Computational Methods of Optimization

Author: Wenxiao Yang

Institute: Department of Mathematics, University of Illinois at Urbana-Champaign

Date: 2022

All models are wrong, but some are useful.

Contents

Chapter 1 Math Foundations

1.1 Strongly Convexity

1.1.1 μ-Strongly Convex: $\langle \nabla f(w) - \nabla f(v), w - v \rangle \geq \mu \|w - v\|^2$

Definition: We say $f : C \to \mathbb{R}$ is a μ-strongly convex function in a convex set C if f is differentiable and

$$\langle \nabla f(w) - \nabla f(v), w - v \rangle \geq \mu \|w - v\|^2, \quad \forall w, v \in C.$$

1.1.2 μ-strongly convex $\Leftrightarrow \nabla^2 f(x) \succeq \mu I \Leftrightarrow$"$f(x) - \frac{m}{2}\|x\|^2$ is convex"

If f is twice differentiable, then f is μ-strongly convex iff

$$\nabla^2 f(x) \succeq \mu I, \quad \forall x \in C.$$

Definition 1.1

A twice continuously differentiable function is strongly convex if

$$\exists m > 0 \; s.t. \; \nabla^2 f(x) \succeq mI \quad \forall x$$

which is also called $m-$strongly convex.

(alternative): *"$f(x) - \frac{m}{2}\|x\|^2$ is convex" is also an equivalent definition for $f(x)$ is $m-$strongly convex.*

Namely, all eigenvalues of the Hessian at any point is at least μ.

if $f(w)$ is convex, then $f(w) + \frac{\mu}{2}\|w\|^2$ is μ-strongly convex.

- In machine learning, easy to change a convex function to a strongly convex function: just add a regularizer

1.1.3 Lemma: Strongly convexity $\Rightarrow$ Strictly convexity

Lemma 1.1

Strongly convexity $\Rightarrow$ Strictly convexity.

Proof 1.1

$$\begin{aligned}\nabla^2 f(x) \succeq mI &\Rightarrow \nabla^2 f(x) - mI \succeq 0 \\ &\Rightarrow \forall z \neq 0 \quad z^T(\nabla^2 f(x) - mI)z \geq 0 \\ &\Rightarrow z^T \nabla^2 f(x) z \geq m z^T z > 0\end{aligned}$$

Note: converse is not true: e.g. $f(x) = x^4$ is strictly convex but $\nabla^2 f(0) = 0$

1.1.4 Lemma: $\nabla^2 f(x) \succeq mI \Rightarrow f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2$

Lemma 1.2

$\nabla^2 f(x) \succeq mI \quad \forall x$

$$\Rightarrow f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2$$

Proof 1.2

By Taylor's Theorem,

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f((1-\beta)x+\beta y)(y-x), \quad \textit{for some } \beta \in [0,1] \\ &\geq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T m(y-x) \\ &\geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2 \end{aligned}$$

1.2 Lipschitz Gradient (L-Smooth)

Definition 1.2 (Lipschitz Continunity)

A function $g : \mathbb{R}^n \to \mathbb{R}^m$ is called Lipschitz (continuous) if $\exists L > 0$ s.t.

$$\|g(y) - g(x)\| \leq L\|y-x\|, \forall x, y \in \mathbb{R}^n$$

L is Lipschitz constant.

Definition 1.3 (Lipschitz Gradient)

$\nabla f(x)$ is Lipschitz if $\exists L > 0$ s.t.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|, \forall x, y \in \mathbb{R}^n$$

We can say f is L-Smooth.

Example 1.1

1. $f(x) = \|x\|^4, \nabla f(x) = 4\|x\|^2 x$

   Test $\|\nabla f(x) - \nabla f(-x)\| \leq L\|2x\|, 8\|x\|^2\|x\| \leq 2L\|x\|$ which doesn't hold when $\|x\|^2 > \frac{L}{4}$.
2. If f is twice continuously differentiable with $\nabla^2 f(x) \succeq -MI$ and $\nabla^2 f(x) \preceq MI$ then $\|\nabla f(x) - \nabla f(y)\| \leq M\|x-y\|, \forall x, y \in \mathbb{R}^n$. ($A \succeq B$ means $A - B \succeq 0$, $A \preceq B$ means $A - B \preceq 0$)

1.2.1 Theorem: $-MI \preceq \nabla^2 f(x) \preceq MI \Rightarrow f$ is M-smooth

Theorem 1.1

$-MI \preceq \nabla^2 f(x) \preceq MI, \forall x \Rightarrow \|\nabla f(x) - \nabla f(y)\| \leq M\|x-y\|, \forall x, y$ ♡

Proof 1.3

For symmetric A,

1. $x^TAx \leq \lambda_{\max}(A)\|x\|^2$
2. $\lambda_i(A^2) = \lambda_i^2(A)$
3. $-MI \preceq A \preceq MI \Rightarrow \lambda_{\min}(A) \geq -M, \lambda_{\max}(A) \leq M$

Define $g(t) = \frac{\partial f}{\partial x_i}(x + t(y-x))$. *Then*

$$g(1) = g(0) + \int_0^1 g'(s)ds$$

$$\Rightarrow \frac{\partial f(y)}{\partial x_i} = \frac{\partial f(x)}{\partial x_i} + \int_0^1 \sum_{j=1}^n \frac{\partial^2 f(x+s(y-x))}{\partial x_i \partial x_j}(y_j - x_j)ds$$

$$\nabla f(y) = \nabla f(x) + \int_0^1 \nabla^2 f(x+s(y-x))(y-x)ds$$

$$\begin{aligned}
\|\nabla f(y) - \nabla f(x)\| &= \|\int_0^1 \nabla^2 f(x+s(y-x))(y-x)ds\| \\
&\leq \int_0^1 \|\nabla^2 f(x+s(y-x))(y-x)\|ds \\
&= \int_0^1 \sqrt{(y-x)^T[\nabla^2 f(x+s(y-x))]^2(y-x)}ds \\
&(\text{Set } H = \nabla^2 f(x+s(y-x))) \\
&\leq \int_0^1 \sqrt{\lambda_{\max}(H^2)\|y-x\|^2}ds \\
&\leq M\|y-x\|
\end{aligned}$$

1.2.2 Descent Lemma: f is L-smooth $\Rightarrow f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2$

Lemma 1.3 (Descent Lemma)

Let $f: \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable with a Lipschitz gradient with Lipschitz constant L. Then

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}L\|y-x\|^2$$ ♡

Proof 1.4

Let $g(t) = f(x+t(y-x))$. *Then* $g(0) = f(x)$ *and* $g(1) = f(y)$, $g(1) = g(0) + \int_0^1 g'(t)dt$.

Where $g'(t) = \nabla f(x+t(y-x))^T(y-x)$

$$\begin{aligned}
\Rightarrow f(y) &= f(x) + \int_0^1 \nabla f(x+t(y-x))^T(y-x)dt \\
&= f(x) + \int_0^1 (\nabla f(x+t(y-x)) - \nabla f(x))^T(y-x)dt + \nabla f(x)^T(y-x) \\
&\le f(x) + \int_0^1 \|\nabla f(x+t(y-x)) - \nabla f(x)\|\|y-x\|dt + \nabla f(x)^T(y-x) \\
&\le f(x) + L\int_0^1 \|t(y-x)\|\|y-x\|dt + \nabla f(x)^T(y-x) \\
&= f(x) + \frac{1}{2}L\|y-x\|^2 + \nabla f(x)^T(y-x)
\end{aligned}$$

1.2.3 Co-coercivity Condition: $(\nabla f(x) - \nabla f(y))^T(x-y) \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$

Theorem 1.2 (Co-coercivity Condition)

Let f be convex and continuously differentiable. Let f be L-smooth. Then

$$(\nabla f(x) - \nabla f(y))^T(x-y) \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$$

♡

Proof 1.5

Let $y \in \mathbb{R}^n$, and define $g(x) = f(x) - \nabla f(y)^T x$. Then $\nabla g(y) = \nabla f(y) - \nabla f(y) = 0$ and $\nabla^2 g(y) = \nabla^2 f(y) \succeq 0$, i.e. y minimize g. Because $g(y) \le g(\cdot)$, $g(y) \le g(x - \frac{1}{L}\nabla g(x))$ According to the descent lemma,

$$\begin{aligned}
g(x - \frac{1}{L}\nabla g(x)) &= f(x - \frac{1}{L}\nabla g(x)) - \nabla f(y)^T(x - \frac{1}{L}\nabla g(x)) \\
&\le f(x) + \frac{L}{2}\| - \frac{1}{L}\nabla g(x)\|^2 + \nabla f(x)^T(-\frac{1}{L}\nabla g(x)) - \nabla f(y)^T(x - \frac{1}{L}\nabla g(x)) \\
&\le f(x) + \frac{1}{2L}\|\nabla g(x)\|^2 - (\nabla f(x) - \nabla f(y))^T\frac{1}{L}\nabla g(x) - \nabla f(y)^T x \\
&= f(x) - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 - \nabla f(y)^T x \\
&= g(x) - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2
\end{aligned}$$

Then,

$$\begin{aligned}
g(y) \le g(x - \frac{1}{L}\nabla g(x)) = g(x) - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \\
\Rightarrow g(y) - g(x) = f(y) - \nabla f(y)^T y - f(x) - \nabla f(y)^T x \le -\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2
\end{aligned}$$

We can interchange x, y,

$$\begin{cases}
f(y) - \nabla f(y)^T y - f(x) - \nabla f(y)^T x \le -\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \\
f(x) - \nabla f(x)^T x - f(y) - \nabla f(x)^T y \le -\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2
\end{cases}$$

Add these two inequalities together,

$$(\nabla f(x) - \nabla f(y))^T(x-y) \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$$

Chapter 2 (Unconstrained Optimization) Gradient Methods

Definition 2.1 (Iterative Descent)

Start at some point x_0, and successively generate $x_1, x_2, ..$ s.t.

$$f(x_{k+1}) < f(x_k) \quad k = 0, 1, \ldots$$

♣

Definition 2.2 (General Gradient Descent Algorithm)

Assume that $\nabla f(x_k) \neq 0$. Then

$$x_{k+1} = x_k + \alpha_k d_k$$

where d_k is s.t. d_k has a positive projection along $-\nabla f(x_k)$,

$$\nabla f(x_k)^T d_k < 0 \equiv -\nabla f(x_k)^T d_k > 0$$

♣

- If $d_k = -\nabla f(x_k)$ we get **steepest descent**.
- Often d_k is constructed using matrix $D_k \succ 0$

$$d_k = -D_k \nabla f(x_k)$$

2.1 Steepest Descent

We want the x_k that decreases the function most.

Proposition 2.1

$-\nabla f(x_k)$ is the direction deceases the function most.

♠

Proof 2.1

Suppose the direction is $v \in \mathbb{R}^n, v \neq 0$.

$$f(x + \alpha v) = f(x) + \alpha v^T \nabla f(x) + O(\alpha)$$

The rate of change of f along direction v:

$$\lim_{\alpha \to 0} \frac{f(x + \alpha v) - f(x)}{\alpha} = v^T \nabla f(x)$$

By Cauchy-schwarz inequality,

$$|v^T \nabla f(x)| \leq \|v\| \|\nabla f(x)\|$$

Equation holds when $v = \beta \nabla f(x)$. Hence, $-\nabla f(x)$ is the direction decreases the function most.

Definition 2.3 (Steepest Descent Algorithm)

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

α_k is the step size, which need to choose carefully. ♣

2.2 Methods for Choosing Step Size α_k

(1): Fixed step size: $\alpha_k = \alpha$ (can have issue with *convergence*)

(2): **Optimal Line Search**: choose α_k to optimize the value of next iteration, i.e. solve

$$\min_{\alpha \geq 0} f(x_k + \alpha d_k)$$

(may be *difficult in practice*)

(3): **Armijo's Rule** (successive step size reduction):

$$f(x_k + \alpha_k d_k) = f(x_k) + \alpha_k \nabla f(x_k)^T d_k + O(\alpha_k)$$

Since $\nabla f(x_k)^T d_k < 0$, f decreases when α_k is sufficiently small. But we also don't want α_k to be too small (slow).

2.2.1 Optimal (Exact) Line Search

Example 2.1 (False $\times$) The gradient descent algorithm with an exact line search always finds the minimum of a strictly convex quadratic function in exactly one iteration.

Note: the moving direction is restricted to the gradient.

Counterexample: False. It is not necessary that the gradient at x_0 towards the exact solution. For example, let $f(x) = \frac{1}{2}x^\top Qx + x^\top b$ where $Q = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ and $b = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Clearly we have $x^* = \begin{pmatrix} -1/2 \\ 1 \end{pmatrix}$. If we start with $x_0 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, by using exact line search, the step size $\alpha = \arg\min f\left(x_0 - \alpha \nabla f\left(x_0\right)\right) = 10/19$. Hence $x_1 = x_0 - \alpha \nabla f\left(x_0\right) = \begin{pmatrix} -11/19 \\ 28/19 \end{pmatrix} \neq x^*$.

2.2.2 Armijo's Rule

(i) Initialize $\alpha_k = \tilde{\alpha}$. Let $\sigma, \beta \in (0, 1)$ be prespecified paramenters.

(ii) If $f(x_k) - f(x_k + \alpha_k d_k) \geq -\sigma \alpha_k \nabla f(x_k)^T d_k$, stop.

(Which shows $f(x_k + \alpha_k d_k)$ is at least smaller than $f(x_k)$ in a degree that correlated with $\nabla f(x_k)^T d_k$)

(iii) Else, set $\alpha_k = \beta\alpha_k$ and go back to step 2. (use a smaller α_k)

Termination at smallest integer m s.t.

$$f(x_k) - f(x_k + \beta^m \tilde{\alpha} d_k) \geq -\sigma\beta^m\tilde{\alpha}\nabla f(x)^T d_k$$

In Bersekas's book: $\sigma \in [10^{-5}, 10^{-1}], \beta \in [\frac{1}{10}, \frac{1}{2}]$.

As σ, β are smaller, the algorithm is quicker.

2.2.3 Armijo's Rule for Steepest Descent

$\alpha_k = \tilde{\alpha}\beta^{m_k}$, where m_k is smallest m s.t.

$$f(x_k) - f(x_k - \tilde{\alpha}\beta^m\nabla f(x_k)) \geq \sigma\tilde{\alpha}\beta^m\|\nabla f(x_k)\|^2$$

Proposition 2.2

Assume $\inf_x f(x) > -\infty$. Then every limit point of $\{x_k\}$ for steepest descent with Armijo's rule is a stationary point of f.

Proof 2.2

Assume that $\bar{x}$ is a limit point of $\{x_k\}$ s.t. $\nabla f(\bar{x}) \neq 0$.

- *Since $\{f(x_k)\}$ is monotonically non-increasing and bounded below, $\{f(x_k)\}$ converges.*
- *f is continuous $\Rightarrow$ $f(\bar{x})$ is a limit point of $\{f(x_k)\}$ $\Rightarrow$ $\lim_{k\to\infty} f(x_k) = f(\bar{x}) \Rightarrow f(x_k) - f(x_{k+1}) \to 0$*
- *By definition of Armijo's rule:*

$$f(x_k) - f(x_{k+1}) \geq \sigma\alpha_k\|\nabla f(x_k)\|^2$$

Hence, $\sigma\alpha_k\|\nabla f(x_k)\|^2 \to 0$.

