence of the iterates to a local minimum.
- **Applications**
  - Steepest descent, proximal gradient
  - Variable metric method, trust-region method
  - Gauss-Seidel method
  - Sequential quadratic programming

2 Motivation: Continuous-Time Steepest Descent Trajectories

In this section, we discuss concepts in continuous-time optimization models. We start by a brief review of convex analysis, and then move on to definitions and properties of slope for non-smooth functions and steepest descent trajectories. We finally state Brezis's Theorem for existence and uniqueness of steepest descent trajectories, and provide some ideas for relating such continuous trajectories to discrete algorithms like the proximal point method.

2.1 Preliminaries from Convex Analysis

This section reviews some basic notions from convex analysis.

Basic Settings and Definitions

- **Setting**: Consider a Hilbert Space $\mathcal{H}$ and the set of extended real-numbers $\overline{\mathbb{R}} = [-\infty, \infty]$, and let $f : \mathcal{H} \to \overline{\mathbb{R}}$. We will consider the unconstrained optimization problem $\min_{x \in \mathcal{H}} f(x)$.
- **Epigraph**: $epi(f) = \{(x, r) \in \mathcal{H} \times \mathbb{R} : r \geq f(x)\}$
- **Closed Function**: f is closed if $epi(f)$ is closed.
- **Convex Function**: f is convex if $epi(f)$ is convex.
- **Proper Function**: f is proper if f is never $-\infty$ and everywhere finite.
- **Positively Homogeneous Function**: f is positively homogeneous if $f(tx) = tf(x)$ for all $t \geq 0$, $x \in \mathcal{H}$ (so we must have $f(0) = 0$)
- **Domain of the Function**: $dom(f) = \{x : f(x) < +\infty\}$.

Subgradient

- **Definition of Subgradient**: $y \in \partial f(x)$ means for $x \in dom(f)$, $f(z) \geq f(x) + \langle y, z - x \rangle$ for all $z \in \mathcal{H}$.
- **Unique Shortest Subgradient**: If $\partial f(x)$ is nonempty, there is a unique shortest subgradient $\partial^0 f(x)$. (Uniqueness guaranteed by convexity of $\partial f(x)$.)
- **Theorem 1 (Local Lipschitz Property of Convex Function)**: For proper convex $f : \mathcal{H} \to \overline{\mathbb{R}}$ and $x \in int(dom(f))$, then f is locally Lipschitz around x.
- **Theorem 2 (Non-Emptiness and Boundedness of Subdifferential for Convex Function)**: Moreover, if the above Lipschitz constant is L, then we have and $\emptyset \neq \partial f(x) \subseteq LB$, where B is the closed unit ball.
- **Idea for Proof of Theorem 1**: First prove this for points $\overline{x}$ such that f has upper bound on some neighborhood of $\overline{x}$. Then use a simplex (which has extreme values on vertices from convexity) to provide a neighborhood where such an upper bound exist.
- **Idea for Proof of Theorem 2**: Non-emptiness comes from constructing a vector using supporting hyperplane theorem. Boundedness comes from both Theorem 1 and an inequality from supporting hyperplane theorem.

Duality and Fenchel Conjugate

- **Definition of Fenchel conjugate**: For a function $f : \mathcal{H} \to \overline{\mathbb{R}}$, its Fenchel conjugate $f^* : \mathcal{H}^* \to \overline{\mathbb{R}}$ is defined by $f^*(y) = \sup_x \{\langle x, y \rangle - f(x)\}$.
- **Theorem (Biconjugation)**: For closed proper convex f, $f^{**} = f$.
- **Equivalent Definitions of Subdifferential**: For proper closed convex $f : \mathcal{H} \to \overline{\mathbb{R}}$, we have

$$y \in \partial f(x) \iff x \in \partial f^*(y) \iff \text{ equality holds in } f(x) + f^*(y) \geq \langle x, y \rangle$$

Example: Indicator Function:

- **Definition**: For $C \subseteq \mathcal{H}$, $\delta_C(x) = \begin{cases} 0 \text{ , if } x \in C \\ +\infty \text{ , if } x \notin C \end{cases}$
- **Theorem (Characterization of Indicator Function)**: δ_C is a closed, proper and convex function if and only if C is a strict subset of $\mathcal{H}$ that is nonempty, closed and convex.
- **Fenchel Conjugate as Support Function**: $\delta_C^*(y) = \sigma_C(y) = \sup_{x \in C} \langle x, y \rangle$ is the support function.
- **Theorem (Characterization of Support Function)**: A function $f : \mathcal{H} \to \overline{\mathbb{R}}$ is closed, convex, and positively homogeneous if and only if $f = \delta_C^*$ for C a strict subset of $\mathcal{H}$ that is nonempty, closed and convex.
  - Proof of $\Leftarrow$ is easy.
  - Proof of $\Rightarrow$:
    * We will construct a concrete the set C explicitly. Specifically, let $C = \partial f(0)$.
    * Since $f(0) = 0$, we know that for $y \in C$, we have $f(x) \geq \langle x, y \rangle$ for all x. In this case, $f^*(y) = \sup_x \{\langle x, y \rangle - f(x)\} = 0$ (achieved at $x = 0$.)
    * For $y \notin C$, $\exists x'$ with $f(x') < \langle x', y \rangle$, so $f^*(y) = \sup_x \{\langle x, y \rangle - f(x)\} \geq \sup_{t \geq 0} \{\langle tx', y \rangle - f(tx')\} = +\infty$ when $t \to \infty$.
    * This implies $f^* = \delta_C$, and since f is closed and convex, $f = f^{**} = \delta_C^*$.
    * Finally, C is not empty, as otherwise we will have $\sup_x \{\langle x, y \rangle - f(x)\} = f^*(y) = \delta_C(y) \equiv +\infty$, and picking $y = 0$ would indicate f achieving $-\infty$ at some point, a contradiction.
- **Subgradient**
  - **Subgradient of indicator function**: $\partial \delta_C(x) = N_C(x)$ is the normal cone, defined by $N_C(x) = \{y \in \mathcal{H} : \langle y, z - x \rangle \leq 0 \text{ for any } z \in S\}$.
  - **Subgradient of support function**: $\partial \delta_C^*(y) = \partial \sigma_C(y) = \arg\max_{x \in C} \langle x, y \rangle$.

Max Formula and Steepest Descent

- **Directional Derivative**: For proper $f : \mathcal{H} \to \overline{\mathbb{R}}$ and point $x \in int(domf)$, direction $v \in \mathcal{H}$, we define the directional derivative of f at x with direction v to be $f'(x; v) = \lim_{t \to 0} \frac{f(x+tv) - f(x)}{t}$.
- **Properties of Directional Derivative**: If f is convex, then we have the following:
  - (1) **Existence and finiteness of directional derivative**: $f'(x; v)$ exists and is finite for all $x \in int(dom(f))$ and $v \in H$.
  - (2) **Convexity and positive homogeneity of $v \to f'(x; v)$**: $f'(x; v)$ as a function of direction v is convex and positive homogeneous.
  - (3) **Max formula**: We have $f'(x, v) = \max_{g \in \partial f(x)} \langle g, v \rangle$.
  - (4) **Shortest subgradient = - steepest descent direction**: If x is not a minimizer, then $\max_{v \in B} -f'(x; v) = dist(0, \partial f(x))$ and equality is attained uniquely by $v = -\frac{g_0}{\|g_0\|}$, where $g_0 = \partial^0 f(x)$.
- **Proof**:
  - (1) **Existence and finiteness**:
    * From convexity of f, we know that $g(t) = \frac{f(x+tv) - f(x)}{t}$ is an increasing function of t on $\mathbb{R}$.
    * Thus, we can rewrite $f'(x; v) = \lim_{t \to 0} \frac{f(x+tv) - f(x)}{t} = \inf_{t > 0} \frac{f(x+tv) - f(x)}{t}$, and the infimum is lower bounded by finite value $g(-\epsilon) = \frac{f(x) - f(x - \epsilon v)}{\epsilon}$ (where we pick ϵ such that $x - \epsilon v$ is still in $dom(f)$), so infimum is achieved, and thus $f'(x; v)$ exists and is finite.
  - (2) **Convexity and positive homogeneity**: Follows directly by plugging in the definitions.
  - (3) **Max formula**:

* (1) The direction $f'(x,v) \geq \max_{g \in \partial f(x)} \langle g, v \rangle$ is simple, as for any $y \in \partial f(x)$, we have

$$f'(x,v) = \lim_{t \to 0} (f(x+tv) - f(x)) \geq \lim_{t \to 0} \langle g, v \rangle = \langle g, v \rangle$$

where the inequality simply comes from definition of subgradient.

* (2) We now show that $f'(x,v) \leq \max_{g \in \partial f(x)} \langle g, v \rangle$.
  - (2-1) Let us denote $h(v) = f'(x;v)$. We know from above that h is convex and positive homogeneous, so for any $d \in \mathcal{H}$, $\alpha \geq 0$, and $g \in \partial h(v)$, we have

$$\alpha f'(x;d) = f'(x, \alpha d) = h(\alpha d) \geq h(v) + \langle g, \alpha d - v \rangle$$

$$\Rightarrow \alpha(f'(x,d) - \langle g, d \rangle) \geq f'(x,v) - \langle g, v \rangle$$

  - (2-2) Since the above holds for all $\alpha > 0$, we know that for any $d \in \mathcal{H}$ and $g \in \partial h(v)$, we must have

$$f'(x,d) - \langle g, d \rangle \geq 0$$

otherwise picking $\alpha \to \infty$ would make the LHS in (2-1) approach $-\infty$, a contradiction. Now given this inequality, we pick $\alpha = 0$ and this implies

$$f'(x,v) \leq \langle g, v \rangle$$

  - (2-3) Now from convexity of f, we know that for any $y \in \mathcal{H}$, we have

$$f'(x; y-x) = \lim_{t \to 0} \frac{f(x + t(y-x)) - f(x)}{t} \leq f(y) - f(x)$$

Thus, from first inequality in (2-2), we know that for any $y \in \mathcal{H}$ and $g \in h(v)$, we have

$$f(y) \geq f(x) + f'(x; y-x) \geq f(x) + \langle g, y - x \rangle$$

and thus $g \in \partial f(x)$. This combined with second inequality in (2-2) implies

$$f'(x,v) \leq \langle g, v \rangle \leq \max_{g \in \partial f(x)} \langle g, v \rangle$$

– (4) **Steepest descent direction**

* From max formula, we know that

$$\max_{v \in B} -f'(x;v) = \max_{v \in B} \left\{ - \max_{g \in \partial f(x)} \langle g, v \rangle \right\} = \max_{v \in B} \min_{g \in \partial f(x)} -\langle g, v \rangle = \max_{v \in B} \min_{g \in \partial f(x)} \langle g, v \rangle$$

where last equality comes from $v \in B$ being symmetric with respect to the origin, so we can pick $-v$ instead of v.

* Now from a duality argument (verify this!), we can swap the max and min to get

$$\max_{v \in B} -f'(x;v) = \min_{g \in \partial f(x)} \max_{v \in B} \langle g, v \rangle = \min_{g \in \partial f(x)} \|g\| = dist(0, \partial f(x))$$

where the second equality comes from the Cauchy inequality, with the maximum achieved when $v = \frac{g}{\|g\|}$, and the third equality is achieved when $g = g_0 = \partial^0 f(x)$. Moreover, since we've done a transform $v \to -v$ in the above step, we know that $v = -\frac{g_0}{\|g_0\|}$ achieves the maximum.

2.2 Slope, Steepest Descent, and the Chain Rule

In this section, we first introduce the notion of slope for non-smooth functions, and describe its relation to the steepest descent direction. We then describe the chain rule for computing the slope of a composite function, which will be useful in later sections.

Definition and Properties of Slope

- **Idea**:
  - In the previous lecture, we have shown that we can pick the steepest descent direction by considering all possible directional derivatives.
  - Here we want to relate the notion of steepest descent direction to the slope of the function (which simply maximizes our descent rate without looking at the directions first).
- **Definition of Slope**: For proper function $f : \mathcal{H} \to \mathbb{R}$, and $x \in dom(f)$ not a local minimizer, we define the slope of f at x to be
$$|\nabla f|(x) = \limsup_{z \to x} \frac{f(x) - f(z)}{\|x - z\|}$$
If $x \in dom(f)$ is a local minimizer, then we define $|\nabla f|(x) = 0$.
- **Existence of Slope**: Since x is not a local minimizer, there is some z such that $-A = \frac{f(x)-f(z)}{\|x-z\|} < 0$. Thus, the lim sup is taken over a bounded sequence between $[-A, 0]$, and thus exists.
- **Correspondence of Slope with Gradient for Smooth Functions**
  - **Statement**: Suppose f is C^1 on a neighborhood of x, then we have $|\nabla f|(x) = \|\nabla f(x)\|$.
  - **Proof of $\geq$**:
    * Since f is C^1 in a neighborhood around x, we can write down its first-order Taylor approximation
$$f(z) = f(x) + \nabla f(x)^\top (z - x) + o(\|z - x\|) \text{ as } z \to x.$$
    * Then we have
$$\begin{aligned} |\nabla f|(x) &= \limsup_{z \to x} \frac{f(x) - f(z)}{\|x - z\|} = \limsup_{z \to x} \frac{f(x) - (f(x) + \nabla f(x)^\top (z - x) + o(|z - x|))}{\|x - z\|} \\ &= \limsup_{z \to x} \frac{-\nabla f(x)^\top (z - x) - o(|z - x|)}{\|x - z\|} \end{aligned}$$
    * Now we can pick $z = x + \epsilon \frac{\nabla f(x)}{\|\nabla f(x)\|}$ for small $\epsilon > 0$, and thus we have
$$\begin{aligned} |\nabla f|(x) &= \limsup_{z \to x} \frac{-\nabla f(x)^\top (z - x) - o(\|z - x\|)}{\|x - z\|} \geq \lim_{\epsilon \to 0} \frac{-\epsilon \frac{\nabla f(x)^\top \nabla f(x)}{\|\nabla f(x)\|} - o(\epsilon)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \|\nabla f(x)\| = \|\nabla f(x)\| \end{aligned}$$
  - **Proof of $\leq$**:
    * Suppose that we have $|\nabla f|(x) > \|\nabla f(x)\|$. This means
$$\limsup_{z \to x} \frac{f(x) - f(z)}{\|x - z\|} > \|\nabla f(x)\|$$
    * Thus, we can find some sequence $\{z_n\}$ such that $\lim_{n \to \infty} z_n = x$, with
$$\lim_{n \to \infty} \frac{f(x) - f(z_n)}{\|x - z_n\|} > \|\nabla f(x)\|$$
    * Since f is C^1 around x, we know that for n sufficiently large, we have the Taylor expansion
$$f(z_n) = f(x) + \nabla f(x)^\top (z_n - x) + o(\|z_n - x\|) \text{ as } z \to x$$
    * Plugging this into the above gives
$$\lim_{n \to \infty} \frac{\|\nabla f(x)\| \cdot \|x - z_n\| + \nabla f(x)^\top (z_n - x) + o(\|z_n - x\|)}{\|x - z_n\|} < 0$$

* Thus, we have

$$\lim_{n\to\infty} \frac{\|\nabla f(x)\| \cdot \|x - z_n\| + \nabla f(x)^\top (z_n - x)}{\|x - z_n\|} < 0$$

* However, from Cauchy's inequality, we have

$$|\nabla f(x)^\top (z_n - x)| \leq \|\nabla f(x)\| \cdot \|x - z_n\|$$

* This means we must have

$$\|\nabla f(x)\| \cdot \|x - z_n\| + \nabla f(x)^\top (z_n - x) > 0$$

so the limit above must be greater than or equal to 0, a contradiction. Thus, we must have $|\nabla f|(x) = \|\nabla f(x)\|$.

Steepest Descent and Slope

- **Preliminary: Sum Rule of Subdifferential**: For proper convex $f, g : \mathcal{H} \to \overline{\mathbb{R}}$ with $dom(f) \cap int(dom(g)) \neq \emptyset$, we have $\partial(f + g)(x) = \partial f(x) + \partial g(x)$.
- **Theorem (Correspondence between Slope and Steepest Descent)**: For proper convex functions $f : \mathcal{H} \to \overline{\mathbb{R}}$ and $x \in dom(f)$ not a minimizer, we have

$$|\nabla f|(x) = -\inf_{\|v\|=1} f'(x; v) = dist(0, \partial f(x))$$

- **Note**: This second equality in this theorem looks similar to the max formula theorem, though it's actually stronger: here we only require $x \in dom(f)$, while in the max formula we require $x \in int(dom(f))$.
- **Proof**:
  - **(3) $\geq$ (1)**: We know that for any $y \in \partial f(x)$, we have

$$\frac{f(x) - f(z)}{\|x - z\|} \leq \frac{\langle y, x - z\rangle}{\|x - z\|} \leq \|y\|$$

where first inequality comes from subgradient definition, and second inequality from Cauchy inequality. Thus, taking $\limsup_{x\to z}$ on both sides gives

$$\forall y \in \partial f(x), \|\nabla f\|(x) \leq \|y\| \Rightarrow dist(0, \partial f(x)) \geq |\nabla f|(x)$$

  - **(1) $\geq$ (2)**: From definition of slope, for any unit vector $v \in \mathbb{R}$, we have

$$|\nabla f|(x) \geq \lim_{t\to 0} \frac{f(x) - f(x + tv)}{t} = -f'(x; v)$$

Thus, taking the supremum on both sides gives

$$|\nabla f|(x) \geq -\inf_{\|v\|=1} f'(x; v)$$

  - **(2) $\geq$ (3)**:
    * Consider any $\sigma \in (0, dist(0, \partial f(x)))$. We want to show that $-\inf_{\|v\|=1} f'(x; v) > \sigma$ for all such σ.
    * From definition of σ, we have $\partial f(x) \cap \sigma B = \emptyset$, so

$$0 \notin \partial f(x) + \sigma B = \partial f(x) + \partial(\sigma\|\cdot - x\|)(x) = \partial(f + \sigma\|\cdot - x\|)(x)$$

where last step comes from the sum rule of subdifferential.

    * Thus, we know that x does not minimize $f + \sigma\|\cdot - x\|$, i.e. there exists $z \in \mathcal{H}$ with

$$f(z) + \sigma\|z - x\| < f(x) + \sigma\|x - x\| = f(x)$$

* This means if we pick $v = \frac{z-x}{\|z-x\|}$, then

$$f(x;v) = \frac{1}{\|z-x\|}\lim_{t\to 0}\frac{f(x+t(z-x))-f(x)}{t} \leq \frac{1}{\|z-x\|}(f(z)-f(x)) < -\sigma$$

where the first inequality comes from convexity, and second comes from our obtained inequality on the last step.

* Thus, we know that $-\inf_{\|v\|=1} f'(x;v) > \sigma$, and this proves the theorem.

Chain Rule for Slope of Composite Functions

- **Theorem (Chain Rule for Slope of Proper Bounded Functions)**: For proper $f : H \to \overline{\mathbb{R}}$ and C^1 function $\phi : (\alpha, \beta) \to \overline{\mathbb{R}}$, then if $\alpha < f(x) < \beta$, then

$$|\nabla(\phi \circ f)|(x) = \phi'(f(x))|\nabla f|(x)$$

- **Proof of Chain Rule**:
  - $\leq$:
    * If we already have $|\nabla f|(x) = +\infty$ or $|\nabla(\phi \circ f)|(x) = 0$, then there is nothing to prove.
    * Now suppose we have $|\nabla f|(x) < +\infty$ and $|\nabla(\phi \circ f)|(x) > 0$.
    * We know that $|\nabla(\phi \circ f)|(x) = \lim_{n\to\infty} \frac{\phi(f(x))-\phi(f(x_n))}{\|x-x_n\|}$ for some sequence $x_n \to x$.
    * Since $|\nabla(\phi \circ f)|(x) > 0$, we know that starting from some index N, we will have $\phi(f(x)) \geq \phi(f(x_n))$ for all $n \geq N$. Since $\phi' > 0$, we know that ϕ is increasing, so $f(x) \geq f(x_n)$ for all $n \geq N$.
    * Also, since $|\nabla f|(x) = \limsup_{z\to x} \frac{f(x)-f(z)}{\|x-z\|} < +\infty$, we know that $f(x_n) \to f(x)$. [Otherwise the numerator would be constant while denominator goes to 0, and the limsup would be $+\infty$.]
    * Thus, we have

$$\begin{aligned}|\nabla(\phi \circ f)|(x) &= \lim_{n\to\infty}\frac{\phi(f(x))-\phi(f(x_n))}{\|x-x_n\|} = \lim_{n\to\infty}\frac{\phi(f(x))-\phi(f(x_n))}{f(x)-f(x_n)}\frac{f(x)-f(x_n)}{\|x-x_n\|} \\ &= \phi'(f(x))\lim_{n\to\infty}\frac{f(x)-f(x_n)}{\|x-x_n\|} \leq \phi'(f(x))|\nabla f|(x)\end{aligned}$$

  - $\geq$:
    * Similarly, if we already have $|\nabla f|(x) = 0$ or $|\nabla(\phi \circ f)|(x) = +\infty$, then there is nothing to prove.
    * Now suppose we have $|\nabla f|(x) > 0$ and $|\nabla(\phi \circ f)|(x) < +\infty$.
    * We can apply a similar argument as above to $\phi \circ f$ and $\phi^{-1}(\phi \circ f) = f$:

$$|\nabla f|(x) = |\nabla(\phi^{-1} \circ (\phi \circ f))|(x) \leq (\phi^{-1})'(\phi(f(x)))|\nabla(\phi \circ f)| = \frac{1}{\phi'(f(x))}|\nabla(\phi \circ f)|(x)$$

and this proves this direction.

2.3 Steepest Descent Trajectory, Brezis's Theorem, and Relation to Proximal Point Method

In this section, we introduce steepest descent trajectories as curves generated by following the steepest descent method. We prove that this trajectory does not depend on reparametrization, and has finite length given that the KL inequality holds. Moreover, we describe Brezis's Theorem for the existence and uniqueness of such steepest descent trajectories. The above tools are then used to give some intuition about how we may apply the continuous-time ideas to the proximal point method.

