University of Illinois Fall 2018

ECE 544NA: Pattern Recognition
Lecture 4: Sep. 6, Sep. 18

Lecturer: Alexander Schwing Scribe: Jiaying Wu

1 Review

1.1 Optimal Problems that we have learned

1. Linear Regression:
$$\min_w \frac{1}{2} \sum_{(x^{(i)},y^{(i)})\in D} (y^{(i)} - \phi(x^{(i)})^T w)^2$$

2. Logistic Regression:
$$\min_w \sum_{(x^{(i)},y^{(i)})\in D} log(1 + exp(-y^{(i)} w^T \phi(x^{(i)})))$$

3. Ways to Find Optimum
   - Analytic Solution: take derivative
   - Gradient Descent

2 General Optimization Problem

Example: $\min_w f_0(w)$ such that$f_i \leq 0\ \forall i \in \{1, 2 \cdots c\}$ In this example, we want to find w that can minimize $f_0(w^*)$ among all values that satisfy c different constrains.

2.1 When can we find the optimum

1. General Optimization Problem: Difficult to solve
2. Solvable Optimization Problem: Least square, linear, convex programs
   - Least Square Program: $\min_w \frac{1}{2} \sum_{(x^{(i)},y^{(i)})\in D} (y^{(i)} - \phi(x^{(i)})^T w)^2$
   - Linear Program: $\min_w c^T w$ such that $Aw \leq b$
   - Convex Program: $\min_w f_0(w)$ such that$f_i \leq 0\ \forall i \in \{1, 2 \cdots c\}$ when all f_i convex

3 Convexity

3.1 Convex Set

1. Definition: A set is convex if for any two points w_1, w_2 in the set, the line segment $w_1 + w_2(1 - \lambda)$ for $\lambda \in [0, 1]$ also lies in the set.

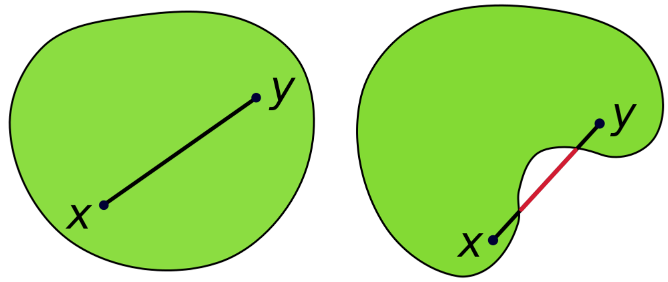

Figure 1: Convex and None-Convex Set Graph [6]

2. Special Case: Any empty set and a single point set are convex set.
3. Example: Polyhedron is convex.$P = \{w|Aw \leq b, Cw = d\}$

   Proof: (from [7])
   Given $P = \{w|Aw \leq b, Cw = d\}$, where $\lambda \in [0, 1]$ and w_1, $w_2 \in$ P

$$A((1-\lambda)w_1 + \lambda w_2) = (1-\lambda)Aw_1 + \lambda Aw_2 \leq (1-\lambda)b + \lambda b = b$$

$$C((1-\lambda)w_1 + \lambda w_2) = (1-\lambda)Cw_1 + \lambda Cw_2 \leq (1-\lambda)d + \lambda d = d$$

3.2 Convex Function

1. Definition: A function is convex if its domain is a convex set and for any point w_1, w_2 in the domain and any $\lambda \in [1, 0]$:

$$f((1-\lambda)w_1 + \lambda w_2 \leq (1-\lambda)f(w_1) + \lambda f(w_2)$$

2. Lemmas to Recognize Convex Function

   Lemma 1 *If f is differentiable, then f is convex if and only if its domain is convex and for all w_1,w_2 in the domain*

