Unconstrained convex optimization through first-order approximation methods

Sang Jun Moon

August 29, 2020

Statistics, University of Seoul

Contents

Introduction

Convex optimization problem

- An convex optimization problem is one of the form

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & x \in \mathcal{X} \end{aligned}$$

  where $x \in \mathbb{R}^p$, $f : \mathbb{R}^p \mapsto \mathbb{R}$ is convex function, and $\mathcal{X} \subseteq \mathbb{R}^p$ is convex set.
- We call f the objective function, $\mathcal{X}$ the feasible set, and $x \in \mathcal{X}$ the constraint.
- An optimal value p^* is defined as

$$p^* = \inf\{f(x) : x \in \mathcal{X}\}.$$

- In addition, if $x^* \in \mathcal{X}$ and $f(x^*) = p^*$, then x^* is called optimal.

Unconstrained convex optimization problem

- If there are no constraints, we say the problem (1) is unconstrained convex optimization problem:

$$\min_{x} \quad f(x)$$

- In this presentation, we will deal with the algorithms for solving unconstrained convex optimization problem.
- Among various algorithms, the algorithms based on the first-order approximation method and its convergence properties are explained.

Representative algorithms

- The algorithms will be introduced can be written as

$$x^{(k+1)} = x^{(k)} + \eta_k \Delta x^{(k)} \quad (1)$$

  in common where $\eta_k > 0$ is called a step size or learning rate and $\Delta x^{(k)}$ is called a direction.
- The algorithm varies depending on the type of direction.
- The convergence properties of each algorithm can be seen when an appropriate step size.
- The following two types are considered in the selection of step size:
  - Fixed constant step size: $\eta_k = \eta$
  - Diminishing step size: η_k satisfying

$$\sum_{k=1}^{\infty} \eta_k^2 < \infty, \quad \sum_{k=1}^{\infty} \eta_k = \infty$$

Objective function f

To prove convergence property, it is assumed that the objective function f should satisfy one or more of the following conditions:

- Lipschitz continuous gradient.
- Strong convexity.
- Lipschitz continuous.

Lipschitz continuous gradient condition (A1)

- A differentiable function f is L-Lipschitz continuous gradient iff

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for all } x, y \tag{2}$$

for some $L > 0$ where

$$\nabla f(x) = \Big(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \cdots, \frac{\partial f(x)}{\partial x_p}\Big)^T.$$

- The L-Lipschitz continuous gradient condition guarantees that

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2. \tag{3}$$

Strong convexity condition (A2)

- A differentiable f is S-strongly convex iff

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{S}{2}\|y - x\|_2^2 \tag{4}$$

- If f is S-strongly convex, then

$$\|\nabla f(x)\|_2^2 \geq 2S(f(x) - f(x^*)) \tag{5}$$

Lipschitz continuous condition (A3)

- A function f is C-Lipschitz continuous iff

$$|f(x) - f(y)| \leq C\|x - y\|_2 \quad \text{for all } x, y \tag{6}$$

  for some $s > 0$.

- If f is differentiable and C-Lipschitz continuous, then

$$\|\nabla f(x)\|_2 \leq C \quad \text{for all } x$$

Gradient descent method (GD)

Descent methods

- The descent methods satisfy

$$f(x^{(k+1)}) < f(x^{(k)}) \tag{7}$$

for $x^{(k)} \neq x^*$ with $x^{(k+1)} = x^{(k)} + \eta_k \Delta x^{(k)}$.

- From convexity, we know that

$$\begin{aligned} f(x^{(k+1)}) - f(x^{(k)}) &\geq \nabla f(x^{(k)})^T (x^{(k+1)} - x^{(k)}) \quad (8) \\ &= \eta_k \nabla f(x^{(k)})^\top \Delta x^{(k)} \quad (9) \end{aligned}$$

- Since $\nabla f(x^{(k)})^T \Delta x^{(k)} \geq 0$ implies $f(x^{(k+1)}) \geq f(x^{(k)})$,

$$\nabla f(x^{(k)})^T \Delta x^{(k)} < 0 \tag{10}$$

is necessary condition for descent methods.