Steepest Descent Trajectory

- **Idea**:

- Now we would like to generate curves of descent for a function, where we always follow the direction of steepest descent.
- Intuitively, the **rate of change of our function value** should be the product of **slope of the function** and the **speed we move along the curve**.

- **Definition (Steepest Descent Trajectory)**:
  - Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be a proper function.
  - A curve $\chi : \mathbb{R}_+ \to \mathcal{H}$ is a **steepest descent trajectory** for f if both χ and $f \circ \chi$ are absolutely continuous, and for all $t \geq 0$, we have

$$\frac{d}{dt}^+ f(\chi(t)) = -|\nabla f|(\chi(t)) \cdot \left\| \frac{d}{dt}^+ \chi(t) \right\|$$

  - Here $\frac{d}{dt}^+$ means the derivative approached from above. (e.g. $\frac{d}{dt}^+ \chi(t) = \lim_{u \to 0+} \frac{\chi(t+u)-\chi(t)}{u}$)
  - Absolutely continuous means we can write the function in terms of the integral of its derivative, as in the form of Fundamental Theorem of Calculus. (e.g. $\chi(t) = \int_0^t \dot{\chi}(u)du$ for some integrable function $\dot{\chi} : \mathbb{R}_+ \to \mathcal{H}$)

- **Note (Sharp Functions)**:
  - The reason that we consider derivative from above is that we often encounter examples of functions where the slope is different if we approach from different directions (such functions are called "sharp functions", as defined below).
  - In such cases, we may favor the consideration of one particular direction (e.g. the KL property, as discussed later, would consider approaching from direction with positive slope).
  - **Definition of sharp function**: For proper function $f : H \to \overline{\mathbb{R}}$, a minimizer $x \in dom(f)$ is **weak sharp** if $\exists \epsilon > 0$ such that $|\nabla f|(x) \geq \epsilon$ for all z close to x with $f(z) \geq f(x)$.
  - **Example (ReLU activation)**: $f(x) = \begin{cases} 0 \text{ , if } x < 0 \\ x \text{ , otherwise} \end{cases}$ is weak sharp at $x = 0$.

- **Theorem (Reparametrization Preserves Trajectory)**
  - **Statement**: For proper $f : \mathcal{H} \to \overline{\mathbb{R}}$ and $\phi : \mathbb{R} \to \mathbb{R}$ a C^1 function with $\phi' > 0$, the steepest descent trajectories for f and $\phi \circ f$ are the same.
  - **Proof**: Let $\chi : \mathbb{R}^+ \to \mathcal{H}$ be a steepest descent trajectory for f. Then we have the following:

$$\begin{aligned} \frac{d}{dt}^+ \phi(f(\chi(t))) &= \phi'(f(\chi(t))) \cdot \frac{d}{dt}^+ f(\chi(t)) \\ &= -\phi'(f(\chi(t))) \cdot |\nabla f|(\chi(t)) \cdot \left\| \frac{d}{dt}^+ \chi(t) \right\| \\ &= -|\nabla(\phi \circ f)|(\chi(t)) \cdot \left\| \frac{d}{dt}^+ \chi(t) \right\| \end{aligned}$$

    where first equality comes from the chain rule for functions on $\mathbb{R}$, second equality comes from χ being a steepest descent trajectory for f, and third equality comes from chain rule for slope.

Conditions for Finite Length Trajectory

- **Idea**: The main theorem we prove here is that if we have the condition called the Kurdyka-Łojasiewicz inequality, we are guaranteed to have a steepest descent trajectory of finite length.

- **Theorem (Finite Length Trajectory under Kurdyka-Lojasiewicz Inequality)**:

- Suppose that $f : \mathcal{H} \to \overline{\mathbb{R}}$ satisfies the Kurdyka-Łojasiewicz inequality: There exists some $\phi : \mathbb{R} \to \mathbb{R}$ that is C^1 with $\phi' > 0$, so that $|\nabla(\phi \circ f)| \geq 1$ along a trajectory $\chi : \mathbb{R}^+ \to \mathcal{H}$.
- Then we have the following upper bound on length of the trajectory:

$$\text{length}(\chi(\cdot), r, s) \leq \phi(f(\chi(r))) - \phi(f(\chi(s)))$$

- Moreover, if we have f bounded below, then the trajectory is of finite length, so it converges to some point.

- **Proof**
  - **Length formula**
    * Let $\chi : \mathbb{R}^+ \to \mathcal{H}$ be one such steepest descent trajectory for f. Then from reparametrization property above, we know that χ is also a steepest descent trajectory for $\phi \circ f$.
    * Since we have $|\nabla(\phi \circ f)| \geq 1$ along χ from KL inequality, we know that

$$-\frac{d}{dt}^+ \phi(f(\chi(t))) = |\nabla\phi \circ f|(\chi(t)) \left\| \frac{d}{dt}^+ \chi(t) \right\| \geq \left\| \frac{d}{dt}^+ \chi(t) \right\|$$

      Now plugging this into the arc length formula gives

$$\text{length}(\chi(\cdot), r, s) = \int_r^s \left\| \frac{d}{dt}^+ \chi(t) \right\| dt \leq \int_r^s \left[-\frac{d}{dt}^+ \phi(f(\chi(t))) \right] dt = \phi(f(\chi(r))) - \phi(f(\chi(s)))$$

      where last equality comes from the Fundamental Theorem of Calculus (applied on absolutely continuous $\phi \circ f$.)
  - **Convergence**
    * Since f is bounded below, we know that $\text{length}(\chi(t), r, s) \leq \phi(f(0)) - \phi(\inf f)$, and this gives a finite upper bound of the length of the trajectory, thus the length is finite.
    * We know that all such finite trajectories must converge to a final endpoint (since you can pick points on the curve that satisfy conditions of a Cauchy sequence). Thus, the steepest descent trajectory converges to some point.

Uniqueness of Steepest Descent Trajectory and Convergence to Minimizer

- **Preliminary: Moreau Envelope**
  - **Definition**: For $f : \mathcal{H} \to \mathbb{R}$ proper closed and convex, we define the **Moreau envelope of f at x** as

$$e_f(x) = \inf_y \left\{ f(y) + \frac{1}{2}\|y - x\|^2 \right\}$$

  - **Properties**:
    * The infimum is attained uniquely at $y = prox_f(x) = (I + \partial f)^{-1}(x)$.
    * e_f is everywhere finite, convex, and C^1-smooth.
    * $\nabla e_f(x) = x - y = x - prox_f(x)$, so proximal point iteration for f is just gradient descent for e_f.
    * $I - prox_f$, $prox_f$ and $2prox_f - I$ are all non-expansive operators with Lipschitz constant 1.
    * f and $e_{\lambda f}$ have the same minimizers, as

$$\frac{1}{\lambda} e_{\lambda f}(x) = \min_y \left\{ f(y) + \frac{1}{2\lambda}\|y - x\|^2 \right\} \leq f(x) \text{ and } \frac{1}{\lambda} e_{\lambda f}(x) \uparrow f(x) \text{ as } \lambda \downarrow 0$$

- **Brezis Theorem (1973)** [5]:
  - **Setting**: Consider proper closed convex function $f : \mathcal{H} \to \overline{\mathbb{R}}$ and starting point $x_0 \in cl(dom(f))$.
  - **Existence and uniqueness of steepest descent trajectory**: Then there exists a unique steepest descent trajectory $\chi : \mathbb{R}^+ \to \mathcal{H}$ with $\chi(0) = x_0$.

- **Objective value decreasing and convergent to infimum**: Along this trajectory, we have $f(\chi(t)) \downarrow \inf f$ converging to the infimum of f.
- **Distance to any minimizer strictly decreasing and convergence to a minimizer**: For any minimizer $\overline{x}$ of f, we have $|\chi(t) - \overline{x}|$ strictly decreasing. Moreover, $\chi(t)$ will converge to a minimizer x^* of f.

- **Idea for Proof**
  - Recall that in steepest descent trajectories, the condition that we need to satisfy has the form
  $$\dot{\chi}(t) \in -\partial f(\chi(t))$$
  - Let $e_{\lambda f}$ denote the Moreau envelope, then from last property above, we have
  $$\lim_{\lambda \to 0} \frac{1}{\lambda} e_{\lambda f}(x) = f(x)$$
  - Thus, we can approximate f with the differentiable $\frac{1}{\lambda} e_{\lambda f}$ to get:
  $$\dot{\chi}(t) = -\nabla \left(\frac{1}{\lambda} e_{\lambda f} \right) (\chi(t)) = -\frac{1}{\lambda}(I - prox_{\lambda f})(\chi(t)) = \frac{1}{\lambda}((I + \lambda \partial f)^{-1} - I)(\chi(t))$$
  so we have replaced the original condition into an ordinary differential equation (where the original set contain is replaced by the equal sign here).
  - Most of the details of the proof revolves around analyzing the dynamics of this ODE under the limit $\lambda \to 0$. A full proof can be seen in [1, Chapter 17]

- **Note**: The convergence in this theorem is the weak convergence in Hilbert Space [6]. Moreover, we cannot obtain results better than weak convergence. Counterexample for no convergence in norm is given by [2].

- **Combining Brezis with KL Inequality**:
  - **Idea**: If we can combine the two theorems above (KL inequality and Brezis Theorem), we can obtain a guarantee for the steepest descent trajectory to be finite (from KL inequality), unique and convergent to a minimizer (from Brezis).
  - **Statement**:
    * Consider proper closed convex function $f : \mathcal{H} \to \overline{\mathbb{R}}$ and $x_0 \in cl(dom(f))$.
    * Suppose KL inequality holds, i.e. There exists some $\phi : \mathbb{R} \to \mathbb{R}$ that is C^1 with $\phi' > 0$, so that $|\nabla(\phi \circ f)|(z) \geq 1$ for any z satisfying $\|z - \overline{x}\| \leq \|x_0 - \overline{x}\|$ and $f(z) \geq \inf f$.
    * Then we have a unique steepest descent trajectory $\chi : \mathbb{R}^+ \to \mathcal{H}$ with $\chi(0) = x_0$ that is of finite length, and it converges strongly to a minimizer.

Motivating Example: Steepest Descent Trajectory and Proximal Point Method

- **Idea**: We would like to identify ways to relate our previous discussion of continuous methods to discrete algorithms.

- **Motivating Example: Steepest Descent Trajectory on Quadratic Form**
  - **Objective function**: $f(x) = \frac{1}{2} x^\top A x$, $A \in \mathbb{R}^{n \times n}$ is symmetric positive definite.
  - **Steepest descent trajectory**:
    * $\chi : \mathbb{R}^+ \to \mathbb{R}^n$ satisfies $\chi(0) = x_0$, $\dot{\chi}(t) \in -\partial f(\chi(t))$.
    * In this case, the subdifferential only contains the gradient, so $\frac{d\chi}{dt} = -A\chi(t)$. This means $\chi(t) = e^{-tA} x_0$ can be written in terms of the matrix exponential.

- **General Case**
  - If we want $\dot{\chi}(t) \in -\partial f(\chi(t))$, a natural analogue is to write $\frac{d\chi}{dt} = -\partial f(\chi(t))$.
  - **Interpretation 1**:

* Informally, we can write the solution in terms of the operator exponential, and perform the following Taylor approximation

$$\chi(t) \in e^{-t\partial f} x_0 = \lim_{m\to\infty} \left(\left(I + \frac{t}{m}\partial f \right)^{-1} \right)^m x_0$$

* Then, we can discretize this algorithm by writing

$$x_{k+1} \in \left(I + \frac{t}{m}\partial f \right)^{-1} x_k \text{ for } k = 0, \cdots, m-1$$

and note that this is exactly the proximal point method step $x_{k+1} = prox_{\frac{t}{m}f}(x_k)$.

* **Theorem (Convergence of proximal point method to steepest descent trajectory)**: The proximal point updates $\{x_k\}$ evaluated by $x_{k+1} = prox_{\frac{t}{m}f}(x_l)$ satisfy $\lim_{m\to\infty} x_m = \chi(t)$, so it converges pointwise to the steepest descent trajectory.

– **Interpretation 2**:

* We can discretize the derivative as

$$\frac{\chi(t) - \chi(t-\delta)}{\delta} \in -\partial f(\chi(t)) \text{ , where } \delta = \frac{t}{m}$$

and this can be seen as the backward discretization in Numerical Analysis.

3 KL Convergence Analysis on Convex Problems

In this section, we consider discrete analogs of steepest descent trajectories - slope descent sequences, and we consider how we may use the KL inequality to derive conditions for convergence. We then describe ways to apply this type of analysis to different optimization frameworks.

3.1 Slope Descent Sequences

In this section, we define slope descent sequences as analogs of steepest descent trajectories in discrete optimization algorithms. We then provide several examples of algorithms that generate slope descent sequences.

3.1.1 Motivation and Definition

- **Motivation**:
  - We would like to mimic the key properties of steepest descent trajectories:
    - (1) **Speed = Slope**: $\left\| \frac{d}{dt}^+ \chi(t) \right\| = |\nabla f|(\chi(t))$
    - (2) **Objective rate of decrease = Speed^2**: $-\frac{d}{dt}^+ f(\chi(t)) = \left\| \frac{d}{dt}^+ \chi(t) \right\|^2$.
- **Slope Descent Sequence**
  - **Definition**: $\{x_k\}_{k\in\mathbb{N}}$ is a **slope descent sequence** for function $f : \mathcal{H} \to \overline{\mathbb{R}}$ if there exists constants $\alpha, \beta \in \mathbb{R}$ such that the following conditions hold:
    - (A) $f(x_k) - f(x_{k+1}) \geq \alpha \|x_k - x_{k+1}\|^2$.
    - (B) $\|x_k - x_{k+1}\| \geq \beta |\nabla f|(x_{k+1})$
  - **Intuition**:
    - Condition (A) corresponds to idea (2) above; this can also be seen as a sufficient descent condition, where we want the improvement of the objective value in each step to be lower bounded by the square of the step size.
    - Condition (B) corresponds to idea (1) above; we can think of this as saying that we don't want to make our steps too small, in case where a potential large improvement (from a large slope) can be made.

3.1.2 Examples of Slope Descent Sequences

In this section, we consider some examples of optimization algorithms and show that the sequence of iterates they produce are indeed slope descent sequences.

Example 1: Proximal Point Method

- **Setting**:
  - Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be a convex function.
- **Method**:
  - Let x_{k+1} minimize the model

$$x_{k+1} = \arg\min_{z\in\mathcal{H}} \left\{ f(z) + \frac{1}{2}\|z - x_k\|^2 \right\}$$

  - From optimality conditions, this is equivalent to $0 \in \partial f(x_{k+1}) + x_{k+1} - x_k$.
- **Proof of Slope Descent Sequence**
  - **Condition (A)**: Since x_{k+1} is a minimizer, we have $f(x_{k+1}) + \frac{1}{2}\|x_{k+1} - x_k\|^2 \leq f(x_k)$, and rearranging terms gives $f(x_k) - f(x_{k+1}) \geq \frac{1}{2}\|x_{k+1} - x_k\|^2$, so $\alpha = \frac{1}{2}$ satisfies the condition.

– **Condition (B)**: Since $0 \in \partial f(x_{k+1}) + x_{k+1} - x_k$, we know that $x_k - x_{k+1} \in \partial f(x_{k+1})$. Since the slope $|\nabla f|(x_{k+1})$ is defined to be the shortest subgradient at x_{k+1}, we know that $\|x_k - x_{k+1}\| \geq |\nabla f|(x_{k+1})$, so $\beta = 1$ satisfies the condition.

Example 2: Variable Metric Method

- **Setting**:
  - Consider differentiable convex function $f : \mathcal{H} \to \mathbb{R}$ with ∇f Lipschitz with constant $L > 0$.
  - Let A_k be matrices that are positive definite and self-adjoint. Let the smallest and largest eigenvalues of A_k be $\underline{\lambda}_k$ and $\overline{\lambda}_k$, and suppose $\inf_k \underline{\lambda}_k \geq \frac{L}{2}$ and $\sup_k \overline{\lambda}_k < \infty$.
- **Method**:
  - Let x_{k+1} minimize the quadratic model

$$x_{k+1} = \arg\min_{z \in \mathcal{H}} \left\{ f(x_k) + \langle \nabla f(x_k), z - x_k \rangle + \frac{1}{2} \langle A(z - x_k), (z - x_k) \rangle \right\}$$

  - Concretely, we have

$$x_{k+1} = x_k - A_k^{-1} \nabla f(x_k)$$

- **Concrete Examples**:
  - **Gradient descent**: Set $A_k = \frac{L}{2} I$, then we have

$$x_{k+1} = x_k - \frac{2}{L} \nabla f(x_k)$$

  - **Newton's method**: Set $A_k = \nabla^2 f(x_k)$, then we have

$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

  - **Quasi-Newton method**: Set A_k to be an approximation of the Hessian $\nabla^2 f(x_k)$.
- **Proof of Slope Descent Sequence**
  - **Condition (A)**: We first show that $f(x_k) - f(x_{k+1}) \geq \alpha \|x_k - x_{k+1}\|^2$ for some constant $\alpha > 0$.
    - Since ∇f is C^1, we have the quadratic upper bound

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^\top (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2$$

    - Also, our iteration step gives

$$x_{k+1} = x_k - A_k^{-1} \nabla f(x_k) \Rightarrow \nabla f(x_k) = A_k (x_k - x_{k+1})$$

      Plugging this into the above and rearranging gives

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \nabla f(x_k)^\top (x_k - x_{k+1}) - \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &= (x_k - x_{k+1})^\top A_k (x_k - x_{k+1}) - \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &\geq \underline{\lambda}_k \|x_{k+1} - x_k\|^2 - \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &\geq \left(\inf_k \underline{\lambda}_k - \frac{L}{2} \right) \|x_{k+1} - x_k\|^2 \end{aligned}$$

      where the second inequality comes from the lower bound on the Rayleigh quotient $\frac{x^\top A x}{x^\top x} \geq \lambda_{min}(A)$.
    - Thus, we can pick $\alpha = \inf_k \underline{\lambda}_k - \frac{L}{2}$ that satisfies this condition.

- **Condition (B)**: We then show that $\|x_k - x_{k+1}\| \geq \beta |\nabla f|(x_{k+1})$ for some constant $\beta > 0$.
  - Since our f is C^1, we know that
  $$|\nabla f|(x_k) = \|\nabla f(x_k)\| = \|A_k(x_k - x_{k+1})\| \leq \|A_k\| \|x_k - x_{k+1}\|$$
  - Also, we know that ∇f is L-Lipschitz, so we have
  $$\|\nabla f(x_k) - \nabla f(x_{k+1})\| \leq L\|x_k - x_{k+1}\|$$
  - Summing the above two formulas up and applying the triangle inequality gives
  $$\|\nabla f(x_{k+1})\| \leq \|\nabla f(x_k)\| + \|\nabla f(x_k) - \nabla f(x_{k+1})\| \leq \big(\|A_k\| + L\big) \|x_k - x_{k+1}\|$$
  - Thus, we have
  $$\|x_k - x_{k+1}\| \geq \frac{1}{\|A_k\| + L} \|\nabla f(x_{k+1})\| \geq \frac{1}{\overline{\lambda}_k + L} \|\nabla f\|(x_{k+1}) \geq \frac{1}{\sup\limits_k \overline{\lambda}_k + L} \|\nabla f\|(x_{k+1})$$
  - Thus we can pick $\beta = \frac{1}{\sup\limits_k \overline{\lambda}_k + L}$ that satisfies the condition.

Example 3: Trust Region Method

- **Setting**:
  - Let $f : \mathcal{H} \to \mathbb{R}$ be a differentiable convex function.
- **Method**:
  - We now consider an algorithm, where each step is defined by finding the minimizer x_{k+1} of a linear approximation in some region around the previous point x_k.
  $$\begin{aligned} \min_z \quad & f(x_k) + \langle \nabla f(x_k), z - x_k \rangle \\ \text{s.t.} \quad & \|z - x_k\| \leq \rho_k \end{aligned}$$
  - In this case, the solution to the above is $\hat{x} = x_k - \rho_k \nabla f(x_k)$.
  - Now we choose to perform the update based on the descent of the algorithm:
    - If $f(x_k) - f(\hat{x}) \geq -\frac{1}{2} \nabla f(x_k)^\top (\hat{x} - x_k)$ (enough descent), then $x_{k+1} = \hat{x}$ and $\rho_{k+1} = 2\rho_k$.
    - Otherwise (not enough descent), $x_{k+1} = x_k$ and $\rho_{k+1} = \frac{1}{2}\rho_k$
- **Note**:
  - We can also replace the linear approximation with a quadratic optimization, which is more commonly seen in papers.
  - This can also be thought of minimizing the Lagrangian that takes the constraint into account:
  $$L(x, z, \lambda) = f(x) + \langle \nabla f(x), z - x \rangle + \lambda \|z - x\|^2$$
- **Proof of Slope Descent Sequence**: The proof is similar to the variable metric method, though with considerably more details due to the conditional update step, and thus is omitted here.

Example 4: Proximal Gradient Method

- **Setting**:
  - Let $g : \mathcal{H} \to \overline{\mathbb{R}}$ be a proper closed convex function.
  - Let $h : \mathcal{H} \to \overline{\mathbb{R}}$ be a convex function with ∇h being L-Lipschitz.
  - We would like to minimize $f(x) = g(x) + h(x)$.