$$f(w_2) \geq f(w_1) + \triangledown f(w_1)^T(w_2 - w_1)$$

   Proof: (from [9])
   When n=1

$$\begin{aligned}
f((1-\lambda)w_1 + \lambda w_2) &\leq (1-\lambda)f(w_1) + \lambda f(w_2) \\
f(w_1 - \lambda w_1 + \lambda w_2) &\leq f(w_1) - \lambda f(w_1) + \lambda f(w_2) \\
f(w_1 + \lambda(w_2 - w_1)) &\leq f(w_1) + \lambda(f(w_2) - f(w_1)) \\
\frac{f(w_1 + \lambda(w_2 - w_1)) - f(w_1)}{\lambda} &\leq f(w_2) - f(w_1) \\
\frac{(f(w_1 + \lambda(w_2 - w_1)) - f(w_1))(w_2 - w_1)}{\lambda(w_2 - w_1)} &\leq f(w_2) - f(w_1) \\
\triangledown f(w_1)(w_2 - w_1) &\leq f(w_2) - f(w_1) \qquad (3.2.1)
\end{aligned}$$

To show sufficiency, assume the function satisfies equation 3.2.1. For all x and y in domain of f, choose any $w_1 \neq w_2$, $\lambda \in [0,1]$ and $z = \lambda w_1 + (1-\lambda)w_2$.

$$\begin{aligned} f(w_1) &\geq f(z) + \triangledown f(z)(w_1 - z) \\ \lambda f(w_1) &\geq \lambda(f(z) + \triangledown f(z)(w_1 - z) \end{aligned} \tag{3.2.2}$$

$$\begin{aligned} f(w_2) &\geq f(z) + \triangledown f(z)(w_2 - z) \\ (1-\lambda) f(w_2) &\geq (1-\lambda)(f(z) + \triangledown f(z)(w_2 - z)) \end{aligned} \tag{3.2.3}$$

Let 3.2.1+3.2.2 get $\lambda f(w_1) + (1-\lambda)f(w_2) \geq f(z)$

General Case: Consider f restricted to the line passing through them.

$$\begin{aligned} g(\lambda) &= f(\lambda w_1 + (1-\lambda)w_2) \\ \triangledown g(\lambda) &= \triangledown f(\lambda w_1 + (1-\lambda)w_2)^T(w_1 - w_2) \end{aligned}$$

Assume the f is convex, which implies g is convex, so by argument above, we have $g(1) \geq g(0) + \triangledown g(0)$, which implies:

$$f(w_1) \geq f(w_2) + \triangledown f(w_2)^T(w_1 - w_2)$$

Now assume that this inequality holds for any x and y, so if $\lambda w_1 + (1-\lambda)w_2 \in$ domain of f, and $\widetilde{\lambda} w_1 + (1-\widetilde{\lambda})w_2 \in$ domain of f, we have

$$f(\lambda w_1 + (1-\lambda)w_2) \geq f(\widetilde{\lambda} w_1 + (1-\widetilde{\lambda})w_2) + \triangledown f(\widetilde{\lambda} w_1 + (1-\widetilde{\lambda})w_2)^T(w_1 - w_2)(\lambda - \widetilde{\lambda})$$

so $g(\lambda) \leq g(\widetilde{\lambda}) + \triangledown g(\widetilde{\lambda})(\lambda - \widetilde{\lambda})$ which implies g is convex.

Lemma 2 *If f is differentiable, then f is convex if and only if its domain is convex and $\forall w_1, w_2$ in the domain*

$$(\triangledown f(w_1) - \triangledown f(w_2))^T(w_1 - w_2) \geq 0$$

Proof:
(from [1]) *If f is convex, then:*

$$f(w_1) \geq f(w_2) + \triangledown f(w_2)(w_1 - w_2) \tag{3.2.3}$$

$$f(w_2) \geq f(w_1) + \triangledown f(w_1)(w_2 - w_1) \tag{3.2.4}$$

Use 3.2.3 + 3.2.4, we get:

$$\begin{aligned} f(w_1) + f(w_2) &\geq f(w_2) + f(w_1) + \triangledown f(w_2)(w_1 - w_2) + \triangledown(w_1)(w_2 - w_1) \\ 0 &\geq (\triangledown f(w_2) - \triangledown f(w_1))^T(w_1 - w_2) \\ 0 &\leq (\triangledown f(w_1) - \triangledown f(w_2))^T(w_1 - w_2) \end{aligned}$$