Gradient descent method

- The negative gradient, $-\nabla f(x^{(k)})$, is the most easily conceived direction for descent methods.
- This is because, for an arbitrary unit descent direction v, the change of $f(x^{(k)})$ is given by

$$\frac{\partial}{\partial \eta_k} f(x^{(k)} + \eta_k v)\bigg|_{\eta_k = 0} = \nabla f(x^{(k)})^\top v.$$

which implies that the direction of steepest descent is

$$v = -\nabla f(x^{(k)}) / \|\nabla f(x^{(k)})\|_2.$$

- Thus, the update rule of GD,

$$x^{(k+1)} = x^{(k)} - \eta_k \nabla f(x^{(k)}),$$

  intuitively makes sense.
- From now on, we will show the convergence properties which is defined as the upper bound of $f(x^{(K)}) - p^*$ where K is the number of iterations.

Theorem 1

Under (A1) and fixed step size $0 < \eta < 1/L$, the following property holds

$$f(x^{(K+1)}) - p^* \leq \frac{\|x^{(1)} - x^*\|_2^2}{2\eta K}.$$

- We need $O(1/\epsilon)$ iterations to get $f(x^{(K)}) - p^* \leq \epsilon$.

- For simplicity, denote update rule as $x^+ = x - \eta \nabla f(x)$.
- From the assumption (A1), (3) holds as follows:

$$f(x^+) \leq f(x) + \nabla f(x)^T (x^+ - x) + \frac{L}{2}\|x^+ - x\|_2^2 \quad (11)$$

$$= f(x) - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla f(x)\|_2^2 \quad (12)$$

- The range of step size makes upper bound on

$$-\eta \left(1 - \frac{L\eta}{2}\right) \|\nabla f(x)\|_2^2 < -\frac{\eta}{2}\|\nabla f(x)\|_2^2 \quad (13)$$

which implies that the algorithm is a descent method.

- Thus, we can obtain that

$$\begin{aligned} f(x^+) &\leq f(x) - \frac{\eta}{2}\|\nabla f(x)\|_2^2 & (14) \\ &\leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{\eta}{2}\|\nabla f(x)\|_2^2 & (15) \\ &= f(x^*) + \frac{1}{2\eta}(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2). & (16) \end{aligned}$$

where the second inequality is due to convexity of f.

- By summing both sides of (14) from $k = 1$ to K, it follows that

$$\begin{aligned} \sum_{k=1}^{K}\Big(f(x^{(k+1)}) - f(x^*)\Big) &\leq \frac{1}{2\eta}(\|x^{(1)} - x^*\|_2^2 - \|x^{(K+1)} - x^*\|_2^2) \\ &\leq \frac{1}{2\eta}\|x^{(1)} - x^*\|_2^2. \end{aligned}$$

- Since it is a decent method,

$$K(f(x^{(K+1)}) - f(x^*)) \leq \sum_{k=1}^{K} \Big(f(x^{(k)}) - f(x^*) \Big)$$

holds. Therefore,

$$f(x^{(K+1)}) - f(x^*) \leq \frac{\|x^{(1)} - x^*\|_2^2}{2\eta K}$$

Theorem 2

Under (A1), (A2), and fixed step size $0 < \eta < \frac{1}{L+S}$,

$$f(x^{(K+1)}) - p^* \leq (1 - \eta S)^K \left(f(x^{(1)}) - p^*\right)$$

- We need $O(\log(1/\epsilon))$ iterations to get $f(x^{(K)}) - p^* \leq \epsilon$.