- **Method**:
  - Consider the quadratic upper bound for $h(x)$:

$$\begin{aligned} f(z) &\leq g(z) + h(x) + \langle \nabla h(x), z - x \rangle + \frac{1}{2\lambda}\|z - x\|^2 \\ &= \left(g(z) + \frac{1}{2\lambda}\|z - x + \lambda \nabla h(x)\|^2 \right) + \left(h(x) - \frac{\lambda}{2}\|\nabla h(x)\|^2 \right) \end{aligned}$$

    which is true if $\lambda > 0$ is sufficiently small.
  - Now if we want to minimizer RHS with respect to z, this would be in the form of the proximal point method:

$$x_{k+1} = \min_z \left\{ g(z) + \frac{1}{2\lambda}\|z - x_k + \lambda \nabla h(x_k)\|^2 \right\}$$

    and we would use the following update:

$$x_{k+1} = prox_{\lambda g}(x_k - \lambda \nabla h(x_k))$$

- **Proof of Slope Descent Sequence for** $\lambda \in (0, \frac{2}{L})$:
  - **Condition (A)**: We first show that $f(x_k) - f(x_{k+1}) \geq \alpha \|x_k - x_{k+1}\|^2$ for some constant $\alpha > 0$.
    - Since x_{k+1} is the optimizer, we know (from taking gradient) that

$$0 \in \partial g(x_{k+1}) + \frac{1}{\lambda}\left(x_{k+1} - x_k + \nabla h(x_k)\right)$$

      Rearranging this gives

$$x_k - x_{k+1} - \lambda \nabla h(x_k) \in \partial(\lambda g)(x_{k+1})$$

    - Now we apply definition of subgradient to get (plug in $z = x_k$)

$$\begin{aligned} (\lambda g)(x_k) - (\lambda g)(x_{k+1}) &\geq \langle x_k - x_{k+1} - \lambda \nabla h(x_k), x_k - x_{k+1} \rangle \\ &= \|x_k - x_{k+1}\|^2 + \lambda \langle \nabla h(x_k), x_k - x_{k+1} \rangle \end{aligned}$$

    - Thus, we can plug in the quadratic upper bound to get

$$\begin{aligned} g(x_k) - g(x_{k+1}) &\geq \frac{1}{\lambda}\|x_k - x_{k+1}\|^2 + \langle \nabla h(x_k), x_k - x_{k+1} \rangle \\ &\geq \frac{1}{\lambda}\|x_k - x_{k+1}\|^2 + h(x_{k+1}) - h(x_k) - \frac{L}{2}\|x_k - x_{k+1}\|^2 \end{aligned}$$

      Thus, we have

$$f(x_k) - f(x_{k+1}) \geq \left(\frac{1}{\lambda} - \frac{L}{2}\right)\|x_k - x_{k+1}\|^2$$

      and we can pick $\alpha = \frac{1}{\lambda} - \frac{L}{2}$ that satisfies this condition.
  - **Condition (B)**: We then show that $\|x_k - x_{k+1}\| \geq \beta |\nabla f|(x_{k+1})$ for some constant $\beta > 0$.
    - We start with the same property as above:

$$x_k - x_{k+1} - \lambda \nabla h(x_k) \in \partial(\lambda g)(x_{k+1})$$

    - Rearranging terms on both sides gives

$$\frac{1}{\lambda}(x_k - x_{k+1}) + (\nabla h(x_{k+1}) - \nabla h(x_k)) \in \partial f(x_{k+1})$$

    - Thus, as long as x_{k+1} is not the minimizer, we can apply the theorem on slope above

$$\begin{aligned} |\nabla f|(x_{k+1}) = dist(0, \partial f(x_{k+1})) &\leq \|\frac{1}{\lambda}(x_k - x_{k+1}) + (\nabla h(x_{k+1}) - \nabla h(x_k))\| \\ &\leq \left(\frac{1}{\lambda} + L\right)\|x_k - x_{k+1}\| \end{aligned}$$

      and we can pick $\beta = \frac{1}{\lambda} + L$ that satisfies the condition.

3.2 Convergence Theorem from KL Property

We then describe the KL property, and we prove a general theorem that gives the convergence rates of slope descent sequences given certain KL properties.

3.2.1 KL Property and Convergence Theorem

- **KL Property on a Set**:
  - **Setting**: Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be bounded below with $\inf f = 0$. (For general functions, we can shift it by $f_0 = f - \inf f$ to obtain $\inf f_0 = 0$.)
  - **KL property**: We say that the **KL property holds on set** $X \subseteq \mathcal{H}$ if there exists $\phi : [0, \rho) \to \overline{\mathbb{R}}$ satisfying the following conditions:
    - (1) **Continuity**: ϕ is continuous on $[0, \rho)$.
    - (2) **Concavity**: ϕ is concave on $[0, \rho)$.
    - (3) **Origin**: $\phi(0) = 0$
    - (4) **Strictly Increasing**: ϕ is C^1 on $(0, \rho)$ with $\phi' > 0$ (thus ϕ is strictly increasing)
    - (5) **KL Inequality**: $\forall x \in X$ with $0 < f(x) < \rho$, we have $|\nabla(\phi \circ f)|(x) \geq 1$.

    If such a ϕ exists, we call ϕ a **deregularizer** of f.
- **Lemma (Subdifferential Continuity)**:
  - **Setting**:
    - Let $f : \mathcal{H} \to \mathbb{R}$ be proper, closed and convex.
    - Let $\{x_k\}$ be a sequence with $x_k \to x^*$ (strong convergence).
    - Let $\{x_k\}$ be a sequence with $x_k \to x^*$ (weak convergence).
    - Assume that $y_k \in \partial f(x_k)$ for all k.
  - **Conclusion**: In this case, we have $y^* \in \partial f(x^*)$ (i.e. ∂f is closed), and $f(x_k) \to f(x^*)$.
- **Theorem (Slope Descent Sequence + KL Property Guarantees Convergence)**
  - **Setting and assumptions**
    - Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be closed, proper and convex, with $\inf f = 0$.
    - Suppose $\{x_k\}_{k \in \mathbb{N}}$ is a slope descent sequence for f. This implies
      - (A) $f(x_k) - f(x_{k+1}) \geq \alpha \|x_k - x_{k+1}\|^2$.
      - (B) $|\nabla f|(x_{k+1}) \leq \beta \|x_k - x_{k+1}\|$
    - Suppose for some set $X \subseteq \mathcal{H}$, the KL property holds, and let the deregularizer be $\phi : [0, \rho) \to \mathcal{H}$.
  - **Conclusions**
    - (i) **Convergence of sequence**: x_k converges to some minimizer x^*.
    - (ii) **Decreasing objective value**: $f(x_k)$ is strictly decreasing and converges to $\inf f = 0$.
    - (iii) **Bound on distance to minimizer**: The distance of x_k to the minimizer x^* is bounded by

      $$\|x_k - x^*\| \leq \frac{\beta}{\alpha} \phi(f(x_k)) + \sqrt{\frac{f(x_{k-1}) - f(x_k)}{\alpha}}$$

      for all k.
- **Proof of Theorem**
  - (1) **Fixed point implies minimizer**
    - If at some iteration we have $x_{k+1} = x_k$, then from condition (B), we know that the slope $\|\nabla f\|(x_{k+1}) = 0$, so x_{k+1} is the minimizer, by definition of slope.
    - Then from condition (A), we know that the $x_{k+2} = x_{k+1}$, and so forth, and thus the sequence will stay at the minimizer $x^* = x_{k+1}$. Thus, there is nothing to prove here.
  - (2) **Lower bound on descent of** $\phi(f(x))$: Now suppose $x_{k+1} \neq x_k$ for any k.

* Our goal is to give an lower bound on the difference $\phi(f(x_k)) - \phi(f(x_{k+1}))$, which is used as a proxy to the original $f(x_k)) - f(x_{k+1})$, with the additional KL inequality.
* We can use the chain rule to evaluate the KL inequality:

$$1 \leq |\nabla(\phi \circ f)|(x_k) = \phi'(f(x_k))|\nabla f|(x_k)$$

* We can then compute the following:

$$\phi(f(x_k)) - \phi(f(x_{k+1})) \geq \phi'(f(x_k))(f(x_k) - f(x_{k+1})) \geq \frac{f(x_k) - f(x_{k+1})}{|\nabla f|(x_k)} \geq \frac{\alpha\|x_k - x_{k+1}\|^2}{\beta\|x_{k-1} - x_k\|}$$

where the first inequality comes from concavity of ϕ, second inequality comes from the KL inequality (as simplified above), and third inequality comes from condition (A) and (B) in slope descent conditions.

* Now note that

$$0 \leq \|x_{k-1} - x_{k+1}\|^2 \leq \|x_{k-1} - x_k\|^2 + \|x_k - x_{k+1}\|^2 + 2\|x_{k-1} - x_k\| \cdot \|x_k - x_{k+1}\|$$

plugging this into the above inequality gives

$$\begin{aligned}\phi(f(x_k)) - \phi(f(x_{k+1})) &\geq \frac{\alpha(2\|x_{k-1} - x_k\| \cdot \|x_k - x_{k+1}\| - \|x_{k-1} - x_k\|^2)}{\beta\|x_{k-1} - x_k\|} \\ &= \frac{\alpha}{\beta}(2\|x_k - x_{k+1}\| - \|x_{k-1} - x_k\|)\end{aligned}$$

– (3) **Proof of convergence of sequence**

* Let us define $\lambda_k = \frac{\beta}{\alpha}\phi(f(x_k)) + \|x_{k-1} - x_k\|$. We know that $\lambda_k \geq 0$, since $\phi(0) = 0$ and ϕ is increasing, and $\|x_{k-1} - x_k\| \geq 0$.
* Then the inequality in (2) simplifies to $\lambda_k - \lambda_{k+1} \geq \|x_k - x_{k+1}\|$.
* This implies $\sum_{k=1}^{\infty} \|x_k - x_{k+1}\| = \lambda_1 < \infty$, and thus $\{x_k\}$ is a Cauchy sequence, and thus converges.

– (4) **Proof of decreasing objective value**:

* From Condition (B), we know that there exists some sequence $\{y_k\}$ with $y_k \in \partial f(x_k)$ and $\|y_k\| \leq \beta\|x_k - x_{k+1}\|$.
* From (3), we know that $\|x_k - x_{k+1}\| \to 0$, so $\|y_k\| \to 0$.
* Thus, we know that $x_k \to x^*$ strongly, $y_k \to 0$ weakly, and $y_k \in \partial f(x_k)$. From the subdifferential continuity lemma, we know that $0 \in \partial f(x^*)$ and $f(x_k) \to f(x^*)$. This implies x^* is a minimizer of f, and $f(x_k)$ decreases (from condition (A)) and converges to the minimizer $f(x^*)$.

– (5) **Proof of bound on distance to minimizer**:

* From the inequality in (2), we know that

$$\|x_k - x^*\| \leq \sum_{i=k}^{m-1} \|x_i - x_{i+1}\| + \|x_m - x^*\| \leq \lambda_k - \lambda_m + \|x_m - x^*\|$$

* Thus, if we take $m \to \infty$, we have

$$\|x_k - x^*\| \leq \lambda_k = \frac{\beta}{\alpha}\phi(f(x_k)) + \|x_{k-1} - x_k\| \leq \frac{\beta}{\alpha}\phi(f(x_k)) + \sqrt{\frac{f(x_{k-1}) - f(x_k)}{\alpha}}$$

where last inequality again applies condition (A).

3.2.2 Convergence Comparison to Proximal Point Iteration

- **Idea**

– While the above theorem guarantees convergence and an upper bound on distance to minimizer, this bound by itself does not give us much intuition about the convergence rate of the sequence.

- Here, we would like to relate the convergence rate of slope descent sequences to proximal point method.
- This involves the clever idea of changing from considering properties of f to considering the inverse of the deregularizer $\psi = \phi^{-1}$.

- **Theorem (Convergence of Slope Descent Sequences Compared to Proximal Point Iteration)**
  - **Setting**:
    - Suppose that $f : \mathcal{H} \to \overline{\mathbb{R}}$ is proper, closed and convex, with $\inf f = 0$.
    - Suppose that f satisfies the KL property with desingularizer $\phi : [0, \rho) \to \mathbb{R}^+$.
      - This includes ϕ being C^1, concave, $\phi(0) = 0$, $\phi' > 0$, and ϕ satisfying the KL inequality
      $$|\nabla(\phi \circ f)|(x) = \phi'(f(x))|\nabla f|(x) \geq 1$$
    - Let $\psi = \phi^{-1} : [0, \mu) \to \mathbb{R}^+$.
      - From inverse function theorem, we automatically get ψ being C^1, convex, $\psi(0) = 0$, and the formula
      $$\psi'(t) = \frac{1}{\phi'(\phi^{-1}(t))} = \frac{1}{\phi'(\psi(t))} > 0$$
      - We will further assume that $\psi'(0) = 0$ and ψ' being L-Lipschitz continuous.
    - Let $\{x_k\}$ be a slope descent sequence, i.e. for some constants α, β, we have
      - (A) $f(x_k) - f(x_{k+1}) \geq \alpha \|x_k - x_{k+1}\|^2$.
      - (B) $\|x_k - x_{k+1}\| \geq \beta |\nabla f|(x_{k+1})$
  - **Statement**:
    - Let f, ϕ, ψ and $\{x_k\}$ be defined to satisfy the above conditions (with constants L, α, β known).
    - Let $t_0 = \phi(f(x_0))$, and $t_k = prox^{-}_{\overline{\lambda}\psi}(t_{k-1})$, where $\overline{\lambda} = \sqrt{\frac{2\alpha}{L\beta^2} + \frac{1}{L^2}} - \frac{1}{L}$.
    - Then for some minimizer x^*, we have
    $$\|x_k - x^*\| \leq \frac{\beta}{\alpha} t_k + \sqrt{\frac{\psi(t_{k-1})}{\alpha}}$$

- **Proof of Theorem**
  - (1) **Setting up the proximal iteration**:
    - Let us define τ_k such that $\tau_k = \phi(f(x_k))$, i.e. $\psi(\tau_k) = f(x_k)$.
    - Now define $\lambda_k = \frac{\tau_{k-1} - \tau_k}{\psi'(t_k)}$. This can also be interpreted the definition of τ_k in terms of proximal point method:
    $$\lambda_k = \frac{\tau_{k-1} - \tau_k}{\psi'(t_k)} \Rightarrow \tau_k = (I + \lambda_k \partial\psi)^{-1}(\tau_{k-1}) = prox_{\lambda_k \psi}(\tau_{k-1})$$
    - The remaining goals are two-fold: obtaining a uniform lower bound for the proximal parameter λ_k, and relating τ_k to the original proximal sequence t_k.
  - (2) **Finding a uniform lower bound for λ_k**:
    - (2-1) **Combining slope descent conditions and KL inequality**: We first combine the conditions (A) and (B) in slope descent sequences to get:
    $$\frac{\alpha}{\beta^2} \leq \frac{f(x_{k-1}) - f(x_k)}{|\nabla f|(x_k)^2} \leq \frac{f(x_{k-1}) - f(x_k)}{1/(\phi'(f(x_k)))^2}$$
    where second inequality is given by the KL inequality.
    - (2-2) **Applying the quadratic upper bound**: Recall that for convex functions with Lipschitz gradient, we have the quadratic upper bound
    $$0 \leq f(z) - (f(x) + \langle \nabla f(x), z - x \rangle) \leq \frac{L}{2}\|z - x\|^2$$
    Plugging this into the numerator of above inequality gives
    $$\frac{\alpha}{\beta^2} \leq \frac{\psi(\tau_{k-1}) - \psi(\tau_k)}{\psi'(\tau_k)^2} \leq \frac{\psi'(\tau_k)(\tau_{k-1} - \tau_k) + \frac{L}{2}(\tau_{k-1} - \tau_k)^2}{\psi'(\tau_k)^2}$$

* (2-3) **Plugging in the definition of λ_k**: Putting the definition of λ_k into play, this would reduce to

$$\frac{\alpha}{\beta^2} \leq \lambda_k + \frac{L}{2}\lambda_k^2 \Rightarrow \lambda_k \geq \sqrt{\frac{2\alpha}{L\beta^2} + \frac{1}{L^2}} - \frac{1}{L} = \overline{\lambda}$$

So we have $\overline{\lambda}$ being a lower bound for all λ_k's.

– (3) **Relating τ_k to t_k**: Since each $\lambda_k \geq \overline{\lambda}$, we know from properties of proximal operator that

$$prox_{\overline{\lambda}\psi}(t) \geq prox_{\lambda_k}(t)$$

thus, from induction, we know that $t_0 = \tau_0$ and $t_k \geq \tau_k$ for each $k \in \mathbb{N}$.

– (4) **Upper bound on distance to minimizer**: Finally, we apply the theorem from the last lecture, and obtain

$$\|x_k - x^*\| \leq \frac{\beta}{\alpha}\phi(f(x_k)) + \sqrt{\frac{f(x_{k-1}) - f(x_k)}{\alpha}} \leq \frac{\beta}{\alpha}\tau_k + \sqrt{\frac{\psi(\tau_{k-1})}{\alpha}} \leq \frac{\beta}{\alpha}t_k + \sqrt{\frac{\psi(t_{k-1})}{\alpha}}$$

where second inequality comes from $f(x_{k-1}) - f(x_k) \leq f(x_{k-1}) = \psi(\tau_{k-1})$, and third inequality comes from ϕ increasing.

- **Implication on Convergence Complexity**

– We know that the function $\psi(t)$ has the minimum 0, achieved at $t = 0$. Thus, the sequence t_k would converge to the minimum 0, under the order of convergence of the proximal method applied on ψ.

– In reality, we most often choose $\phi(t) = kx^{1-\theta}$ with $\theta \in [\frac{1}{2}, 1)$, and this would give $\psi(t) = k'x^{\frac{1}{1-\theta}}$ with $\frac{1}{1-\theta} \in (1, 2]$. In this case, the second term $\sqrt{\frac{\psi(t_{k-1})}{\alpha}}$ in the above inequality is also at most of order $O(t_{k-1})$, and thus $\{x_k\}$ would have the same order of convergence as the proximal sequence $\{t_k\}$.

– For other choices of ϕ, we can also deduce separate order of convergence based on the growth of ψ.

3.2.3 Deriving Concrete Convergence Rates from KL Desingularizer

- **Idea**:

– In the previous sections, we have proven the convergence of slope descent sequences, and related this convergence to iterates in a proximal point iteration with desingularizer ψ.

– Our eventual goal is to compute and prove a convergence rate of such slope descent sequences. In the ideal case, we would want our desingularizer ψ to be prox-friendly, so that we can directly compute the convergence rate of proximal iteration on ψ and extend it to slope descent sequences.

– Specifically, for semi-algebraic functions, we can always pick $\psi(s) = ks^\theta$ for some θ, and this guarantees a simple form for the desingularizer.

- **KL Property around a Point**

– **Setting**: Suppose $\min f = 0$ and the minimum is attained at some point x^*.

– **Recall: KL property on set X**: We say that the **KL property holds on set** $X \subseteq \mathcal{H}$ if there exists continuous concave $\phi : [0, \rho) \to \overline{\mathbb{R}}$ with $\phi(0) = 0$, $\phi' > 0$ continuous on $(0, \rho)$, and $|\nabla(\phi \circ f)| \geq 1$ in X.

– **KL property around some point $\overline{x}$**: We say that the KL property holds around $\overline{x}$ if we can pick some set $X = \{x : 0 < f(x) < \alpha, \|x - \overline{x}\| < \epsilon\}$, such that there exists continuous concave $\phi : [0, \rho) \to \mathbb{R}$ with $\phi(0) = 0$, and $\phi' > 0$ continuous on $(0, \rho)$, so that $|\nabla(\phi \circ f)| \geq 1$ in X.

- **Theorem (Existence of Desingularizers for Semi-Algebraic Functions)** [3]

– **Setting**: Suppose that f is **semi-algebraic**, i.e. $epi(f)$ is a finite union of sets defined by finitely many polynomial inequalities.

– **Conclusion**: Then there exists ϕ of the form $\phi(s) = ks^{1-\theta}$ for some $\theta \in (0, 1)$, such that KL property holds uniformly over bounded subsets of $\arg\min_x f(x)$.

- **Error Bound Condition Guaranteeing KL Inequality**
  - **Idea**:
    - Given the existence theorem above, our goal is now to find some concrete $\phi = ks^{1-\theta}$ such that the KL property holds. This would correspond to a concrete convergence rate for our sequence.
    - While the KL inequality $|\nabla(\phi \circ f)| \geq 1$ looks simple, this is not very concrete if we want to check if some ϕ satisfies this condition.
    - Thus, we would like to have a sufficient condition to the KL inequality, which is more concrete for us to check.
  - **Motivation from continuous setting**:
    - **Setting**:
      - Suppose f is proper convex, $\min f = 0$ and the minimum is attained at some point.
      - Suppose KL property holds around some point $\overline{x}$ (this can be a minimizer, or some point close to a minimizer). Let X be the set around $\overline{x}$ in which the KL inequality holds.
      - We would like to see what condition we may derive from the KL property in the continuous setting (as its discrete analog may serve as a part of the sufficient condition).
    - **Derivation**:
      - Consider the steepest descent trajectory ($x_0 \in X$)

$$\begin{cases} \chi(0) = x_0 \\ \chi(t) \in -\partial f(\chi(t)) \text{ a.e.} \end{cases}$$

      - We know from Brezis Theorem that χ will stay in X (as both $f(x)$ and $\|x - x^*\|$ decreases monotonously), and it will converge to some minimizer x^*.
      - Furthermore, the length of the trajectory is upper bounded by

$$length(\chi(\cdot), s, t) \leq \phi(f(\chi(s))) - \phi(f(\chi(t)))$$

      - Taking $s = 0$ and $t \uparrow \infty$ gives

$$\|x_0 - x^*\| \leq length(\chi(\cdot), 0, +\infty) \leq \phi(f(x_0)) - \phi(\min f) = \phi(f(x_0))$$

        and this implies

$$dist_{\arg\min f}(x_0) \leq \phi(f(x_0)) \text{ for } x_0 \text{ around } \overline{x}$$

    - **Error bound condition**:
      - **Condition**: There exists some function ϕ such that

$$dist_{\arg\min f}(x_0) \leq \phi(f(x_0)) \text{ for } x_0 \text{ around } \overline{x}$$

      - **Note**: This is easier to check than the KL inequality - specifically we can try to bound the distance on the LHS using a geometric argument, or algebraically given the structure of our optimization problem.
  - **Theorem (Error bound condition and KL inequality)**
    - **Idea**: Now we would like to do the opposite, where we are given an error bound, and we would like to derive a desingularizer such that KL inequality holds.
    - **Assumptions**:
      - **Homogeneity**: Suppose $\min f = 0$ and the minimum is attained at some point.
      - **Error bound condition**: Suppose that we have the error bound condition

$$dist_{\arg\min f}(x) \leq \phi(f(x))$$

        for some function ϕ on some set X.
      - **Polynomial increase**: Suppose there exists some $c > 0$ such that $s \mapsto \frac{\phi(s)}{s^c}$ is non-decreasing for small $s > 0$.