Assume that $\triangledown f$ is momotone: Define $A = \{w | f(w) \leq a\}$
If A is not convex, then there are $w_1, w_2 \in A$ such that

$$\triangledown f(w_1)(w_2 - w_1) > 0$$

$$\triangledown f(w_2)(w_1 - w_2) > 0$$

Since $\triangledown f$ is monotone, sign($\triangledown f(w_1)$)=sign($\triangledown f(w_2)$)
So we have $(\triangledown f(w_1) - \triangledown f(w_2))^T(w_2 - w_1) > 0$ which has contradiction.

Lemma 3 *If f is twice differentiable, then f is convex if and only if its domain is convex and* $\nabla^2 f(w) \geq 0$ *in domain*

Proof: (from [2]) *From definition of convex, we have:*

$$f(\lambda w_1 + (1-\lambda)w_2) \leq \lambda f(w_1) + (1-\lambda) f(w_2) \forall \lambda \in [0,1]$$

Let $\lambda = \frac{1}{2}$ *and* w *be the midpoint between* w_1 *and* w_2*, so we have:*

$$w_1 = w + h, w_2 = w - h$$

$$\begin{aligned} f(\lambda w_1 + (1-\lambda)w_2) &\leq \lambda f(w_1) + (1-\lambda) f(w_2) \forall \lambda \in [0,1] \\ f(\frac{1}{2}(w+h) + \frac{1}{2}(w-h)) &\leq \frac{1}{2} f(w+h) + \frac{1}{2} f(w-h) \\ f(w) &\leq \frac{1}{2} f(w+h) + \frac{1}{2} f(w-h) \\ 2f(w) &\leq f(w+h) + f(w-h) \\ f(w+h) + f(w-h) - 2f(w) \leq 0 \end{aligned}$$

The second derivative can be written in to single limit form:

$$f \nabla^2(w) = \lim_{h \to 0} \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

Since we get $f(w+h) + f(w-h) - 2f(w) \leq 0$ *and* h^2 *is greater than 0,* $f\nabla^2(w) \geq 0$

Example of convex functions:

- Exponential function:

  $$y = e^{ax}, x \in R, \forall a \in R$$

  Proof:

  $$y\nabla^2 = a^2 e^{ax} > 0$$

  From Lemma3, we can know exponential function is convex.

- Negative Logarithm:

  $$y = -log(x), x \in R_{++}$$

  Proof:

  $$y\nabla^2 = \frac{1}{x^2} > 0$$

  From Lemma3, we can know negative logarithm function is convex.

- Negative Entropy:

  $$-H(x) = xlog(x), x \in R_{++}$$

  Proof: Let $y = xlog(x)$

  $$y\nabla^2 = \frac{1}{x} > 0, \forall x \in R_{++}$$

  From Lemma3, we can know y is convex, so negative entropy function is convex.

- Norms:

  $$y = ||w||_p, p \geq 1$$

  Proof: (from [3]) When p=1, by triangle inequality and defintion of convex

  $$||\lambda w_1 + (1-\lambda)w_2|| \leq ||\lambda w_1|| + ||(1-\lambda)w_2|| = \lambda ||w_1|| + (1-\lambda)||w_2||$$