- For simplicity, we denote update rule by $x^+ = x - \eta \nabla f(x)$.
- As in Theorem 1, $-\eta + \frac{\eta^2 L}{2} < -\frac{\eta}{2}$ holds because the step size η is always smaller than $1/L$.
- Thus, by (A2), we can get

$$
\begin{aligned}
f(x^+) &\leq f(x) - \frac{\eta}{2}\|\nabla f(x)\|_2^2 && (17)\\
&\leq f(x) - \eta S(f(x) - p^*). && (18)
\end{aligned}
$$

- By subtracting p^* from both side, it is obtained that

$$f(x^+) - p^* \leq (1 - \eta S)\,(f(x) - p^*). \tag{19}$$

Therefore,

$$f(x^{(K+1)}) - p^* \leq (1 - \eta S)^K\,(f(x^{(1)}) - p^*) \tag{20}$$

Stochastic gradient descent method (SGD)

Stochastic gradient

- Often, it is hard to apply the GD to some cases.
- Suppose that the objective function can be decomposed as

$$f(x) = \sum_{i=1}^{n} f_i(x).$$

- The following are simple cases where it is difficult to apply GD.
  - Case 1: n is very large such that computing $\nabla f(x)$ is intractable.
  - Case 2: $f(x)$ is not fixed since f_i is observed on-line.

- A stochastic gradient is the gradient calculated by some of $f_i(x)$.
- Let $\xi = (\xi_1, \cdots, \xi_n)^T \in \mathbb{R}^n$ where $\xi_i \in \{0, 1\}$ for $i = 1, \cdots, n$ be random sampled value.
- The stochastic gradient, $g(x, \xi)$, is defined as

$$g(x, \xi) = \sum_{i=1}^{n} \xi_i \nabla f_i(x) \tag{21}$$

- The update rule of SGD is

$$x^{(k+1)} = x^{(k)} - \eta_k g(x^{(k)}, \xi^{(k)}) \tag{22}$$

in which $\xi^{(k)}$ is randomly sampled at each k-th iteration.

Additional assumptions for SGD

- There exist $\mu_G \geq \mu \geq 0$ such that, for all $k \in \mathbb{N}$,

$$\begin{aligned} \nabla f(x^{(k)})^T \mathrm{E}_{\xi^{(k)}}(g(x^{(k)}, \xi^{(k)})) &\geq \mu \|\nabla f(x^{(k)})\|_2^2 \\ \|\mathrm{E}_{\xi^{(k)}}(g(x^{(k)}, \xi^{(k)}))\|_2 &\leq \mu_G \|\nabla f(x^{(k)})\|_2 \end{aligned} \tag{23}$$

- There exists $M, M_V \geq 0$ such that

$$\mathrm{E}_{\xi^{(k)}}(\|g(x^{(k)}, \xi^{(k)})\|_2^2) \leq M + (M_V + \mu_G^2)\|\nabla f(x^{(k)})\|_2^2. \tag{24}$$

Here, we let $M_G = M_V + \mu_G^2$.

Lemma 1

Under (A1), (A2), (23), and (24), the following holds

$$\mathrm{E}(f(x^{(k+1)}) - p^*) \leq (1 - \eta_k \mu S)\mathrm{E}(f(x^{(k)}) - p^*) + \frac{L}{2}\eta_k^2 M \qquad (25)$$

for $0 < \eta_k < \mu / L M_G$.

- For ease of notation, we denote update rule as $x^+ = x - \eta g(x, \xi)$.
- From Assumption (A1), it follows that

$$\begin{aligned} f(x^+) &\leq f(x) + \nabla f(x)^T (x^+ - x) + \frac{L}{2}\|x^+ - x\|_2^2 \\ &= f(x) - \eta \nabla f(x)^T g(x, \xi) + \frac{L}{2}\eta^2 \|g(x, \xi)\|_2^2 \end{aligned}$$

- Taking expectations with respect to ξ, we can obtain

$$\begin{aligned}\mathrm{E}_\xi(f(x^+) - f(x)) &\leq -\eta \nabla f(x)^T \mathrm{E}_\xi(g(x,\xi)) + \frac{L}{2}\eta^2 \mathrm{E}_\xi(\|g(x,\xi)\|_2^2) \\ &\leq -\eta\mu\|\nabla f(x)\|_2^2 + \frac{L}{2}\eta^2(M + M_G\|\nabla f(x)\|_2^2) \\ &= -\eta\left(\mu - \frac{L}{2}\eta M_G\right)\|\nabla f(x)\|_2^2 + \frac{L}{2}\eta^2 M.\end{aligned}$$

by Assumption (23) and (24).

