EE6550 Machine Learning

Lecture Eleven – Stochastic Gradient Descent

Chung-Chin Lu

Department of Electrical Engineering

National Tsing Hua University

May 22, 2017

Gradient Descent

- $f : S \to \mathbb{R}$: a real-valued function defined on a subset S of $\mathbb{R}^d$ which is differentiable at an interior point $\boldsymbol{a}$ of S, i.e.,

$$f(\boldsymbol{b}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})(\boldsymbol{b} - \boldsymbol{a}) + o(\|\boldsymbol{b} - \boldsymbol{a}\|)$$

  for all $\boldsymbol{b}$ in a neighborhood $B(\boldsymbol{a}; r)$ of $\boldsymbol{a}$ in S, where $\nabla f(\boldsymbol{a}) = (\partial f(\boldsymbol{a})/\partial x_1, \partial f(\boldsymbol{a})/\partial x_2, \ldots, \partial f(\boldsymbol{a})/\partial x_d)$ is the gradient of f at $\boldsymbol{a}$.

- $\tilde{f}(\boldsymbol{b}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})(\boldsymbol{b} - \boldsymbol{a})$ is a linear approximation of $f(\boldsymbol{x})$ in the neighborhood $B(\boldsymbol{a}; r)$ of $\boldsymbol{a}$.
  - If $(\boldsymbol{b} - \boldsymbol{a})$ is in the opposite direction of the gradient $\nabla f(\boldsymbol{a})$, $\tilde{f}$ has the greatest rate of decrease from the point $\boldsymbol{a}$.
  - But we have to control the length $\|\boldsymbol{b} - \boldsymbol{a}\|$, otherwise $\tilde{f}$ will not be a good approximation of f.

- A minimization problem:

$$\begin{aligned} \text{Minimize} \quad F(\boldsymbol{b}) &= \tfrac{1}{2}\|\boldsymbol{b}-\boldsymbol{a}\|^2 + \eta \tilde{f}(\boldsymbol{b}) \\ &= \tfrac{1}{2}\|\boldsymbol{b}-\boldsymbol{a}\|^2 + \eta(f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})(\boldsymbol{b}-\boldsymbol{a})) \\ \text{Subject to} \quad & \boldsymbol{b} \in B(\boldsymbol{a}; r). \end{aligned}$$

  - The first term $\frac{1}{2}\|\boldsymbol{b}-\boldsymbol{a}\|^2$ is the regularization term.
  - The parameter $\eta > 0$ controls the tradeoff between the two terms.

- $\boldsymbol{b}^* = \boldsymbol{a} - \eta \nabla f(\boldsymbol{a})^T$: the optimal $\boldsymbol{b}$ which minimizes the object function $F(\boldsymbol{b})$, since

$$F(\boldsymbol{b}) = \frac{1}{2}\|(\boldsymbol{b}-\boldsymbol{a}) + \eta \nabla f(\boldsymbol{a})^T\|^2 + \eta f(\boldsymbol{a}) - \frac{\eta^2}{2}\|\nabla f(\boldsymbol{a})^T\|^2$$

  achieves the minimum value if and only if $\boldsymbol{b} - \boldsymbol{a} = -\eta \nabla f(\boldsymbol{a})^T$. Here we assume that $\eta < \frac{r}{\|\nabla f(\boldsymbol{a})\|}$.

- Gradient descent algorithm: With an initial interior point $\boldsymbol{x}^{(1)}$, the recursive update rule is

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \eta \nabla f(\boldsymbol{x}^{(t)})^T, \ \forall \ t \geq 1.$$

  - Assume that $\eta < \frac{r^{(t)}}{\|\nabla f(\boldsymbol{x}^{(t)})\|}$ for all $t \geq 1$, where $B(\boldsymbol{x}^{(t)}; r^{(t)})$ is a neighborhood of $\boldsymbol{x}^{(t)}$ in S.

- Output of the gradient descent (GD) algorithm : After T iterations, the algorithm outputs the averaged vector,

$$\bar{\boldsymbol{x}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}^{(t)}.$$

  - The output could also be the last vector, $\boldsymbol{x}^{(T)}$, or the best performing vector, $\arg\min_{\boldsymbol{x}^{(t)}, 1 \leq t \leq T} f(\boldsymbol{x}^{(t)})$.

Comments

- $f : S \to \mathbb{R}$: a real-valued function defined on an open convex subset S of $\mathbb{R}^d$ such that f and all its first-order partial derivatives $\partial f / \partial x_i$, $1 \leq i \leq d$, are differentiable at each point of S.