  When p¿1, by Minkowski Inequality, triangle inequality still holds

- Log-Sum-Exp:
$$y = log(e^{w_1} + e^{w_2} + \cdots + e^{w_d})$$

  Proof: (from [5]) The Hessian of log-sum-exp function is:
$$\nabla^2 f(w) = \frac{1}{(e^T z)^2}((e^T z)diag(z) - zz^T),$$
  where $e = [1, \cdots, 1]^T$ and $z = [e^{w_1}, e^{w_2}, \cdots e^{w_n}]$. To verify that $\nabla^2 f(w) \geq 0$, we must show that $v^T \nabla^2 f(w) v \geq 0$ for all v, But
$$v^T \nabla^2 f(w) v = \frac{1}{(e^T z)^2}((\sum_{i=1}^{n} z_i)(\sum_{i=1}^{n} {v_i}^2 z_i) - (\sum_{i=1}^{n} v_i z_i)^2) \geq 0$$
  by Cauchy-Schwarz inequality.

3.3 Convex Operation

1. Non-negative weighted sum:
   for all $\alpha_i \geq 0$, if f_i is convex $\forall i$, so is
$$g = \alpha_1 f_1 + \alpha_2 f_2 + \cdots + \alpha_n f_n$$

   Proof: (from [4])
   Let $f_1, \cdots f_n$ be convex functions, $\alpha_1, \alpha_2, \cdots, \alpha_n \geq 0, w_1, w_2 \in R^n$, and $\lambda \in [0, 1]$. Then:
$$\begin{aligned} f(\lambda w_1 + (1-\lambda) w_2) &= \alpha_1 f_1(\lambda w_1 + (1-\lambda) w_2)) + \cdots + \alpha_n f_n(\lambda w_1 + (1-\lambda) w_2)) \\ &\leq \alpha_1 \cdot (\lambda f_1(w_1) + (1-\lambda) f_1(w_2)) + \cdots + \alpha_n \cdot (\lambda f_n(w_1) + (1-\lambda) f_n(w_2)) \\ &= \lambda(\alpha_1 f_1(w_1) + \cdots + \alpha_n f_n(w_1)) + (1-\lambda)(\alpha_1 f_1(w_2) + \cdots + \alpha_n f_n(w_2)) \\ &= \lambda f(w_1) + (1-\lambda) f(w_2) \end{aligned}$$
   where the second line is obtained using convexity of $f_1, \cdots, f_n$ and the fact that the inequalities preserved as $\alpha_1, \cdots, \alpha_n$ are non-negative.

2. Composition with an affine mapping: if f is convex, so is
$$g(w) = f(Aw + b)$$

   Proof: (from [4])
   Let $w_1, w_2 \in R^n$ and $\lambda \in [0, 1]$. Then:
$$\begin{aligned} g(\lambda w_1 + (1-\lambda) w_2)) &= f(A(\lambda w_1 + (1-\lambda) w_2) + b) \\ &= f(\lambda \cdot Aw_1 + (1-\lambda) \cdot Aw_2 + \lambda b + (1-\lambda) b) \\ &= f(\lambda \cdot (Aw_1 + b) + (1-\lambda) \cdot (Aw_2 + b)) \\ &\leq \lambda f(Aw_1 + b) + (1-\lambda) f(Aw_2 + b) \\ &= \lambda g(w_1) + (1-\lambda) g(w_2) \end{aligned}$$
   So g is convex.

3. If f_1, f_2 are convex, so is
$$g(w) = max\{f_1(w), f_2(w)\}$$
Proof: (from [9])
Let $0 \leq \lambda \leq 1$ and $w_1, w_2 \in$ domain of f, then:
$$\begin{aligned} f(\lambda w_1 + (1-\lambda)w_2)) &= max\{f_1(\lambda w_1 + (1-\lambda)w_2)), f_2(\lambda w_1 + (1-\lambda)w_2))\} \\ &\leq max\{\lambda f_1(w_1) + (1-\lambda)f_1(w_2), \lambda f_2(w_1) + (1-\lambda)f_2(w_2)\} \\ &\leq \lambda max\{f_1(w_1), f_2(w_1)\} + (1-\lambda)max\{f_1(w_2), f_2(w_2)\} \\ &= \lambda f(w_1)_(1-\lambda)f(w_2) \end{aligned}$$
which establishes convexity of f. It is easily shown than if $f_1, f_2, \cdots, f_m$ are convex, then pointwise maximum
$$f(x) = max\{f_1(w), f_2(w), \cdots, f_m(w)\}$$
is also convex.