Since $\nabla f(\bar{x}) \neq 0$, $\lim_{k\to\infty}\alpha_k = 0$

$$ln\alpha_k = ln(\tilde{\alpha}\beta^{m_k}) = ln\tilde{\alpha} + m_k ln\beta \Rightarrow m_k = \frac{ln\alpha_k - ln\tilde{\alpha}}{ln\beta} \Rightarrow \lim_{k\to\infty} m_k = \infty$$

Exist $\bar{k}$ s.t. $m_k > 1, \forall k > \bar{k}$

$$f(x_k) - f(x_k - \frac{\alpha_k}{\beta}\nabla f(x_k)) < \sigma\frac{\alpha_k}{\beta}\|\nabla f(x_k)\|^2, \forall k > \bar{k}$$

By Taylor's Theorem,

$$f(x_k - \frac{\alpha_k}{\beta}\nabla f(x_k)) = f(x_k) - \nabla f(x_k - \frac{\bar{\alpha}_k}{\beta}\nabla f(x_k))^T\frac{\alpha_k}{\beta}\nabla f(x_k)$$

for some $\bar{\alpha}_k \in (0, \alpha_k)$

Hence,

$$\nabla f(x_k - \frac{\bar{\alpha}_k}{\beta}\nabla f(x_k))^T \frac{\alpha_k}{\beta}\nabla f(x_k) < \sigma\frac{\alpha_k}{\beta}\|\nabla f(x_k)\|^2$$

$$\nabla f(x_k - \frac{\bar{\alpha}_k}{\beta}\nabla f(x_k))^T \nabla f(x_k) < \sigma\|\nabla f(x_k)\|^2, \forall k > \bar{k}$$

$$\text{As } \alpha_k \to 0 \Rightarrow \bar{\alpha}_k \to 0$$

$$\|\nabla f(x_k)\|^2 < \sigma\|\nabla f(x_k)\|^2$$

Which contradicts to $\sigma < 1$.

2.3 Algorithm Convergence

(1) **Linear convergence:** A minimization algorithm converges linearly if

$$\lim_{n\to\infty} \sup \frac{e_{n+1}}{e_n} = \beta \in (0,1)$$

This is obtained if $e_n \le c\beta^n$.

(2) **Superlinear convergence:** A minimization algorithm converges superlinearly if

$$\lim_{n\to\infty} \sup \frac{e_{n+1}}{e_n} = 0$$

(3) **Quadratic convergence:** A minimization algorithm converges quadratically if

$$\lim_{n\to\infty} \sup \frac{e_{n+1}}{e_n^2} = \beta \in (0,1)$$

2.4 Convergence of The Steepest Descent with Fixed Step Size

2.4.1 Theorem: f is L-smooth $\Rightarrow \{x_k\}$ converges to stationary point

Theorem 2.1

Consider the GD algorithm

$$x_{k+1} = x_k - \alpha\nabla f(x_k), \quad k = 0, 1, ...$$

Assume that f *has Lipschitz gradient with a Lipschitz gradient with Lipschitz constant* L. *Then if* α *is sufficiently small* ($\alpha \in (0, \frac{2}{L})$) *and* $f(x) \ge f_{\min}$ *for all* $x \in \mathbb{R}^n$,

(1). $f(x_{k+1}) \le f(x_k) - \alpha(1 - \frac{L\alpha}{2})\|\nabla f(x_k)\|^2$

(2). $\sum_{k=0}^{N} \|\nabla f(x_k)\|^2 \le \frac{f(x_0) - f_{\min}}{\alpha(1-\frac{L\alpha}{2})}$

(3). every limit point of $\{x_k\}$ *is a stationary point of* f.

♡

Proof 2.3

Applying the descent lemma,

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&= f(x_k) - \alpha \nabla f(x_k)^T \nabla f(x_k) + \frac{L}{2}\alpha^2 \|\nabla f(x_k)\|^2 \\
&= f(x_k) + \alpha(\frac{L\alpha}{2} - 1)\|\nabla f(x_k)\|^2 \\
\Rightarrow \alpha(1 - \frac{L\alpha}{2})\|\nabla f(x_k)\|^2 &\leq f(x_k) - f(x_{k+1}) \\
\alpha \sum_{k=0}^{N}(1 - \frac{L\alpha}{2})\|\nabla f(x_k)\|^2 &\leq f(x_0) - f(x_{N+1}) \\
&\leq f(x_0) - f_{\min}
\end{aligned}$$

If $\alpha \in (0, \frac{2}{L})$, i.e. $\alpha(1 - \frac{L\alpha}{2})$,

$$\begin{aligned}
&\sum_{k=0}^{N} \|\nabla f(x_k)\|^2 \leq \frac{f(x_0) - f_{\min}}{\alpha(1 - \frac{L\alpha}{2})} < \infty, \forall N \\
&\Rightarrow \lim_{k \to \infty} \nabla f(x_k) = 0
\end{aligned}$$

If $\bar{x}$ is a limit point of $\{x_k\}$, $\lim_{k \to \infty} x_k = \bar{x}$.

By continunity of ∇f, $\nabla f(\bar{x}) = 0$

Example 2.2 $f(x) = \frac{1}{2}x^2, x \in \mathbb{R}$, $\nabla f(x) = x$, Lipschitz with $L = 1$.

$$\begin{aligned}
x_{k+1} &= x_k - \alpha \nabla f(x_k) \\
&= x_k(1 - \alpha)
\end{aligned}$$

$0 < \alpha < \frac{2}{L} = 2$ is needed for convergence.

Test (1) $\alpha = 1.5$ Then $x_{k+1} = x_k(-0.5)$,

$$\Rightarrow x_k = x_0(-0.5)^k \to 0 \text{ as } k \to \infty$$

Test (2) $\alpha = 2.5$ Then $x_{k+1} = x_k(-1.5)$.

$$\Rightarrow x_k = x_0(-1.5)^k \Rightarrow |x_k| \to \infty$$

Test (3) $\alpha = 2$ Then $x_{k+1} = -x_k$.

$$\Rightarrow x_k = (-1)^k x_0 \Rightarrow \text{ oscillation between } -x_0, x_0$$

Example 2.3 What if gradient is not Lipschitz? e.g. $f(x) = x^4, x \in \mathbb{R}$, $\nabla f(x) = 4x^3$, $x = 0$ is the only stationary point (global-min)

$$x_{k+1} = x_k - 4\alpha x_k^3 = x_k(1 - 4\alpha x_k^2)$$

- $|x_1| = |x_0|$, then $|x_k| = |x_0|$ for all k, and $\{x_k\}$ stays bounded away from 0, except if $x_0 = 0$

- $$\begin{aligned} |x_1| < |x_0| &\Leftrightarrow |x_0||1-4\alpha x_0^2| < |x_0| \\ &\Leftrightarrow -1 < 1-4\alpha x_0^2 < 1 \\ &\Leftrightarrow 0 < x_0^2 < \frac{1}{2\alpha} \Leftrightarrow 0 < |x_0| < \frac{1}{\sqrt{2\alpha}} \end{aligned}$$
- Therefore, if $|x_1| < |x_0|$, then $|x_1| < |x_0| < \frac{1}{\sqrt{2\alpha}} \Rightarrow |x_2| < |x_1|, ..., |x_{k+1}| < |x_k|, \forall k \Rightarrow \{|x_k|\}$convergences
- And if $|x_1| > |x_0|$, then $|x_{k+1}| > |x_k|$ for all k and $\{x_k\}$ stays bounded away from 0.

Proposition 2.3

$0 < |x_0| < \frac{1}{\sqrt{2\alpha}} \Rightarrow |x_k| \to 0$

Proof 2.4

Suppose $|x_k| \to c > 0$. *Then* $\frac{|x_{k+1}|}{|x_k|} \to 1$

But $\frac{|x_{k+1}|}{|x_k|} = |1-4\alpha x_k^2| \to |1-4\alpha c^2|$. *Thus* $|1-4\alpha c^2| = 1 \Rightarrow c = \frac{1}{\sqrt{2\alpha}}$, *which contradicts to* $c < |x_0| < \frac{1}{\sqrt{2\alpha}}$, *hence* $c = 0$

2.4.2 Theorem: f is convex and L-smooth $\Rightarrow f(x_k)$ converges to global-min value with rate $\frac{1}{k}$

Theorem 2.2

Consider the GD algorithm

$$x_{k+1} = x_k - \alpha\nabla f(x_k), \quad k = 0, 1, ...$$

Assume that f has Lipschitz gradient with Lipschitz constant L. Further assume that

(a) *f is a convex function.*

(b) $\exists x^*$ *s.t.* $f(x^*) = \min f(x)$

Then for sufficiently small α:

(i) $\lim_{k\to\infty} f(x_k) = \min f(x) = f(x^*)$

(ii) *$f(x_k)$ converges to $f(x^*)$ at rate $\frac{1}{k}$.*

Proof 2.5

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - \alpha\nabla f(x_k) - x^*\|^2 \\ &= \|x_k - x^*\|^2 + \alpha^2\|\nabla f(x_k)\|^2 - 2\alpha\nabla f(x)^T(x_k - x^*) \end{aligned}$$

By convexity,

$$f(x^*) \geq f(x_k) + \nabla f(x_k)^T(x^* - x_k)$$

$$\Rightarrow \nabla f(x_k)^T(x^* - x_k) \leq f(x^*) - f(x_k)$$

Thus,

$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2+\alpha^2\|\nabla f(x_k)\|^2+2\alpha(f(x^*)-f(x_k))$$

$$\Rightarrow 2\alpha(f(x_k)-f(x^*)) \leq \|x_k-x^*\|^2-\|x_{k+1}-x^*\|^2+\alpha^2\|\nabla f(x_k)\|^2$$

$$2\alpha\sum_{k=0}^{N}(f(x_k)-f(x^*)) \leq \|x_0-x^*\|^2-\|x_{N+1}-x^*\|^2+\alpha^2\sum_{k=0}^{N}\|\nabla f(x_k)\|^2$$

$$\leq \|x_0-x^*\|^2+\alpha^2\sum_{k=0}^{N}\|\nabla f(x_k)\|^2$$

According to previous theorm, if $\alpha \in (0, \frac{2}{L})$, $\sum_{k=0}^{N}\|\nabla f(x_k)\|^2 \leq \frac{f(x_0)-f(x^*)}{\alpha(1-\frac{L\alpha}{2})}$ *and*

$$f(x_{k+1})-f(x_k) \leq -\alpha(1-\frac{L\alpha}{2})\|\nabla f(x_k)\|^2 \leq 0$$

$$\Rightarrow f(x_N) \leq f(x_k), \quad \forall k=0,1...,N$$

$$\Rightarrow \sum_{k=0}^{N}(f(x_k)-f(x^*)) \geq (N+1)(f(x_N)-f(x^*))$$

$$f(x_N)-f(x^*) \leq \frac{1}{N+1}\sum_{k=0}^{N}(f(x_k)-f(x^*))$$

$$\leq \frac{1}{2\alpha(N+1)}(\|x_0-x^*\|^2+\alpha^2\frac{f(x_0)-f(x^*)}{\alpha(1-\frac{L\alpha}{2})})$$

$$\to 0 \text{ as } N \to \infty$$

The rate of convergence is $\frac{1}{N}$.

To make $f(x_N)-f(x^*)<\varepsilon$, *we need* $N \sim O(\frac{1}{\varepsilon})$.

Note: Armijo's rule also convergencesat rate $\frac{1}{N}$ if ∇f is Lipschitz, without priot knowledge of L. But need $r \in [\frac{1}{2}, 1)$

2.4.3 Theorem: f is strongly convex and $L-$smooth $\Rightarrow \{x_k\}$ converges to global-min geometrically

Strong convexity with parameter m, along with $M-$Lipschitz gradient assumption (with $M \geq m$) According to the lemmas we proved before

$$\frac{m}{2}\|y-x\|^2 \leq f(y)-f(x)-\nabla^T f(x)(y-x) \leq \frac{M}{2}\|y-x\|^2$$

Theorem 2.3

If f has Lipschitz gradient with Lipschitz constant M and strongly convex with parameter m, $\{x_k\}$ converges to x^* ***geometrically***. ♡

$$\begin{aligned}
\|x_{k+1}-x^*\|^2 &= \|x_k-\alpha\nabla f(x_k)-x^*\|^2 \\
(\nabla f(x^*)=0) \qquad &= \|(x_k-x^*)-\alpha(\nabla f(x_k)-\nabla f(x^*))\|^2 \\
&= \|x_k-x^*\|^2+\alpha^2\|\nabla f(x_k)-\nabla f(x^*)\|^2-2\alpha(x_k-x^*)^T(\nabla f(x_k)-0) \\
(\nabla f \text{ is M-Lipschitz}) \qquad &\le \|x_k-x^*\|^2+\alpha^2M^2\|x_k-x^*\|^2+2\alpha(x^*-x_k)^T\nabla f(x_k) \\
(\text{Strong convexity with } m) \qquad &\le \|x_k-x^*\|^2+\alpha^2M^2\|x_k-x^*\|^2+2\alpha(f(x^*)-f(x_k)-\frac{m}{2}\|x^*-x_k\|^2) \\
&= (1+\alpha^2M^2-\alpha m)\|x_k-x^*\|^2+2\alpha(f(x^*)-f(x_k))
\end{aligned}$$

By strong convexity of f

$$\begin{aligned}
f(x_k) &\ge f(x^*)+\nabla^T f(x^*)(x_k-x^*)+\frac{m}{2}\|x_k-x^*\|^2 \\
&= f(x^*)+\frac{m}{2}\|x_k-x^*\|^2 \\
\Rightarrow f(x^*)-f(x_k) &\le -\frac{m}{2}\|x_k-x^*\|^2
\end{aligned}$$

Then,

$$\begin{aligned}
\|x_{k+1}-x^*\|^2 &\le (1+\alpha^2M^2-\alpha m)\|x_k-x^*\|^2+2\alpha(-\frac{m}{2}\|x_k-x^*\|^2) \\
&\le (1+\alpha^2M^2-2\alpha m)\|x_k-x^*\|^2 \\
&\le (1+\alpha^2M^2-2\alpha m)^{k+1}\|x_0-x^*\|^2 \\
\Rightarrow \|x_N-x^*\|^2 &\le (1+\alpha^2M^2-2\alpha m)^N\|x_0-x^*\|^2
\end{aligned}$$

If $\alpha\in(0,\frac{2m}{M^2})$, $1+\alpha^2M^2-2\alpha m<1$. Then $x_N\to x^*$ **geometrically** as $N\to\infty$.

Note: Just having $0<\alpha<\frac{2}{M}$ doesn't guarantee geometric convergence to x^*. e.g. $\alpha=\frac{1}{M}\Rightarrow 1+\alpha^2M^2-2m\alpha=2(1-\frac{m}{M})\ge 1$ if $\frac{m}{M}\le 0.5$

To get the highest convergence rate:

$$\begin{aligned}
1+\alpha^2M^2-2m\alpha &= (\alpha M)^2-2\alpha M\frac{m}{M}+1 \\
&= (\alpha M-\frac{m}{M})^2+1-\frac{m^2}{M^2}
\end{aligned}$$

Which is minimized by setting

$$\alpha=\alpha^*=\frac{m}{M^2}$$

$$\min_{\alpha>0} 1+\alpha^2M^2-2m\alpha=1-\frac{m^2}{M^2}\in[0,1)$$

Since $M>m$, $\alpha^*=\frac{m}{M^2}<\frac{1}{M}<\frac{2}{M}$.

With $\alpha=\alpha^*$,

$$\|x_N-x^*\|^2\le(1-\frac{m^2}{M^2})^N\|x_0-x^*\|^2$$

$\frac{M}{m}$ is called the **condition number**.