- Taylor's formula: for all points $\boldsymbol{b}$ and $\boldsymbol{a}$ in S, we have

$$f(\boldsymbol{b}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})(\boldsymbol{b} - \boldsymbol{a}) + \frac{1}{2}(\boldsymbol{b} - \boldsymbol{a})^T H_f(\boldsymbol{z})(\boldsymbol{b} - \boldsymbol{a})$$

  for some $\boldsymbol{z}$ on the line segment $[\boldsymbol{a}, \boldsymbol{b}]$ joining the two points $\boldsymbol{a}$ and $\boldsymbol{b}$, where and $H_f(\boldsymbol{z}) = [\partial f(\boldsymbol{z}) / \partial x_i \partial x_j]$ is the Hessian matrix of f at $\boldsymbol{z}$.

- f is convex on S if and only if the Hessian matrix $H_f(\boldsymbol{x})$ of f at every point $\boldsymbol{x}$ in S is positive semi-definite, i.e., $\boldsymbol{v}^T H_f(\boldsymbol{x}) \boldsymbol{v} \geq 0$ for all $\boldsymbol{v} \in \mathbb{R}^d$.

- If f is convex on an open convex subset S of $\mathbb{R}^d$ such that f and all its first-order partial derivatives $\partial f / \partial x_i$, $1 \leq i \leq d$, are differentiable at each point of S, then we have

$$f(\boldsymbol{b}) \geq f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})(\boldsymbol{b} - \boldsymbol{a}).$$

- The inequality in above is generally true even for a non-differentiable convex function f.
- We will generalize the concept of gradient $\nabla f(\boldsymbol{a})$ in the following.

Epigraphs and Convexity

- $f : S \to \mathbb{R}$: a real-valued function defined on a subset S of $\mathbb{R}^n$.
- $\{(\boldsymbol{x}, f(\boldsymbol{x})) \mid \boldsymbol{x} \in S\}$: the graph of the function f, which is a subset of $\mathbb{R}^{n+1}$.
- $\{(\boldsymbol{x}, y) \mid \boldsymbol{x} \in S, y \in \mathbb{R}, y \geq f(\boldsymbol{x})\}$: the epigraph of the function f.
- $\{(\boldsymbol{x}, y) \mid \boldsymbol{x} \in S, y \in \mathbb{R}, y \leq f(\boldsymbol{x})\}$: the hypograph of the function f.

Theorem 1: Let $f : S \to \mathbb{R}$ be a real-valued function defined on a convex subset S of $\mathbb{R}^n$. Then f is convex if and only if the epigraph epi f of f is a convex subset of $\mathbb{R}^{n+1}$.

Proof.

”$\Rightarrow$” Let $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2)$ be in the epi f, i.e., $f(\boldsymbol{x}_1) \leq y_1$ and $f(\boldsymbol{x}_2) \leq y_2$. Consider any point $\lambda(\boldsymbol{x}_1, y_1) + (1-\lambda)(\boldsymbol{x}_2, y_2) = (\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2, \lambda y_1 + (1-\lambda)y_2)$ on the line segment $[(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2)]$, $0 \leq \lambda \leq 1$. Then we have

$$f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \leq \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2) \leq \lambda y_1 + (1-\lambda) y_2,$$

which shows that the point $\lambda(\boldsymbol{x}_1, y_1) + (1-\lambda)(\boldsymbol{x}_2, y_2)$ is in the epigraph of f.

”$\Leftarrow$” For any two points $\boldsymbol{x}_1, \boldsymbol{x}_2$ of S, $(\boldsymbol{x}_1, f(\boldsymbol{x}_1))$ and $(\boldsymbol{x}_2, f(\boldsymbol{x}_2))$ are in epi f. Since epi f is convex, $\lambda(\boldsymbol{x}_1, f(\boldsymbol{x}_1)) +$ $(1-\lambda)(\boldsymbol{x}_2, f(\boldsymbol{x}_2)) = (\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2, \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2))$ are also in epi f for all $\lambda \in [0,1]$, which implies that

$$f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \leq \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2) \ \forall\ \lambda \in [0,1].$$

This shows that f is convex. □

- Similarly, a real-valued function $f : S \to \mathbb{R}$ defined on a convex subset S of $\mathbb{R}^n$ is concave if and only if the hypograph hypo f of f is a convex subset of $\mathbb{R}^{n+1}$.