Note: Pointwise minimum of two convex functions may not be convex.(from [4])

3.4 Optimality of Convex Optimization

1. Local Optimal and Global Optimal
   - A point w^* is locally optimal if $f(w^*) \leq f(w) \forall w$ in a neighborhood of w^*
   - A pointw^* is globally optimal if $f(w^*) \leq f(w) \forall w$
   - To find a local optimum of f, $\triangledown f(w^*) = 0$ is sufficient, but it is hard to find global optimum.
2. Optimal of Convex function
   - Convex Optimization is special
   - For convex problems, global optimality follows directly from local optimality
   - For convex function f, $\triangledown f(w^*) = 0$ is sufficient for global optimality.

4 Algorithm to Search Optimum

4.1 Overview of Descent Method

Algorithm 1: iterative algorithm

start with some guess w, $i = 0$;
while $\triangledown f(w) \neq 0$,*iterate k times* **do**
  select direction d_k;
  select step size α_k;
  $w \longleftarrow w + \alpha_k \cdot d_k$;
end while

4.2 Descent Direction

The direction of gradient vector is the greatest increase of f, so d_k must be negative gradient direction, which is $\nabla f(w)^T d_k < 0$. Some options of d_k are listed below

1. Steepest descent: $d_k = -\nabla f(w)$
2. Scaled gradient: $d_k = -D_k \nabla f(w)$ for some constant $D_k > 0$

4.3 Descent Step Size

1. Exact: $\alpha_k = argmin_{\alpha \geq 0} f(w_k + \alpha d_k)$
2. Constant: $\alpha_k = \frac{1}{L}$ for some suitable L
3. Diminishing: $\alpha_k \to 0$ but $\sum_k \alpha_k = \infty$ i.e.,$\alpha_k = \frac{1}{k}$
4. Armijo Rule: Start with initial step size s, and continue with $\alpha = \beta s$, $\alpha = \beta^2 s$, $\alpha = \beta^m s$ till$\alpha = \beta^m s$ falls within the set of α with

   $$f(w_k + \alpha d_k) - f(w_k) \leq \sigma \alpha \nabla f(w_k)^T d_k$$

   where $\sigma \in (0, 1)$ Armijo rule allows us to choose suitable step size that makes every next step to be sufficient decrease such that every next step is smaller than $\sigma \alpha \nabla f(w_k)^T d_k$.

   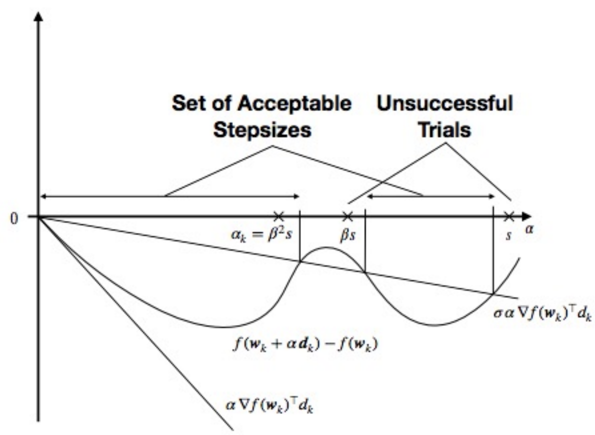

   Figure 2: Constrain of Armijo Rule

   However, Armijo rule is not sufficient to ensure every next step is a reasonable progress, the curvature condition bound the lower bound of every step size, so the step size will not be unacceptably short:

   $$\nabla f(w_k + \alpha_k d_k)^T d_k \geq \epsilon \nabla f_k{}^T d_k$$

   where $\epsilon \in (\sigma, 1)$ [8].