* **Statement**: Under the above assumptions, we have $\left|\nabla\left(\frac{1}{c}\phi \circ f\right)\right|(x) \leq 1$ for $x \in X$.
* **Proof**:
  · Given any $x \in X$, let $y = proj_{\arg\min f}(x)$. Then we have

$$0 = f(y) \geq f(x) + \langle \partial^0 f(x), y - x\rangle \geq f(x) - \|\partial^0 f(x)\| \cdot \|y - x\|$$

  Since $\|\partial^0 f(x)\| = |\nabla f|(x)$, we can rearrange the above as

$$f(x) \leq |\nabla f(x)| \cdot \|y - x\| = |\nabla f(x)| \cdot dist_{\arg\min f}(x) \leq |\nabla f|(x) \cdot \phi(f(x))$$

  · Since $s \mapsto \frac{\phi(s)}{s^c}$ is non-decreasing for small $s > 0$, we can take the derivative to get

$$\forall s > 0 \text{ small}, \ \phi'(s)s^c - \phi(s)cs^{c-1} \geq 0 \Rightarrow \frac{\phi(s)}{s} \leq \frac{\phi'(s)}{c}$$

  Plugging this into the above gives

$$1 \leq |\nabla f|(x) \cdot \frac{\phi(f(x))}{f(x)} \leq |\nabla f|(x) \cdot \frac{\phi'(f(x))}{c} = \left|\nabla\left(\frac{1}{c}\phi \circ f\right)\right|(x)$$

  and thus $\frac{1}{c}\phi$ satisfies the KL inequality.
* **Note**: For the case where we found ϕ to be of the form $\phi(s) = ks^{1-\theta}$ (which is common in later examples), we can just pick any $c \in (0, 1-\theta]$, so the actual desingularizer for the problem still has the form $\frac{1}{c}\phi = k's^{1-\theta}$ for the same exponent θ, though for some different constant $k' \in \mathbb{R}$.

- **Summary: General Steps of Deriving Complexity Results from KL Property**:
  - **Step 1: Deriving an error bound**: We start by deriving an error bound in the form

$$dist_{\arg\min f}(x_0) \leq \phi(f(x_0)) \text{ for } x_0 \text{ around } \overline{x}$$

    This step may involve arguments using geometry of the problem, or analysis properties like metric regularity.
  - **Step 2: Finding a KL desingularizer**: Now from derivations in the last lecture, we know that given certain conditions, $\varphi = \frac{1}{c}\phi$ is a valid desingularizer for some constant $c > 0$.
  - **Step 3: Deriving the complexity**: From Main Theorem in Lecture 6, the convergence of the slope descent sequence $\{x_k\}$ has the same speed as the convergence of the proximal point method applied on $\psi = \varphi^{-1}$, under certain assumptions on φ. Thus, as long as we know the complexity result for proximal point applied on ψ, we have the same complexity result for our slope descent sequence $\{x_k\}$.

- **Note**:
  - We observe that Step 1 (Error bound) and Step 2 (KL desingularizer) are based only the property of the optimization problem f, while Step 3 (Slope descent sequences) depends on the specific optimization algorithm and the sequence we produce.
  - Thus, this provides a nice decoupling of the analysis of optimization algorithms: analysis on the global property of the optimization problem is separated from analysis on the specific algorithm applied to the problem.

3.3 Examples of KL Analysis for Convex Problems

This section describes how we may apply a KL-type analysis above to specific convex optimization problems. We will consider two specific optimization algorithms, alternating projection and the LASSO problem. Both problems are solved using the proximal gradient method, which is known to generate slope descent sequences as what we've shown above. Thus, the key focus here is how we may derive an error bound for these two problems.

3.3.1 Alternating Projection Method on Convex Feasibility Problem

- **Optimization Problem**:
  - For convex sets C, D, we would like to find $x \in C$, $y \in D$ that minimizes the distance $\|x - y\|$.
  - In the special case where $C \cap D \neq \emptyset$, this reduces to the convex feasibility problem of finding a point $x \in C \cap D$.
- **Alternating Projection as Proximal Gradient Method**
  - **Reformulation**: We rewrite the problem as
$$\min_x \left\{ \delta_C(x) + \frac{1}{2} dist(x, D)^2 \right\}$$
  - **Method**:
    - Take $g(x) = \delta_C(x)$ and $h(x) = \frac{1}{2} dist(x, D)^2$
    - Then we have $\nabla h(x) = x - proj_D(x)$, and Lipschitz constant is $L = 1$.
    - Thus, the proximal gradient method becomes
$$x_{k+1} = prox_{\lambda g}(x_k - \lambda \nabla h(x_k)) = proj_C(x - \lambda(x - proj_D(x))) = proj_C(\lambda proj_D(x) + (1 - \lambda)x)$$
      where the RHS is also known as the "relaxed projection". When $\lambda = 1$, we recover the alternating projection method.
  - **Slope descent sequence**:
    - From previous analysis of proximal gradient method, we know that this algorithm generates a slope descent sequence as long as $\lambda \in (0, 2)$.
    - In particular, we can compute the slope descent sequence parameters $\alpha = \frac{1}{\lambda} - \frac{1}{2}$ and $\beta = \frac{1}{\lambda} + 1$.
  - **Note**: For those who don't know, the idea of the proximal operator is to generalize the projection operator (which is a special case of proximal operator taken on indicator functions). This gives the intuition of why proximal gradient method on this problem gives a generalization of the alternating projection method.
- **Error Bound on Convex Feasibility Problem**
  - **Assumption**:
    - **Distance assumption**: We assume that $dist_D(x) \geq dist_C(x)$. (Otherwise, we can apply the same analysis below to the symmetric problem $f_{alt}(x) = \delta_D(x) + \frac{1}{2} dist_C^2(x)$.)
    - **Non-degenerate intersection**: Let us assume that the intersection contains a small ball around the origin: $\delta B \subseteq C \cap D$. (If 0 is not contained in the intersection, we can shift the two sets to contain the origin.)
  - **Theorem (Error bound on convex feasibility problem)**: We have the following error bound
$$dist_{C \cap D}(x) \leq \left(1 + \frac{2\|x\|}{\delta}\right) \max\{dist_C(x), dist_D(x)\}$$
  - **Proof of error bound**
    - (1) **Rewriting the error bound**: Assume that $dist_C(x), dist_D(x) \leq \frac{d}{2}$ for some constant d. Then we just need to show that
$$dist_{C \cap D}(x) \leq \left(\frac{1}{2} + \frac{\|x\|}{\delta}\right) d$$
    - (2) **Deriving properties of projection**: Let $u = P_C(x)$, and $v = P_D(u)$. Then we have
$$\|u - v\| = \|u - P_D(u)\| \leq \|u - P_D(x)\| \leq \|u - x\| + \|x - P_D(x)\| = dist_C(x) + dist_D(x) \leq d$$
      where first inequality comes from definition of projection, and second inequality is from the triangle inequality.

* (3) **Geometric construction of point in the intersection**: Now define $z = \frac{\delta}{d+\delta}u$. We can show that $z \in C \cap D$:
  - Obviously, we have $z = \frac{d}{d+\delta}u \cdot 0 + \frac{\delta}{d+\delta}u \in C$.
  - Also, we have $z = \frac{\delta}{d+\delta}v + \frac{\delta}{d+\delta}(u-v) = \frac{\delta}{d+\delta}v + \frac{d}{d+\delta}\left(\frac{\delta}{d}(u-v)\right) \in D$, where we have $\frac{\delta}{d}(u-v) \in D$ because $\|u-v\| \leq d$, and this implies $\frac{\delta}{d}(u-v) \in \delta B \in C \cap D \subseteq D$.
* (4) **Deriving error bound based on point in intersection**: Thus, we know that

$$dist_{C\cap D}(x) \leq \|x-z\| \leq \|x-u\| + \|u-z\| \leq \frac{d}{2} + \frac{d}{d+\delta}\|u\| \leq \left(\frac{1}{2} + \frac{\|x\|}{d+\delta}\right)d \leq \left(\frac{1}{2} + \frac{\|x\|}{\delta}\right)d$$

  where third inequality comes from plugging in definition of z, and fourth inequality comes from $\|u\| \leq \|x\|$ from u being a projection.

- **Completing the KL Analysis**
  - From the above theorem, we know that for x in some bounded set, we can pick $k = 1 + \frac{2\|x\|}{\delta} < \infty$ such that

$$dist_{\arg\min f}(x) \leq k\sqrt{f(x)}$$

  - Thus, we can pick the KL desingularizer $\phi(x) = k'x^{\frac{1}{2}}$ for some constant k'.
  - Then the slope descent sequence converge in the same complexity as the proximal point method applied to $\psi(x) = k''x^2$, which is known to have **linear convergence**.

3.3.2 ISTA Method on the LASSO Problem

- **LASSO Problem**:
  - We're given $A \in \mathbb{R}^{n\times m}$ with $n \geq m$, $b \in \mathbb{R}^n$, and $\mu > 0$.
  - Let $f(x) = \frac{1}{2}\|Ax-b\|_2^2 + \mu\|x\|_1$, and we would like to find $\min\limits_{x\in\mathbb{R}^m} f(x)$.
- **ISTA Method as Proximal Gradient Method**:
  - **ISTA (Iterative Shrinkage-Thresholding Algorithm)**:
    * Take $g(x) = \rho\|x\|_1$ and $h(x) = \frac{1}{2}\|Ax-b\|_2^2$.
    * Then we have $\nabla h(x) = A^\top(Ax-b)$, and the Lipschitz constant for h is given by $L = \|A^\top A\|_2$.
    * Thus, the proximal gradient method becomes

$$x_{k+1} = prox_{\lambda g}(x_k - \lambda\nabla h(x_k)) = prox_{\mu\|\cdot\|_1}(x - \lambda A^\top(Ax-b))$$

      which is the iterating step we take in the ISTA algorithm
  - **Slope descent sequence**:
    * From previous analysis of proximal gradient method, we know that this algorithm generates a slope descent sequence as long as $\lambda \in (0, \frac{2}{\|A^\top A\|_2})$.
    * In particular, we can compute the slope descent sequence parameters $\alpha = \frac{1}{\lambda} - \frac{\|A^\top A\|_2}{2}$ and $\beta = \frac{1}{\lambda} + \|A^\top A\|_2$.
- **Properties of the Minimizer Set of LASSO**: $X^* = \arg\min\limits_{x\in\mathbb{R}^m} f$ is nonempty, convex and compact.
  - **Boundedness**: Note that if we plug in $x = 0$, we have $\min f \leq \frac{\|b\|^2}{2}$. Then we must have $\mu\|x\|_1 \leq \frac{\|b\|^2}{2}$, so $\|x\|_1 \leq \frac{\|b\|^2}{2\mu}$. This ensures that $\|x\|_2 \leq \|x\|_1^2 \leq \frac{\|b\|^4}{4\mu^2}$ is bounded.
  - **Closure**: Since f is continuous, picking a limit of minimizers will converge to another minimizer.
  - **Nonemptiness**: f is continuous on a compact set, thus must achieve a minimum at some point.
  - **Convexity**: This follows directly from the LASSO objective function being convex.

- **Rewriting the Problem**: If we introduce additional variable y for the 1-norm, we can rewrite the problem as follows:

$$\begin{array}{ll}\min\limits_{x\in\mathbb{R}^m} & \frac{1}{2}\|Ax-b\|_2^2+\mu y \\ \text{s.t.} & \|x\|_1 \le y \\ & y \le \dfrac{\|b\|^2}{2\mu}\end{array} \iff \begin{array}{ll}\min\limits_{z} & \frac{1}{2}\|Cz-d\|_2^2+\langle e,z\rangle \\ \text{s.t.} & z\in \text{polytope } K\end{array}$$

  where we write $z=\begin{bmatrix}x\\y\end{bmatrix}$ and $C=\begin{bmatrix}A\\\mathbf{0}\end{bmatrix}$, $d=\begin{bmatrix}b\\0\end{bmatrix}$, $e=\begin{bmatrix}\mathbf{0}\\\mu\end{bmatrix}$, and define K to encode all constraints. Let us denote $g(z)=\frac{1}{2}\|Cz-d\|_2^2+\langle e,z\rangle$.

- **Theorem (Error Bound on the LASSO Problem)**: We would like to find an error bound in the form

$$\forall z\in K,\ dist_{\arg\min_K g} \le \nu^2\lambda\sqrt{g(z)-\min_k g}$$

  from which we could deduce linear convergence of ISTA.

- **Idea of Proof of Error Bound**
  - We note that the above problem is further equivalent to

$$\begin{array}{ll}\min\limits_{z} & \frac{1}{2}\|Cz-d\|_2^2+\langle e,z\rangle \\ \text{s.t.} & z\in \text{polytope } K\end{array} \iff \begin{array}{ll}\min\limits_{x\in\mathbb{R}^m} & \frac{1}{2}\|u-d\|_2^2+v \\ \text{s.t.} & Cx=u \\ & \langle e,z\rangle = v \\ & z\in K\end{array} \iff \begin{array}{ll}\min\limits_{x\in\mathbb{R}^m} & \frac{1}{2}\|u-d\|_2^2+v \\ \text{s.t.} & z\in K\cap \begin{bmatrix}C\\e\end{bmatrix}^{-1}\begin{bmatrix}u\\v\end{bmatrix}\end{array}$$

  - To derive an error bound, we will show the metric regularity property

$$dist_{K\cap \begin{bmatrix}C\\e\end{bmatrix}^{-1}\begin{bmatrix}u\\v\end{bmatrix}}(z) \le k\left\|\begin{bmatrix}Cz-u\\\langle e,z\rangle - v\end{bmatrix}\right\|_2$$

    for some constant k.
  - A key ingredient that is involved in proving this metric regularity property is a theorem by Hoffman, which is described nicely in [4]:
    - Let $P_b=\{x: Ax\le b\}$. Hoffman showed that there exists some constant κ only depending on A (but not b), such that
$$dist_{P_b}(x) \le \kappa\|(Ax-b)_+\|$$
    - Note that if we write $f(x)=\|(Ax-b)_+\|$, then $f^{-1}(0)=P_b$, so the above is equivalent to
$$\forall x, dist_{f^{-1}(0)}(x) \le \kappa\cdot dist_0(f(x))$$
      And we see that this is exactly a metric regularity property.
    - Thus, we can think of the Hoffman inequality as a stronger version of metric regularity (for all x,b) under a specific choice of function f (with a polyhedral structure).
  - The details in this proof is actually not that important, and thus is omitted from the lecture.

4 KL Convergence Analysis on Non-Convex Problems: Proximal Alternating Linearized Minimization (PALM)

In this section, we consider a similar KL-type analysis on algorithms for non-convex optimization problems. To analyze these algorithms, we need additional tools from variational analysis.

4.1 Proximal Alternating Linearized Minimization (PALM)

This section describes the Proximal Alternating Linearized Minimization (PALM framework). We first motivate such alternating schemes by considering one of the earliest optimization algorithms that uses an alternating procedure, the coordinate descent. We then describe the entire framework of the PALM algorithm.

Motivation for Alternating Schemes (Coordinate Descent)

- **Idea**: Coordinate descent only works for strictly convex functions.
- **Counterexample (Powell, 1973)**:
  - **Problem**:
    - Let $\phi(x,y,z) = (x-1)_+^2 + (-x-1)_+^2 - yz$.
    - We define $f(x,y,z) = \phi(x,y,z) + \phi(y,z,x) + \phi(z,x,y)$. This is C^1, non-convex, and unbounded below.
  - **Process**: Starting from a point near $(1,1,-1)$, the process of applying $\min_x \min_y \min_z f(x,y,z)$ repeatedly would result in a cycle that never converges (gradient is never small).
- **Goal**: We would like to analyze the specific conditions we need to have for such alternating schemes to work. Here, we will consider how KL inequality enables us to prove convergence for one particular class of alternating schemes.

Proximal Alternating Linearized Minimization (PALM)

- **Setting**: Consider the optimization problem

$$\min_{x,y}\{f(x) + g(y) + H(x,y)\}$$

- **Assumptions**
  - (1) **Conditions on f and g**: The functions $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^m \to \mathbb{R}$ are closed, proper, bounded below, and prox-friendly (but not necessarily convex)
  - (2) **Conditions on H**: The function $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a C^2 function.
  - (3) **Optimization problem bounded below**: The optimization problem is bounded below, i.e. $\inf_{x\in\mathbb{R}^n, y\in\mathbb{R}^m}\{f(x)+ g(y) + H(x,y)\} > -\infty$.
  - (4) **Upper bound on Lipschitz constant**: For each (x_k, y_k), we can pick some $c_k, d_k \in \mathbb{R}$ such that $\frac{c_k}{L_k} \geq \gamma > 1$ and $\frac{d_k}{L_k'} \geq \gamma > 1$, where L_k and L_k' are Lipschitz constants for $\nabla_x H(\cdot, y_k)$ and $\nabla_y H(x_k, \cdot)$, respectively.
- **Algorithm**:
  - **Repeat**:
    - $x_{k+1} \in prox_{\frac{1}{c_k} f}\left(x_k - \frac{1}{c_k}\nabla_x H(x_k, y_k)\right)$
    - $y_{k+1} \in prox_{\frac{1}{d_k} g}\left(y_k - \frac{1}{d_k}\nabla_y H(x_{k+1}, y_k)\right)$
  - **End**

4.2 Preliminaries from Variational Analysis

This section discusses the essentials of variation analysis that are used to analyze non-convex optimization problems.

Preliminary for Analyzing PALM: Basics of Non-Convex Analysis

- **Idea**:
  - Since we do not assume that f and g are convex, we need analog of notions in non-convex analysis to deal with these functions.
  - Specifically, we will generalize the notions of subgradient and slope to limiting subgradient and limiting slope, and demonstrate how these relate to metric regularity.
  - Finally, we use the above notions to provide a definition of KL property for non-convex functions.
- **Regular Subgradient**:
  - **Setting**: Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be finite at $x \in \mathbb{R}^n$.
  - **Definition**: We say that y **is a regular subgradient of** f **at** x (or $y \in \hat{\partial} f(x)$) if we have
    $$f(x+z) \geq f(x) + \langle y, z \rangle + o(z)$$
    where $\lim_{z \to 0} \frac{o(z)}{\|z\|} = 0$.
  - **Note**: We can think of this as a one-sided first-order approximation of f (which is weaker than the first-order lower bound in the subgradient definition for convex functions).
- **Limiting Subgradient**:
  - **Definition**: We say that y **is a limiting subgradient of** f **at** x if there exists sequence $(x_k, y_k) \to (x, y)$ with $f(x_k) \to f(x)$, and $y_k \in \hat{\partial} f(x_k)$ for all k.
  - **Relation to usual subgradient/gradient definition**:
    - If f is closed, proper and convex, then the limiting subgradient is the same as the subgradient definition for convex functions.
    - If f is C^1 around x, then we have $\partial f(x) = \hat{\partial} f(x) = \{\nabla f(x)\}$ being the same as the gradient for differentiable functions.
  - **Properties**:
    - For smooth function h, we have $\partial(f + h)(x) = \partial f(x) + \nabla h(x)$.
- **Critical Point**: We say x **is a critical point** of f if $0 \in \partial f(x)$, i.e. 0 is a limiting subgradient of f at x.
  - Specifically, if x is a minimizer of f, then we have $0 \in \hat{\partial} f(x) \subseteq \partial f(x)$, so x must be a critical point.
- **Limiting Slope**
  - **Setting**: Assume that $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is finite at x.
  - **Slope**: Recall that we defined the slope of f at x to be $|\nabla f|(x) = \limsup_{y \to x} \frac{(f(x) - f(y))_+}{\|x - y\|}$.
  - **Limiting slope**: We define the limiting slope of f at x to be $\overline{|\nabla f|}(x) = \liminf_{\substack{x \to \overline{x} \\ f(x) \to f(\overline{x})}} |\nabla f|(x)$
  - **Idea**:
    - Often times, the function that we consider will have different descent rate in different directions.
    - The slope consider a single direction that gives us the steepest descent, while the limiting slope consider all directions and pick the smallest slope.
  - **Property (Limiting slope and distance of** 0 **to limiting subgradient)**:
    - **Statement**: For closed proper f, we have $\overline{|\nabla f|}(\overline{x}) = dist_{\partial f(\overline{x})}(0)$.
    - **Note**: We have seen the convex version of this theorem in Section 2.2, and this is a generalization of that theorem to non-convex case.

- **Metric Regularity**
  - **Setting**: Suppose $F : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz.
  - **Definition**: We say F is **metrically regular** at $(\overline{x}, F(\overline{x}))$ if we have $\frac{dist_{F^{-1}(y)}(x)}{dist_{\{y\}}(F(x))} \leq K$ locally bounded for any (x, y) in some neighborhood of $(\overline{x}, F(\overline{x}))$.
  - **Equivalent Definition**: The above definition is equivalent to saying that $\inf_{z \in \mathbb{R}^m, \|z\|=1} \overline{|\nabla(z^\top F)|}(\overline{x}) > 0$.

- **Examples of Metric Regularity**
  - **Example 1**:
    - **Setting**: Consider $h : \mathbb{R} \to \mathbb{R}$, $h(x) = \begin{cases} x \text{ , if } x \leq 0 \\ 0 \text{ , if } x > 0 \end{cases}$. We would like to show that this is not metrically regular at $\overline{x} = 0$.
    - **Proof via original definition**:
      - We will directly apply the metric regularity definition on $(0, 0)$.
      - Pick $(x, y) = (\epsilon, \epsilon)$ for some small $\epsilon > 0$, in the neighborhood of $(\overline{x}, h(\overline{x})) = (0, 0)$.
      - Then we have $h^{-1}(y) = h^{-1}(\epsilon) = \emptyset$, and $h(x) = h(\epsilon) = 0$.
      - Thus, the denominator is $dist_{\{y\}}(h(x)) = dist(\epsilon, 0) = \epsilon$, while the numerator $dist_{h^{-1}(y)}(x) = dist(\emptyset, \epsilon) = +\infty$, so their ratio is not locally bounded.
    - **Proof via equivalent definition**
      - We have slope $|\nabla h|(0) = 1$ (as steepest descent is given by left side).
      - We have limiting slope $\overline{|\nabla h|}(0) = 0$ (as approaching from right side would given slope of 0).
      - Thus, $\inf_{z=\pm 1} \overline{|\nabla(z^\top h)|}(0) = 0$, so h is not metrically regular at 0.