Supporting Hyperplanes of a Set at Boundary Points

Let E be a nonempty subset of $\mathbb{R}^n$ and $\bar{\boldsymbol{x}} \in \partial E$, the boundary of E. A hyperplane $H = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{w} \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}) = 0\}$ in $\mathbb{R}^n$ with weight vector $\boldsymbol{w}$ is called a supporting hyperplane of E at $\bar{\boldsymbol{x}}$ if either $E \subseteq H^+$ such that $\boldsymbol{w} \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0$ for all $\boldsymbol{x} \in S$ or $E \subseteq H^-$ such that $\boldsymbol{w} \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq 0$ for all $\boldsymbol{x} \in S$.

Existence of Supporting Hyperplanes of a Convex Set at Boundary Points

Theorem 2: Let E be a nonempty convex subset of $\mathbb{R}^n$ and $\bar{\boldsymbol{x}} \in \partial E$. There exists a hyperplane that supports E at $\bar{\boldsymbol{x}}$, i.e., there is a nonzero vector $\boldsymbol{w}$ in $\mathbb{R}^n$ such that $\boldsymbol{w} \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq 0$ for all $\boldsymbol{x} \in \mathrm{cl}E$, the closure of E.

Proof.

- Since $\bar{\boldsymbol{x}} \in \partial E$, there exists a sequence $\{\boldsymbol{y}_k\}$ not in cl$E$ such that $\boldsymbol{y}_k \to \bar{\boldsymbol{x}}$.
- Since cl$E$ is a closed convex set, there is a unique point $\bar{\boldsymbol{x}}_k$ in cl$E$ with minimum distance to $\boldsymbol{y}_k$ and $\bar{\boldsymbol{x}}_k$ is the minimizing point if and only if $(\boldsymbol{x} - \bar{\boldsymbol{x}}_k) \cdot (\boldsymbol{y}_k - \bar{\boldsymbol{x}}_k) \leq 0$ for all $\boldsymbol{x}$ in cl$E$. [a]
- Let $\boldsymbol{w}_k \triangleq (\boldsymbol{y}_k - \bar{\boldsymbol{x}}_k)/\|\boldsymbol{y}_k - \bar{\boldsymbol{x}}_k\|$. Then we have $\boldsymbol{w}_k \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}_k) \leq 0$ which implies that $\boldsymbol{w}_k \cdot \boldsymbol{x} \leq \boldsymbol{w}_k \cdot \bar{\boldsymbol{x}}_k \triangleq \alpha$ for all $\boldsymbol{x}$ in cl$E$. Also $\boldsymbol{w}_k \cdot \boldsymbol{y}_k - \alpha = \boldsymbol{w}_k \cdot (\boldsymbol{y}_k - \bar{\boldsymbol{x}}_k) = \|(\boldsymbol{y}_k - \bar{\boldsymbol{x}}_k\| > 0$ which implies that $\boldsymbol{w}_k \cdot \boldsymbol{y}_k > \alpha$.
- Since $\{\boldsymbol{w}_k\}$ is bounded, it has a convergent subsequence $\{\boldsymbol{w}_{k_i}\}$ with limit $\boldsymbol{w}$ which is also a unit vector.

[a] See Theorem 2.4.1 in M.S. Bzaraa, H.D. Sherali, and C.M. Shetty, *Nonlinear Programming: Theory and Algorithm,* 3rd edn., John Wiley and Sons, 2006, pp. 50-51.

- For all i, we have $\boldsymbol{w}_{k_i} \cdot \boldsymbol{y}_{k_i} > \boldsymbol{w}_{k_i} \cdot \boldsymbol{x}$ for all $\boldsymbol{x}$ in cl$E$.
- Fix an $\boldsymbol{x}$ in cl$E$ and let $i \to \infty$. We have $\lim_{i\to\infty} \boldsymbol{w}_{k_i} = \boldsymbol{w}$, $\lim_{i\to\infty} \boldsymbol{y}_{k_i} = \bar{\boldsymbol{x}}$ and then $\boldsymbol{w} \cdot \bar{\boldsymbol{x}} \geq \boldsymbol{w} \cdot \boldsymbol{x}$.
- Now it is true that $\boldsymbol{w} \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq 0$ for all $\boldsymbol{x}$ in cl$E$. □

Subgradients of a Convex Function

Let $f : S \to \mathbb{R}$ be a convex function defined on a convex subset S of $\mathbb{R}^n$. A vector $\boldsymbol{w}$ in $\mathbb{R}^n$ is called a subgradient of f at a point $\boldsymbol{a}$ in S if

$$f(\boldsymbol{x}) \geq f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{a}) \ \forall\ \boldsymbol{x} \in S.$$

- $y = f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{a})$, i.e., $(\boldsymbol{w}, -1)((\boldsymbol{x}, y) - (\boldsymbol{a}, f(\boldsymbol{a}))) = 0$ is a supporting hyperplane of the epigraph of f at $(\boldsymbol{a}, f(\boldsymbol{a}))$ in $\mathbb{R}^{n+1}$.