4.4 Convergence Rate

1. Lipschitz Continuous Gradient
   Definition: The gradient of f is Lipschitz continuous with parameter $L > 0$ if

   $$\|\nabla f(w_1) - \nabla f(w_2)\|_2 \leq L \|w_1 - w_2\|_2$$

for all $w_1, w_2 \in dom(f)$
Intuition: if have Lipschitz continuous gradient, then $g(w) = \frac{L}{2}\|w\|_2^2 - f(w)$ is convex **Proof:**

$$\begin{aligned}(\nabla f(w_1) - \nabla f(w_2))^T(w_1 - w_2) &\leq \|\nabla f(w_1) - \nabla f(w_2)\|_2 \|w_1 - w_2\|_2 \\ &\leq L\|w_1 - w_2\|_2^2 \end{aligned} \tag{4.4.1}$$

Since $\nabla g(w) = L(w) - \nabla f(w)$

$$\begin{aligned}(\nabla g(w_1) - \nabla g(w_2))^T(w_1 - w_2) &= (L(w_1 - w_2) - (\nabla f(w_1) - \nabla f(w_2)))^T(w_1 - w_2) \\ &= L\|w_1 - w_2\|_2^2 - (\nabla f(w_1) - \nabla f(w_2))^T(w_1 - w_2) \\ &\geq 0 \end{aligned} \tag{4.4.2}$$

We can get 4.4.2 from 4.4.1. Since

$$(\nabla g(w_1) - \nabla g(w_2))^T(w_1 - w_2) \geq 0$$

, by lemma2 of convex function we get $g(w)$ is convex
Intuition: if $g(w) = \frac{L}{2}\|w\|_2^2 - f(x)$ is convex, then:

$$\begin{aligned} g(w_2) &\geq g(w_1) + \nabla g(w_1)^T(w_2 - w_1) \\ \frac{L}{2}\|w_2\|_2^2 - f(w_2) &\geq \frac{L}{2}\|w_1\|_2^2 - f(w_1) + (L \cdot w_1 - \nabla f(w_1))(w_2 - w_1) \\ \frac{L}{2}\|w_2 - w_1\|_2^2 - f(w_2) &\geq -f(w_1) - \nabla f(w_1)^T(w_2 - w_1) \\ f(w_2) &\leq f(w_1) + \nabla f(w_1)^T(w_2 - w_1) + fracL2\|w_2 - w_1\|_2^2 \forall w_1, w_2 \end{aligned} \tag{1}$$

2. Strong Convexity

$$f(w_2) \geq f(w_1) + \nabla f(w_1)^T(w_2 - w_1) + \frac{\sigma}{2}\|w_2 - w_1\|_2^2 \;\; \forall w_1, w_2$$

3. Convergence Rate of Lipschitz Continuous Rradient with Strong Convexity

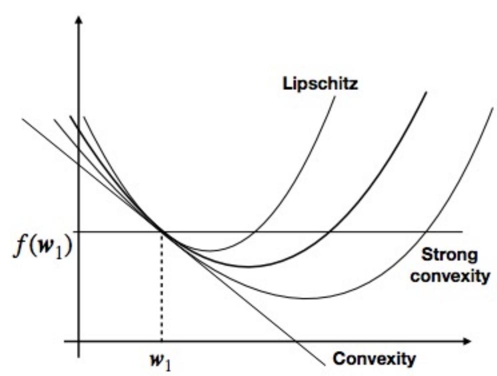

Figure 3: Lipschitz Continuous and Strong Convexity Bound

As shown in the picture, the strong convexity and Lipschitz continuous gradient work as lower bound and upper bound of f respectively. Note if f is twice differentiable,

$$\sigma I < \nabla^2 f(w) < LI \;\forall w$$

To know the convergence rate, we need to know how many iterations k such that

$$f(w_k) - f(w^*) \leq \epsilon$$

for $w_{k+1} = w_k + \alpha d_k$, where the chosen αd_k can minimize w.r.t w_{k+1} right-hand-side of upper bound.