- If $\frac{M}{m} >> 1$, then $1 - \frac{m^2}{M^2}$ is close to 1 and convergence is slow.
- If $\frac{M}{m} = 1$, $\alpha^* = \frac{1}{M}$, and $x_N = x^*, \forall N \geq 1$. (Convergence in one step.)

Note that since $\nabla f(x^*) = 0$,

$$\begin{aligned} f(x_N) - f(x^*) &\leq \frac{M}{2}\|x_N - x^*\|^2 \\ &\leq (1 - \frac{m^2}{M^2})^N \frac{M}{2}\|x_0 - x^*\|^2 \end{aligned}$$

To make $f(x_N) - f(x^*) < \varepsilon$,

$$\begin{aligned} (1 - \frac{m^2}{M^2})^N \frac{M}{2}\|x_0 - x^*\|^2 &\sim \varepsilon \\ (1 - \frac{m^2}{M^2})^{-N} &\sim \frac{1}{\varepsilon} \\ -N\log(1 - \frac{m^2}{M^2}) &\sim \log\frac{1}{\varepsilon} \\ N &\sim \log\frac{1}{\varepsilon} \end{aligned}$$

we only need $N \sim O(log\frac{1}{\varepsilon})$ - called ”linear” convergence.

Example 2.4 $f(x) = \frac{1}{2}x^TQx + b^Tx + c, \quad Q \succ 0, \nabla^2 f(x) = Q.$

Let $\lambda_{\min}$ and $\lambda_{\max}$ be the min and max eigenvalue of Q. Then we know

$$\lambda_{\min}\|z\|^2 \leq z^TQz \leq \lambda_{\max}\|z\|^2$$

Thus for all $z \in \mathbb{R}^n$

$$z^T(Q - \lambda_{\min}I)z \geq 0 \Rightarrow Q \succeq \lambda_{\min}I$$

Similarly, $Q \preceq \lambda_{\max}I$. Thus

$$\lambda_{\min}I \preceq \nabla^2 f(x) \preceq \lambda_{\max}I$$

$\lambda_{\min}I \preceq \nabla^2 f(x) \Leftrightarrow f$ is $\lambda_{\min}$-strongly convex; $\nabla^2 f(x) \preceq \lambda_{\max}I$ is a sufficient condition for f is $\lambda_{\max}$-smooth.

The condition number $= \frac{\lambda_{\max}}{\lambda_{\min}}$

Special Case: $Q = \mu I, \quad \mu > 0, \lambda_{\min} = \lambda_{\max} = \mu = m = M.$

$f(x) = \frac{\mu}{2}\|x\|^2 + b^Tx + c$, $\nabla f(x) = \mu x + b$, $x^* = -\frac{b}{\mu}$, $\alpha^* = \frac{m}{M^2} = \frac{1}{\mu}$,

$$x_1 = x_0 - \alpha^*\nabla f(x_0) = x_0 - \frac{1}{\mu}(\mu x_0 + b) = -\frac{b}{\mu} = x^*$$

Convergence in one step!

2.5 Convergence of Gradient Descent on Smooth Strongly-Convex Functions

Still consider the constant stepsize gradient method

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

Lemma 2.1

Suppose the sequences $\{\xi_k \in \mathbb{R}^p : k = 0, 1, \ldots\}$ *and* $\{u_k \in \mathbb{R}^p : k = 0, 1, 2, \ldots\}$ *satisfy* $\xi_{k+1} = \xi_k - \alpha u_k$. *In addition, assume there is a martix* M*, the following inequality holds for all* k

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \geq 0$$

If there exist $0 < \rho < 1$ *and* $\lambda \geq 0$ *such that*

$$\begin{bmatrix} \left(1 - \rho^2\right) I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} + \lambda M$$

is a negative semidefinite matrix, then the sequence $\{\xi_k : k = 0, 1, \ldots\}$ *satisfies* $\|\xi_k\| \leq \rho^k \|\xi_0\|$. ♡

Proof 2.6

The key relation is

$$\|\xi_{k+1}\|^2 = \|\xi_k - \alpha u_k\|^2 = \|\xi_k\|^2 - 2\alpha(\xi_k)^T u_k + \alpha^2\|u_k\|^2 = \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top \begin{bmatrix} I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}$$

Since $\begin{bmatrix} \left(1 - \rho^2\right) I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} + \lambda M$ *is negative semidefinite, we have*

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top \left(\begin{bmatrix} \left(1 - \rho^2\right) I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} + \lambda M \right) \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0$$

Expand the inequality,

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top \begin{bmatrix} I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} + \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top \begin{bmatrix} -\rho^2 I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} + \lambda \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0$$

Applying the key relation

$$\|\xi_{k+1}\|^2 - \rho^2\|\xi_k\|^2 + \lambda \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0$$

$$\|\xi_{k+1}\|^2 - \rho^2\|\xi_k\|^2 \leq -\lambda \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0$$

Hence, $\|\xi_{k+1}\| \leq \rho\|\xi_k\|$ *for all* k*. Therefore, we have* $\|\xi_k\| \leq \rho^k \|\xi_0\|$.

Theorem 2.4

Suppose f is L-smooth and m-strongly convex. Let x^* be the unique global min. Given a stepsize α,if there exists $0 < \rho < 1$ and $\lambda \geq 0$ such that

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix}$$

is a negative semidefinite matrix, then the gradient method satisfies

$$\|x_k - x^*\| \leq \rho^k \|x_0 - x^*\|$$

♡

Proof 2.7

We set f is L-smooth and m-strongly convex,

According to the definition of m-strongly convex

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq m\|x-y\|^2$$

And the co-coercivity condition, if f is $L-$smooth,

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$$

Set $g(x) = f(x) - \frac{m}{2}\|x\|^2$, $\nabla g(x) = \nabla f(x) - mx$.

$$\begin{aligned} f \text{ is } L\text{-smooth} &\Leftrightarrow \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\| \\ &\Leftrightarrow \|\nabla g(x) - \nabla g(y)\| \leq (L-m)\|x-y\| \\ &\Leftrightarrow g \text{ is } L-m\text{-smooth} \end{aligned}$$

Hence,

$$(\nabla g(x) - \nabla g(y))^T(x-y) \geq \frac{1}{L-m}\|\nabla g(x) - \nabla g(y)\|^2$$

$$(\nabla f(x) - \nabla f(y) - m(x-y))^T(x-y) \geq \frac{1}{L-m}\|\nabla f(x) - \nabla f(y) - m(x-y)\|^2$$

$$(L-m)[(\nabla f(x) - \nabla f(y))^T(x-y) - m\|x-y\|^2]$$

$$\geq \|\nabla f(x) - \nabla f(y)\|^2 + m^2\|(x-y)\|^2 - 2m(\nabla f(x) - \nabla f(y))^T(x-y)$$

$$(L+m)(\nabla f(x) - \nabla f(y))^T(x-y) \geq mL\|x-y\|^2 + \|\nabla f(x) - \nabla f(y)\|^2$$

$$\Rightarrow (\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{mL}{m+L}\|x-y\|^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|^2$$

Which can be rewritten as

$$\begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^T \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \geq 0$$

Let $y = x^*$ and $\nabla f(y) = \nabla f(x^*) = 0$

$$\begin{bmatrix} x - x^* \\ \nabla f(x) \end{bmatrix}^T \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} x - x^* \\ \nabla f(x) \end{bmatrix} \geq 0$$

Set $\xi_k = x_k - x^*$ and $u_k = \nabla f(x_k)$. And $\xi_{k+1} = x_{k+1} - x^* = x_k - \alpha \nabla f(x_k) - x^* = \xi_k - \alpha u_k$

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^T \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \geq 0$$

Choose $M = \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix}$. Then prove by previous lemma.

Now we apply the theorem to obtain the convergence rate ρ for the gradient method with various stepsize choices.

- Case 1: If we choose $\alpha = \frac{1}{L}, \rho = 1 - \frac{m}{L}$, and $\lambda = \frac{1}{L^2}$, we have

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} = \begin{bmatrix} -\frac{m^2}{L^2} & \frac{m}{L^2} \\ \frac{m^2}{L^2} & -\frac{1}{L^2} \end{bmatrix} = \frac{1}{L^2} \begin{bmatrix} -m^2 & m \\ m & -1 \end{bmatrix}$$

  The right side is clearly negative semidefinite due to the fact that $\begin{bmatrix} a \\ b \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -m^2 & m \\ m & -1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = -(ma-b)^2 \leq 0$. Therefore, the gradient method with $\alpha = \frac{1}{L}$ converges as

$$\|x_k - x^*\| \leq \left(1 - \frac{m}{L}\right)^k \|x_0 - x^*\|$$

- Case 2: If we choose $\alpha = \frac{2}{m+L}, \rho = \frac{L-m}{L+m}$, and $\lambda = \frac{2}{(m+L)^2}$, we have

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

  The zero matrix is clearly negative semidefinite. Therefore, the gradient method with $\alpha = \frac{2}{m+L}$ converges as

$$\|x_k - x^*\| \leq \left(\frac{L-m}{L+m}\right)^k \|x_0 - x^*\|$$

Notice $L \geq m > 0$ and hence $1 - \frac{m}{L} \geq \frac{L-m}{L+m}$. This means the gradient method with $\alpha = \frac{2}{m+L}$ converges slightly faster than the case with $\alpha = \frac{1}{L}$. However, m is typically unknown in practice. The step choice of $\alpha = \frac{1}{L}$ is also more robust. The most popular choice for α is still $\frac{1}{L}$.

We can further express ρ as a function of α. To do this, we need to choose λ carefully for a given α. If we choose λ reasonably, we can show the best value for ρ that we can find is $\max\{|1 - m\alpha|, |L\alpha - 1|\}$.

2.6 From convergence rate to iteration complexity

The convergence rate ρ naturally leads to an iteration number T guaranteeing the algorithm to achieve the so-called **ε-optimality**, i.e. $\|x_T - x^*\| \leq \varepsilon$.

To guarantee $\|x_T - x^*\| \leq \varepsilon$, we can use the bound $\|x_T - x^*\| \leq \rho^T \|x_0 - x^*\|$. If we choose T such that $\rho^T \|x_0 - x^*\| \leq \varepsilon$, then we guarantee $\|x_T - x^*\| \leq \varepsilon$. Denote $c = \|x_0 - x^*\|$. Then $c\rho^k \leq \varepsilon$ is equivalent to

$$\log c + k \log \rho \leq \log(\varepsilon)$$

Notice $\rho < 1$ and $\log \rho < 0$. The above inequality is equivalent to

$$k \geq \log\left(\frac{\varepsilon}{c}\right) / \log \rho = \log\left(\frac{c}{\varepsilon}\right) / (-\log \rho)$$

So if we choose $T = \log\left(\frac{c}{\varepsilon}\right) / (-\log \rho)$, we guarantee $\|x_T - x^*\| \leq \varepsilon$. Notice $\log \rho \leq \rho - 1 < 0$ (this can be proved using the concavity of $\log$ function and we will talk about concavity in later lectures), so $\frac{1}{1-\rho} \geq -\frac{1}{\log \rho}$ and we can also choose $T = \log\left(\frac{c}{\varepsilon}\right) / (1-\rho) \geq \log\left(\frac{c}{\varepsilon}\right) / (-\log \rho)$ to guarantee $\|x_T - x^*\| \leq \varepsilon$.

Another interpretation for $T = \log\left(\frac{c}{\varepsilon}\right) / (1-\rho)$ is that a first-order Taylor expansion of $-\log \rho$ at $\rho = 1$ leads to $-\log \rho \approx 1 - \rho$. So $\log\left(\frac{c}{\varepsilon}\right) / (-\log \rho)$ is roughly equal to $\log\left(\frac{c}{\varepsilon}\right) / (1-\rho)$ when ρ is close to 1 .

Clearly the smaller T is, the more efficient the optimization method is. The iteration number T describes the ”ε-optimal iteration complexity” of the gradient method for smooth strongly-convex objective functions.

- For the gradient method with $\alpha = \frac{1}{L}$, we have $\rho = 1 - \frac{m}{L} = 1 - \frac{1}{\kappa}$ and hence $T = \log\left(\frac{c}{\varepsilon}\right) / (1-\rho) = \kappa \log\left(\frac{c}{\varepsilon}\right) = O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$ .2 Here we use the big O notation to highlight the dependence on κ and ε and hide the dependence on the constant c.
- For the gradient method with $\alpha = \frac{2}{L+m}$, we have $\rho = \frac{\kappa-1}{\kappa+1} = 1 - \frac{2}{\kappa+1}$ and hence $T = \log\left(\frac{c}{\varepsilon}\right) / (1-\rho) = \frac{\kappa+1}{2} \log\left(\frac{c}{\varepsilon}\right)$. Although $\frac{\kappa+1}{2} \leq \kappa$, we still have $\frac{\kappa+1}{2} \log\left(\frac{c}{\varepsilon}\right) = O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$. Therefore, the stepsize $\alpha = \frac{2}{m+L}$ can only improve the constant C hidden in the big O notation of the iteration complexity. People call this “improvement of a constant factor”.
- In general, when ρ has the form $\rho = 1 - 1/(a\kappa + b)$, the resultant iteration complexity is always $O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$.

There are algorithms which can significantly decrease the iteration complexity for unconstrained optimization problems with smooth strongly-convex objective functions. For example, Nesterov’s method can decrease the iteration complexity from $O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$ to $O\left(\sqrt{\kappa} \log\left(\frac{1}{\varepsilon}\right)\right)$. Momentum is used to accelerate optimization as:

$$x_{k+1} = x_k - \alpha \nabla f\left((1+\beta)x_k - \beta x_{k-1}\right) + \beta\left(x_k - x_{k-1}\right).$$

Chapter 3 (Unconstrained Optimization) Gradient Projection Methods

3.1 Projection onto Closed Convex Set

3.1.1 Def: Projection $[z]^{\&}$

Definition 3.1

Let $\&$ be a closed convex subset of $\mathbb{R}^n$. Then, for $z \in \mathbb{R}^n$, the projection of z on $\&$ is denoted by $[z]^{\&}$ and is given by

$$[z]^{\&} = \arg\min_{y \in \&} \|z - y\|^2$$

♣

i.e. Find the min distance from $\&$ to z

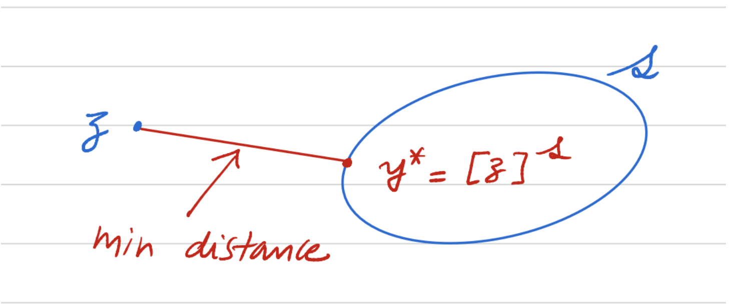

Figure 3.1: Projection onto Closed Convex Set

Note: $[z]^{\&}$ exists and is unique in convex $\&$, however, when $\&$ is not convex, $[z]^{\&}$ may not be unique.