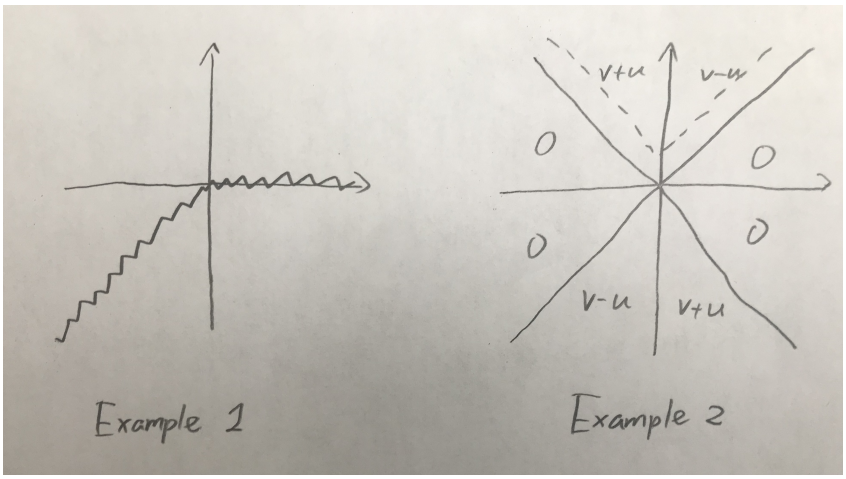

Figure 1: Metric Regularity Examples

  - **Example 2**:
    - **Setting**: Consider $f : \mathbb{R}^2 \to \mathbb{R}$, with values given in the picture above. We will show that f is not metrically regular at $(u, v) = (0, 0)$.
    - **Proof via original definition**:
      - Consider $x = (\epsilon, 0)$, $y = \epsilon^2$ for some small $\epsilon > 0$, in the neighborhood of $\overline{x} = (0, 0)$ and $f(\overline{x}) = 0$.
      - We have $f(x) = f(\epsilon, 0) = 0$, and $f^{-1}(y) = f^{-1}(\epsilon^2)$, consisting of points on the dotted lines above.
      - Then we have $dist_{f^{-1}(y)}(x) > \frac{\epsilon}{\sqrt{2}}$, and $dist_{\{y\}}(f(x)) = \epsilon^2$.
      - Thus, we have $\frac{dist_{f^{-1}(y)}(x)}{dist_{\{y\}}(f(x))} > \frac{1}{\epsilon\sqrt{2}} \to \infty$ if we pick $\epsilon \to 0$, so the ratio is not locally bounded.

* **Proof via equivalent definition**:
  · We have slope $|\nabla f|(0,0) = 1$ (if we go in negative v-axis), and $|\nabla(-f)|(0,0) = 1$ (if we go in positive v-axis)
  · We have limiting slope $\overline{|\nabla f|}(0,0) = \overline{|\nabla(-f)|}(0,0) = 0$ (if we go in u-axis)
  · Thus, $\inf_{z=\pm 1} \overline{|\nabla z^\top f|}(0,0) = 0$, so f is not metrically regular at $(0,0)$.

- **Redefining the KL Property for Non-Convex Case**
  - **Upper slice**: For $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, finite at $\overline{u} \in \mathbb{R}^n$, an **upper slice** of f at $\overline{u}$ is a set of form $\{u | f(\overline{u}) < f(u) < f(\overline{u}) + \rho\}$ for some $\rho > 0$.
  - **KL property on non-convex functions**: We say that KL property holds at $\overline{u}$ if there exists some continuous concave $\phi : [0, \rho) \to \mathbb{R}$ with $\phi(0) = 0$, $\phi' > 0$ and continuous on $(0, \rho)$ such that

$$\forall u \in \{u | f(\overline{u}) < f(u) < f(\overline{u}) + \rho\}, \phi'(f(u) - f(\overline{u})) dist_{\partial f(u)}(0) \geq 1$$

  - **Intuition for the definition**:
    * Recall that for the convex case, the KL property is written as $|\nabla(\phi \circ f)| \geq 1$.
    * Also recall how we previously applied chain rule to the slope: $|\nabla(\phi \circ f)|(x) = \phi'(f(x))|\nabla f|(x) \geq 1$.
    * Here, instead of using original slope definition (which may not be lower-semicontinuous, as noted in [7]), we would use the limiting slope

$$\overline{|\nabla f|}(\overline{u}) = \liminf_{\substack{u \to \overline{u} \\ f(u) \to f(\overline{u})}} |\nabla f|(u)$$

      As stated in the above properties, we know that under certain conditions, we have $\overline{|\nabla f|}(\overline{u}) = dist_{\partial f(\overline{u})}(0)$, where ∂f represents the limiting subdifferential.
    * Also recall that in the original KL property, we requested that f to have a global infimum of 0. Here, to enforce this condition on an upper slice, we would take $f(\cdot) - f(\overline{u}) \in (0, \rho)$.
    * Thus, if we want to write a KL property in the form of $\overline{|\nabla \left(\phi \circ (f - f(\overline{u}))\right)|}(u) \geq 1$, we could simplify it first by chain rule into $\phi'(f(u) - f(\overline{u}))\overline{|\nabla f|}(u) \geq 1$, then transform the limiting slope into the distance to limiting subdifferential: $\phi'(f(u) - f(\overline{u})) dist_{\partial f(u)}(0) \geq 1$. This gives the same expression as above.

4.3 KL Analysis of PALM Algorithm

In this section, we perform the KL analysis on the PALM algorithm. This requires a bit more machinery than our previous KL analysis on convex functions.

4.3.1 Convergence Theorem and Overview of Proof Steps

- **Notations**
  - Let $z = (x, y)$ denote the combined input, and correspondingly $z_k = (x_k, y_k)$.
  - Let $\Phi(z) = \Phi(x, y) = f(x) + g(y) + H(x, y)$.
  - Let Z be the set of limit points of the sequence $\{z_k\}$.

- **Recall: PALM Algorithm**
  - **Repeat**:
    * $x_{k+1} \in prox_{\frac{1}{c_k} f}\left(x_k - \frac{1}{c_k}\nabla_x H(x_k, y_k)\right)$
    * $y_{k+1} \in prox_{\frac{1}{d_k} g}\left(y_k - \frac{1}{d_k}\nabla_y H(x_{k+1}, y_k)\right)$
  - **End**

- **Theorem (Convergence of PALM to a Critical Point)**:

- **Recall: Original assumptions in PALM**
  - (1) **Conditions on f and g**: The functions $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^m \to \mathbb{R}$ are closed, proper, bounded below, and prox-friendly (but not necessarily convex)
  - (2) **Conditions on H**: The function $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a C^2 function.
  - (3) **Optimization problem bounded below**: The optimization problem is bounded below, i.e. $\inf_{x \in \mathbb{R}^n, y \in \mathbb{R}^m} \{f(x) + g(y) + H(x, y)\} > -\infty$.
  - (4) **Upper bound on Lipschitz constant**: For each (x_k, y_k), we can pick some $c_k, d_k \in \mathbb{R}$ such that $\frac{c_k}{L_k} \geq \gamma > 1$ and $\frac{d_k}{L'_k} \geq \gamma > 1$, where L_k and L'_k are Lipschitz constants for $\nabla_x H(\cdot, y_k)$ and $\nabla_y H(x_k, \cdot)$, respectively.
- **Further assumptions beyond what is given in PALM**
  - (5) **Bounded iterates**: Assume that the sequence (z_k) is bounded.
  - (6) **KL property**: Assume that the KL Property (non-convex version) holds at every critical point of Φ. (This holds when Φ is semi-algebraic.)
- **Conclusion**: With the above assumptions, we have (z_k) converging to a critical point of Φ.

- **Outline of Proof**
  - **Step 1: We first show two conditions similar to what we've done for slope descent sequences.**
    - **Proposition 1 (Sufficient descent condition)**: For the sequence $\{z_k\}$, there exists $\rho > 0$ such that $\Phi(z_k) - \Phi(z_{k+1}) \geq \rho \|z_k - z_{k+1}\|^2$.
      - Proof comes from bounds on proximal gradient method that is used to define z_k.
    - **Proposition 2 (Bound on regular subgradient)**: Given that $\{z_k\}$ is bounded, then $\exists \beta > 0$ such that $\forall k$, $\exists w_k \in \hat{\partial}\Phi(z_{k+1})$ with $\|w_k\| \leq \beta \|z_k - z_{k+1}\|$.
      - Proof constructs w_k from optimality conditions for proximal operator.
  - **Step 2: We then show additional properties linking limit points to critical points. This is specific to our problem being non-convex.**
    - **Proposition 3 (Limit points are critical)**: Given that $\{z_k\}$ is bounded, then any limit point $z^* \in Z$ of the sequence is a critical point, i.e. $0 \in \partial\Phi(z^*)$, where $\partial\Phi$ denote the limiting subgradient of Φ.
      - Proof Idea: Proposition 1 ensures that $\|z_k - z_{k+1}\|$ has limit 0 as $k \to \infty$, so $\|w_k\| \to 0$ for Proposition 2. This would allow us to reach the critical point definition (where limiting subgradient goes to 0).
    - **Proposition 4 (Φ is constant on set of limit points Z)**: $\Phi(z^*) = v$ is constant for any limit point $z^* \in Z$.
      - From compactness of Z, which ensures existence of minimizer, then apply convergence definition of limit points.
      - Note that this holds automatically for convex functions (which have all minimizers with same value).
  - **Step 3: We finally apply the KL Property to prove convergence of the algorithm.**
    - **Proposition 5 (KL Property holds at all points)**: KL Property holds at all non-critical points, and thus holds at all points (from assumption that it holds at critical points).
    - **Proposition 6 (KL Property implies desingularizer)**: Suppose that KL Property holds at all points $u \in U$ compact on which σ is constant. Then $\exists \phi$ as a desingularizer, such that KL inequality holds for all u near U on some upper slice.
    - **Final Proof Step**: Assuming existence of desingularizer ϕ, we then have (z_k) converging to a critical point.

4.3.2 Detailed Convergence Proof

- **Lemma [for Proposition 1] (Bound for Update of Proximal Gradient Method)**:
  - **Setting**:

* Let $h : \mathbb{R}^d \to \mathbb{R}$ be C^1 with ∇h being L-Lipschitz.
* Let $\sigma : \mathbb{R}^d \to \mathbb{R}$ be closed, proper, and bounded below.

– **Theorem**: Let $u^+ \in prox_{\frac{1}{t}\sigma}\left(u - \frac{1}{t}\nabla h(u)\right)$, then we have

$$(h(u) + \sigma(u)) - (h(u^+) + \sigma(u^+)) \geq \frac{t - L}{2}\|u^+ - u\|^2$$

– **Proof**:

* From definition of proximal operator, we have

$$\sigma(u^+) + \frac{t}{2}\left\|u^+ - u + \frac{1}{t}\nabla h(u)\right\|^2 \leq \sigma(u) + \frac{t}{2}\left\|\frac{1}{t}\nabla h(u)\right\|^2$$

rearranging this gives

$$\begin{aligned}\sigma(u) - \sigma(u^+) &\geq \frac{t}{2}\left\|u^+ - u + \frac{1}{t}\nabla h(u)\right\|^2 - \frac{t}{2}\left\|\frac{1}{t}\nabla h(u)\right\|^2 \\ &= \frac{t}{2}\|u^+ - u\|^2 + \langle u^+ - u, \nabla h(u)\rangle\end{aligned}$$

* Also, since ∇h is L-Lipschitz, we have the quadratic upper bound

$$h(u) - h(u^+) \geq -\langle u^+ - u, \nabla h(u)\rangle - \frac{L}{2}\|u^+ - u\|^2$$

* Summing the above two inequalities give the requested result.

– **Note**: For convex function σ, we have the stronger condition where we replace $\frac{t-L}{2}$ by $t - \frac{L}{2}$.

- **Proposition 1 (Sufficient Descent Condition)**

– **Setting**: Let $\Phi(z) = \Phi(x, y) = f(x) + g(y) + H(x, y)$.

– **Statement**: There exists $\rho > 0$ such that $\Phi(z_k) - \Phi(z_{k+1}) \geq \rho\|z_k - z_{k+1}\|^2$.

– **Proof**:

* We know that $\frac{c_k}{\gamma} \geq L_k$ is a Lipschitz constant for $\nabla_x H(\cdot, y_k)$. From Lemma above, we have

$$f(x_k) + H(x_k, y_k) - f(x_{k+1}) - H(x_{k+1}, y_k) \geq \frac{c_k - \frac{c_k}{\gamma}}{2}\|x_k - x_{k+1}\|^2$$

where in LHS the terms $g(y_k)$ cancel with each other.

* Similarly, we know that $\frac{d_k}{\gamma} \geq L'_k$ is a Lipschitz constant for $\nabla_y H(x_{k+1}, \cdot)$, so we have

$$g(y_k) + H(x_{k+1}, y_k) - g(y_{k+1}) - H(x_{k+1}, y_{k+1}) \geq \frac{d_k - \frac{d_k}{\gamma}}{2}\|y_k - y_{k+1}\|^2$$

* Summing the above two inequalities gives

$$\begin{aligned}\Phi(z_k) - \Phi(z_{k+1}) &\geq \frac{1}{2}\left(1 - \frac{1}{\gamma}\right)\left[c_k\|x_k - x_{k+1}\|^2 + d_k\|y_k - y_{k+1}\|^2\right] \\ &\geq \frac{1}{2}\left(1 - \frac{1}{\gamma}\right)\min\left\{\inf_k c_k, \inf_k d_k\right\}\|z_k - z_{k+1}\|\end{aligned}$$

* Thus, we can pick $\rho = \frac{1}{2}\left(1 - \frac{1}{\gamma}\right)\min\left\{\inf_k c_k, \inf_k d_k\right\} > 0$, and this satisfies the condition.

– **Corollary**: Since Φ is bounded below, we know that $\|z_k - z_{k+1}\| \to 0$.

- **Proposition 2 (Bound on Regular Subgradient)**

– **Statement**: Assume that $\{z_k\}$ is bounded, then $\exists\beta > 0$ such that $\forall k$, $\exists w_k \in \hat{\partial}\Phi(z_{k+1})$ with $\|w_k\| \leq \beta\|z_k - z_{k+1}\|$.

- **Proof**:
  - (1) **We first try to construct some w_k within the subdifferential set.**
    - From definition, we know that x_{k+1} minimize $\frac{1}{c_k}f + \frac{1}{2}\|\cdot - (x_k - \frac{1}{c_k}\nabla_x H(x_k, y_k))\|$
    - From optimality conditions, we have $0 \in \hat{\partial} f(x_{k+1}) + c_k\left[x_{k+1} - \left(x_k - \frac{1}{c_k}\nabla_x H(x_k, y_k)\right)\right]$.
    - Rearranging the above gives $c_k(x_{k+1} - x_k) - \nabla_x H(x_k, y_k) \in \hat{\partial} f(x_{k+1})$.
      Symmetrically, we also have $d_k(y_{k+1} - y_k) - \nabla_y H(x_k, y_k) \in \hat{\partial} g(y_{k+1})$.
      Moreover, we have
      $$\begin{bmatrix}\nabla_x H(x_{k+1}, y_{k+1})\\ \nabla_y H(x_{k+1}, y_{k+1})\end{bmatrix} = \nabla H(x_{k+1}, y_{k+1}) = \nabla H(z_{k+1}) = \hat{\partial} H(z_{k+1})$$
    - Summing the above up, we have
      $$w_k = \begin{bmatrix}c_k(x_{k+1} - x_k) + \nabla_x H(x_{k+1}, y_{k+1}) - \nabla_x H(x_k, y_k)\\ d_k(y_{k+1} - y_k) + \nabla_y H(x_{k+1}, y_{k+1}) - \nabla_y H(x_k, y_k)\end{bmatrix} \in \hat{\partial}\Phi \begin{bmatrix}x_{k+1}\\ y_{k+1}\end{bmatrix}$$
  - (2) **Now we want to show that $\|w_k\| \le \beta\|z_k - z_{k+1}\|$.**
    - Now for any compact set containing (z_k), we know that both ∇_x and ∇_y are continuous, and thus bounded on that set. This means the Lipschitz constants L_k, L_k' or these two functions are uniformly bounded by some constant N. Then picking $c_k = \gamma L_k$, $d_k = \gamma L_k'$ also ensures that c_k and d_k are uniformly bounded by constant $M = \gamma N$.
    - Since c_k, d_k are uniformly bounded by M, we know that
      $$c_k^2\|x_{k+1} - x_k\|^2 + d_k^2\|x_{k+1} - x_k\|^2 \le M^2\|z_{k+1} - z_k\|^2$$
    - Moreover, since Lipschitz constants of $\nabla_x H$ and $\nabla_y H$ are uniformly bounded by N, we have
      $$\begin{cases}\|\nabla_x H(x_{k+1}, y_{k+1}) - \nabla_x H(x_k, y_k)\| \le N\|z_k - z_{k+1}\|\\ \|\nabla_y H(x_{k+1}, y_{k+1}) - \nabla_y H(x_{k+1}, y_k)\| \le N\|z_k - z_{k+1}\|\end{cases}$$
    - Combining the above, we know that
      $$\|w_k\|^2 \le (M^2 + 4MN + 2N^2)\|z_k - z_{k+1}\|^2$$
      So we can pick $\beta = \sqrt{M^2 + 4MN + 2N^2}$.

- **Proposition 3 (Limit Points are Critical)**
  - **Statement**: Assume that $\{z_k\}$ is bounded, then any limit point z^* of the sequence is a critical point, i.e. $0 \in \partial\Phi(z^*)$, where $\partial\Phi$ denote the limiting subgradient of Φ.
  - **Proof**:
    - Suppose some subsequence of $\{z_k\}$ converges to z^*, denoted by $z_k \xrightarrow{K} z^*$.
    - (1) **We first show that $\liminf_{k \xrightarrow{K} \infty} f(x_k) \ge f(x^*)$.** This is obvious since f is closed with $x_k \xrightarrow{K} x^*$, and this inequality directly comes from lower-semicontinuity.
    - (2) **We then show that $\limsup_{k \xrightarrow{K} \infty} f(x_k) \le f(x^*)$.**
      - We know that x_k minimizes $\frac{1}{c_{k-1}}f + \frac{1}{2}\|\cdot - (x_{k-1} - \frac{1}{c_{k-1}}\nabla_x H(x_{k-1}, y_{k-1}))\|$, so we have
        $$\begin{aligned}&\frac{1}{c_{k-1}}f(x_k) + \frac{1}{2}\left\|x_k - (x_{k-1} - \frac{1}{c_{k-1}}\nabla_x H(x_{k-1}, y_{k-1}))\right\|\\ &\le \frac{1}{c_{k-1}}f(x^*) + \frac{1}{2}\left\|x^* - (x_{k-1} - \frac{1}{c_{k-1}}\nabla_x H(x_{k-1}, y_{k-1}))\right\|\end{aligned}$$
      - Rearranging terms of this gives
        $$f(x_k) + \frac{c_{k-1}}{2}\|x_k - x_{k-1}\|^2 \le f(x^*) + \frac{c_{k-1}}{2}\|x^* - x_{k-1}\|^2 + \langle x_k - x^*, \nabla_x H(x_{k-1}, y_{k-1})\rangle$$

· Now if we take the $\limsup_{k \xrightarrow{K} \infty}$ on both sides, then $\|z_k - z_{k+1}\|$ and $\|z_k - z^*\|$ would shrink to 0, while $\nabla_x H$ would be bounded (as proven in Proposition 3). This implies $\limsup_{k \xrightarrow{K} \infty} f(x_k) \leq f(x^*)$.

* (3) **The results in (1) and (2) combined gives us $f(x_k) \xrightarrow{K} f(x^*)$.** (Note that this is not obvious as f is not assumed to be continuous.)
* (4) **Applying regular subgradient convergence in Proposition 2**: From Proposition 2, we know that $\exists w_k \in \hat{\partial}\Phi(z_{k+1})$ with $\|w_k\| \xrightarrow{K} 0$. This combined with (3) gives exactly the definition of limiting subgradient $0 \in \partial\Phi(z^*)$.

- **Lemma [for Proposition 4] (Properties of Set of Limit Points)**:
  - **Statement**: For any bounded sequence (z_k) with $\|z_k - z_{k+1}\| \to 0$, the set of limit points Z of $\{z_k\}$ is nonempty, compact, connected, and $dist_Z(z_k) \to 0$.
  - **Proof**
    * (1) **We first show that Z is nonempty.**
      · This is straightforward from Bolzano-Weierstrass Theorem: Since (z_k) is bounded, it has a convergent subsequence converging to some limit $z^* \in Z$, so $Z \neq \emptyset$.
    * (2) **We then show that Z is compact.**
      · Since (z_k) is bounded, we know that $\exists M > 0$ s.t. $\|z_k\| \leq M$. Since $B_M(x) = \{x : \|x\| \leq M\}$ is a closed set, any limit of the sequence of points in $B_M(x)$ will also be in $B_M(x)$, and this means any limit point z^* must also satisfy $\|z^*\| \leq M$, thus is bounded.
      · To show that Z is closed, we pick any convergent sequence of limit points (t_k) with $t_k \in Z$, $t_k \to z^*$.
      · Then from definition of limit points, for each t_k we can find a subsequence $z_{i_{k,1}}, z_{i_{k,2}}, \cdots$ with $z_{i_{k,l}} \to t_k$ as $l \to \infty$.
      · Now form the diagonal sequence $\{z_{i_{k,k}}\}$, and remove terms such that the index is increasing. Since each $\|z_{i_{k,k}} - z^*\| \leq \|z_{i_{k,k}} - t_k\| + \|t_k - z^*\| \to 0$ as $k \to 0$, we know that $z_{i_{k,k}} \to z^*$, so $z^* \in Z$. This proves that Z is closed.
      · Thus, Z is closed and bounded, thus compact.
    * (3) **We now show that Z is connected.**
      · Assume otherwise that Z is disconnected. Then we can find two nonempty, disjoint open sets $X, Y \subset \mathbb{R}^d$, such that we can write $Z = A \cup B$ where $A = Z \cap X$, $B = Z \cap Y$, and $A \cap B = \emptyset$.
      · Since X, Y are disjoint, let the shortest distance between points in sets X and Y be $\delta > 0$. Since $\|z_k - z_{k+1}\| \to 0$, we know that there exists N such that $\forall k \geq N$, we have $\|z_k - z_{k+1}\| < \delta$.
      · We now show that after some step N', we must have $z_k \in X$ for all $k \geq N'$ or $z_k \in Y$ for all $k \geq N'$. In that case, we know that either of A or B would be empty, a contradiction.
      · Suppose otherwise (so we jump infinitely many times between the sets X and Y), then we know that after the N-th step, each path from set X to set Y (and vice versa) must contain one point in the sequence that is neither in X or Y. Since we jumped infinitely many times, we can pick a subsequence $\{z_{k_l}\}$ with $z_{k_l} \notin X \cup Y$. Since all z_{k_l}'s are bounded, there is a subsequence of this sequence, convergent to some limits point $z' \notin X \cup Y$, a contradiction to all limit points being in X or Y.
      · Thus, in all cases, we have a contradiction, and that means Z must be connected.
    * (4) **We finally show that $d_Z(z_k) \to 0$ as $k \to \infty$.**
      · Suppose this is not true. Then we have some subsequence $\{z_{k_l}\}_l$ such that for some $\epsilon > 0$, $d_Z(z_{k_l}) \geq \epsilon$ for all l.
      · Then since z_{k_l} is bounded, we can pick a convergent subsequence $z_{k_l} \xrightarrow{L} z^*$. As $z^* \in Z$, we know that $d_Z(z_{k_l}) = 0$, a contradiction to each $d_Z(z_{k_l}) \geq \epsilon$.
      · Thus, we must have $d_Z(z_k) \to 0$ as $k \to \infty$.