- We can take expectation and apply same technique in Theorem 2 as follows:

$$\mathrm{E}(f(x^+) - p^*) \leq (1 - \eta\mu S)\mathrm{E}(f(x) - p^*) + \frac{L}{2}\eta^2 M,$$

since if f is S-strongly convex, then $\|\nabla f(x)\|_2^2 \geq 2S(f(x) - f(x^*))$, and $0 < \eta < \mu/LM_G$.

Theorem 3

Under (A1), (A2), (23), and (24), for fixed step size satisfying $0 < \eta < \frac{\mu}{LM_G}$, the following inequality holds:

$$\mathrm{E}(f(x^{(K+1)}) - p^*) - \frac{L\eta M}{2\mu S} \leq (1 - \eta\mu S)^K \left(f(x^{(1)}) - p^* - \frac{L\eta M}{2\mu S} \right)$$

which implies that

$$\lim_{k\to\infty} \mathrm{E}(f(x^{(k)}) - p^*) \leq \frac{L\eta M}{2\mu S}.$$

- From Lemma 1, we can obtain

$$\mathrm{E}(f(x^{(k+1)}) - p^*) \leq (1 - \eta\mu S)\mathrm{E}(f(x^{(k)}) - p^*) + \frac{L}{2}\eta^2 M.$$

- Let $t = \frac{L\eta M}{2\mu S}$, then

$$\mathrm{E}(f(x^{(k+1)}) - p^*) - t \leq (1 - \eta\mu S)(\mathrm{E}(f(x^{(k)}) - p^*) - t)$$

  hold. Therefore, the following holds:

$$\mathrm{E}(f(x^{(K+1)}) - p^*) - t \leq (1 - \eta\mu S)^K (f(x^{(1)}) - p^* - t).$$

- Also, since $L \geq S, M_G = \mu_G^2 + M_V \geq \mu^2$,

$$0 < \eta\mu S < \frac{\mu^2 S}{L M_G} \leq 1.$$

Theorem 4

Under (A1), (A2), (23), (24), and for decreasing step size $\eta_k = \frac{\beta}{\gamma+k}$ for some $\beta > \frac{1}{\mu S}$ and $\gamma > 0$ such that $\eta_1 \leq \frac{\mu}{LM_G}$, expected optimality gap satisfies the following inequality.

$$\mathrm{E}(f(x^{(K)}) - p^*) \leq \frac{\nu}{\gamma + K} \tag{26}$$

where

$$\nu = \max\left\{\frac{\beta^2 LM}{2(\beta\mu S - 1)}, (\gamma+1)(f(x^{(1)}) - p^*)\right\} \tag{27}$$

- Since the step size is decreases as

$$\eta_k \leq \eta_1 \leq \frac{\mu}{LM_G},$$

  we can obtain

$$\mathrm{E}(f(x^{(k+1)}) - p^*) \leq (1 - \eta_k \mu S)\mathrm{E}(f(x^{(k)}) - p^*) + \frac{L}{2}\eta_k^2 M,$$

  from Lemma 1.
- Then, Theorem can be shown by induction.
- It is obvious that Theorem holds when $k = 1$.