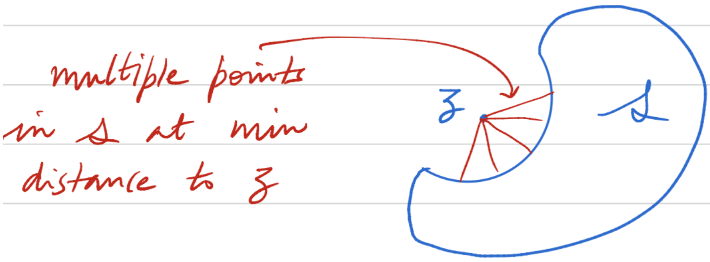

Figure 3.2: Projection onto Closed non-Convex Set

3.1.2 Prop: unique projection $[z]^{\&}$ on closed convex subset of $\mathbb{R}^n$

Proposition 3.1 (Existence and Uniqueness of Projection)

Let & be a closed convex subset of $\mathbb{R}^n$. Then, for every $z \in \mathbb{R}^n$, there exists a unique $[z]^{\&}$.

Proof 3.1

Nee to show that $\min_{y\in\&} \|z-y\|^2$ *exists and is unique.*

Let x be some element of &. Then

$$\text{minimizing } \|z-y\|^2 \text{ over all } y \in \&$$

$$\equiv \text{minimizing } \|z-y\|^2 \text{ over the set } A = \{y \in \& : \|z-y\|^2\}$$

$g(y) = \|z-y\|^2$ is strictly convex on set & $\Rightarrow$ A is a convex set and g is convex on A.

Also g is continuous $\Rightarrow$ A is closed.

Finally, $y \in A \Rightarrow \|y\|^2 = \|y-z+z\|^2 \leq \|y-z\|^2 + \|z\|^2 \leq \|z-x\|^2 + \|z\|^2 \Rightarrow$ A is bounded.

Thus, $g(y) = \|z-y\|^2$ is strictly convex over set A, which is compact.

Therefore, $\min_{y\in\&} \|\& - y\|^2 = \min_{y\in A} \|\& - y\|^2$ exists (Weierstrass' Theorem) and is unique (strict convexity).

3.1.3 Obtuse Angle Criterion: $x = [z]^{\&}$ is projection on closed convexsubset of $\mathbb{R}^n \Leftrightarrow$ $(z-x)^T(y-x) \leq 0, \forall y \in \&$

When solving the closest point problem for a vector space $S = \{\vec{y} : A\vec{y} = \vec{0}\}$ or affine subspace $S = \{\vec{y} : A\vec{y} = \vec{b}\}$, we use **perpendicularity condition** (orthogonality condition) -i.e., that $\vec{x}^* \cdot \vec{y} = 0, \forall \vec{y} \in S$

A weaker form of that condition is the obtuse angle criterion, **which holds for any convex set**:

Theorem 3.1 (Obtuse angle criterion)

Let C be a convex set and let $\vec{y}$ be a point outside C. $\vec{x}^*$ is the closest point in C to $\vec{y}$ if and only if

$$(\vec{y} - \vec{x}^*) \cdot (\vec{x} - \vec{x}^*) \leq 0, \quad \forall \vec{x} \in C$$

Why is this the obtuse angle criterion? Because it says that the angle between $\mathbf{y} - \mathbf{x}^*$ (the vector pointing from $\mathbf{x}^*$ to $\mathbf{y}$) and $\mathbf{x} - \mathbf{x}^*$ (the vector pointing from $\mathbf{x}^*$ to $\mathbf{x}$) is a right angle or an obtuse angle. See the diagram below:

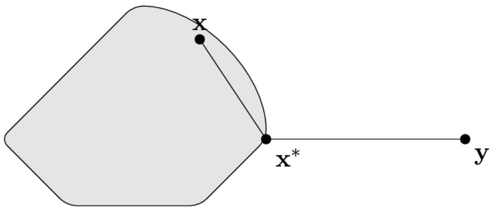

Figure 3.3: Obtuse angle criterion

In Projection Form

Proposition 3.2 (Necessary and Sufficient Condition for Projection)

Let $\&$ be a closed conex subset of $\mathbb{R}^n$. Then,

$$[z]^{\&} = y^* \Leftrightarrow (y^* - z)^T (y - y^*) \geq 0, \quad \forall y \in \&.$$
$$\Leftrightarrow (z - y^*)^T (y - y^*) \leq 0, \quad \forall y \in \&.$$

Proof 3.2

$[z]^{\&} = \mathrm{argmin}_{y \in \&}\, g(y)$, *with* $g(y) = \|z - y\|^2$ *(which is strictly convex),* $\nabla g(y) = 2(y - z)$.

By the optimality conditions,

$$y^* \textit{ is the unique minimizer of } g(y) \textit{ over } \&$$
$$\Leftrightarrow \nabla g(y^*)^T (y - y^*) \geq 0 \quad \forall y \in \&$$
$$\Leftrightarrow (y^* - z)^T (y - y^*) \geq 0, \quad \forall y \in \&.$$
$$\Leftrightarrow (z - y^*)^T (y - y^*) \leq 0, \quad \forall y \in \&.$$

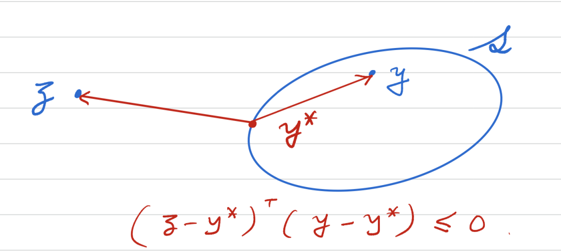

Figure 3.4: Necessary and Sufficient Condition for Projection

3.1.4 Prop: Projection is non-expansive $\|[x]^{\&} - [z]^{\&}\| \leq \|x - z\|, \forall x, z \in \mathbb{R}^n$

Proposition 3.3 (Projection is non-expansive)

Let $\&$ be a closed convex subset of $\mathbb{R}^n$. Then for $x, z \in \mathbb{R}^n$

$$\|[x]^{\&} - [z]^{\&}\| \leq \|x - z\| \quad \forall x, z \in \mathbb{R}^n$$

Proof 3.3

From previous theorem, we know

$$(1). \quad ([x]^{\&} - x)^T (y - [x]^{\&}) \geq 0, \quad \forall y \in \&.$$
$$(2). \quad ([z]^{\&} - z)^T (y - [z]^{\&}) \geq 0, \quad \forall y \in \&.$$

set $y = [z]^{\&}$ in (1) and $y = [x]^{\&}$ in (2), and adding,

$$([z]^{\&} - [x]^{\&})^T([x]^{\&} - x + z - [z]^{\&}) \geq 0$$

$$\Rightarrow([z]^{\&} - [x]^{\&})^T(z - x) \geq \|[z]^{\&} - [x]^{\&}\|^2$$

Applying Cauchy-schwary inequality,

$$\|[z]^{\&} - [x]^{\&}\|^2 \leq \|[z]^{\&} - [x]^{\&}\|\|z - x\|$$

$$\|[z]^{\&} - [x]^{\&}\| \leq \|z - x\|$$

3.2 Projection on (Linear) Subspaces of $\mathbb{R}^n$

3.2.1 Orthogonality Principle in subspaces of $\mathbb{R}^n$: $(z - y^*)^T x = 0, \forall x \in \&$

Suppose $\&$ is a linear subspace of $\mathbb{R}^n$, any linear combination of points in $\&$ is also in $\&$. Note that $\&$ is closed and convex.

Then, for $z \in \mathbb{R}^n$, $[z]^{\&} = y^*$ satisfies:

$$(z - y^*)^T(y - y^*) \leq 0, \quad \forall y \in \&.$$

According to the property of subsapce, we can infer that

$$(z - y^*)^T x \leq 0, \quad \forall x \in \&.$$

$-x$ also in $\&$, $-x \in \& \Rightarrow$

$$(z - y^*)^T x \geq 0, \quad \forall x \in \&.$$

Then we can infer that

$$(z - y^*)^T x = 0, \quad \forall x \in \&.$$

which is called orthogonality principle.

3.3 Gradient Projection Method

$\min_{x \in \&} f(x)$, $\&$ is convex and closed.

$$x_{k+1} = [x_k + \alpha_k d_k]^{\&}$$

Special Case: Fixed step-size, steepest descent

$$x_{k+1} = [x_k - \alpha \nabla f(x_k)]^{\&} \tag{3.1}$$

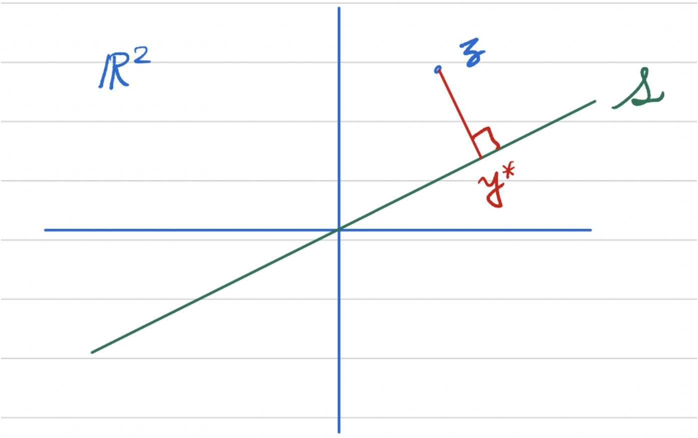

Figure 3.5: Point from $\mathbb{R}^2$ to $\mathbb{R}$

3.3.1 Def: fixed point in fixed step-size steepest descent method, $\tilde{x} = [\tilde{x} - \alpha \nabla f(\tilde{x})]^{\&}$

Definition 3.2

$\tilde{x}$ is a fixed (stationary) point of iteration in (1) if

$$\tilde{x} = [\tilde{x} - \alpha \nabla f(\tilde{x})]^{\&}$$

3.3.2 Prop: $L-$smooth, $0 < \alpha < \frac{2}{L} \Rightarrow$ limit point is a fixed point (in fixed step-size steepest descent method)

Proposition 3.4

If f has $L-$Lipschitz gradient and $0 < \alpha < \frac{2}{L}$, every limit point of (1) is a fixed point of (1).

Proof 3.4

By the Descent Lemma,

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2 \tag{3.2}$$

By the necessary and sufficient condition for projection,

$$(x_k - \alpha \nabla f(x_k) - x_{k+1})^T (x - x_{k+1}) \leq 0, \quad \forall x \in \&$$

Set $x = x_k$ above

$$\Rightarrow \alpha \nabla f(x_k)^T (x_{k+1} - x_k) \leq -\|x_k - x_{k+1}\|^2 \tag{3.3}$$

According to (2) and (3),

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

where $\frac{L}{2} - \frac{1}{\alpha} < 0$

If $\{x_k\}$ *has limit point* $\bar{x}$, $LHS \overset{k\to\infty}{\longrightarrow} 0$

$$\|x_{k+1} - x_k\| \overset{k\to\infty}{\longrightarrow} 0 \Rightarrow [\bar{x} - \alpha\nabla f(\bar{x})]^{\&} = \bar{x}$$

3.3.3 Prop: x is minimizer in convex func $\Leftrightarrow$ fixed point (in fixed step-size steepest descent method)

Proposition 3.5

If f *is convex, then* x^* *is a minimizer of* f *over* $\&$ $\Leftrightarrow x^* = [x^* - \alpha\nabla f(x^*)]^{\&}$ *(i.e.,* x^* *is a fixed point of (1))* ♠

Proof 3.5

$$\begin{aligned} x^* \text{ is minimizer of convex } f \text{ over } \& &\Leftrightarrow \nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in \& \\ &\Leftrightarrow -\alpha\nabla f(x^*)^T(x - x^*) \leq 0, \forall x \in \& \\ &\Leftrightarrow (x^* - \alpha\nabla f(x^*) - x^*)^T(x - x^*) \leq 0, \forall x \in \& \\ \text{(By Projection Theorem)} &\Leftrightarrow [x^* - \alpha\nabla f(x^*)]^{\&} = x^* \end{aligned}$$

3.3.4 Thm: Convergence of Gradient Projection: Convex, $L-$smooth, $0 < \alpha < \frac{2}{L} \Rightarrow f(x_k) \to f(x^*)$ at rate $\frac{1}{k}$

Theorem 3.2

If f *is convex and* $L-$*Lipschitz gradient, it can be shown that for* $0 < \alpha < \frac{2}{L}$

$$f(x_k) \to f(x^*) \text{ at rate } \frac{1}{k} \text{(same as unconstrainted)}$$

♡

3.3.5 Thm: Strongly convex, Lipschitz gradient $\Rightarrow \{x_k\}$ converges to x^* geometrically

Theorem 3.3

If f *has Lipschitz gradient with Lipschitz constant* M *and strongly convex with parameter* m, $\{x_k\}$ *converges to* x^* ***geometrically***. ♡

Proof 3.6

$M-$*smooth* $\Rightarrow$

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|, \quad \forall x, y \in \&$$

$m-$*strongly convex* $\Rightarrow$

$$\nabla^2 f(x) \succeq mI, \quad \forall x \in \&$$

$$(x-y)^T(\nabla f(x)-\nabla f(y)) \geq m\|x-y\|^2 \quad \forall x,y \in \&$$

Let x^* be the (unique) min of f over $\&$

$$\begin{aligned}
\|x_{k+1}-x^*\|^2 &= \|[x_k-\alpha\nabla f(x_k)]^{\&}-x^*\|^2 \\
(x^* \text{ is fixed point}) \quad &= \|[x_k-\alpha\nabla f(x_k)]^{\&}-[x^*-\alpha\nabla f(x^*)]^{\&}\|^2 \\
(\text{non-expansive}) \quad &\leq \|(x_k-\alpha\nabla f(x_k))-(x^*-\alpha\nabla f(x^*))\|^2 \\
&= \|(x_k-x^*)-\alpha(\nabla f(x_k)-\nabla f(x^*))\|^2 \\
&= \|x_k-x^*\|^2+\alpha^2\|\nabla f(x_k)-\nabla f(x^*)\|^2-2\alpha(x_k-x^*)^T(\nabla f(x_k)-\nabla f(x^*)) \\
(\nabla f \text{ is M-Lipschitz}) \quad &\leq \|x_k-x^*\|^2+\alpha^2M^2\|x_k-x^*\|^2-2\alpha(x_k-x^*)^T(\nabla f(x_k)-\nabla f(x^*)) \\
(m-\text{strong convexity}) \quad &\leq \|x_k-x^*\|^2+\alpha^2M^2\|x_k-x^*\|^2-2\alpha m\|x_k-x^*\|^2 \\
&= (1+\alpha^2M^2-2\alpha m)\|x_k-x^*\|^2 \\
&\|x_{k+1}-x^*\|^2 \leq (1+\alpha^2M^2-2\alpha m)\|x_k-x^*\|^2
\end{aligned}$$

If $|1+\alpha^2M^2-2\alpha m|<1$. Then $x_N \to x^*$ ***geometrically*** *as $N\to\infty$. (Same as unconstrained case)*

Chapter 4 (Unconstrained Optimization) Sub-gradient Methods

Gradient descent methods require ∇f exists. What if ∇f doesn't exist at some point?

Recall that when ∇f exists

f is convex on $\&$ $\Leftrightarrow f(y) \geq f(x) + \nabla f(x)^T(y-x), \forall x, y \in \&$ (the inequality is strict for strict convexity)

4.1 Sub-gradient

Definition 4.1

*For convex f on $\mathbb{R}^n$, g is called a **sub-gradient** of f at $x \in \mathbb{R}^n$ if*

$$f(y) \geq f(x) + g^T(y-x), \quad \forall y \in \mathbb{R}^n$$

♣

Properties of Sub-gradient

1) Sub-gradient always exist at any point for convex functions.
2) If ∇f exists at a point x for convex f, sub-gradient is unique and $= \nabla f(x)$
3) Some definition for sub-gradient can be applied for non-convex f, but sub-gradient may not exist.