- **Proposition 4 (Φ is Constant on Set of Limit Points Z)**
  - **Statement**: $\Phi(z^*) = v$ is constant for any limit point $z^* \in Z$.

- **Proof**:
  - We know that $\Phi(z_k)$ is a decreasing sequence bounded below, so $\Phi(z_k)$ must converge to some value v.
  - For any limit point $z^* \in Z$, we can find some subsequence such that $z_k \xrightarrow{K} z^*$. From our proof of Proposition 3, we know that $\Phi(z_k) \xrightarrow{K} \Phi(z^*)$. This means $\Phi(z^*) = v$.

- **Proposition 5 (KL Property Holds at All Points)**
  - **Statement**: KL Property holds at all non-critical points, and thus holds at all points (from assumption that it holds at critical points).
  - **Proof**:
    - Let $\overline{u}$ be a non-critical point, so $0 \notin \partial f(\overline{u})$. Let $u \in \{u : f(\overline{u}) < f(u) < f(\overline{u}) + \rho\}$ be in the upper slice of $\overline{u}$ with threshold ρ, with u close to $\overline{u}$.
    - (1) **We first show that the limiting subdifferential set $\partial f(\overline{u})$ is bounded away from 0, i.e. $dist_{\partial f(\overline{u})}(0) > \epsilon$ for some $\epsilon > 0$.**
      - Suppose this is not the case, then for any k, there exists some $u_k \in \partial f(\overline{u})$ with $\|u_k\| < \frac{1}{k}$. Then we know that there is a sequence of points $\{u_k\}$ in the limiting subdifferential converging to 0. From properties of limiting subdifferential sets, we know that they are closed, so we have the limit $0 \in \partial f(\overline{u})$, a contradiction.
      - Thus, we must have $dist_{\partial f(\overline{u})}(0) > \epsilon$ for some $\epsilon > 0$.
    - (2) **We then show that there exists some ρ such that $\partial f(u)$ is bounded away from 0 for all u in the upper slice of $\overline{u}$ with threshold ρ, i.e. $dist_{\partial f(u)}(0) > \epsilon'$ for some $\epsilon' > 0$.**
      - Similar to the above, suppose that this is not the case. Then we know that for each $\rho_k = \frac{1}{k}$, there exists some u_k in the upper slice of $\overline{u}$ with threshold ρ_k, and $\exists v_k \in \partial f(u_k)$ with $\|v_k\| < \frac{1}{k}$.
      - Since we assumed that each u_k is close to $\overline{u}$, we can find a compact set containing all the u_k's, and thus it contains some convergent subsequence with $u_k \xrightarrow{K} u'$. Then since u_k is becoming closer to $\overline{u}$, we know that $u_k \xrightarrow{K} \overline{u}$. Note that we also have $\rho_k \to 0$, so $f(u_k) \xrightarrow{K} f(u') = f(\overline{u})$.
      - Then we have $(u_k, v_k) \xrightarrow{K} (\overline{u}, 0)$, $f(u_k) \xrightarrow{K} f(\overline{u})$, and $v_k \in \partial f(u_k)$. From definition of limiting subgradient, we know that $0 \in \partial f(\overline{u})$, a contradiction. Thus, we must have $dist_{\partial f(u)}(0) > \epsilon$ for any u in the upper slice of $\overline{u}$.
    - (3) **We now show that the KL property holds at $\overline{u}$.**
      - To satisfy the KL property, we need to have
        $$\phi'(\sigma(u) - \sigma(\overline{u}))dist_{\partial f(u)}(0) \geq 1$$
        for all u in an upper slice of $\overline{u}$.
      - Since we know that $dist_{\partial f(u)}(0) > \epsilon$, we just need to have
        $$\phi'(\sigma(u) - \sigma(\overline{u})) \geq \frac{1}{\epsilon}$$
      - Now consider functions of the form $\phi(x) = kx^{1-\theta}$, we know that $\phi'(x) = k(1-\theta)x^{-\theta}$, so we just need
        $$k(1-\theta)(\sigma(u) - \sigma(\overline{u}))^{-\theta} \geq \frac{1}{\epsilon} \Rightarrow k \geq \frac{\rho^\theta}{\epsilon(1-\theta)}$$
        where ρ is the threshold we picked for the upper slice.
      - Thus, if we pick some $\theta \in (0, 1)$ and k satisfying the above inequality, then this ϕ would ensure that KL property holds at $\overline{u}$.

- **Proposition 6 (KL Property Implies Global Desingularizer)**:
  - **Idea**: We previously proved KL property locally for each point. Now we would like to extend this to prove a global KL inequality on some compact set.
  - **Statement**: Suppose that KL Property holds at all points $u \in U$ compact on which σ is constant. Then $\exists \phi$ as a desingularizer, such that KL inequality holds for all u near U on some upper slice.

- **Proof sketch**:
  - Since KL property holds at each $u \in U$, we know that at each $u \in U$, there exists open neighborhood N_u (contained in some upper slice of u with threshold ρ_u) where we can find a desingularizer ϕ_u such that KL property holds for all points in N_u.
  - Now note that U is compact, so the open cover $\{N_u : u \in U\}$ of U admits a finite subcover N_{u_i} with $i = 1, \cdots, k$.
  - Thus, we can pick $\rho = \min_i \rho_{u_i} \geq 0$ with $\phi = \sum_{i=1}^{k} \phi_{u_i} : [0, \rho) \to \mathbb{R}$, such that KL inequality holds on $\bigcup_i N_{u_i} \cap \{\text{upper slice for } \rho\}$.

- **Final Step**:
  - **Statement**: Assuming (z_k) is bounded, (c_k, d_k) satisfy given assumptions, and KL holds at every critical point, we then have (z_k) converging to a critical point.
  - **Proof**
    - From Proposition 4, we know that Φ is constant on the set of limit points Z. Then we can apply Proposition 6, and we know that we can find desingularizer ϕ such that KL inequality holds on for all z near Z on some upper slice. From Proposition for Proposition 4, we know that $dist_Z(z_k) \to 0$, the sequence $\{z_k\}$ eventually converges to the upper slice of Z (where KL inequality holds), so we have for large enough k,
    $$\phi'(\Phi(z_k)) dist_{\partial \Phi(z_k)}(0) \geq 1$$
    - The following argument is similar to the corresponding proof of the convex case, described in Section 3.2.1: Since ϕ is concave, we know that
    $$\phi'\left(\Phi(z_k)\right)\left(\Phi(z_{k+1}) - \Phi(z_k)\right) \geq \phi(\Phi(z_{k+1})) - \phi(\Phi(z_k))$$
    - Applying the KL inequality and the slope descent conditions, we have
    $$\begin{aligned}\Phi(z_{k+1}) - \Phi(z_k) &\geq \frac{1}{dist_{\partial \Phi(z_k)}(0)}\left(\phi(\Phi(z_{k+1})) - \phi(\Phi(z_k))\right) \geq \frac{\alpha \|z_k - z_{k+1}\|^2}{2\beta \|z_{k-1} - z_k\|} \\ &\geq \frac{\alpha\left(\|z_k - z_{k+1}\|^2 - (\|z_k - z_{k+1}\| - \|z_{k-1} - z_k\|)^2\right)}{2\beta \|z_{k-1} - z_k\|} \\ &= \frac{\alpha}{2\beta}(2\|z_k - z_{k+1}\| - \|z_{k-1} - z_k\|)\end{aligned}$$
    - Now if we define
    $$\lambda_k = \phi(\Psi(z_k)) + \frac{\alpha}{2\beta}\|z_{k-1} - z_k\| \geq 0$$
    Then we have
    $$\lambda_k - \lambda_{k+1} \geq \frac{\alpha}{2\beta}\|z_k - z_{k+1}\| \Rightarrow \sum_{k=1}^{\infty} \|z_k - z_{k+1}\| \leq \frac{2\beta}{\alpha}\lambda_1 < \infty$$
    and hence $\{z_k\}$ is a Cauchy sequence, and thus converges to some point z^*. From Proposition 3, we know that z^* must be a critical point, and this proves the theorem.

4.4 Examples of PALM Algorithm

In this section, we consider two applications of PALM algorithm: Alternating Projection algorithm (revisited) and matrix factorization.

4.4.1 Alternating Projection

We first show that we can illustrate the alternating projection algorithm as a specific instance of PALM algorithm. This simple examples shows that we can interpret one algorithm in terms of various frameworks of different generality.

- **Problem**: For convex sets C, D, we would like to find $x \in C$, $y \in D$ that minimizes the distance $\|x - y\|$.

- **PALM Formulation**:
  - Pick $f = \delta_C$, $g = \delta_D$, $H(x, y) = \frac{1}{2}\|x - y\|^2$
  - Then $\nabla_x H(\cdot, y)$ and $\nabla_y H(x, \cdot)$ has Lipschitz constant 1, so we can pick $c_k = d_k = \frac{1}{1-\lambda} > 1$ for $\lambda \in (0, 1]$.
  - Plugging this into the original algorithm gives the following:
  - **Repeat**:
    - $x_{k+1} = proj_C\left(\delta x_k + (1-\delta)y_k\right)$
    - $y_{k+1} = proj_D\left((1-\delta)x_{k+1} + \delta y_k\right)$
  - **End**
  - Here, we see that if we pick $\delta = 0$, we recover the original alternating projection algorithm.

4.4.2 Nonnegative Matrix Factorization

This section discusses a more complicated example of PALM: Nonnegative Matrix Factorization.

- **Motivation (Singular Value Decomposition)**
  - **Problem**: Given $A \in \mathbb{R}^{m\times n}$, we would like to find the solution to
$$\min_{X,Y}\left\{\frac{1}{2}\|A - XY\|_F^2 : X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{r\times n}\right\}$$
  - **SVD Solution**:
    - This problem has a known solution from the singular value decomposition (SVD) in linear algebra.
    - Let us write the SVD Factorization $A = \sum_{i=1}^{n} \sigma_i u_i v_i^\top$ where $\{u_i\}, \{v_i\}$ are orthonormal and $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$.
    - Then we can just choose X with columns $\sigma_i u_i$, and Y with rows $v_i^\top$, and this gives a valid solution to the problem.
- **Nonnegative Matrix Factorization**
  - **Problem**: Now say that we want our matrix to have nonnegative entries, i.e. the problem becomes
$$\min_{X,Y}\left\{\frac{1}{2}\|A - XY\|_F^2 : X \in \mathbb{R}_+^{m\times r}, Y \in \mathbb{R}_+^{r\times n}\right\}$$
  - **PALM Formulation**:
    - We can rewrite the above problem as
$$\min_{X,Y}\left\{\delta_{\mathbb{R}_+^{m\times r}}(X) + \delta_{\mathbb{R}_+^{r\times n}}(Y) + \frac{1}{2}\|A - XY\|_F^2\right\}$$
    - Thus, we can let $f(X) = \delta_{\mathbb{R}_+^{m\times r}}(X)$, $g(Y) = \delta_{\mathbb{R}_+^{r\times n}}(Y)$, and $H(X, Y) = \frac{1}{2}\|A - XY\|_F^2$.
  - **Algorithm Derivation**:
    - The above implies $\nabla_X H(X, Y) = (XY - A)Y^\top$ and $\nabla_Y H(X, Y) = X^\top(XY - A)$.
    - Moreover, we have $\nabla_X H(X, Y)$ being $\|Y\|_F^2$-Lipschitz, and $\nabla_Y H(X, Y)$ being $\|X\|_F^2$-Lipschitz.
    - Thus, picking any $\gamma \geq 1$, we have the PALM updates as follows:
    - **Repeat**:
    - $X_{k+1} = \left(X_k - \frac{1}{\gamma\|X_k\|_F}(X_k Y_k - A)Y_k^\top\right)_+$
    - $Y_{k+1} = \left(Y_k - \frac{1}{\gamma\|X_{k+1}\|_F}X_{k+1}^\top(X_{k+1}Y_k - A)\right)_+$
    - **End**
  - **Convergence Guarantee**: If we have (X_k, Y_k) being uniformly bounded, then (X_k, Y_k) converges to a critical point.

- **Sparse Nonnegative Matrix Factorization**
  - **Problem**:
$$\min_{X,Y} \left\{ \frac{1}{2}\|A - XY\|_F^2 : X \in \mathbb{R}_+^{m\times r}, Y \in \mathbb{R}_+^{r\times n}, \|X\|_0 \le \lambda, \|Y\|_0 \le \mu, \right\}$$
    where $\|X\|_0 \le \lambda$ represents that the number of nonzero entries of X is at most λ, and similar for Y.
  - **Idea for Modification**
    - **Notation on index set**: For $X \in \mathbb{R}^I$ with some index set I and subset $J \subseteq I$, let us define X^J by
$$(X^J)_i := \begin{cases} X_i \text{ , if } i \in J \\ 0 \text{ , otherwise} \end{cases}$$
      i.e. this is the restriction of X on index set J.
    - **Lemma (Projection onto the space with sparsity constraint)**:
$$proj_{\{X:\|X\|_0\le\lambda\}}(Z) = \{Z^J : J \subseteq I, |J| \le \lambda, |Z|_j \ge |Z|_i \forall j \in J, i \notin J\}$$
      and also
$$proj_{\{X\ge 0:\|X\|_0\le\lambda\}}(Z) = proj_{\{X:\|X\|_0\le\lambda\}}(Z_+)$$
    - **Algorithm**: The above steps are modified such that after taking positive parts, we zero out the positive entries starting from the smallest entry, until no more than λ (or μ) entries remain for X_{k+1} (or Y_{k+1}).

5 Modified KL Convergence Analysis: The Majorization-Minimization Framework

In this section, we analyze the majorization-minimization framework using KL Analysis. While this problem is also non-convex as for the PALM algorithm in the last section, it requires a modified version of our previous KL Analysis.

5.1 Motivation: Composite Optimization

We first discuss composite optimization as a motivation for such majorization-minimization framework. We introduce the prox-linear algorithm for solving this problem, and write down several examples of this algorithm applied to different problems. Later we'll use KL analysis to show that such algorithms are guaranteed to converge to a critical point.

- **Problem Formulation**: Consider the optimization problem $\inf_{x\in\mathbb{R}^n} g(H(n))$ with $H:\mathbb{R}^n\to\mathbb{R}^m$ being C^2-smooth and $g:\mathbb{R}^m\to\mathbb{R}$ convex, finite, and "simple" (i.e. proximal step below is easy to solve).
- **Prox-Linear Algorithm**: Given $\lambda>0$, we perform the update

$$x_{k+1}=\arg\min_{x\in\mathbb{R}^n}\left\{g\left(H(x_k)+\nabla H(x_k)^\top(x-x_k)\right)+\frac{\lambda}{2}\|x-x_k\|^2\right\}$$

- **Example 1 (Splitting Problem)**:
  - **Problem**: $\inf_{x\in\mathbb{R}^n}\{p(x)+q(x)\}$ with p convex, q smooth.
  - **Composite optimization reformulation**: Define $H(x)=\begin{pmatrix}x\\q(x)\end{pmatrix}$ and $g\begin{pmatrix}x\\t\end{pmatrix}=p(x)+t$.
  - **Prox-linear equivalent to proximal gradient method**:

$$\begin{aligned}
x_{k+1}&=\arg\min_{x\in\mathbb{R}^n}\left\{g\left(\begin{matrix}x\\ q(x_k)+\nabla q(x_k)^\top(x-x_k)\end{matrix}\right)+\frac{\lambda}{2}\|x-x_k\|^2\right\}\\
&=\arg\min_{x\in\mathbb{R}^n}\left\{p(x)+q(x_k)+\nabla q(x_k)^\top(x-x_k)+\frac{\lambda}{2}\|x-x_k\|^2\right\}\\
&=\arg\min_{x\in\mathbb{R}^n}\left\{p(x)+\frac{\lambda}{2}\left\|x-\left(x_k-\frac{1}{\lambda}\nabla q(x_k)\right)\right\|^2+q(x_k)-\frac{1}{\lambda}\|\nabla q(x_k)\|^2\right\}\\
&=\arg\min_{x\in\mathbb{R}^n}\left\{p(x)+\frac{\lambda}{2}\left\|x-\left(x_k-\frac{1}{\lambda}\nabla q(x_k)\right)\right\|^2\right\}\\
&=prox_{\frac{1}{\lambda}p}\left(x_k-\frac{1}{\lambda}\nabla q(x_k)\right)
\end{aligned}$$

    so the prox-linear algorithm is equivalent to the proximal gradient method.
- **Example 2 (Classical Nonlinear Programming)**
  - **Original problem**:

$$\begin{aligned}\min_{x\in\mathbb{R}^n}\quad & p(x)\\ \text{s.t.}\quad & q_i(x)\le 0 \text{ for } i=1,\dots,k\end{aligned}$$

    where p, q_i are all smooth functions.
  - **Exact penalty function formulation**: We can show that above problem is equivalent to

$$\min_{x\in\mathbb{R}^n}\left\{p(x)+\gamma\sum_{i=1}^k q_i^+(x)\right\}$$

with $q_i^+(x) = \max\{q_i(x), 0\}$. Specifically, if we pick γ sufficiently large (dependent on Lipschitz constants of p, q_i), then this formulation has the same local and global minimizers as the original problem.

- **Composite optimization reformulation**:
  - Let $H(x) = \begin{pmatrix} p(x) \\ q(x) \end{pmatrix}$ and $g\begin{pmatrix} u \\ v \end{pmatrix} = u + \gamma \sum_{i=1}^{k} v_i^+$, where $q(x) = \begin{pmatrix} q_1(x) & \cdots & q_k(x) \end{pmatrix}^\top$
  - Then the prox-linear algorithm has the form of

$$x_{k+1} = \arg\min_x \left\{ p(x_k) + \nabla p(x_k)^\top (x - x_k) + \sum_{i=1}^{k} \Big(q_i(x_k) + \nabla q(x_k)^\top (x - x_k) \Big)^+ + \frac{\lambda}{2}\|x - x_k\|^2 \right\}$$

  and we can write this in the form of a simple quadratic program

$$\min_x \left\{ a^\top x + b + \gamma \sum_{i=1}^{k} (a_i^\top x + b_i)^+ + \frac{\lambda}{2}\|x - \overline{x}\|^2 \right\} \iff \begin{array}{ll} \min\limits_x & a^\top x + b + \gamma \sum\limits_{i=1}^{k} t_i + \frac{\lambda}{2}\|x - \overline{x}\|^2 \\ \text{s.t.} & t_i \geq a_i^\top x + b_i, t_i \geq 0 \end{array}$$

  - This is one of the ideas underlying the KNITRO software for solving nonlinear programming.

- **Example 3 (Min-max Problem)**
  - **Original Problem**: $\min\limits_{x \in \mathbb{R}^n} \left\{ \max\limits_{i=1,\ldots,m} f_i(x) \right\}$ where each $f_i : \mathbb{R}^n \to \mathbb{R}$ is convex and C^2.
  - **Composite Optimization Formulation**:
    - Let us define

$$g\begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \max_{i=1,\cdots,m} x_i, H(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix}$$

    - Then the prox-linear algorithm simplifies to

$$\begin{aligned} x_{k+1} &= \arg\min_x \left\{ g\begin{pmatrix} f_1(x_k) + \nabla f_1(x_k)^\top (x - x_k) \\ \vdots \\ f_m(x_k) + \nabla f_m(x_k)^\top (x - x_k) \end{pmatrix} + \frac{\lambda}{2}\|x - x_k\|^2 \right\} \\ &= \arg\min_x \left\{ \max_{i=1,\cdots,m} \left\{ f_i(x_k) + \nabla f_i(x_k)^\top (x - x_k) \right\} + \frac{\lambda}{2}\|x - x_k\|^2 \right\} \end{aligned}$$

    - We can then write this in terms of the quadratic program

$$\begin{array}{ll} \min\limits_x & t + \frac{\lambda}{2}\|x - x_k\|^2 \\ \text{s.t.} & t \geq f_i(x_k) + \nabla f_i(x_k)^\top (x - x_k) \text{ for } i = 1, \cdots, m \end{array}$$

5.2 Majorization-Minimization Framework and KL Analysis

In this section, we introduce the majorization-minimization framework, and use KL analysis to prove the convergence of such algorithms to a critical point.

5.2.1 Majorize-Minimize Framework and Overview of Convergence Proof

Majorize-Minimize Framework

- **Problem**: We would like to minimize a closed, proper, continuous function $f : \mathbb{R}^n \to \mathbb{R}$.
- **Majorizing Model**: Assume that we have a **majorizing model** $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ continuous and satisfying
  - **Majorizing Property**: $\forall x, f(y) \leq h(x, y) = f(y) + o(\|x - y\|^2)$ as $y \to x$.