- Assume that it holds for some $k > 1$ which is

$$\mathrm{E}(f(x^{(k)}) - p^*) \leq \frac{v}{\hat{k}} \quad \text{where} \quad \hat{k} = \gamma + k, \tag{28}$$

then since $\eta_k = \beta/\hat{k}$,

$$\begin{aligned}
\mathrm{E}(f(x^{(k+1)}) - p^*) &\leq (1 - \eta_k \mu S)\mathrm{E}(f(x^{(k)}) - p^*) + \frac{L}{2}\eta_k^2 M \\
&\leq \left(1 - \frac{\beta\mu S}{\hat{k}}\right)\frac{\nu}{\hat{k}} + \frac{\beta^2 LM}{2\hat{k}^2} \qquad (29) \\
&= \left(\frac{1}{\hat{k}} - \frac{1}{\hat{k}^2}\right)\nu - \frac{(\beta\mu s - 1)\nu}{\hat{k}^2} + \frac{\beta^2 cM}{2\hat{k}^2}
\end{aligned}$$

holds.

- From the definition of ν, $\nu \geq \frac{\beta^2 LM}{2(\beta\mu S - 1)}$ holds which implies that

$$\frac{(\beta\mu S - 1)\nu}{\hat{k}^2} \geq \frac{\beta^2 LM}{2\hat{k}^2}$$

- Therefore, it holds for $k+1$ as follows:

$$\begin{aligned} \mathrm{E}(f(x^{(k+1)}) - p^*) &\leq \left(\frac{1}{\hat{k}} - \frac{1}{\hat{k}^2}\right)\nu - \frac{(\beta\mu S - 1)\nu}{\hat{k}^2} + \frac{\beta^2 LM}{2\hat{k}^2} \\ &\leq \left(\frac{\hat{k} - 1}{\hat{k}^2}\right)\nu \leq \frac{1}{\hat{k} + 1}\nu. \end{aligned}$$

Subgradient method (SM)

- The above two algorithms are used when the objective function could be differentiated.
- The subgradient method (SM) is an algorithm that can be used when the objective function cannot be differentiated.

- A vector $g \in \mathbb{R}^n$ is subgradient at x iff

$$f(y) \geq f(x) + g^\top (y - x), \quad \text{for all} \quad y. \tag{30}$$

- The set of all subgradient of f at x is called the subdifferential

$$\partial f(x) = \{g \in \mathbb{R}^n : f(y) \geq f(x) + g^T (y - x)\}. \tag{31}$$

- For convex function f,

$$f(x^*) = \min_x f(x) \iff 0 \in \partial f(x^*) \tag{32}$$

which is called subgradient optimality condition.

- The update rule of SM is that

$$x^{(k+1)} = x^{(k)} - \eta_k g^{(k)},$$

  where $g^{(k)} \in \partial f(x^{(k)})$.

- The main difference between SM and GD is that the subgradient method saves the updated solutions, $x^{(k)}$, and selects the solution that makes the objective function the smallest among them as follows:

$$\widehat{x}^{(K)} = \underset{k=1,\cdots,K}{\operatorname{argmin}} f(x^{(k)}),$$

  because it is not one of descent method.

Lemma 2

Under (A3), the following holds:

$$f(\hat{x}^{(K)}) - p^* \leq \frac{R^2 + C^2 \sum_{k=1}^{K} \eta_k^2}{2 \sum_{k=1}^{K} \eta_k} \tag{33}$$

where $R = \|x^{(1)} - x^*\|_2$ and $\eta_k > 0$ is an arbitrary step size.

- From the definition of subgradient, we set $y = x + g$ as

$$f(x+g) \geq f(x) + g^\top (x+g-x) = f(x) + \|g\|_2^2$$

which implies that

$$\|g\|_2^2 \leq |f(x+g) - f(x)| \leq C\|g\|_2 \Rightarrow \|g\|_2 \leq C$$

under (A3).