Example 4.1 $f(x) = |x|, x \in \mathbb{R}$

For $x \neq 0$, ∇f exists and $=$ sub-gradient.

For $x = 0$, any $g \in [-1, 1]$ is a sub-gradient.

Proof 4.1

(1) For $y > 0$, $f(y) = y \geq f(0) + gy = gy, \forall g \in [-1, 1]$

(2) For $y < 0$, $f(y) = -y \geq f(0) + gy = gy, \forall g \in [-1, 1]$

4.2 Sub-differential

Definition 4.2

*Set of all sub-gradient at x is called **sub-differential** at x, denoted $\partial f(x)$.*

♣

Example 4.2 For $f(x) = |x|$,

$$\partial f(x) = \begin{cases} -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Example 4.3 For $f(x) = \max\{1, |x| - 1\}$. (Note: $f(x)$ is convex since $1, |x| - 1$ are both convex.)

$$\partial f(x) = \begin{cases} -1 & \text{if } x < -2 \\ [-1, 0] & \text{if } x = -2 \\ 0 & \text{if } -1 < x < 2 \\ [0, 1] & \text{if } x = 2 \\ 1 & \text{if } x > 2 \end{cases}$$

When $x = 2$,

$$f(y) = \max\{1, |y| - 1\} \geq f(2) + g(y - 2) = 1 + g(y - 2)$$

(1) $\underline{y \geq 0}$: $0 \geq g(y - 2)$ or $0 \geq (g - 1)(y - 2)$. If $y > 2$, $g \leq 1$; If $y = 2$, $\forall g$; If $0 \leq y < 2$, $g \geq 0$. $\Rightarrow g \in [0, 1]$

(2) $\underline{y < 0}$: $0 \geq g(y - 2)$ or $-y - 2 \geq g(y - 2)$, i.e. $g \geq 0$ or $g \leq \frac{2+y}{2-y}$ (satisifed by $g \in [0, 1]$)

4.3 More examples

Example 4.4 $f(x) = \|x\| = \sqrt{x^T x}$

- f is convex (by Triangle Inequality: $\|x\| + \|y\| \geq \|x + y\|$)
- For $x \neq 0$, $\nabla f(x)$ exists and

$$\partial f(x) = \nabla f(x) = \frac{1}{2\sqrt{x^T x}} \cdot 2x = \frac{x}{\|x\|}$$

- If $x = 0$, $\nabla f(x)$ doesn't exist.

Proposition 4.1

$\partial f(0) = \{g \in \mathbb{R}^n : \|g\| \leq 1\}$ ♠

Proof 4.2

Need to show that for $\|g\| \leq 1$ and $\forall y \in \mathbb{R}^n$,

$$f(y) = \|y\| \geq f(0) + g^T(y - 0) = g^T y$$

But by Cauchy-Schwarz inequality, for $\|g\| \leq 1$,

$$g^T y \leq \|g\|\|y\| \leq \|y\|, \forall y \in \mathbb{R}^n$$

To estabilish the converse, suppose $\|g\| > 1$.

Then, setting $y = \frac{g}{\|g\|} \Rightarrow \|y\| = 1$ *but* $g^T y = \|g\| > 1 = \|y\|$

Example 4.5 $f(x) = |x_1 - x_2| \leftarrow$ convex

If $x_1 > x_2$, $|x_1 - x_2| = x_1 - x_2$, and ∇f exists and $(1, -1)$

If $x_1 < x_2$, $|x_1 - x_2| = x_2 - x_1$, and ∇f exists and $(-1, 1)$

Proposition 4.2

If $x_1 = x_2$, $\partial f(x) = \{(a, b) : a = -b, |a| \leq 1\}$ ♠

Proof 4.3

Suppose $x_1 = x_2 = c$. *Then we need to show* $\forall y \in \mathbb{R}^2$, (a, b) *s.t.* $a = -b$, $|a| \leq 1$

$$|y_1 - y_2| \geq f(c, c) + [a\ b] \begin{bmatrix} y_1 - c \\ y_2 - c \end{bmatrix} = ay_1 + by_2 - c(a + b) = a(y_1 - y_2)$$

Since $|a| < 1$, *this inequality holds* $\forall y \in \mathbb{R}^2$

To show the converse,

1. *Suppose* $a \neq -b$.

   If $c(a + b) < 0$, *setting* $y_1 = y_2 = 0$. $\Rightarrow |y_1 - y_2| = 0$, *and* $ay_1 + by_2 - c(a + b) = -c(a + b) > 0 = |y_1 - y_2|$, *above inequality fails to hold.*

   If $c(a + b) > 0$, *setting* $y_1 = y_2 = 2c$. $\Rightarrow |y_1 - y_2| = 0$, *and* $ay_1 + by_2 - c(a + b) = c(a + b) > 0 = |y_1 - y_2|$, *above inequality fails to hold.*

   If $c = 0$, *setting* $y_1 = y_2 = (a + b)$. $\Rightarrow |y_1 - y_2| = 0$, *and* $ay_1 + by_2 - c(a + b) = (a + b)^2 > 0 = |y_1 - y_2|$, *above inequality fails to hold.*

2. *Suppose* $a = -b$ *with* $|a| > 1$.

   If $a > 1$, *setting* $y_1 = y_2 + 1$; *If* $a < -1$, *setting* $y_1 = y_2 - 1$.

4.4 First-order necessary conditions for optimality in terms of subgradient

Proposition 4.3

For convex f, $f(x^*) = \min_x f(x) \Leftrightarrow 0 \in \partial f(x^*)$ ♠

Proof 4.4

x^* *is a minimizer* $\Leftrightarrow f(x^*) \leq f(y), \forall y \in \mathbb{R}^n \Leftrightarrow f(x^*) + 0^T(y - x^*) \leq f(y), \forall y \in \mathbb{R}^n \Leftrightarrow 0 \in \partial f(x^*)$

4.5 Properties of Subgradients

Let f, f_1, f_2 be convex functions.

(a) **Scaling:** For scalar $a > 0$, $\partial(af) = a\partial f$, i.e., g is a subgradient of f at x if and only if ag is a subgradient of af at x.

(b) **Addition:** If g_1 is a subgradient of f_1 at x, and g_2 is a subgradient of f_2 at x, then $g_1 + g_2$ is subgradient of $f_1 + f_2$ at x.

(c) **Affine Combination:** Let $h(x) = f(Ax + b)$, with A being a square, invertible matrix. Then $\partial h(x) = A^T \partial f(Ax + b)$, i.e., g is a subgradient of f at $Ax + b$ if and only if $A^T g$ is a subgradient of h at x.

4.6 Sub-gradient Descent for Unconstrained Optimization

Assumptions:

(i) f is convex on $\mathbb{R}^n$.

(ii) $f^* = \inf_{x \in \mathbb{R}^n} f(x)$ exists and there exists an x^* s.t. $f(x^*) = f^*$.

(iii) For all $x \in \mathbb{R}^n$ and for all $g \in \partial f(x)$, $\|g\| \leq a$.

Subgradient Descent with constant step-size:

$$x_{k+1} = x_k - \alpha g_k, \quad g_k \in \partial f(x_k)$$

Analysis:

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - \alpha g_k - x^*\|^2 \\ &= \|x_k - x^*\|^2 + \alpha^2 \|g_k\|^2 - 2\alpha g_k^T (x_k - x^*) \\ &\leq \|x_k - x^*\|^2 + \alpha^2 a^2 - 2\alpha g_k^T (x_k - x^*)\end{aligned}$$

By the definition of g_k,

$$f(x_k) + g_k^T (x^* - x_k) \leq f(x^*) = f^*$$

$$\Rightarrow \|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + \alpha^2 a^2 + 2\alpha(f^* - f(x_k))$$

$$f(x_k) - f^* \leq \frac{\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha^2 a^2}{2\alpha}$$

Define $f_N^* = \min\{f(x_0), f(x_1), ..., f(x_{N-1})\}$

$$\sum_{k=0}^{N-1} (f(x_k) - f^*) \geq \sum_{k=0}^{N-1} (f_N^* - f^*) = N(f_N^* - f^*)$$

Then,

$$\begin{aligned} N(f_N^* - f^*) &\leq \sum_{k=0}^{N-1} \frac{\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha^2 a^2}{2\alpha} \\ &= \frac{\|x_0 - x^*\|^2 - \|x_N - x^*\|^2 + N\alpha^2 a^2}{2\alpha} \\ \Rightarrow \quad f_N^* &\leq f^* + \frac{1}{2\alpha N}\|x_0 - x^*\|^2 + \frac{\alpha a^2}{2} \\ \lim_{N\to\infty} f_N^* &\leq f^* + \frac{\alpha a^2}{2} \end{aligned}$$

For α samll enough and N large enough f_N^* can be made as close to f^* as desired.

Note: $-$subgradient is not necessarily a descent direction

i.e., if g_k is a subgradient of f at x_k. Then

$$f(x_k - \alpha g_k) \text{ may be } \geq f(x_k), \quad \forall \alpha > 0$$

for some g_k.

Example 4.6 $f(x) = |x_1| + \frac{1}{2}x_2^2$

Suppose $x_k = (0, 1)$, then it is easy to show: $\partial f(0, 1) = ([-1, 1], 1)$

Consider $g_k = (-1, 1) \in \partial f(0, 1)$

$$f(x_k - \alpha g_k) = f(0 + \alpha, 1 - \alpha) = \frac{1}{2}(1 + \alpha^2) > \frac{1}{2} = f(x_k), \forall \alpha > 0$$

i.e., $-g_k$ is not a descent direction.

> If f is convex, there is some $g_k \in \partial f(x_k)$ for which $-g_k$ is a descent direction (usually the one with **the smallest norm**), but finding such g_k may be difficult in high-dimentional settings.
>
> This means we cannot use back-tracking algorithms (Armijo’s Rule) for adopting step-size.

4.7 (Revised) Sub-gradient ”descent” with diminishing stepsize

Assumptions:

(i) f is convex on $\mathbb{R}^n$.

(ii) $f^* = \inf_{x\in\mathbb{R}^n} f(x)$ exists and there exists an x^* s.t. $f(x^*) = f^*$.

(iii) For all $x \in \mathbb{R}^n$ and for all $g \in \partial f(x)$, $\|g\| \leq a$.

Subgradient Descent with constant step-size:

$$x_{k+1} = x_k - \alpha_k g_k, \quad g_k \in \partial f(x_k)$$

Analysis:

$$\begin{aligned}\|x_{k+1}-x^*\|^2 &= \|x_k-\alpha_k g_k - x^*\|^2\\ &= \|x_k-x^*\|^2+\alpha_k^2\|g_k\|^2-2\alpha_k g_k^T(x_k-x^*)\\ &\leq \|x_k-x^*\|^2+\alpha_k^2 a^2-2\alpha_k g_k^T(x_k-x^*)\end{aligned}$$

By the definition of g_k,

$$f(x_k)+g_k^T(x^*-x_k)\leq f(x^*)=f^*$$

$$\begin{aligned}\Rightarrow \|x_{k+1}-x^*\|^2 &\leq \|x_k-x^*\|^2+\alpha_k^2a^2+2\alpha_k(f^*-f(x_k))\\ &\leq \left(\|x_{k-1}-x^*\|^2+\alpha_{k-1}^2a^2+2\alpha_{k-1}(f^*-f(x_{k-1}))\right)+\alpha_k^2a^2+2\alpha_k(f^*-f(x_k))\\ &\dots\\ \Rightarrow \|x_N-x^*\|^2 &\leq \|x_0-x^*\|^2+a^2\sum_{k=0}^{N-1}\alpha_k^2+2\sum_{k=0}^{N-1}\alpha_k(f^*-f(x_k))\end{aligned}$$

Define $f_N^*=\min\{f(x_0),f(x_1),...,f(x_{N-1})\}$

$$\|x_N-x^*\|^2\leq\|x_0-x^*\|^2+a^2\sum_{k=0}^{N-1}\alpha_k^2+2(f^*-f_N^*)\sum_{k=0}^{N-1}\alpha_k$$

Then,

$$\begin{aligned}f_N^*-f^* &\leq \frac{\|x_0-x^*\|^2-\|x_N-x^*\|^2+a^2\sum_{k=0}^{N-1}\alpha_k^2}{2\sum_{k=0}^{N-1}\alpha_k}\\ &\leq \frac{\|x_0-x^*\|^2+a^2\sum_{k=0}^{N-1}\alpha_k^2}{2\sum_{k=0}^{N-1}\alpha_k}\end{aligned}$$

Suppse $\{\alpha_k\}$ is such that $\lim_{N\to\infty}\frac{\sum_{k=0}^{N-1}\alpha_k^2}{\sum_{k=0}^{N-1}\alpha_k}=0$, then $\lim_{N\to\infty}f_N^*=f^*$

Example of $\{\alpha_k\}$ and convergence rate

1) $\alpha_k=\frac{1}{k+1}, k=0,1,...$

$$\begin{aligned}\sum_{k=0}^{N-1}\alpha_k^2=\sum_{k=1}^{N}\frac{1}{k^2}\to\frac{\pi^2}{6}\\ \sum_{k=0}^{N-1}\alpha_k=\sum_{k=1}^{N}\frac{1}{k}>\log N\\ \Rightarrow(f_N^*-f^*)\sim O(\frac{1}{\log N})\end{aligned}$$

2) $\alpha_k = \frac{1}{\sqrt{k+1}}, k = 0, 1, ...$

$$\sum_{k=0}^{N-1} \alpha_k^2 = \sum_{k=1}^{N} \frac{1}{k} < \log N + 1$$

$$\sum_{k=0}^{N-1} \alpha_k = \sum_{k=1}^{N} \frac{1}{\sqrt{k}} > 2\sqrt{N} - 2$$

$$\Rightarrow (f_N^* - f^*) \sim O(\frac{\log N}{\sqrt{N}})$$

Both worse than gradient descent (GD) $O(\frac{1}{N})$.

Chapter 5 (Unconstrained Optimization) Newton's Method

5.1 Classical Newton's Method

Definition 5.1 (Classical Newton's Method)

One dimensional:

Finding solution to non-linear equation:

$$g(x^*) = 0$$

with $g : \mathbb{R} \to \mathbb{R}$*. Given* x_k*, find* x_{k+1} *to solve* x^**.*

$$0 = g(x_{k+1}) \approx g(x_k) + g'(x_k)(x_{k+1} - x_k)$$

Assuming $g'(x_k) \neq 0$*, set*

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}$$

♣

Similarly, when consider the minimizing problem of $f : \mathbb{R}^n \to \mathbb{R}$. The Newton's method is

$$\vec{x}_{k+1} = \vec{x}_k - \nabla g(\vec{x}_k)^{-1} g(\vec{x}_k)$$

5.2 Variants of Newton's Method

5.2.1 Generalization to Optimization

In optimization, the goal is to get to x s.t. $\nabla f(x) = 0$.

Given x_k, we want to find x_{k+1} s.t. $\nabla f(x_{k+1}) = 0$.

Taylor's Approx:

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

Set

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

which can be viewed as GD with $\alpha_k = 1$ and $d_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$

If $\nabla^2 f(x_k) \succeq 0$, then $\nabla f(x_k)^T d_k \geq 0$.