- **Strong convexity**: For all x, $h_x(\cdot) = h(x, \cdot)$ is μ-strongly convex, i.e. $h_x - \frac{\mu}{2}\|\cdot\|^2$ is convex.

- **Iteration Map**: We define the **iteration map** $p : \mathbb{R}^n \to \mathbb{R}^m$ to be $p(x) = \arg\min h_x$.
  - This is well-defined since strongly convex functions have a unique minimizer.

- **Surrogate Function**: Instead of considering the original function $f(x)$, we will consider the surrogate function $F(x) = \min h_x = h(x, p(x))$.

- **Further Assumptions**
  - **Bounded Sequence**: Assume that the iteration sequence (x_k) is bounded.
  - **Subdifferential Properties of F**: Assume that F is subdifferentially continuous with ∂f closed.
  - **KL Property**: Assume that the KL Property (non-convex version) holds at every critical point of F. (This holds when F is semi-algebraic, or for h being semi-algebraic.)
  - **Model Qualification Assumption**: Given any $\alpha > 0$, $\exists K > 0$ so that for any $x, y \in \mathbb{R}^n$ with $\|x\|, \|y\| < \alpha$, we have $h^y = h(\cdot, y)$ being locally Lipschitz around x with $(\text{lip } h^y)(x) = K\|x - y\|$.

Overview of Convergence Proof of Majorize-Minimize Framework

- **Modification from Analysis of PALM Algorithm**
  - **Modifying Slope Descent Conditions**: Since the prox-linear algorithm does not satisfy the usual slope descent conditions, we will apply a variant of the slope descent condition, where the indices are shifted by 1.
  - **Applying KL on Surrogate Function**: The above modification of slope descent conditions do not necessarily hold on the original objective function f. Thus, we will prove these conditions on the surrogate function F, and then deduce a KL property on F instead.

- **Step 1 (Variant of Slope Descent Conditions)**: We first show the analogous slope descent properties that hold on the surrogate function F. This guarantees our application of KL property.
  - **Proposition 1 (Sufficient Descent Condition Variant)**: $F(x_{k-1}) - F(x_k) \geq \lambda\|x_k - x_{k+1}\|^2$
  - **Proposition 2 (Slope Property Variant)**: $dist_{\partial F(x_{k-1})}(0) \leq \mu\|x_{k-1} - x_k\|$

- **Step 2 (Iterations Converge to Critical Points of f)**: We now show that whenever the iteration converges, it will converge to a critical point of f.
  - **Proposition 3 (Continuity of Iteration Map / Convergence to Fixed Point)**: The iteration map p is continuous. Thus, if we have $x_{k+1} = p(x_k)$ converging to x, then x is a fixed point of p.
  - **Proposition 4 (Fixed Points are Critical Points)**: Fixed points of the iteration map p are critical points of the objective function f.

- **Step 3 (Applying KL Property to Prove Convergence)**
  - Now we just need to prove convergence of the sequence from the KL property.
  - **Proposition 5 (KL Property and Existence of Desingularizer)**: Assuming KL Property holds on all critical points, we can then find a desingularizer ϕ satisfying the standard properties and the KL inequality.
  - **Theorem (Convergence of Sequence to Critical Point)**: Given the desingularizer ϕ and (x_k) satisfying the variant of slope descent conditions above, the sequence (x_k) converges. Thus, from Step 2 we know that (x_k) converges to a critical point of f.

- **Extra Step (Extending Semi-Algebraic Properties)**
  - **Proposition 6 (Extension of Semi-Algebraic Property)**: If we have $h(x, y)$ being semi-algebraic, then F is also semi-algebraic, and thus we can apply the KL argument above.

5.2.2 Detailed Convergence Proof

- **Proposition 1 (Sufficient Descent Condition Variant)**
  - **Setting**: Recall that we defined $p(x) = \arg\min_x h_x$, and $F(x) = \min_y h_x(y) = h(x, p(x))$.
  - **Assumption**: We assume that h_x is μ-strongly convex.
  - **Theorem**: We have $F(x_{k-1}) - F(x_k) \geq \frac{\mu}{2}\|x_k - x_{k+1}\|_2^2$.
  - **Proof**:
    - We know that for any $x \in \mathbb{R}^n$, the majorizing property implies

      $$f(x) \leq h(x, x) \leq f(x) + o(\|x - x\|^2) \Rightarrow h(x, x) = f(x)$$

      and thus we have

      $$F(x) = h(x, p(x)) \geq f(p(x)) = h(p(x), p(x))$$

      where the inequality also comes from the majorizing property.
    - Also, from definition of F applied at $p(x)$, we have

      $$F(p(x)) = h(p(x), p(p(x)))$$

      and thus we have

      $$F(x) - F(p(x)) \geq h(p(x), p(x)) - h(p(x), p(p(x))) \geq \frac{\mu}{2}\|p(x) - p(p(x))\|^2$$

      where last inequality comes from the μ-strong convexity of h_x. Thus, we have

      $$F(x_{k-1}) - F(x_k) \geq \frac{\mu}{2}\|x_k - x_{k+1}\|_2^2$$

- **Proposition 2 (Slope Property Variant)**
  - **Setting**:
    - Assume that we have the **Model Qualification Assumption**: Given any $\alpha > 0$, $\exists K > 0$ so that for any $x, y \in \mathbb{R}^n$ with $\|x\|, \|y\| < \alpha$, we have $h^y(x) = h(x, y)$ being locally Lipschitz around each x with Lipschitz constant $K\|x - y\|$.
  - **Recall: Facts from Variational Analysis**
    - **Fact 1**: If q is L-Lipschitz, then for any x, we have $\hat{\partial} q(x) \subseteq LB$ where B is the unit ball.
    - **Fact 2**: If F is continuous, then $\hat{\partial} F(x)$ is nonempty at a dense set of points. i.e. For any (x, y) satisfying $y \in \partial F(x)$, we can find some sequence $\{(x_k, y_k)\} \to (x, y)$ such that $y_k \in \hat{\partial} F(x_k)$.
  - **Theorem**: We have $dist_{\partial F(x_{k-1})}(0) \leq \mu\|x_{k-1} - x_k\|$.
  - **Proof**:
    - (1) We know from Proposition 3 (proven below) that p is continuous, so for any compact set X containing iterates x, the corresponding $p(X)$ is also compact. Thus, both sets can be bounded by some ball of radius $\alpha > 0$.
    - (2) Since we have assumed that $h^y(x)$ is locally Lipschitz with $L = K\|x - y\|$ (on some compact set $X \subseteq \alpha B$ containing all iterates and corresponding compact $p(X)$), we know from Fact 1 that

      $$\forall y \in \mathbb{R}^n, \hat{\partial} h^y(x) \subseteq K\|x - y\|B \Longrightarrow \hat{\partial} h^{p(x)}(x) \subseteq K\|x - p(x)\|B$$

    - (3) Now note that $h^{p(x)}(x) = h(x, p(x)) = F(x)$ and for all z, $h^{p(x)}(z) = h(z, p(x)) \geq F(z)$. Thus, for any $y \in \hat{\partial} F(x)$, we have

      $$h^{p(x)}(x + y) \geq F(x + y) \geq F(x) + \langle y, z\rangle + o(z) = h^{p(x)}(x) + \langle y, z\rangle + o(z)$$

      and thus $y \in \hat{\partial} h^{p(x)}(x)$, so we have $\hat{\partial} F(x) \subseteq \hat{\partial} h^{p(x)}(x)$, and thus

      $$\hat{\partial} F(x) \subseteq \hat{\partial} h^{p(x)}(x) \subseteq K\|x - p(x)\|B$$

      holds on some compact set $X \subseteq \alpha B$ containing all iterates.

* (4) Now for any $x \in X \subseteq \alpha B$ and $y \in \partial F(x)$, from Fact 2, we know that there exists $\{(x_k, y_k)\}$ with $(x_k, y_k) \to (x, y)$ and $y_k \in \hat{\partial} F(x_k)$. Eventually, $\|x_k\| < \alpha$, and thus from (3) we have $\|y_k\| \leq K\|x_k - p(x_k)\|$. Thus, taking the limit on both sides, we know that $\|y\| \leq K\|x - p(x)\|$.
* (5) Since the process of (4) holds for any pair (x, y) satisfying $y \in \partial F(x)$, we know that there $\exists K > 0$ with

$$dist_{\partial F(x)}(0) \leq K\|x - p(x)\|$$

on some compact set X containing all iterates, and this proves the requested inequality.

- **Lemma [for Proposition 3] (Quadratic Growth in Model)**:
  - **Assumption**: We assume that h_x is μ-strongly convex.
  - **Theorem**: We have for any $y \in \mathbb{R}^n$,

$$h_x(y) - \min h_x \geq \frac{\mu}{2}\|y - p(x)\|^2$$

  - **Proof**:
    * Since h_x is μ-strongly convex, we know from definition that $h_x(y) - \frac{\mu}{2}\|y\|^2$ is convex.
    * This implies $h_x - \frac{\mu}{2}\|\cdot - p(x)\|^2$ is also convex, since it is the above expression minus a linear function.
    * Now since $p(x) = \arg\min_x h_x$, we know from optimality conditions that $0 \in \partial h_x(p(x))$.
    * Since 0 also minimizes $\frac{\mu}{2}\|y - p(x)\|^2$, we know from sum rule that $0 \in \partial(h_x - \frac{\mu}{2}\|\cdot - p(x)\|^2)(p(x))$.
    * Thus, we know that $p(x)$ minimizes $h_x - \frac{\mu}{2}\|\cdot - p(x)\|^2$, and this implies the requested result.
- **Proposition 3 (Continuity of Iteration Map / Limit as Fixed Point)**
  - **Theorem**: We have the iteration map p being continuous.
  - **Proof**:
    * Suppose that we have some convergent sequence $x_k \to x$, where we assume that $p(x)$ is bounded.
    * **(1) We first show that $\{p(x_k)\}$ is bounded.**
      · From Lemma above, we know that at $x = x_k$ and $y = 0$, we have

$$h(x_k, 0) \geq \min h_{x_k} + \frac{\mu}{2}\|p(x_k)\|^2 \geq \inf f + \frac{\mu}{2}\|p(x_k)\|^2$$

      · Thus, taking the limsup on both sides gives

$$h(x, 0) \geq \inf f + \frac{\mu}{2}\limsup_k \|p(x_k)\|^2$$

      since h is continuous. Since $h(x, 0) \leq o(\|x\|^2)$ and f is proper (i.e. $\inf f > -\infty$), we know that $(p(x_k))$ is bounded.
    * **(2) We then show that $p(x_k) \to p(x)$.**
      · Suppose $p(x_k) \not\to p(x)$. Since $p(x_k)$ is bounded, we know that there is a convergent subsequence that does not converge to $p(x)$. WLOG let $p(x_k)$ itself be the sequence, so $p(x_k) \to y \neq p(x)$.
      · Then from Lemma above with $x = x_k$ and $y = p(x)$, we have

$$h(x_k, p(x)) - h(x_k, p(x_k)) \geq \frac{\mu}{2}\|p(x) - p(x_k)\|^2$$

      where we used $\min h_x = h(x, p(x))$.
      · Now we take the limit as $k \to \infty$ to get

$$h(x, p(x)) - h(x, y) \geq \frac{\mu}{2}\|p(x) - y\|^2 > 0 \Rightarrow h(x, y) < h(x, p(x))$$

      and this contradicts with the fact that $p(x)$ is the minimizer of h_x.
      · Thus, we must have $p(x_k) \to p(x)$, so p is continuous.

- **Corollary (Limit as fixed point)**: If the iteration map $x_{k+1} = p(x_k)$ converges to x^*, then x^* is a fixed point of p.

- **Proposition 4 (Fixed Points are Critical Points)**:
  - **Theorem**: Fixed points x of the iteration map p are critical points of the objective function f.
  - **Proof**:
    - **(1) We first show that $\hat{\partial} f(x) = \hat{\partial} h_x(x)$.**
      - (1-1) On one hand, let $y \in \hat{\partial} f(x)$. Then we have for all z,

        $$h_x(x+z) = h(x, x+z) \geq f(x+z) \geq f(x)+\langle y, z\rangle+o(z) = h(x,x)+\langle y, z\rangle+o(z) = h_x(x)+\langle y, z\rangle+o(z)$$

        and this implies $y \in \hat{\partial} h_x(x)$, so $\hat{\partial} f(x) \subseteq \hat{\partial} h_x(x)$.
      - (1-2) On the other hand, let $y \in \hat{\partial} h_x(x)$, then we have for all z,

        $$f(x+z) + o(\|z\|^2) \geq h_x(x+z) \geq h_x(x) + \langle y, z\rangle + o(z) \geq f(x) + \langle y, z\rangle + o(z)$$

        and thus we can combine the first-order small terms and get $y \in \hat{\partial} f(x)$, so $\hat{\partial} h_x(x) \subseteq \hat{\partial} f(x)$.
    - (2) Now since $p(x)$ minimizes h_x, we will have $0 \in \hat{\partial} h_x(p(x))$. Since x is a fixed point, we have $p(x) = x$, so $0 \in \hat{\partial} h_x(x) = \hat{\partial} f(x) \subseteq \partial f(x)$, so x is a critical point of f.

- **Proposition 5 (KL Property and existence of desingularizer)**
  - **Notation**:
    - Let us define the sequence of iterations by $x_{k+1} = F(x_k)$.
    - Let X be the set of limit points of the sequence $\{x_k\}$.
  - **Assumption**:
    - Assume that the sequence (x_k) is bounded.
    - Assume that F satisfies the KL property on all critical points. This holds when F is semi-algebraic, for example.
    - Assume that F is subdifferentially continuous with ∂f closed.
  - **Statements**:
    - **Properties of Set of Limit Points**: X is nonempty, compact, connected, with $dist_X(x_k) \to 0$.
    - **F is constant on set of limit points**: F has constant value on all of X.
    - **Existence of desingularizer**: There exists a single desingularizer $\phi : [0, \rho) \to \mathbb{R}^+$ that is concave, continuous, and $\phi' > 0$ on $(0, \rho)$ such that $\phi'(F(x)) dist_{\partial F(x)}(0) \geq 1$ for all x close to X with $F(x) > 0$ small enough.
  - **Proof**: This is similar to what we have done in the PALM algorithm analysis, thus omitted here.

- **Theorem (KL Property and Convergence of Sequence to Critical Point)**
  - **Statement**: Given the existence of KL desingularizer ϕ for F, we can show that the iteration sequence (x_k) is convergent.
  - **Proof**:
    - Since ϕ is concave, we know that

      $$\phi(F(x_k)) - \phi(F(x_{k-1})) \leq \phi'(F(x_{k-1}))(F(x_k) - F(x_{k-1}))$$

    - Then reversing the sign and plugging in the KL inequality $\phi'(F(x_{k-1})) dist_{\partial F(x_{k-1})}(0) \geq 1$ to get

      $$\begin{aligned}
      \phi(F(x_{k-1})) - \phi(F(x_k)) &\geq \frac{F(x_{k-1}) - F(x_k)}{dist_{\partial F(x_{k-1})}(0)} \geq \frac{\lambda \|x_k - x_{k+1}\|^2}{\mu \|x_{k-1} - x_k\|} \\
      &\geq \frac{\lambda \left[\|x_k - x_{k+1}\|^2 - (\|x_k - x_{k+1}\| - \|x_{k-1} - x_k\|)^2\right]}{\mu \|x_{k-1} - x_k\|} \\
      &= \frac{\lambda \left(2\|x_{k-1} - x_k\| \cdot \|x_k - x_{k+1}\| - \|x_{k-1} - x_k\|^2\right)}{\mu \|x_{k-1} - x_k\|} \\
      &= \frac{\lambda}{\mu} \left(2\|x_k - x_{k+1}\| - \|x_{k-1} - x_k\|\right)
      \end{aligned}$$

where the second inequality applies the slope descent property variants in Proposition 1 and 2, third equality subtracts a strictly positive term in the numerator, and the follows are given by algebra.

* Now if we define

$$\alpha_k = \phi(F(x_{k-1})) + \frac{\lambda}{\mu}\|x_{k-1} - x_k\| \geq 0$$

then we have

$$\begin{aligned}\alpha_k - \alpha_{k+1} &= \big(\phi(F(x_{k-1})) - \phi(F(x_k))\big) + \frac{\lambda}{\mu}\left(\|x_{k-1} - x_k\| - \|x_k - x_{k+1}\|\right) \\ &\geq \frac{\lambda}{\mu}\left(2\|x_k - x_{k+1}\| - \|x_{k-1} - x_k\|\right) + \frac{\lambda}{\mu}\left(\|x_{k-1} - x_k\| - \|x_k - x_{k+1}\|\right) = \frac{\lambda}{\mu}\|x_k - x_{k+1}\|\end{aligned}$$

* This implies

$$\sum_{k=0}^{\infty}\|x_k - x_{k+1}\| \leq \frac{\mu}{\lambda}\sum_{k=0}^{\infty}(\alpha_k - \alpha_{k+1}) \leq \frac{\mu\alpha_0}{\lambda}$$

and thus (X_k) is a Cauchy sequence, and it converges.

– **Remark**: This combined with Proposition 4 shows that under the above assumptions, (x_k) converges to a critical point of f.

- **Proposition 6 (Extension of Semi-Algebraic Property)**
  - **Idea**: Often times, we don't necessarily have an explicit expression for $F(x)$, and thus it is unclear why it is semi-algebraic. We will show that having semi-algebraic property for $h(x, y)$ guarantees the semi-algebraic property for $F(x)$.
  - **Statement**: If $h(x, y)$ is semi-algebraic, then $F(x)$ is also semi-algebraic.
  - **Proof**:
    * We know from definition that $F(x) = \inf_y h(x, y)$, so $\rho > F(x)$ is equivalent to $\exists y$ s.t. $\rho > h(x, y)$.
    * Thus, we can write the graph of F as projection of graph of h onto the y-axis:

$$\{(x, \rho) : \rho > F(x)\} = proj_y(\{(x, y, \rho)|\rho > h(x, y)\})$$

    * Then from Tarski-Seidenberg Theorem, we know that the projection is also semi-algebraic, and this proves the statement.
  - **Note**:
    * The nice property of closure under projection is one of the reasons that we use semi-algebraic sets.
    * The same property does not hold for algebraic sets, for instance consider $proj_{\{y=0\}}(\{x^2+y^2 = 1\}) = [-1, 1]$, which is not an algebraic set (since all algebraic sets on $\mathbb{R}$ are finite.)

5.3 Examples of Algorithms under Majorization-Minimization Framework

This section provides several examples of optimization algorithms that can be interpreted to fit into the majorization-minimization framework.

5.3.1 Proximal Point Method

- **Problem**: We would like to minimize a closed convex function $f : \mathbb{R}^n \to \mathbb{R}$.
- **Majorizing Model**: $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $h(x, y) = f(y) + \frac{\mu}{2}\|x - y\|^2$.
- **Proximal Point Iteration Map**: $x_{k+1} = p(x_k) = \arg\min_y h_{x_k} = \arg\min_y\{f(y) + \frac{\mu}{2}\|x_k - y\|^2\}$
- **Model Qualification Assumption**: We have

$$h^y(x) - h^y(x') = \frac{\mu}{2}(\|x - y\|^2 - \|x' - y\|^2) \leq \frac{\mu}{2}(\|x - y\| + \|x' - y\|)\|x - x'\|$$

Thus, for x' sufficiently close to x, we have lip $h^y(x) = (\mu + \epsilon)\|x - y\|$ for arbitrary small $\epsilon > 0$, so model qualification assumption holds.

5.3.2 Gradient Descent

- **Problem**: We would like to minimize a C^2 function $f : \mathbb{R}^n \to \mathbb{R}$, with ∇f being L-Lipschitz.
- **Majorizing Model**: $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $h(x, y) = f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2$.
- **Gradient Descent Iteration Map**:

$$x_{k+1} = p(x_k) = \arg\min_y h_{x_k} = \arg\min_y \left\{ \nabla f(x)^\top (y - x_k) + \frac{\mu}{2}\|y - x_k\|^2 \right\} = x_k - \frac{1}{\mu}\nabla f(x_k)$$

- **Model Qualification Assumption**: We know that for C^2-functions, local Lipschitz constant is just the norm of the derivative, so

$$(\text{lip } h^y)(x) = \|\nabla h^y(x)\| = \|(\nabla^2 f(x) - \mu I)(y - x)\| \le \|\nabla^2 f(x) - \mu I\| \cdot \|y - x\|$$

  where $\|\nabla^2 f(x) - \mu I\| \le K$ on bounded set around x.