- For simplification, the update rule is denoted by $x^+ = x - \eta g$.
- Thus, it follows that

$$\begin{aligned} \|x^+ - x^*\|_2^2 &= \|x - x^*\|_2^2 - 2\eta g^T(x - x^*) + \eta^2 \|g\|_2^2 \\ &\leq \|x - x^*\|_2^2 - 2\eta(f(x) - f(x^*)) + \eta^2 C^2 \end{aligned}$$

in which the second inequality holds from the definition of subgradient.

- By summing both sides for $k = 1$ to K,

$$\|x^{(K+1)} - x^*\|_2^2 \leq \|x^{(1)} - x^*\|_2^2 - 2\sum_{k=1}^{K} \eta_k (f(x^{(k)}) - p^*) + C^2 \sum_{k=1}^{K} \eta_k^2 \quad (34)$$

- Since $\|x^{(K+1)} - x^*\|_2^2 \geq 0$ and $R = \|x^{(1)} - x^*\|_2$,

$$2\sum_{k=1}^{K} \eta_k (f(x^{(k)}) - p^*) \leq R^2 + C^2 \sum_{k=1}^{K} \eta_k^2 \quad (35)$$

holds.

- In addition,

$$2(f(\hat{x}^{(K)}) - p^*) \sum_{k=1}^{K} \eta_k \leq 2 \sum_{k=1}^{K} \eta_k (f(x^{(k)}) - p^*) \tag{36}$$

  holds because $f(\hat{x}^{(K)}) = \min_{k=1,\cdots,K} f(x^{(k)})$.

- Therefore,

$$f(\hat{x}^{(K)}) - p^* \leq \frac{R^2 + C^2 \sum_{k=1}^{K} \eta_k^2}{2 \sum_{k=1}^{K} \eta_k} \tag{37}$$

Theorem 5

Under (A3) and for fixed step size η,

$$f(\hat{x}^{(K)}) - p^* \leq \frac{R^2}{2K\eta} + \frac{\eta C^2}{2} \tag{38}$$

where $R = \|x^{(1)} - x^*\|_2^2$ which implies that

$$\lim_{k\to\infty} f(\hat{x}^{(k)}) \leq p^* + \frac{\eta C^2}{2}.$$

- For making right hand side of above inequality less than ϵ, we can choose

$$\eta = \frac{\epsilon}{C^2}, \quad K = \frac{R^2}{\eta\epsilon} = \frac{C^2R^2}{\epsilon^2}.$$

- That is, we need $O(1/\epsilon^2)$ iterations to get $f(\widehat{x}^{(K)}) - p^* \leq \epsilon$.

Theorem 6

Under (A3) and diminishing step size η_k,

$$f(\hat{x}^{(K)}) - p^* \leq O\left(\frac{1}{\sum_{k=1}^{K} \eta_k}\right)$$

holds which implies that

$$\lim_{k \to \infty} f(\hat{x}^{(k)}) = p^*$$

Proximal gradient method (PG)

- Like the SM, the proximal gradient method (PG) is a method that can be used when the objective function cannot be differentiated, but unlike the SM, suppose that f can be decomposed into

$$f(x) = g(x) + h(x)$$

  where g is convex and differentiable and h is convex but non-differentiable.
- The motivation for PG is to approximate the differentiable function g at $x = x^{(k)}$ as follows:

$$g(z) \approx g(x^{(k)}) + \nabla g(x^{(k)})^T (z - x^{(k)}) + \frac{1}{2\eta_k} \|z - x^{(k)}\|_2^2 := \tilde{g}(z)$$

- The update rule of PG is as follows:

$$\begin{aligned} x^{(k+1)} &= \underset{z}{\text{argmin}} \;\; \tilde{g}(z) + h(z) \\ &= \underset{z}{\text{argmin}} \;\; g(x^{(k)}) + \nabla g(x^{(k)})^T (z - x^{(k)}) + \frac{1}{2\eta_k} \|z - x^{(k)}\|_2^2 \\ &\qquad + h(z) \\ &= \underset{z}{\text{argmin}} \;\; \frac{1}{2\eta_k} \|z - (x^{(k)} - \eta_k \nabla g(x^{(k)}))\|_2^2 + h(z) \end{aligned} \tag{39}$$