- The collection of all subgradients of a convex function f at a point $\boldsymbol{a}$ in a convex set S is a convex subset of $\mathbb{R}^n$ and is called the differential set of f at $\boldsymbol{a}$, denoted as $\partial f(\boldsymbol{a})$.

Existence of Subgradients of a Convex Function at Interior Points of Its Defining Convex Set

Theorem 3: Let $f : S \to \mathbb{R}$ be a convex function defined on a nonempty convex subset S of $\mathbb{R}^n$. Then for each interior point $\boldsymbol{a}$ of S, there exists a vector $\boldsymbol{w}$ in $\mathbb{R}^n$ such that the hyperplane

$$H = \{(\boldsymbol{x}, y) \in \mathbb{R}^{n+1} \mid y = f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{a})\}$$

supports epi$f$ at $(\boldsymbol{a}, f(\boldsymbol{a}))$. That is, $\boldsymbol{w}$ is a subgradient of f at $\boldsymbol{a}$, i.e.,

$$f(\boldsymbol{x}) \geq f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{a}) \ \forall \ \boldsymbol{x} \in S.$$

Proof.

- By Theorem 1, the epigraph epi$f$ of f is a convex set in $\mathbb{R}^{n+1}$.
- For a point $\boldsymbol{a}$ in S, $(\boldsymbol{a}, f(\boldsymbol{a}))$ is on the boundary of epi$f$ so that there is a supporting hyperplane of epi$f$ at the boundary point $(\boldsymbol{a}, f(\boldsymbol{a}))$ in $\mathbb{R}^{n+1}$ by Theorem 2.
- That is, there is a vector $\boldsymbol{w}'$ in $\mathbb{R}^n$ and a scalar ζ, not both zero, such that

$$(\boldsymbol{w}', \zeta)\cdot((\boldsymbol{x}, y)-(\boldsymbol{a}, f(\boldsymbol{a})) = \boldsymbol{w}'\cdot(\boldsymbol{x}-\boldsymbol{a})+\zeta(y-f(\boldsymbol{a})) \leq 0 \,\forall\, (\boldsymbol{x}, y) \in \mathrm{epi} f.$$

- ζ cannot be positive. Otherwise, by letting $y \to \infty$, the inequality will be violated.
- Suppose $\zeta = 0$. Then $\boldsymbol{w}' \neq \mathbf{0}$ and we have $\boldsymbol{w}'(\boldsymbol{x} - \boldsymbol{a}) \leq 0$ for all $\boldsymbol{x} \in S$.

- Since $\boldsymbol{a}$ is an interior point of S, there is a neighborhood $B(\boldsymbol{a}; r)$ of $\boldsymbol{a}$ in S. Taking $\boldsymbol{x} = \boldsymbol{a} + \epsilon \boldsymbol{w}'$ in this neighborhood with $\epsilon > 0$ sufficiently small, we have

$$\epsilon \|\boldsymbol{w}'\|^2 \leq 0,$$

a contradiction.

- We conclude that $\zeta < 0$. And by dividing $|\zeta|$, we have

$$\frac{\boldsymbol{w}'}{|\zeta|} \cdot (\boldsymbol{x} - \boldsymbol{a}) - (y - f(\boldsymbol{a})) \leq 0 \ \forall \ (\boldsymbol{x}, y) \in \mathrm{epi} f$$

and then

$$y \geq f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{a}) \ \forall \ (\boldsymbol{x}, y) \in \mathrm{epi} f,$$

where $\boldsymbol{w} = \boldsymbol{w}' / |\zeta|$. In particular,

$$f(\boldsymbol{x}) \geq f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{a}) \ \forall \ \boldsymbol{x} \in S,$$

which shows that $\boldsymbol{w}$ is a subgradient of f at $\boldsymbol{a}$. □

A Corollary

Let $f : S \to \mathbb{R}$ be a strictly convex function defined on a nonempty convex subset S of $\mathbb{R}^n$. Then for each interior point $\boldsymbol{a}$ of S, there exists a vector $\boldsymbol{w}$ in $\mathbb{R}^n$ such that

$$f(\boldsymbol{x}) > f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{a}) \;\forall\, \boldsymbol{x} \in S,\; \boldsymbol{x} \neq \boldsymbol{a}.$$

Proof.