5.2.2 The Secant Method

The secant method is a replacement for ordinary Newton's method when we can't compute derivatives. In this case, we replace the derivative $f'(x_k)$ by the approximation $\frac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}}$ using the values of f at the *two* most recent points.

This gives us the iterative step

$$x_{k+1} = x_k - \frac{f\left(x_k\right)}{\frac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}}} = x_k - \frac{f\left(x_k\right)}{f\left(x_k\right)-f\left(x_{k-1}\right)}\left(x_k - x_{k-1}\right)$$

which involves no derivatives. That's the secant method.

As we approach a solution x^*, x_k and x_{k-1} become very close to each other, and therefore $\frac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}}$ is a very good approximation of $f'\left(x_k\right)$ (or of $f'\left(x_{k-1}\right)$). So we expect this method to perform almost as well as Newton's method (in cases where Newton's method does perform well).

We lose some efficiency, of course, because the approximation is not perfect. But the advantage is that we don't have to compute derivatives, so we can deal with a wider class of problems.

5.3 A New Interpretation of Newton's Method

Since $f(x) \approx f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$, at each step k, we can solve a quadratic minimization problem,

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^p} \{f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)\}$$

5.4 Convergence of Newton's Method

5.4.1 Guarantees of Convergence

Theorem 5.1

Suppose that we are using Newton's method to solve $g(\vec{x}) = \vec{0}$ for some $g : \mathbb{R}^n \to \mathbb{R}^n$ (with continuous first partial derivatives). If the sequence of points $\vec{x}_1, \vec{x}_2, \ldots$ that we get converges to some limit $\vec{x}^*$, then $g(\vec{x}^*) = \vec{0}$. ♡

5.4.2 Convergence Rate

Let x^* be s.t. $\nabla f(x^*) = 0$, then

$$\begin{aligned}\|x_{k+1} - x^*\| &= \|x_k - x^* - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)\| \\ &= \|x_k - x^* - (\nabla^2 f(x_k))^{-1}(\nabla f(x_k) - \nabla f(x^*))\|\end{aligned}$$

By Taylor's theorem,

$$\nabla f(x_k) = \nabla f(x^*) + \nabla^2 f(x^* + \beta(x_k - x^*))(x_k - x^*) \text{ for some } \beta \in [0,1]$$

Thus,

$$\begin{aligned}
\|x_{k+1} - x^*\| &= \|x_k - x^* - (\nabla^2 f(x_k))^{-1} \nabla^2 f(x^* + \beta(x_k - x^*))(x_k - x^*)\| \\
&= \|(\nabla^2 f(x_k))^{-1} (\nabla^2 f(x^* + \beta(x_k - x^*)) - \nabla^2 f(x_k))(x_k - x^*)\| \\
&\leq \|(\nabla^2 f(x_k))^{-1}\| \|\nabla^2 f(x^* + \beta(x_k - x^*)) - \nabla^2 f(x_k)\| \|x_k - x^*\|
\end{aligned}$$

We use 1-norm $\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$ here, $\|A\| \geq \frac{\|Ax\|}{\|x\|} \Rightarrow \|Ax\| \leq \|A\|\|x\|$.

Easy to prove, for symmetric $A \succeq 0$, $\|A\| = \lambda_{\max}(A)$, $\|A^{-1}\| = \lambda_{\max}(A^{-1}) = \lambda_{\min}^{-1}(A)$

- Now suppose f **is loacl** m**-strongly convex near** x^*, then

$$\begin{aligned}
\nabla^2 f(x^*) &\succeq mI \text{ with } m > 0 \\
\Rightarrow \lambda_{\min}(\nabla^2 f(x^*)) &\geq m > 0 \\
\Rightarrow \lambda_{\min}^{-1}(\nabla^2 f(x^*)) &\leq \frac{1}{m}
\end{aligned}$$

- When f is not local strongly convex near x^*. Assuming $\nabla^2 f(x)$ is continuous, if $\|x_k - x^*\|$ is small, then $\lambda_{\min}(\nabla^2 f(x_k))$ is close to $\lambda_{\min}(\nabla^2 f(x^*))$ i.e $\lambda_{\min}(\nabla^2 f(x^*))$ should be greater than a constant $\lambda_{\min}(\nabla^2 f(x^*)) \geq \bar{\gamma} > 0$. Then,

$$\|\nabla^2 f(x_k)^{-1}\| = \lambda_{\min}^{-1}(\nabla^2 f(x_k)) \leq \frac{1}{\bar{\gamma}} = \gamma$$

Furthurmore, assume that $\nabla^2 f$ **is L-Lipschitz in a neighborhood** $\&$ **of** x^*, i.e.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\| \quad \forall x, y \in \&$$

Thus,

$$\begin{aligned}
\|x_{k+1} - x^*\| &\leq \|(\nabla^2 f(x_k))^{-1}\| \|\nabla^2 f(x^* + \beta(x_k - x^*)) - \nabla^2 f(x_k)\| \|x_k - x^*\| \\
&\leq \gamma L \|x^* + \beta(x_k - x^*) - x_k\| \|x_k - x^*\| \\
&\leq \gamma L \|(\beta - 1)(x_k - x^*)\| \|x_k - x^*\| \\
\text{(Since } \beta \in [0,1]) \quad &\leq \gamma L \|x_k - x^*\|^2
\end{aligned}$$

Hence,

$$\|x_{k+1} - x^*\| \leq \gamma L \|x_k - x^*\|^2$$

Now suppose x_0 **is close enough to** x^* s.t.

$$\gamma L \|x_0 - x^*\| = \sigma < 1$$

Then,

$$
\begin{aligned}
\|x_1 - x^*\| &\leq \sigma \|x_0 - x^*\| \\
\|x_2 - x^*\| &\leq \gamma L \|x_1 - x^*\|^2 \\
&\leq \gamma L \sigma^2 \|x_0 - x^*\|^2 = \sigma^3 \|x_0 - x^*\| \\
\|x_3 - x^*\| &\leq \gamma L \|x_2 - x^*\|^2 \\
&\leq \gamma L \sigma^6 \|x_0 - x^*\|^2 = \sigma^7 \|x_0 - x^*\| \\
&\dots \\
\|x_N - x^*\| &\leq \sigma^{2^N - 1} \|x_0 - x^*\|
\end{aligned}
$$

Assuming **∇f is M-Lipschitz in neighborhood of x^***,

$$
\begin{aligned}
f(x_N) - f(x^*) &\leq \nabla f(x^*)(x_N - x^*) + \frac{M}{2}\|x_N - x^*\|^2 \\
&\leq \frac{M}{2}\sigma^{(2^{N+1}-2)}\|x_N - x^*\|^2
\end{aligned}
$$

Thus to make $f(x_N) - f(x^*) < \varepsilon$, need $N \sim O(log(log(\frac{1}{\varepsilon})))$

We call it **order-2 or super-linear convergence**.

5.5 Note: Cons and Pros

- Newton's Method is super-fast close to local min if function strongly convex around min.
- If the function is quadratic, Newton's method converges in one step.

  $$
  \begin{aligned}
  f(x) &= \frac{1}{2}x^T Q x + bx + c, \quad Q \succ 0. \\
  \nabla f(x) &= Qx + b, \nabla^2 f(x) = Q. \\
  \text{Global min } x^* \text{ satisfies } Qx^* + b &= 0 \Rightarrow x^* = -Q^{-1}b
  \end{aligned}
  $$

  Newton's method: for any $x_0 \in \mathbb{R}^n$,

  $$
  \begin{aligned}
  x_1 &= x_0 - (\nabla^2 f(x_0))^{-1}\nabla f(x_0) \\
  &= x_0 - Q^{-1}(Qx_0 + b) = -Q^{-1}b = x^*
  \end{aligned}
  $$

  Intuition: when f is a quadratic function, $\nabla^3 f(x) = 0, \forall x$. Hence, $f(x) = f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$, the minimization problem will get the min in one step.
- But Newton's method has several **drawbacks**:
  (1) Newton's method requires the matrix inversion step, and this is quite expensive. So the *per step cost for Newton's method is higher*.
  (2) Newton's method has faster local convergence but may diverge if initialized from some place far from the optimal point.

(3) $\nabla^2 f(x)^{-1}$ may fail to exist, i.e. $\nabla^2 f(x)$ is singular, e.g. linear f.

(4) It is not necessarily a general GD method since $\nabla^2 f(x_k)$ may not be $\succ 0$.

(5) It is not a descent method, $f(x_{k+1})$ may be $> f(x_k)$.

(6) It may stop at local max or saddle points.

5.6 Modifications to ensure global convergence

(a) Try Newton's method. If either $\nabla^2 f(x_k)$ is singular or $f(x_{k+1}) > f(x_k)$ then use (b).

(b) Find δ_k s.t.

$$(\delta_k I + \nabla^2 f(x_k)) \succ 0$$

and

$$\lambda_{\min}(\delta_k I + \nabla^2 f(x_k)) \succeq \Delta > 0$$

so that $\delta_k I + \nabla^2 f(x_k)$ is easily invertible.

Then set $d_k = -(\delta_k I + \nabla^2 f(x_k))^{-1} \nabla f(x_k)$. This ensures that $\nabla^T f(x_k) d_k < 0$.

Then we use $x_{k+1} = x_k + \alpha_k d_k$ with α_k chosen using Armijo's Rule.

If at any point $\nabla^2 f(x_k) \succ 0$, go back to Newton's method and check if $f(x_{k+1}) < f(x_k)$. Continue Newton's method as long as $\nabla^2 f(x_k) \succ 0$ and $f(x_{k+1}) < f(x_k)$.

5.7 Quasi-Newton Methods

Estimating Hessian $\nabla^2 f(x_k)$ is expensive, so we use some simplier matrix H_k instead.

Quasi-Newton method have the iteration form:

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

where H_k is some estimated version of $\nabla^2 f(x_k)$, and the stepsize α_k is typically determined by Armijo rule.

Previously, we approximate $f(x)$ by

$$f(x) \approx f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$$

Now, we define the form by H_k

$$g(x) = f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T H_k (x - x_k)$$

We hope $g(x) \approx f(x)$ and optimize g for this step. We enforce

(1) $\nabla f(x_k) = \nabla g(x_k)$ (Automatically satisfied)

(2) $\nabla f(x_{k-1}) = \nabla g(x_{k-1}) \quad \Leftrightarrow$

$$H_k(x_k - x_{k-1}) = \nabla f(x_k) - \nabla f(x_{k-1})$$

The condition (2) is called the secant equation.

There are infinitely many H_k satisfying this condition. Various choices of H_k lead to different Quasi-Newton methods. We discuss the BFGS method.

5.7.1 BFGS Method

We need H_k to be constructed in a way that it can be efficiently computed.

We want H_k to have two properties:

(1) H_k can be computed by some iterative formula

$$H_k = H_{k-1} + M_{k-1}$$

(2) H_k is positive definite (at least guarantee that the BFGS method is a descent method, i.e. $f(x_{k+1}) \leq f(x_k)$).

We can choose $H_0 > 0$ and then guarantee $M_k \geq 0$.

Rank-2 BFGS Method:

$$H_{k+1} = H_k + a_k v_k v_k^T + b_k u_k u_k^T$$

where $v_k \in \mathbb{R}^p$ and $u_k \in \mathbb{R}^p$ are some vectors. If $H_0 > 0$, the above iterative formula can guarantee H_k to be positive definite.

How can we choose v_k and u_k to guarantee the secant equation $H_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k)$?

Let's denote $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$. The secant equation: $H_{k+1}s_k = y_k$, then substitute it into the above formula,

$$y_k = H_{k+1}s_k = H_k s_k + a_k v_k v_k^T s_k + b_k u_k u_k^T s_k$$

$$\Leftrightarrow y_k - H_k s_k = a_k(v_k^T s_k)v_k + b_k(u_k^T s_k)u_k$$

To let the above equation be satisfied. We let $v_k = y_k$, $u_k = H_k s_k$, $a_k = \frac{1}{y_k^T s_k}$, and $b_k = -\frac{1}{s_k^T H_k s_k}$. Then, the iteration formula becomes

$$H_{k+1} = H_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{H_k s_k s_k^T H_k}{s_k^T H_k s_k}$$

where $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.

This is exactly the BFGS method.

Since we implement the BFGS method as

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

It will be better to compute H_k^{-1} directly instead of H_k.

$$\begin{aligned}
H_{k+1}^{-1} &= \left(H_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{H_k s_k s_k^T H_k}{s_k^T H_k s_k}\right)^{-1} \\
&= \left(H_k + [H_k s_k \ y_k] \begin{bmatrix} -\frac{1}{s_k^T H_k s_k} & 0 \\ 0 & \frac{1}{y_k^T s_k} \end{bmatrix} \begin{bmatrix} s_k^T H_k \\ y_k^T \end{bmatrix}\right)^{-1}
\end{aligned}$$

(by woodbury formula)

$$\begin{aligned}
&= H_k^{-1} - H_k^{-1}[H_k s_k \ y_k] \left(\begin{bmatrix} -\frac{1}{s_k^T H_k s_k} & 0 \\ 0 & \frac{1}{y_k^T s_k} \end{bmatrix}^{-1} + \begin{bmatrix} s_k^T H_k \\ y_k^T \end{bmatrix} H_k^{-1}[H_k s_k \ y_k] \right)^{-1} \begin{bmatrix} s_k^T H_k \\ y_k^T \end{bmatrix} H_k^{-1} \\
&= H_k^{-1} - [s_k \ H_k^{-1} y_k] \begin{bmatrix} 0 & s_k^T y_k \\ y_k^T s_k & y_k^T (s_k + H_k^{-1} y_k) \end{bmatrix}^{-1} \begin{bmatrix} s_k^T \\ y_k^T H_k^{-1} \end{bmatrix} \\
&= H_k^{-1} - [s_k \ H_k^{-1} y_k] \begin{bmatrix} -\frac{y_k^T s_k + y_k^T H_k^{-1} y_k}{y_k^T s_k s_k^T y_k} & \frac{1}{y_k^T s_k} \\ \frac{1}{y_k^T s_k} & 0 \end{bmatrix} \begin{bmatrix} s_k^T \\ y_k^T H_k^{-1} \end{bmatrix} \\
&= H_k^{-1} - \frac{H_k^{-1} y_k s_k^T}{y_k^T s_K} - \frac{s_k y_k^T H_k^{-1}}{y_k^T s_K} + \frac{s_k s_k^T}{y_k^T s_K} + \frac{s_k y_k^T H_k^{-1} y_k s_k^T}{(y_k^T s_k)^2} \\
&= \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) H_k^{-1} \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}
\end{aligned}$$

$$H_{k+1}^{-1} = \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) H_k^{-1} \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}$$

is the iteration computation H_k^{-1} of BFGS method. where $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.

5.8 Trust-Region Method

$$x_{k+1} = \underset{\|x - x_k\| \le \Delta_k}{\operatorname{argmin}} \left\{ f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k) \right\}$$

This method can escape addle points under some assumptions.

5.9 Cubic Regularization

Contain higher order term $\|x - x_k\|^3$ to the quadratic estimation.

Chapter 6 (Constrained Optimization) Barrier Method

6.1 Barrier Method

Computationed method to solve inequality constrained problems.

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & x \in \& \\ & g(x) \leq 0 \end{aligned}$$

where $\&$ is closed set.

Barrier Function

$B(x)$ is a function that is continuous and $\to \infty$ as any $g_j(x) \to 0$

Example 6.1

$$B(x) = -\sum_{j=1}^{r} \ln(-g_j(x))$$
$$B(x) = -\sum_{j=1}^{r} \frac{1}{g_j(x)}$$

Note: that if $g_j(x)$ is convex for all j, then both of these barrier functions are convex.