- Here, the proximal mapping is defined as

$$\text{prox}_{h,\eta_k}(y) = \underset{z}{\text{argmin}} \;\; \frac{1}{2\eta_k} \|z - y\|_2^2 + h(z) \tag{40}$$

- Thus, the update rule of PG can be expressed as follows.

$$\begin{aligned} x^{(k+1)} &= \mathrm{prox}_{h,\eta_k}(x^{(k)} - \eta_k \nabla g(x^{(k)})) \\ &= x^{(k)} - \eta_k G_{\eta_k}(x^{(k)}) \end{aligned}$$

  where

$$G_{\eta_k}(x) = \frac{x^{(k)} - \mathrm{prox}_{h,\eta_k}(x^{(k)} - \eta \nabla g(x^{(k)}))}{\eta_k}$$

- The strength of PG is that the proximal mapping depends only on h not g and can be computed analytically for some h.

Example of proximal gradient descent method: ISTA

- Consider the objective function of Lasso regression

$$f(\beta) = \underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{g(\beta)} + \underbrace{\lambda\|\beta\|_1}_{h(\beta)} \tag{41}$$

for given $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$.

- The proximal mapping is

$$\operatorname{prox}_{h,\eta}(\beta) = \operatorname*{argmin}_z A(z) \tag{42}$$

where $A(z) = \frac{1}{2\eta}\|\beta - z\|_2^2 + \lambda\|z\|_1$.

- By subgradient optimal condition (32), z^* is optimal if

$$0 \in \partial A(z^*) = \frac{1}{\eta}(z^* - \beta) + \lambda \partial \|z^*\|_1 \tag{43}$$

- For some $v \in \partial \|z^*\|_1$,

$$-\frac{1}{\eta}(z^* - \beta) = \lambda v \tag{44}$$

- Choose z^* such that

$$[z^*]_i = [S_{\lambda\eta}(\beta)]_i = \begin{cases} \beta_i - \lambda\eta & \text{if } \beta_i > \lambda\eta \\ 0 & \text{if} - \lambda\eta \leq \beta_i \leq \lambda\eta \\ \beta_i + \lambda\eta & \text{if } \beta_i < -\lambda\eta \end{cases} \tag{45}$$

which satisfying subgradient optimal condition.

- Therefore, proximal mapping is

$$\begin{aligned}\text{prox}_{h,\eta}(\beta) &= \underset{z}{\text{argmin}} \frac{1}{2\eta}\|\beta - z\|_2^2 + \lambda\|z\|_1 \\ &= S_{\lambda\eta}(\beta)\end{aligned}$$

- Since $\nabla g(\beta) = -X^T(y - X\beta)$, update rule is

$$\begin{aligned}\beta^+ &= \text{prox}_{h,\eta}(\beta - \eta\nabla g(\beta)) \qquad (46)\\ &= S_{\lambda\eta}(\beta + \eta X^T(y - X\beta)) \qquad (47)\end{aligned}$$

which is called iterative soft-thresholding algorithm (ISTA).

Lemma 3

$$G_\eta(x) - \nabla g(x) \in \partial h(x^+) \ \text{ where } \ x^+ = x - \eta G_\eta(x)$$

- By definition of proximal mapping and subgradient optimality,

$$u = \underset{z}{\operatorname{argmin}} \frac{1}{2\eta}\|z - x\|_2^2 + h(z) \iff 0 \in \frac{1}{\eta}(u - x) + \partial h(u) \quad (48)$$

holds. In our case, since

$$x^+ = \underset{z}{\operatorname{argmin}} \ \frac{1}{2\eta}\|z - (x - \eta\nabla g(x))\|_2^2 + h(z), \quad (49)$$

Lemma holds as follows:

$$G_\eta(x) - \nabla g(x) \in \partial h(x^+) \quad (50)$$

Lemma 4

Assume that g is L-Lipschitz continuous gradient as (A3) and for fixed $0 < \eta < 1/L$,

$$f(x^+) \leq f(z) + G_\eta(x)^T(x - z) - \frac{\eta}{2}\|G_\eta(x)\|_2^2$$

holds for all z where $x^+ = x - \eta G_\eta(x)$.