$$f(w_{k+1}) \leq f(w_k) + \nabla f(w_k)^T (w_{k+1} - w_k) + \frac{L}{2} \|w_{k+1} - w_k\|_2^2$$

Take deivative can get $\nabla f(w_k) + L(\alpha d_k) = 0$ so $\alpha d_k = -\frac{1}{L}\nabla f(w_k)$.

$$\begin{aligned}
f(w_{k+1}) &\leq f(w_k) + \nabla f(w_k)^T (w_{k+1} - w_k) + \frac{L}{2} \|w_{k+1} - w_k\|_2^2 \\
f(w_{k+1}) &\leq f(w_k) + \nabla f(w_k)^T ((w_k - \frac{1}{L}\nabla f(w_k)) - w_k) + \frac{L}{2} \left\|w_k - \frac{1}{L}\nabla f(w_k) - w_k\right\|_2^2 \\
f(w_{k+1}) &\leq f(w_k) - \frac{1}{2L} \|\nabla f(w_k)\|_2^2 \\
f(w_k) &\leq f(w_{k-1}) - \frac{1}{2L} \|\nabla f(w_{k-1})\|_2^2
\end{aligned}$$

which implies guaranteed progress of step k is $f(w_{k-1}) - \frac{1}{2L} \|\nabla f(w_{k-1})\|_2^2$. Similarly, if we apply strong convexity, we can get the maximum sub-optimality is

$$f(w_*) \geq f(w_k) - \frac{1}{2\sigma} \|\nabla f(w_k)\|_2^2$$

The distance we need to go is

$$f(w_k) - f(w^*) \leq \frac{1}{2\sigma} \left\|\nabla f(w_k)_2^2\right\|)$$

, so $(f(w_k) - f(w^*))2\sigma \leq \left\|\nabla f(w_k)_2^2\right\|)$. Then we have in gauranteed progress:

$$\begin{aligned}
f(w_k) - f(w^*) &\leq f(w_{k-1}) - f(w^*) - \frac{1}{2L} \|\nabla f(w_{k-1})\|_2^2 \\
&\leq f(w_{k-1}) - f(w^*) - \frac{\sigma}{L}(f(w_{k-1}) - f(w^*)) \\
&= (1 - \frac{\sigma}{L})(f(w_{k-1}) - f(w^*)) \\
&\leq (1 - \frac{\sigma}{L})^k (f(w_{k-1}) - f(w^*))
\end{aligned}$$

So we can get:

$$\begin{aligned}
C(1 - \frac{\sigma}{L})^k &\leq \epsilon \\
k &\geq O(log(\frac{1}{\epsilon}))
\end{aligned}$$

4. Lipschitz Continuous Gradient Without Strong Convexity Assumption
   Lipschitz bound and $w_2 = w_1 - \alpha \nabla f(w_1)$ yields

$$f(w_2) \leq f(w_1) - (1 - \frac{L\alpha}{2})\alpha \|\nabla f(w_1)\|_2^2$$

Combined with convexity: $f(w_1) + \nabla f(w_1)^T(w^* - w_1) \leq f(w^*)$

$$f(w_2) \leq f(w^*) + \nabla f(w_1)(w_1 - w^*) - \frac{\alpha}{2}\|\nabla f(w_1)\|_2^2$$

Using $w_2 - w_1 = -\alpha \nabla f(w_1)$ and rearranging terms gives

$$f(w_2) \leq f(w^*) + \frac{1}{2\alpha}(\|w_1 - w^*\|_2^2 - \|w_2 - w^*\|_2^2)$$