In Barrier Method, choose sequence $\{\varepsilon_k\}$ s.t.

$$0 < \varepsilon_{k+1} < \varepsilon_k, \quad k = 0, 1, ...$$

and $\varepsilon_k \to 0$ as $k \to \infty$.

Define feasible set $F = \& \cap \{g_j(x) \leq 0, \forall j\}$. Note F is a closed set since $\&$ and $\{g_j(x) \leq 0, \forall j\}$ are closed.

Let $\vec{x}^{(k)}$ be a solution to

$$\min_{x \in F \cap \text{dom}(B)} f(x) + \varepsilon_k B(x)$$

Since $B(x) \to \infty$ as one $g_j(x) \to 0$ which is on the boundary of F.

$\vec{x}^{(k)}$ must be an interior point of F

$$\Rightarrow \nabla f(\vec{x}^{(k)}) + \varepsilon_k \nabla B(\vec{x}^{(k)}) = 0$$

Therefore, if we have a initial point in the interior of F, we can choose step size of any unconstrained GD method to stay in interior of F for all iterations and solve the ICP. (Because barrier function $B(x)$ will prevent us from reaching boundary)

As $k \to \infty$, $\varepsilon_k \to 0$, and barrier $\varepsilon_k B(x)$ becomes inconsequential, and we expect $\vec{x}^{(k)}$ to approcah minimum of original problem.

Proposition 6.1

Every limit point $\bar{x}$ of $\{\vec{x}^{(k)}\}$ is a global min of the ICP.

Proof 6.1

Let $\bar{x} = \lim_{k\to\infty, k\in\mathcal{K}} \vec{x}^{(k)}$, since $\vec{x}^{(k)} \in F$ forall k, and F is closed, $\bar{x} \in F$.

Suppose x^* is a global min of ICP and x^* is in interior of F, and $f(x^*) < f(\bar{x})$, i.e., $\bar{x}$ is not global min for ICP.

Then, by definition of $\vec{x}^{(k)}$, $f(\vec{x}^{(k)}) + \varepsilon_k B(\vec{x}^{(k)}) \leq f(x^*) + \varepsilon_k B(x^*)$

Taking limit as $k \to \infty$, $k \in \mathcal{K}$,

$$f(\bar{x}) + \lim_{k\to\infty, k\in\mathcal{K}} \varepsilon_k B(\bar{x}) \leq f(x^*) + \lim_{k\to\infty, k\in\mathcal{K}} \varepsilon_k B(x^*) = f(x^*)$$

$$(\textit{Since } |B(x^*)| < \infty, \varepsilon_k \to 0 \textit{ as } k \to \infty)$$

If $\bar{x}$ is in interior of F, then $|B(\bar{x})| < \infty \Rightarrow \lim_{k\to\infty, k\in\mathcal{K}} \varepsilon_k B(\vec{x}^{(k)}) = 0$

If $\bar{x}$ is on boundary of F, then $|B(\bar{x})| \to \infty \Rightarrow \lim_{k\to\infty, k\in\mathcal{K}} \varepsilon_k B(\vec{x}^{(k)}) \geq 0$

Therefore, $f(\bar{x}) < f(x^*)$ is contradiced.

If x^* is not in interior of F, we can assume that $\exists$ an interior point $\bar{x}$ which can be made arbitrarily close to x^*.

6.2 An Exmaple Using KKT or Barrier

Example 6.2

$$\min \quad f(x) = \frac{1}{2}(x_1^2 + x_2^2)$$

$$s.t. \; x_1 \geq 2$$

6.2.1 Solution using KKT conditions

$$\begin{aligned} g(x) &= -x_1 + 2 \\ \nabla g(x) &= (-1, 0) \quad \text{All feasible } x \text{ are regular} \\ \nabla f(x) &= (x_1, x_2) \\ L(x, \mu) &= f(x) + \mu g(x) \\ \nabla L(x, \mu) &= \nabla f(x) + \mu \nabla g(x) = (x_1 - \mu, x_2) \end{aligned}$$

<u>Case</u> 1: constraint inactive, i.e., $\mu = 0$

$$\nabla L(x, \mu) = 0 \Rightarrow x = (0, 0)$$

Doesn't satisfy $x_1 \geq 2$. This case is infeasible.

<u>Case</u> 2: constraint active,

$$\begin{aligned} \nabla L(x, \mu) = 0 &\Rightarrow x_1 - \mu = 0, x_2 = 0 \\ g(x) = 0 &\Rightarrow x_1 = 2 \\ &\Rightarrow x^* = (2, 0), \mu = 2 \end{aligned}$$

It satisfies the first-order KKT condition.

Since $L(x, \mu)$ is strictly convex on $\mathbb{R}^2$, $x^* = (2, 0)$ is the global-min.

6.2.2 Solution using logarithmic barrier

$B(x) = -\ln(-g(x)) = -\ln(x_1 - 2)$

$$\begin{aligned} \text{Set } G^{(k)}(x) &= f(x) + \varepsilon_k B(x) \\ &= \frac{1}{2}(x_1^2 + x_2^2) - \varepsilon_k \ln(x_1 - 2) \\ (G^{(k)}(x) &\text{ is convex in } x \text{ over } \{x : x > 2\}) \\ \nabla G^{(k)}(x) = 0 \Rightarrow \ & x_1 - \frac{\varepsilon_k}{x_1 - 2} = 0, \ x_2 = 0 \\ \Rightarrow \ & \vec{x}^{(k)} = (1 + \sqrt{1 + \varepsilon_k}, 0) \end{aligned}$$

$$\text{as } k \to \infty, \varepsilon_k \to 0 \text{ and } \vec{x}^{(k)} \to (2, 0) = x^*$$

6.3 Penalty Method (For ECP)

Computational method for equality constraints.

$$\begin{aligned} &\min \quad f(x) \\ &s.t.\ x \in \& \\ &\qquad h_i(x) = 0, \quad i = 1, ..., m \end{aligned}$$

Algorithm

(1) Choose an increasing positive sequence $\{c_k\}$ s.t. $c_k \to \infty$ as $k \to \infty$.

(2) Solve for $\vec{x}^{(k)}$ to:

$$\min_{x \in \&} \quad f(x) + c_k \|h(x)\|^2$$

$$\text{Note: } \|h(x)\|^2 = \sum_{i=1}^{m} (h_i(x))^2$$

Proposition 6.2

Every limit point $\bar{x}$ of $\{\vec{x}^{(k)}\}$ is a global min of the ECP if $\&$ is closed.

Proof 6.2

Let $\bar{x} = \lim_{k \to \infty, k \in \mathcal{K}} \vec{x}^{(k)}$

$$\begin{aligned} f^* = \min_{x \in \&, h(x)=0} f(x) &= \min_{x \in \&, h(x)=0} f(x) + c_k \|h(x)\|^2 \\ &\geq \min_{x \in \&} f(x) + c_k \|h(x)\|^2 \\ &= f(\vec{x}^{(k)}) + c_k \|h(\vec{x}^{(k)})\|^2 \end{aligned}$$

$$\Rightarrow c_k \|h(\vec{x}^{(k)})\|^2 \leq f^* - f(\vec{x}^{(k)})$$

By continuity of f, $\lim_{k \to \infty, k \in \mathcal{K}} f(\vec{x}^{(k)}) = f(\bar{x})$.

Thus, as $k \to \infty$, $k \to \mathcal{K}$, $f^* - f(\vec{x}^{(k)}) = f^* - f(\bar{x})$ which is finite.

Since $c_k \to \infty$ as $k \to \infty$, $k \to \mathcal{K}$,

$$\lim_{k \to \infty, k \in \mathcal{K}} \|h(\vec{x}^{(k)})\|^2 = 0$$

By continunity of h,

$$\lim_{k \to \infty, k \in \mathcal{K}} \|h(\vec{x}^{(k)})\|^2 = \|h(\bar{x})\|^2 = 0 \Rightarrow h(\bar{x}) = 0$$

Now, since $\&$ is closed, and $\vec{x}^{(k)} \in \&$ for all k, $\bar{x} \in \&$ as well.

$$
\begin{aligned}
&f^* - f(\vec{x}^{(k)}) \geq c_k \|h(\vec{x}^{(k)})\|^2 \geq 0 \\
\Rightarrow &f(\bar{x}) = \lim_{k \to \infty, k \in \mathcal{K}} f(\vec{x}^{(k)}) \leq f^*
\end{aligned}
$$

Since $\bar{x}$ *is feasible (*$\bar{x} \in \&$ *and* $h(\bar{x}) = 0$*) and* $f(\bar{x}) \leq f^*$*,* $\Rightarrow \bar{x}$ *is a global min of the ECP.*

Chapter 7 Descent Method

We're going to assume that minimizing a single-variable function is easy (after all, you just have to decide to go left or go right on the number line). Minimizing functions of many variables is hard. So we want a method to reduce the many-variable case to the single-variable case.

7.1 Method of Steepest Descent

7.1.1 The Method of Steepest Descent

The method of steepest descent does precisely this. At each iterative step, we

1. Pick a direction to go in.
2. Solve a one-variable problem to determine how far to go in that direction.

Let's focus on a function $f : \mathbb{R}^n \to \mathbb{R}$ with continuous gradient.

Given a unit vector $\vec{u} \in \mathbb{R}^n$, the rate of change of f at $\vec{x}$ in the direction of $\vec{u}$ is the directional derivative given by

$$\nabla f(\vec{x}) \cdot \vec{u}$$

By the Cauchy-Schwarz inequality, we have an upper bound:

$$|\nabla f(\vec{x}) \cdot \vec{u}| \leq \|\nabla f(x)\| \cdot \|\vec{u}\|$$

the upper bound is achieved when $\vec{u}$ is parallel to $\nabla f(\vec{x})$. To make the derivative as negative as possible, we set $\vec{u} = -\frac{\nabla f(\vec{x})}{\|\nabla f(\vec{x})\|}$. Hence, the **steepest descent** is $-\nabla f(\vec{x})$, we implement iteration by

$$\vec{x}^{(k)} - t\nabla f(\vec{x}^{(k)})$$

Definition 7.1 (Method of Steepest Descent)

Given a point $\vec{x}^{(k)} \in \mathbb{R}$ we compute the next point $\vec{x}^{(k+1)}$ as follows:

1. *Compute $\nabla f(\vec{x}^{(k)})$.*
2. *Set $\phi_k(t) = f(\vec{x}^{(k)} - t\nabla f(\vec{x}^{(k)}))$. That is, ϕ_k evaluates f along the line through $\vec{x}^{(k)}$ in the direction of steepest descent.*
3. *Let t_k be the global minimizer of $\phi_k(t)$. This t_k tells us how far along the line we want to go.*
4. *Go that far along the line: set*

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - t_k\nabla f(\vec{x}^{(k)})$$

♣

7.1.2 Properties of steepest descent

Theorem 7.1

If $\vec{x}^{(k)}, \vec{x}^{(k+1)}, \vec{x}^{(k+2)}$ are consecutive iterations of steepest descent, then

$$(\vec{x}^{(k+1)} - \vec{x}^{(k)}) \cdot (\vec{x}^{(k+2)} - \vec{x}^{(k+1)}) = 0$$

♡

Proof 7.1

$\phi_k'(t) = -\nabla f(\vec{x}^{(k)}) \cdot \nabla f(\vec{x}^{(k)} - t\nabla f(\vec{x}^{(k)}))$. *Let* $t = t_k$, $\phi_k'(t_k) = -\nabla f(\vec{x}^{(k)}) \cdot \nabla f(\vec{x}^{(k)} - t_k\nabla f(\vec{x}^{(k)})) = 0$. *Hence,* $(\vec{x}^{(k+1)} - \vec{x}^{(k)}) \cdot (\vec{x}^{(k+2)} - \vec{x}^{(k+1)}) = t_k t_{k+1} \nabla f(\vec{x}^{(k)}) \cdot \nabla f(\vec{x}^{(k+1)}) = t_k t_{k+1} \nabla f(\vec{x}^{(k)}) \cdot \nabla f(\vec{x}^{(k)} - t_k \nabla f(\vec{x}^{(k)})) = 0$

Intuitively, once we pick the direction of steepest descent, we move in that direction for as long as this benefits us: until f starts increasing again. This means that when we stop, the direction we were moving at is useless to us, and the new direction of steepest descent should be orthogonal to the previous one.

Theorem 7.2

The sequence $f(\vec{x}^{(0)}), f(\vec{x}^{(1)}), f(\vec{x}^{(2)}), \cdots$ is strictly decreasing until/unless it reaches a critical point. ♡

Theorem 7.3

If f is ***coercive*** *and has* ***a unique critical point*** *(which must then be a global minimizer) then the method of steepest descent finds it.*

♡

7.2 General Descent Method

7.2.1 Criteria for a descent method

To minimize f, we want:

1. Pick a direction to go in where f is decreasing.
2. Go in that direction far enough to notice, but not too far.

Our general iteration step from a point $\vec{x}^{(k)}$ is to go to a point

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} + t_k \vec{p}^{(k)}$$

Claim 7.1 (Criteria 1)

Criteria 1: *$f(\vec{x}^{(k+1)}) < f(\vec{x}^{(k)})$, equivalent to $\phi_k(t_k) < \phi_k(0)$.*

♠

Claim 7.2 (Criteria 2)

Criteria 2: *$\vec{p}^{(k)} \cdot \nabla f(\vec{x}^{(k)}) < 0$, equivalent to $\phi_k'(0) < 0$. (restrict the decent to the direction that decreases the function)*

Let $\phi_k(t) = f(\vec{x}^{(k)} + t\vec{p}^{(k)})$, $\phi_k'(t) = \vec{p}^{(k)} \cdot \nabla f(\vec{x}^{(k)} + t\vec{p}^{(k)})$ the Criteria 2 asks $\phi_k'(0) < 0$

Claim 7.3 (Criteria 3)

Criteria 3: *For some parameter $\beta \in (0,1)$, we have $\phi_k'(t_k) \geq \beta\phi_k'(0)$. ("don't take steps that are too small".)*

If t_k is close to 0, $\phi_k'(t_k)$ will be close $\beta\phi_k'(0)$ with large negative. We don't want to waste a step at a t_k close to 0.

The value of β is a parameter that measures how much we care about this criterion.

If we set $\beta = 0$, the method goes back to the method of the steepest descent. This criterion would tell us to keep going until we reach (or go past) a critical point of ϕ_k.

If we set $\beta = 1$, the criterion doesn't exist.

Claim 7.4 (Criteria 4)

Criteria 4: *For some parameter $\alpha \in (0, \beta)$, we have $\phi_k(t_k) \leq \phi_k(0) + \alpha t_k \phi_k'(0)$. ("only keep going for as long as you're getting at least a fraction of the promised payoff".)*

With Criteria 2 and 4, we can forget about Criteria 1 (which is a consequence of Criteria 2 and 4).

7.2.2 Wolfe's theorem

Theorem 7.4 (Wolfe's theorem)

Let α, β be real numbers satisfying $0 < \alpha < \beta < 1$. Assume $f : \mathbb{R}^n \to \mathbb{R}$ has continuous ∇f and is bounded below.

Starting from a point $\vec{x}^{(k)}$, whenever $p^{(k)}$ satisfies Criterion 2 (it is a descent direction), then there is a range $[a_k, b_k]$ (with $0 < a_k < b_k$) such that for any $t_k \in [a_k, b_k]$, all four criteria are satisfied by picking direction $\vec{p}^{(k)}$ and step size $t_k \in [a_k, b_k]$.