- By L-Lipschitz condition on g,

$$f(x^+) = g(x^+) + h(x^+) \tag{51}$$

$$\leq \underbrace{g(x) - \eta \nabla g(x)^\top G_\eta(x) + \frac{\eta^2 L}{2} \| G_\eta(x) \|_2^2}_{T1} + \underbrace{h(x^+)}_{T2} \tag{52}$$

holds. Since g is convex and $\eta \leq 1/L$,

$$T1 \leq g(z) + \nabla g(x)^T (x - z) - \eta \nabla g(x)^T G_\eta(x) + \frac{\eta}{2} \| G_\eta(x) \|_2^2$$

holds. In addition, from Lemma 3,

$$T2 \leq h(z) + (G_\eta(x) - \nabla g(x))^T (x^+ - z) \tag{53}$$

holds. Therefore,

$$T1 + T2 \leq f(z) + G_\eta(x)^T (x - z) - \frac{\eta}{2} \| G_\eta(x) \|_2^2 \tag{54}$$

Theorem 7

Assume that g is L-Lipschitz continuous gradient and for fixed step size $0 < \eta < 1/L$,

$$f(x^{(K+1)}) - p^* \leq \frac{\|x^{(1)} - x^*\|_2^2}{2\eta K}$$

holds.

- We need $O(1/\epsilon)$ iterations to make $f(x^{(K+1)}) - p^* \leq \epsilon$.

- For ease of notation, we denote the update rule by $x^+ = x - G_\eta(x)$.
- Remark: (Lemma 4)

$$f(x^+) \leq f(z) + G_\eta(x)^T(x - z) - \frac{\eta}{2}\|G_\eta(x)\|_2^2$$

- Since Lemma 4 is satisfied for all z, we can get

$$f(x^+) \leq f(x) - \frac{\eta}{2}\|G_\eta(x)\|_2^2$$

by substituting $z = x$ which implies that it is descent method.

- Substituting $z = x^*$ into Lemma 4 makes

$$\begin{aligned} f(x^+) - p^* &\leq G_\eta(x)^T(x - x^*) - \frac{\eta}{2}\|G_\eta(x)\|_2^2 && (55) \\ &= \frac{1}{2\eta}(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2) && (56) \end{aligned}$$

- By summing both sides of inequality for $k = 1$ to K, it follows that

$$\begin{aligned} \sum_{k=1}^{K}(f(x^{(k+1)}) - p^*) &\leq \frac{1}{2\eta}(\|x^{(1)} - x^*\|_2^2 - \|x^{(K+1)} - x^*\|_2^2) \\ &\leq \frac{1}{2\eta}\|x^{(1)} - x^*\|_2^2. \end{aligned}$$

Therefore, we can obtain

$$K(f(x^{(K+1)}) - p^*) \leq \frac{1}{2\eta}\|x^{(1)} - x^*\|_2^2.$$

Summary

Summary

	A1	(A1, A2)	A3
GD	c_1/K	c_2^K	$\cdot$
SGD	$\cdot$	$c_3^K + c_4$;f, $c_5/(c_6 + K)$;d	$\cdot$
SM	$\cdot$	$\cdot$	$c_7/K + c_8$
PG	c_9/K	$\cdot$	$\cdot$

Table 1: Convergence properties of each algorithms with respect to assumptions and step size; K is the number of iterations; c_i is some positive constant for $i = 1, \cdots, 9$; ';f' denotes fixed step size and ';d' denotes diminishing step size;