Summing over all iterations

$$\sum_{i=1}^{k}(f(w_i) - f(w^*)) \leq \frac{1}{2\alpha}\|w_0 - w^*\|_2^2$$

$f(w_i)$ non-increasing:

$$f(w_k) - f(w^*) \leq \frac{1}{k}\sum_{i=1}^{k}(f(w_i) - f(w^*)) \leq \frac{1}{2\alpha}\|w_0 - w^*\|_2^2 \leq \epsilon$$

Consequently:

$$k \geq O(\frac{1}{\epsilon})$$

4.5 Acceleration of Optimization

1. The convergence rates in section 4.4 are not optimal, we can add an extra momentum term to accelerate the convergence.
   Intuition: if direction of current gradient step is same as direction of previous step, move further than previous step; if the direction of current gradient step is opposite to previous step direction, move less than previous step.

   - Polyak's method(Aka heavy ball):

   $$w_{k+1} = w_k - \alpha_k \nabla f(w_k) + \beta_k(w_k - wk - 1)$$

   - in deep learning

   $$v_{k+1} = \beta v_k + \nabla f(w_k)$$
   $$w_{k+1} = w_k - \alpha v_{k+1}$$

2. Another aspect we can make the convergence rate better is to reduce the time for computing the gradient. The iteration complexity is linear to the size of dataset, so a large dataset will make gradient computing slow. To deal with this problem, the stochastic gradient descent method is introduced. In stochastic gradient descent, a subset of samples are selected to approximate the gradient based on this batch of data.

   (a) General Steps
   - Select a subset B_k from all samples
   - Gradient Update using Approximation

   $$\nabla f(w) \approx \sum_{(x^i, y^i) \in B_k}$$

   (b) Convergence Rate
   - Lipschitz continuous gradient and strongly convex: $k \geq O(\frac{1}{\epsilon})$
   - Lipschitz continuous gradient: $k \geq O(\frac{1}{\epsilon^2})$

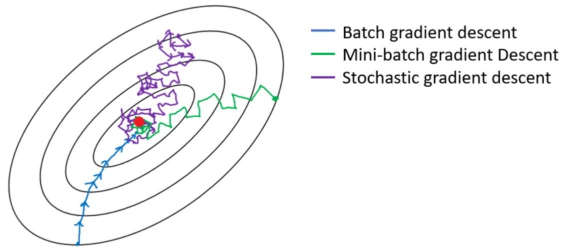

Figure 4: Stochastic,Min batch,Batch gradient descent [7]

References

[1] *Proof of Convexity and Monotone*, 2016. Available at `https://math.stackexchange.com/questions/1717542/a-function-is-convex-if-and-only-if-its-gradient-is-monotone`.

[2] *Convexity and Second Derivative*, 2017. Available at `https://math.stackexchange.com/questions/1224955/proving-that-the-second-derivative-of-a-convex-function-is-nonnegative`.

[3] *Why is every p-norm convex*, 2017. Available at `https://math.stackexchange.com/questions/2280341/why-is-every-p-norm-convex`.

[4] A. A. Ahmadi. *Operations that preserve convexity*, 2014. Available at `http://www.princeton.edu/~amirali/Public/Teaching/ORF363_COS323/F14/ORF363_COS323_F14_Lec6.pdf`.

[5] N. Andrei. *Convex Functions*, 2005. Available at `https://camo.ici.ro/neculai/convex.pdf`.

[6] CheCheDaWaff. *Convex polygon illustration*. Available at `https://commons.wikimedia.org/w/index.php?curid=49543150`.

[7] G. Dahl. *A mini-introduction to convexity*, 2014. Available at `https://www.uio.no/studier/emner/matnat/math/MAT-INF3100/v14/convmat-inf3100.pdf`.

[8] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, second edition, 2006.

[9] L. V. Stephen Boyd. *Convex Optimization)*. Campridge University Press, 2004.