7.2.3 Picking step sizes

One approach that we can use to find an acceptable t_k is:

(1) Make an initial guess. (Maybe $t_k = 1$)

(2) Check if this guess satisfies *Criterion 4*, i.e., if

$$\phi_k(t_k) \leq \phi_k(0) + \alpha t_k \phi_k'(0)$$

(We know this should be satisfied for all ”sufficiently small” t_k.)

- If *Criterion 4* is not satisfied, make t_k smaller and begin step (2) again.
- If *Criterion 4* is satisfied, check if *Criterion 3* is also satisfied, i.e., if

$$\phi_k'(t_k) \geq \beta \phi_k'(0)$$

- If *Criterion 3* is not satisfied, we choose a larger t_k and begin (2) again.

Wolfe’s theorem guarantees a range of acceptable t_k values.

Example 7.1 (One helpful trick) Once we’ve found a value of t_k that is too small, t_s, and a value that is too large t_l, we know

$$[a_k, b_k] \subseteq [t_s, t_e]$$

and we can start just cutting this interval in half, by repeating the following steps:

(1) Guess $t_a = \frac{t_s + t_l}{2}$

(2) If t_a is too large, set $t_l = t_a$ and restart.

(3) If t_a is too small, set $t_s = t_a$ and restart.

(4) If t_a works, we’re done.

7.2.4 Picking descent directions

The only requirement of finding $\vec{p}^{(k)}$ is *Criterion 2*:

$$\vec{p}^{(k)} \cdot \nabla f(\vec{x}^{(k)}) < 0$$

Some ideas/strategies for choosing decent directions:

(1) For any positive definite matrix Q, choosing $\vec{p}^{(k)} = -Q\nabla f(\vec{x}^{(k)})$ will work.

$$\begin{aligned} \nabla f(\vec{x}^{(k)}) \cdot \vec{p}^{(k)} &= \nabla f(\vec{x}^{(k)}) \cdot \left(-Q\nabla f(\vec{x}^{(k)})\right) \\ &= -\nabla f(\vec{x}^{(k)})^T Q \nabla f(\vec{x}^{(k)}) < 0 \end{aligned}$$

Equivalently, we could choose $\vec{p}^{(k)} = -Q^{-1}\nabla f(\vec{x}^{(k)})$. $(Q \succ 0 \Leftrightarrow Q^{-1} \succ 0)$

(2) Let’s look back at another source of inspiration: Newton’s method for minimization. Here, our iterative step is

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \nabla^2 f(\vec{x}^{(k)})^{-1} \nabla f(\vec{x}^{(k)})$$

which is a decent method when $\nabla^2 f(\vec{x}^{(k)})$ is positive definite. Hence, the form will work when f is strictly convex function.

To solve this problem when $\nabla^2 f(\vec{x}^{(k)})$ is not positive definite, we can try to ”fix” the problem by choosing a slightly different Q.

Suppose that the eigenvalues of $\nabla^2 f(\vec{x}^{(k)})$ are $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Then the eigenvalues of $\nabla^2 f(\vec{x}^{(k)}) + \mu I$ are

$$\lambda_1 + \mu \leq \lambda_2 + \mu \leq \cdots \leq \lambda_n + \mu$$

Hence, we can pick $\mu > -\lambda_1$, then $\nabla^2 f(\vec{x}^{(k)}) + \mu I$ will be positive definite, and we can set $\vec{p}^{(k)} = -Q^{-1}\nabla f(\vec{x}^{(k)})$, where $Q = \nabla^2 f(\vec{x}^{(k)}) + \mu I$.

7.2.5 Methods Summary

Altogether, here is the way an iterative step of a descent method might go.

1. Compute $\nabla^2 f(\vec{x}^{(k)})$, and find some μ_k such that $\nabla^2 f(\vec{x}^{(k)}) + \mu_k I \succ 0$.
2. Choose a descent direction $\mathbf{p}^{(k)}$ by solving
$$\left(\nabla^2 f(\vec{x}^{(k)}) + \mu_k I\right) \vec{p}^{(k)} = -\nabla f(\vec{x}^{(k)})$$
3. Choose a step size t_k by trying $t_k = 1$ and then increasing or decreasing t_k until criteria 3 and 4 are both satisfied.
4. Set the next point to $\vec{x}^{(k+1)} = \vec{x}^{(k)} + t_k \vec{p}^{(k)}$.

Chapter 8 Non-differentiable or non-explicitly function: Broyden's Method

Previously, we assumed that computing the gradient $\nabla f(\vec{x})$ or the Hessian matrix $Hf(\vec{x}) = \nabla^2 f(\vec{x})$ was easy and costless. But in practice, there are many situations where that's not true:

1. If $f : \mathbb{R}^n \to \mathbb{R}$ with n large, then computing $Hf(\vec{x})$ requires finding many derivatives, which is expensive even when we can take partial derivatives of f easily.
2. If f is not given to us explicitly, then we cannot compute derivatives of f directly (unless we use some approximation algorithm for derivatives).

In one dimension, we discussed the secant method as a solution to the second of these problems. Here, we used the values at the last two points to estimate the derivative.

Now let's pass to the n-dimensional case: finding the solution to a system of equations $\vec{g}(\vec{x}) = \mathbf{0}$ for a function $\vec{g} : \mathbb{R}^n \to \mathbb{R}^n$. To approximate Newton's method, we need an estimate of the Jacobian $\nabla\vec{g}$; in other words, a linear approximation of $\vec{g}$.

We encounter a problem: just the values of $\vec{g}$ at the last two points are not enough to build a complete linear approximation of $\vec{g}$. (In general, such an approximation would need at least $n+1$ points to define.)

8.1 Rank one updates

Our solution to this problem is to keep around a matrix that approximates the Jacobian, and update it at every step with the new information we learn, instead of throwing it away entirely as the secant method does.

Intuitively, when comparing the new value $\vec{g}\left(\vec{x}^{(k+1)}\right)$ to the previous value $\vec{g}\left(\vec{x}^{(k)}\right)$, we can learn something about the rate of change of $\vec{g}$ in the direction of $\vec{x}^{(k+1)} - \vec{x}^{(k)}$, but nothing about the rate of change of $\vec{g}$ in other directions.

Let's build on the justification behind Newton's method to figure out how to do this. With Newton's method, when we're at a point $\vec{x}^{(k)}$, we make a linear approximation to $\vec{g}$:

$$\vec{g}(\vec{x}) \approx \vec{g}\left(\vec{x}^{(k)}\right) + \nabla\vec{g}\left(\vec{x}^{(k)}\right)\left(\vec{x} - \vec{x}^{(k)}\right).$$

Then we let $\vec{x}^{(k+1)}$ be the point where the linear approximation equals $\mathbf{0}$. Now suppose that instead of computing the Jacobian $\nabla\vec{g}\left(\vec{x}^{(k)}\right)$, we somehow found an approximation D_k : an $n \times n$ matrix that's our best guess at the partial derivatives of $\vec{g}$. Just as with Newton's method, we make a linear approximation

$$\vec{g}(\vec{x}) \approx \vec{g}(\vec{x}^{(k)}) + D_k(\vec{x} - \vec{x}^{(k)}).$$

We let $\vec{x}^{(k+1)}$ be the point where the linear approximation equals 0.

$$\vec{g}(\vec{x}^{(k)}) + D_k(\vec{x}^{(k+1)} - \vec{x}^{(k)}) = \vec{0}$$

In practice, $\vec{g}\left(\vec{x}^{(k+1)}\right)$ doesn't end up being 0: if everything is going well, it should be closer to 0 than $\vec{g}\left(\vec{x}^{(k)}\right)$ was, but the linear approximation is not exact. This is new information, and we should replace D_k by a new matrix D_{k+1} that takes this new information into account.

What properties should D_{k+1} have?

1. We would like D_{k+1} to "fix" the incorrect prediction of D_k. We now know the exact value $\vec{g}(\vec{x}^{(k+1)})$, which our previous approximation though was $\vec{0}$.

   i.e., when we replace D_k with D_{k+1} in our linear approximation, we should get

   $$\vec{g}(\vec{x}^{(k)}) + D_{k+1}(\vec{x}^{(k+1)} - \vec{x}^{(k)}) = \vec{g}(\vec{x}^{(k+1)}) \tag{1}$$

   instead of

   $$\vec{g}(\vec{x}^{(k)}) + D_k(\vec{x}^{(k+1)} - \vec{x}^{(k)}) = \vec{0} \tag{2}$$

2. We should make the same predictions in every direction orthogonal to $\vec{x}^{(k+1)} - \vec{x}^{(k)}$. Because we didn't go in those directions, we didn't learn anything new.

   So, we require that

   $$D_k\vec{y} = D_{k+1}\vec{y}, \quad \forall \vec{y}\ s.t.\ (\vec{x}^{(k+1)} - \vec{x}^{(k)}) \cdot \vec{y} = 0 \tag{3}$$

We get the D_{k+1} by finding the update matrix $u_k = D_{k+1} - D_k$.

Let $\vec{b}_k = \vec{x}^{(k+1)} - \vec{x}^{(k)}$, then $(1) - (2)$ gives:

$$\begin{aligned}(D_{k+1} - D_k)(\vec{x}^{(k+1)} - \vec{x}^{(k)}) &= \vec{g}(\vec{x}^{(k+1)}) \\ u_k\vec{b}_k &= \vec{g}(\vec{x}^{(k+1)})\end{aligned}$$

and (3) can be written as:

$$u_k\vec{y} = 0, \quad \forall \vec{y}\ s.t.\ \vec{b}_k \cdot \vec{y} = 0$$

If u_k exists, it is unique. This is because any $\vec{y}$ can be written as

$$\vec{y} = \vec{y}^{\parallel} + \vec{y}^{\perp}$$

where $\vec{y}^{\parallel}$ is the component parallel to $\vec{b}_k$, i.e., a scalar multiple of $\vec{b}_k$, and $\vec{y}^{\perp}$ is the component orthogonal to $\vec{b}_k$. The equations above define what u_k does to $\vec{y}^{\parallel}$ and $\vec{y}^{\perp}$, and therefore they determine what u_k does to y.

We can find u_k such that

$$u_k = \frac{\vec{g}(\vec{x}^{(k+1)})(\vec{b}_k)^T}{\vec{b}_k \cdot \vec{b}_k}$$

We can then compute D_{k+1} as

$$D_{k+1} = D_k + u_k$$

which is called a *rank-one update*, because the rank of the matrix u_k is 1: its columns are all multiples of $\vec{g}(\vec{x}^{(k+1)})$. (And its rows are all multiples of $(\vec{b}_k)^T$.) This makes it a "minimal" change from D_k to D_{k+1} in some sense, and using a rank-one matrix also has some theoretical benefits.

8.2 Broyden's method

Broyden's method is related to Newton's method, but avoids repeated calculation of Jacobian or Hessian matrices by using a rank one update method.

Definition 8.1 (Broyden's method)

As usual, it begins with an initial point $\vec{x}^{(0)}$ and initial matrix D_0. (IF computing derivatives of $\vec{g}(\vec{x})$ is costly but not impossible, we could set $D_0 = \nabla\vec{g}(\vec{x})$. Otherwise, we could set D_0 equal to something simple like the identity matrix.)

To compute $\vec{x}^{(k+1)}$ and D_{k+1} from $\vec{x}^{(k)}$ and D_k, we:

1. *Solve: $\vec{g}(\vec{x}^{(k)}) + D_k(\vec{x}^{(k+1)} - \vec{x}^{(k)}) = \vec{0}$ for $\vec{x}^{(k+1)}$ and set*

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - D_k^{-1}\vec{g}(\vec{x}^{(k)})$$

2. *Set*

$$D_{k+1} = D_k + \frac{\vec{g}(\vec{x}^{(k+1)})(\vec{b}_k)^T}{\vec{b}_k \cdot \vec{b}_k}$$

where $\vec{b}_k = \vec{x}^{(k+1)} - \vec{x}^{(k)}$ ♣

Example 8.1 Suppose we are solving the system of equations

$$\begin{cases} x + y &= 2 \\ x - y &= 0 \end{cases}$$

and decide that taking the derivatives of these linear functions is too hard for us. So we're going to pick an initial guess $(x_0, y_0) = (0, 0)$ and choose $D_0 = \mathbb{I}_{2\times 2}$. Our function is $g(x, y) = (x + y - 2, x - y)$.

1. Step 1: $g(x_0, y_0) = (-2, 0)$ we set

$$(x_1, y_1) = (0, 0) - (-2, 0) = (2, 0)$$

so the step we took is $\vec{b}_0 = (2,0)$, and $g(x_1, y_1) = (0, 2)$. We can update the formula

$$D_1 = \mathbb{I}_{2\times 2} + \frac{1}{4}\begin{bmatrix} 0 \\ 2 \end{bmatrix} [2\ 0] = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

2. Step 2:

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$

so we set the step $\vec{b}_1 = (0, -2)$ and $\vec{g}(x_2, y_2) = (-2, 4)$. Then,

$$D_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} + \frac{1}{4}\begin{bmatrix} -2 \\ 4 \end{bmatrix} [0\ -2] = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Now, with information at more points, D_2 becomes the correct Jacobian matrix of the linear function $\vec{g}$. In the next step, our linear approximation to $\vec{g}$ will actually be $\vec{g}$, and so (x_3, y_3) will be the correct answer $(1, 1)$.

In general, it is not always true that after a bad guess of D_0, the matrix D_k will always approximate $\nabla\vec{g}(\vec{x})$ after many steps. It is often the case that after enough steps, the matrix D_k will accurately tell us what $\nabla\vec{g}(\vec{x})$ does in the relevant directions: the ones we actually need to use.

8.2.1 The Sherman-Morrison formula

The Sherman-Morrison formula can help us compute D_{k+1}^{-1} directly from D_k^{-1}.

Theorem 8.1 (The Sherman-Morrison formula)

If A is an $n \times n$ invertible matrix and $\vec{u}, \vec{v} \in \mathbb{R}^n$, then $(A + \vec{u}\vec{v}^T)^{-1}$ can be computed from A^{-1} as

$$(A + \vec{u}\vec{v}^T)^{-1} = A^{-1} - \frac{A^{-1}\vec{u}\vec{v}^T A^{-1}}{1 + \vec{v}^T A^{-1}\vec{u}}$$

We can use this to compute D_{k+1}^{-1} directly from D_k^{-1} if we set $A = D_k$, $\vec{u} = \vec{g}(\vec{x}^{(k+1)})$, and $\vec{v} = \frac{\vec{b}_k}{\vec{b}_k \cdot \vec{b}_k}$.

We can re-summarize Broyden's method as:

1. Set $\vec{x}^{(k+1)} = \vec{x}^{(k)} - D_k^{-1}\vec{g}(\vec{x}^{(k)})$.
2. In terms of the vectors $\vec{b}_k = \vec{x}^{(k+1)} - \vec{x}^{(k)} = -D_k^{-1}\vec{g}(\vec{x}^{(k)})$ and $\vec{c}_k = D_k^{-1}\vec{g}(\vec{x}^{(k+1)})$, set

$$D_{k+1}^{-1} = D_k^{-1} - \frac{\vec{c}_k \left((\vec{b}_k)^T D_k^{-1}\right)}{\vec{b}_k \cdot (\vec{b}_k + \vec{c}_k)}$$

Even if the evaluation of $\nabla\vec{g}(\vec{x}^{(k)})$ were costless, this method would be faster than Newton's method when the number of dimensions n is large.