5.3.3 Composite Optimization

- **Problem**: We would like to minimize $f(x) = g(H(x))$ for $g : \mathbb{R}^m \to \mathbb{R}$ finite convex and L-Lipschitz, $H : \mathbb{R}^n \to \mathbb{R}^m$ being C^2.
  - Here $\nabla H(x) = \begin{bmatrix} \nabla H_1(x) & \cdots & \nabla H_m(x) \end{bmatrix}^\top \in \mathbb{R}^{m \times n}$. Assume that $\nabla H(x)$ has Lipschitz constant $\lambda > 0$.
- **Majorizing Model**: $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $h(x, y) = g(H(x) + \nabla H(x)(y - x)) + \frac{\mu}{2}\|y - x\|^2$
  - **Continuity**: This is straightforward as all components here are continuous.
  - **Strong Convexity**: h is μ-strongly convex, as we have

$$h_x(y) - \frac{\mu}{2}\|y\|^2 = g(H(x) + \nabla H(x)(y - x)) - \mu x^\top y + \frac{\mu}{2} x^\top x$$

    where the first half is convex, and second half is linear in y, so the entire function is convex.
  - **Lower Bound**: We now show the lower bound $f(x) \le h(x, y)$.
    - We know that H is C^2, so for any $x, y \in \mathbb{R}^n$ and $i = 1, \cdots, m$, we have from Taylor expansion

$$\exists z \in [x, y] \text{ with } H_i(y) = H_i(x) + \nabla H_i(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 H_i(z)(y - x)$$

    - This means

$$\left| H_i(y) - \left(H_i(x) + \nabla H_i(x)^\top (y - x) \right) \right| \le \frac{1}{2}\|\nabla^2 H_i(z)\|^2 \|y - x\|^2 \le K_i \|y - x\|^2$$

      for some $K_i > 0$, given that x, y are bounded.
    - Thus, taking $K = \sqrt{K_1^2 + \cdots + K_m^2} > 0$ would give us

$$\begin{aligned} h(x, y) - f(y) &= \frac{\mu}{2}\|y - x\|^2 - (f(y) - g(H(x) + \nabla H(x)(y - x))) \\ &\ge \frac{\mu}{2}\|y - x\|^2 - L\|H(y) - (H(x) + \nabla H(x)(y - x))\| \ge \left(\frac{\mu}{2} - KL \right) \|y - x\|^2 \end{aligned}$$

    - So as long as we pick $\mu \ge 2KL$, then the lower bound holds.
  - **Upper Bound**: We finally show the upper bound $h(x, y) = f(y) + o(\|y - x\|^2)$.
    - We can compute

$$h(x, y) - f(y) \le \frac{\mu}{2}\|y - x\|^2 - (g(H(y)) - g(H(x) + \nabla H(x)(y - x)))$$

and thus we have

$$\begin{aligned}|h(x,y)-f(y)| &\le \frac{\mu}{2}\|y-x\|^2 + \big|g(H(y)) - g(H(x)+\nabla H(x)(y-x))\big| \\ &\le \frac{\mu}{2}\|y-x\|^2 + L\|H(y)-H(x)-\nabla H(x)(y-x)\| \\ &\le (\frac{\mu}{2}+\frac{\lambda L}{2})\|y-x\|^2\end{aligned}$$

where second inequality comes from Lipschitz continuity of g, and third inequality comes from Lipschitz continuity of ∇H. We see that RHS is indeed of the form $o(\|y-x\|^2)$, so it satisfies the majorizing property.

- **Model Qualification Assumption**
  - Let $Q(x) = H(x) + \nabla H(x)(y-x)$ with components $Q_i(x) = H_i(x) + \nabla H_i(x)^\top (y-x)$.
  - Then the Lipschitz constant is

$$\text{lip } Q_i(x) = \|\nabla Q_i(x)\| \le K_i\|y-x\|$$

for some constant $K_i > 0$, and thus picking $K = \sqrt{K_1^2 + \cdots + K_m^2}$ gives lip $Q(x) \le K\|y-x\|$, and this implies

$$\text{lip } Q(x) \le K\|y-x\| \Rightarrow \text{lip } (g\circ Q)(x) \le KL\|y-x\| \Rightarrow \text{lip } h_y(x) \le (KL+\mu)\|y-x\|$$

and thus model qualification is satisfied.

5.4 Extended Majorization-Minimization: Sequential Quadratic Programming

This section discusses how we can modify the majorization-minimization framework so that it can be applied to prove the convergence of sequential quadratic programming.

5.4.1 Motivation: Moving Balls Method

We first state a specific instance of sequential quadratic programming that we will consider, the moving balls method. This gives the motivation for how we may want to modify the framework.

- **Problem**: Consider again the classical nonlinear programming problem:

$$\begin{aligned}\min_{x\in\mathbb{R}^n} \quad & p(x) \\ \text{s.t.} \quad & q_i(x) \le 0 \text{ for } i = 1,\dots,k\end{aligned}$$

where p, q_i are C^2 functions, with ∇p and ∇q_i having Lipschitz constants L and L_i respectively.

- **Idea**:
  - Recall that in previous sections, we have formulated this problem as a composite optimization problem by removing the constraints using the penalty function.
  - Here we would like to consider a different approach, where instead of removing the constraints, we use a quadratic approximation for both the objective and the constraint, and solve the corresponding quadratic program recursively. (Thus the name "sequential quadratic programming".)
- **Moving Balls Method (Sequential Quadratic Programming on Nonlinear Programming)**:
  - Following the ideas above, we recursively solve the quadratic programs of the following form:

$$\begin{aligned}\min_{x\in\mathbb{R}^n} \quad & p(x_k) + \nabla p(x_k)^\top (x-x_k) + \frac{L}{2}\|x-x_k\|^2 \\ \text{s.t.} \quad & q_i(x) + \nabla q_i(x_k)^\top (x-x_k) + \frac{L_i}{2}\|x-x_k\|^2 \le 0 \text{ for } i = 1,\dots,k\end{aligned}$$

and in each iteration we set x_{k+1} to be one solution of the above quadratic program.

- **Remark**
  - Here, we are still constructing a majorizing model for the objective function, as what we did for the usual majorization-minimization framework.
  - However, the original framework deals with unconstrained optimization problems, while here we have constraints on our problem.
  - Thus, we would like to extend majorization-minimization framework so that it can deal with constraints as well.

5.4.2 Extended Majorization-Minimization Framework

Motivated by the above section, here we extend the majorization-minimization framework to work with constrained optimization problems.

- **Problem**: Consider the constrained optimization problem $\inf_X f(x)$ with $f : \mathbb{R}^n \to \mathbb{R}$ being closed, proper and continuous, and $X \subseteq \mathbb{R}^n$ being bounded.
- **Majorizing Model**: Similar to the original framework, $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is assumed to be continuous and satisfies
  - **Majorizing property**: $\forall x, f(y) \leq h(x, y) = f(y) + o(\|x - y\|^2)$ as $y \to x$.
  - **Strong convexity**: For all x, $h_x(\cdot) = h(x, \cdot)$ is μ-strongly convex, i.e. $h_x - \frac{\mu}{2}\|\cdot\|^2$ is convex.
- **Constraint Set**: The set-valued function $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is assumed to have the following properties:
  - **Convex and semi-algebraic**: $graph(D) = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : y \in D(x)\}$ is convex and semi-algebraic.
  - **Continuous**: D is continuous, i.e. $graph(D)$ is closed, and D satisfies the inner semi-continuity condition: $x_k \to x$ and $y \in D(x)$, there exists $y_r \in D(x_r)$ with $y_r \to y$.
  - **First-order approximation**: $N_{D(x)}(x) = \hat{N}_X(x)$ has the same normal cone, where $\hat{N}_X(x) = \hat{\partial}\delta_X(x)$ denotes the regular normal cone.
- **Iteration Map**: We define the iteration map $p : \mathbb{R}^n \to \mathbb{R}^n$ to be $p(x) = \arg\min_{D(x)} h_x$.
- **Surrogate Function**: Instead of considering the original function $f(x)$, we will consider the surrogate function $F(x) = \min h_x = h(x, p(x))$.
- **Further Assumptions**
  - **Model Qualification**: Let us denote $\hat{h}(x, y) = \begin{cases} h(x, y) \text{ , if } y \in D(x) \\ +\infty \text{ , otherwise} \end{cases}$. Then we assume that for any bounded X and $x \in X$ and any y, there exists some $K > 0$ such that whenever $(v, 0) \in \hat{\partial}\hat{h}(x, y)$, we have $\|v\| \leq K\|x - y\|$.

Overview of Convergence Proof: What Changed?

- Let us copy the proof steps from our previous discussion of original majorization-minimization framework.
- **Step 1 (Variant of Slope Descent Conditions)**: We first show the analogous slope descent properties that hold on the surrogate function F. This guarantees our application of KL property.
  - **Proposition 1 (Sufficient Descent Condition Variant)**: $F(x_{k-1}) - F(x_k) \geq \lambda\|x_k - x_{k+1}\|^2$
    - This follows from the same argument as before.
  - **Proposition 2 (Slope Property Variant)**: $dist_{\partial F(x_{k-1})}(0) \leq \mu\|x_{k-1} - x_k\|$
    - This is originally proven from the model qualification assumption, and here we use the updated version of this assumption stated above.
    - Here our assumption would guarantee that for bounded X and $x \in X$, there exists $K > 0$ with $\overline{|\nabla F|}(x) \leq K\|x - p(x)\|$, and the rest follows from similar proof pattern.

- **Step 2 (Iterations Converge to Critical Points of f)**: We now show that whenever the iteration converges, it will converge to a critical point of f.
  - **Proposition 3 (Continuity of Iteration Map / Convergence to Fixed Point)**: The iteration map p is continuous. Thus, if we have $x_{k+1} = p(x_k)$ converging to x, then x is a fixed point of p.
    - The proof is also similar to what we have before.
  - **Proposition 4 (Fixed Points are Critical Points)**: Fixed points of the iteration map p are critical points of the objective function f.
    - This step incorporates the constraint set $D(x)$, and thus requires a modified proof.
    - Since we have $p(x) = \arg\min_{D(x)} h_x$, we know that $0 \in \partial(h_x + \delta_{D(x)})(p(x))$.
    - Thus, if x is a fixed point, then we have

$$\begin{aligned} 0 \in \partial(h_x + \delta_{D(x)})(x) = \partial h_x(x) + \partial \delta_D(x) = \hat{\partial} f(x) + N_{D(x)}(x) = \hat{\partial} f(x) + \hat{N}_X(x) \\ = \hat{\partial} f(x) + \hat{\partial} \delta_X(x) \subseteq \hat{\partial}(f + \delta_X)(x) \end{aligned}$$

      where first equality comes from sum rule, latter ones come from evaluating the definitions of normal cone and assumptions on D. This implies $0 \in \hat{\partial}(f + \delta_X)(x)$, and thus x is a critical point for $\min_X f$.

- **Step 3 (Applying KL Property to Prove Convergence)**
  - **Proposition 5 (KL Property and Existence of Desingularizer)**: Assuming KL Property holds on all critical points, we can then find a desingularizer ϕ satisfying the standard properties and the KL inequality.
    - The proof is also similar to what we have before.
  - **Theorem (Convergence of Sequence to Critical Point)**: Given the desingularizer ϕ and (x_k) satisfying the variant of slope descent conditions above, the sequence (x_k) converges. Thus, from Step 2 we know that (x_k) converges to a critical point of f.
    - The proof is also similar to what we have before.

5.4.3 Application to Moving Balls Method

In this section, we check that the assumptions for the extended majorization-minimization framework holds for the moving balls method, and thus we can apply it to show convergence of the method to a critical point.

- **Recall: Nonlinear Programming Problem**:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & p(x) \\ \text{s.t.} \quad & q_i(x) \leq 0 \text{ for } i = 1, \ldots, k \end{aligned}$$

- **Recall: Moving Balls Method**

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & p(x_k) + \nabla p(x_k)^\top (x - x_k) + \frac{L}{2} \|x - x_k\|^2 \\ \text{s.t.} \quad & q_i(x) + \nabla q_i(x_k)^\top (x - x_k) + \frac{L_i}{2} \|x - x_k\|^2 \leq 0 \text{ for } i = 1, \ldots, k \end{aligned}$$

- **Formulation as Extended Majorization-Minimization**
  - **Original Problem**
    - Let $f(x) = p(x)$, and $X = \{x : q_i(x) \leq 0\}$, then our problem is formulated as $\min_X f(x)$.
  - **Moving Balls Method**
    - Let $h(x, y) = p(x) + \nabla p(x)^\top (y - x) + \frac{L}{2} \|y - x\|^2$
    - Let $g_i(x, y) = q_i(x) + \nabla q_i(x)^\top (y - x) + \frac{L_i}{2} \|y - x\|^2$
    - Let $D(x) = \{y : g_i(x, y) \leq 0\}$.

* Then the moving balls method is equivalent to performing $\min_{y \in D(x)} h_x(y)$ in each iteration.

- **Assumption: Mangasarian-Fromowitz Constraint Qualification (MFCQ)**
  - **Idea**: We will assume MFCQ condition on X as the regularity condition in our problem.
  - **Statement**: $\forall x \in X$, $\exists d$ s.t. $\nabla q_i(x)^\top d < 0$ for all active constraints $i \in I(x) = \{i : q_i(x) = 0\}$.
  - **Corollary (Slater Condition)**: There exists some strictly feasible $\hat{y} \in X = \{x : q_i(x) < 0\}$.
  - **Corollary (First-Order Approximation)**: We have

$$N_X(x) = \hat{N}_X(x) = N_{D(x)}(x) = \left\{ \sum_{i \in I(x)} \lambda_i \nabla q_i(x) : \lambda \geq 0 \right\}$$

- **Checking the Assumptions for Convergence**
  - **Basic Assumptions**
    * **Majorizing Model**: The construction $h(x, y) = p(x) + \nabla p(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2$ ensures that it is a majorizing model of $p(x)$.
    * **Convexity of D**: Each $g_i(x, \cdot)$ is convex, and thus the graph of D is the intersection of several convex sets, thus convex as well.
    * **Semi-algebraic property of D**: Assuming that each $q_i(x)$ is a polynomial of x, we know that D is generated by polynomial constraints as well, thus semi-algebraic.
    * **First-order approximation of normal cone**: This directly follows from properties of MFCQ.
    * Now it remains to show the continuity of D and the model qualification condition, which we do in the following propositions.
  - **Proposition 1 (Continuity of $D(x)$)**
    * **Statement**: Given the MFCQ condition on set X, the set-valued map $x \to D(x)$ is continuous.
    * **Proof**:
      · **Closure**: We know that the graph of the map $\{(x, y) \in X^2 : \forall i, g_i(x, y) \leq 0\}$ is closed, since X is closed and g_i is continuous.
      · **Inner-semicontinuity**: It remains to show that if we have $\{x_k\} \subseteq X$ with $x_k \to x$ and $y \in D(x)$, then there exists $\{y_k\}$ with $y_k \in D(x_k)$ and $y_k \to y$.
      · From Slater's condition, we know that $\exists \hat{y}$ with $g_i(x, \hat{y}) < 0$ for all i.
      · Since $g_i(x, \cdot)$ is convex and we have $g_i(x, y) < 0$, we know that for any $r = 1, 2, \ldots$, we have

$$\forall i, g_i\left(x, \left(1 - \frac{1}{r}\right) y + \frac{1}{r}\hat{y}\right) < 0$$

      · From continuity of g_i and $x_k \to x$, we know that given $k \geq k_r$ for some large k_r, we have

$$\forall i, g_i\left(x_k, \left(1 - \frac{1}{r}\right) y + \frac{1}{r}\hat{y}\right) < 0$$

      · Now for the sequence $k_1 \leq k_2 \leq \ldots$, we can define $y_k = \left(1 - \frac{1}{r}\right) y + \frac{1}{r}\hat{y}$, and this would imply $y_k \in D(x_k)$ and $y_k \to y$ as required.
  - **Proposition 2 (Uniformly Bounded Lagrange Multiplier Set)**
    * **Statement**: The set of Lagrange multipliers

$$\Lambda(x) = \left\{ \lambda \geq 0 : \begin{cases} \lambda_i g_i(x, p(x)) = 0 \text{ for all } i \\ \nabla_y h(x, p(x)) + \sum_i \lambda_i \nabla_y g_i(x, p(x)) = 0 \end{cases} \right\}$$

    is nonempty and uniformly bounded as $x \in X$ varies over bounded sets.

* **Proof**:
  - **Nonemptiness** is guaranteed by Slater's condition.
  - **Uniform boundedness**: Now suppose $\Lambda(x)$ is unbounded, so there exists $\{x_k\} \subseteq X$ and $\lambda^k \in \Lambda(x_k)$ with $\|\lambda^k\| \to \infty$.
  - We normalize the vector $\mu^k = \frac{\lambda^k}{\sum_i \lambda_i^k} \geq 0$, so that $\mu^k \in \Delta$ lies in the unit simplex.
  - Since the unit simplex is compact, we know that $\{\mu^k\}$ has a convergent subsequence. WLOG let $\mu^k \to \mu \geq 0$ with $\mu \neq 0$.
  - We know that $p(x) = \min_{y \in D(x)} h(x,y)$, so $\nabla_y h(x, p(x)) = 0$. Note that the Lagrange conditions still hold for each μ^k, since scaling the vector λ^k by a constant does not affect Lagrange's condition. Thus, we have

$$\begin{cases} \mu_i^k g_i(x,p(x)) = 0 \text{ for all } i \\ \sum_i \mu_i^k \nabla_y g_i(x,p(x)) = 0 \end{cases} \Rightarrow \begin{cases} \mu_i g_i(x,p(x)) = 0 \text{ for all } i \\ \sum_i \mu_i \nabla_y g_i(x,p(x)) = 0 \end{cases}$$

    where take the limit as $\mu^k \to \mu$.
  - This implies for $\hat{y}$ generated from Slater's condition, we have

$$0 = \left\langle \hat{y} - p(x), \sum_i \mu_i \nabla_y g_i(x,p(x)) \right\rangle \leq \sum_i \mu_i g_i(x,\hat{y}) - g_i(x,p(x)) = \sum_i \mu_i g_i(x,\hat{y}) < 0$$

    where the first inequality comes from subgradient property, and second inequality comes from Slater's condition. This gives a contradiction, and thus we must have $\Lambda(x)$ being uniformly bounded.

– **Proposition 3 (Extending MFCQ to Approximation Set $D(x)$)**
* **Statement**: Given that $X = \{x : q_i(x) \leq 0\}$ satisfies the MFCQ condition, then $\forall x$, $D(x) = \{y : g_i(x,y) \leq 0\}$ also satisfies the MFCQ condition, i.e. $\forall y \in D(x)$, $\exists d'$ s.t. $\nabla_y g_i(x,y)^\top d' < 0$ for all active constraints $i \in I(x) = \{i : g_i(x,y) = 0\}$.
* **Proof**:
  - **Computing the gradient**: From our definition of $g_i(x,y)$, we have

$$g_i(x,y) = q_i(x) + \nabla q_i(x)^\top (y-x) + \frac{L_i}{2}\|y-x\|^2$$

    and thus its gradient have the form

$$\nabla g_i(x,y) = \begin{pmatrix} \left(L_i I - \nabla^2 q_i(x)\right)(x-y) \\ \nabla q_i(x) + L_i(y-x) \end{pmatrix}$$

  - **Case 1**: If $x = y$, then we have $\nabla g_i(x,y) = \begin{pmatrix} 0 \\ \nabla q_i(x) \end{pmatrix}$, and thus we can pick $d' = \begin{pmatrix} 0 \\ d \end{pmatrix}$ where d is the vector satisfying MFCQ for set X, and this ensures $\nabla g_i(x,y)^\top d' = \nabla q_i(x)^\top d < 0$.
  - **Case 2**: If $x \neq y$, then we can pick L_i such that $L_i \geq \|\nabla^2 q_i(x)\|$, and then we can pick $d' = \begin{pmatrix} y-x \\ 0 \end{pmatrix}$ as the vector satisfying MFCQ for set X, and this ensures $\nabla g_i(x,y)^\top d' = -(x-y)^\top (L_i I - \nabla^2 q_i(x))(x-y) < 0$.

– **Proposition 4 (Model Qualification Condition)**
* **Statement**: For any $x, y \in X$ varying in a bounded set, there exists some $K > 0$ such that whenever $\begin{pmatrix} v \\ 0 \end{pmatrix} \in \hat{\partial}\hat{h}(x,y)$, we have $\|v\| \leq K\|x-y\|$.
* **Proof**:
  - **Optimality condition**: Since $\begin{pmatrix} v \\ 0 \end{pmatrix} \in \hat{\partial}\hat{h}(x,y)$ and $\hat{h}(x,\cdot)$ is strongly convex, we know that if we fix x and only look at the y-component, 0 being inside the subgradient implies y is the minimizer for $\hat{h}(x,y)$, so $y = p(x)$.

- **Construction of Lagrange multiplier**: Thus, if we apply subdifferential calculus and the MFCQ condition on g_i proven in Proposition 3, we know that for some $\lambda \geq 0$, we have

$$\begin{aligned}\begin{pmatrix} v \\ 0 \end{pmatrix} \in \hat{\partial}\hat{h}(x, p(x)) &= \nabla h(x, p(x)) + \sum_{i \in \hat{I}(x)} \lambda_i \nabla g_i(x, p(x)) \\ &= \begin{pmatrix} \left(LI - \nabla^2 f(x)\right)(x - p(x)) \\ L(p(x) - x) + \nabla f(x) \end{pmatrix} + \sum_{i \in \hat{I}(x)} \lambda_i \begin{pmatrix} (L_i I - \nabla^2 q_i(x))(x - p(x)) \\ L_i(p(x) - x) + \nabla q_i(x) \end{pmatrix}\end{aligned}$$

  where $\hat{I}(x)$ denote the active constraints.
- Now the second component of the above equation implies

$$\begin{aligned}0 &= L(p(x) - x) + \nabla f(x) + \sum_{i \in \hat{I}(x)} \lambda_i (L_i(p(x) - x) + \nabla q_i(x)) \\ &= \nabla_y h(x, p(x)) + \sum_{i \in \hat{I}(x)} \lambda_i \nabla_y g_i(x, p(x))\end{aligned}$$

  and thus $\lambda \in \Lambda(x)$ is a Lagrange multiplier,
- **Applying boundedness of Lagrange multiplier set**: From Proposition 2, we know that $\|\lambda\| \leq K'$ is bounded. Now the equation in the first component implies

$$\begin{aligned}\|v\| &\leq \left\| (LI - \nabla^2 f(x)) + \sum_{i \in \hat{I}(x)} \lambda_i (L_i I - \nabla^2 q_i(x)) \right\| \cdot \|x - p(x)\| \\ &\leq \left\| (LI - \nabla^2 f(x)) + K' \sum_{i \in \hat{I}(x)} (L_i I - \nabla^2 q_i(x)) \right\| \cdot \|x - p(x)\| \leq K\|x - y\|\end{aligned}$$

  since all terms in the first norm can be uniformly bounded. This proves the proposition.

6 References

[1] H. Attouch, G. Buttazzo, and G. Michaille. *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*. SIAM, 2014.

[2] J. Baillon. Un exemple concernant le comportement asymptotique de la solution du problème $dudt + \partial\theta(\mu) \ni 0$. *Journal of Functional Analysis*, 28(3):369–376, 1978.

[3] J. Bolte, A. Daniilidis, and A. Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[4] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[5] H. Brezis. *Operateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. Elsevier, 1973.

[6] R. E. Bruck. Asymptotic convergence of nonlinear contraction semigroups in hilbert space. *Journal of Functional Analysis*, 18(1):15–26, 1975.

[7] D. Drusvyatskiy. Slope and geometry in variational mathematics. 2013.