- By Theorem 3, there exists a subgradient vector $\boldsymbol{w}$ in $\mathbb{R}^n$,

$$f(\boldsymbol{x}) \geq f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{a}) \ \forall \ \boldsymbol{x} \in S.$$

- Suppose that there is a point $\boldsymbol{b}$ in S, $\boldsymbol{b} \not\boldsymbol{a}$, such that

$$f(\boldsymbol{b}) = f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{b} - \boldsymbol{a}).$$

- Since f is strictly convex,

$$f(\lambda \boldsymbol{b} + (1-\lambda)\boldsymbol{a}) < \lambda f(\boldsymbol{b}) + (1-\lambda) f(\boldsymbol{a}) = f(\boldsymbol{a}) + \lambda \boldsymbol{w} \cdot (\boldsymbol{b} - \boldsymbol{a}),$$

- Since $\boldsymbol{w}$ is a subgradient vector of f at $\boldsymbol{a}$, we have

$$f(\lambda \boldsymbol{b} + (1-\lambda)\boldsymbol{a}) \geq f(\boldsymbol{a}) + \lambda \boldsymbol{w} \cdot (\boldsymbol{b} - \boldsymbol{a}),$$

  which is a contradiction to the previous inequality.

- Thus we have $f(\boldsymbol{x}) > f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{a})$ for all $\boldsymbol{x}$ in S and $\boldsymbol{x} \neq \boldsymbol{a}$. □

A Converse

Theorem 4: Let $f : S \to \mathbb{R}$ be a function defined on a nonempty convex subset S of $\mathbb{R}^n$. If, for each interior point $\boldsymbol{a}$ of S, there exists a subgradient vector $\boldsymbol{w}$ in $\mathbb{R}^n$ such that

$$f(\boldsymbol{x}) \geq f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{a}) \;\forall\; \boldsymbol{x} \in S,$$

then f is convex on the interior of S.

Proof.

- Since S is convex, the interior int$S$ of S is also convex.
- Let $\boldsymbol{x}_1, \boldsymbol{x}_2$ be in int$S$. Then $\lambda \boldsymbol{x}_1 + (1 - \lambda)\boldsymbol{x}_2$ is also in int$S$ for $\lambda \in (0, 1)$.

- There is a subgradient $\boldsymbol{w}$ of f at $\lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2$ so that

$$\begin{aligned} f(\boldsymbol{x}_1) &\geq f(\lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) + (1-\lambda)\boldsymbol{w}\cdot(\boldsymbol{x}_1 - \boldsymbol{x}_2) \\ f(\boldsymbol{x}_2) &\geq f(\lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) - \lambda\boldsymbol{w}\cdot(\boldsymbol{x}_1 - \boldsymbol{x}_2). \end{aligned}$$

- Multiplying the first inequality by λ and the second by $(1-\lambda)$ and adding them, we have

$$\lambda f(\boldsymbol{x}_1) + (1-\lambda)f(\boldsymbol{x}_2) \geq f(\lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2),$$

which proves that f is convex on the interior of S. □

Comment

- Theorem 4 cannot be extended to show that f is convex on the whole convex set S.
- Example: On $S = \{(x_1, x_2) \mid 0 \leq x_1, x_2 \leq 1\}$, define a function

$$f(x_1, x_2) = \begin{cases} 0, & 0 \leq x_1 \leq 1, 0 < x_2 \leq 1, \\ \frac{1}{4} - (x_1 - \frac{1}{2})^2, & 0 \leq x_1 \leq 1, x_2 = 0. \end{cases}$$

  – f is 0 in the interior of S so that f is convex on int$S$ and f has a sub gradient at each interior point, which is the zero vector.

  – f is not convex on S.

Examples of Subgradients

- Let $f : S \to \mathbb{R}$ be a convex function defined on a nonempty convex subset S of $\mathbb{R}^n$. If f is differentiable at an interior point $\boldsymbol{a}$ of S, then the differential set of f at $\boldsymbol{a}$ is a singleton,

$$\partial f(\boldsymbol{a}) = \{\nabla f(\boldsymbol{a})\}.$$

Proof. Let $\boldsymbol{w}$ be a subgradient of f at $\boldsymbol{a}$, i.e.,

$$f(\boldsymbol{x}) \quad \geq \quad f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{a}), \ \forall \ \boldsymbol{x} \in S.$$

Since f is differentiable at $\boldsymbol{a}$, we have

$$f(\boldsymbol{x}) \quad = \quad f(\boldsymbol{a}) + \nabla f(\boldsymbol{a}) \cdot (\boldsymbol{x} - \boldsymbol{a}) + o(\|\boldsymbol{x} - \boldsymbol{a}\|).$$

Choose a nonzero vector $\boldsymbol{v}$ in $\mathbb{R}^n$. For sufficiently small $\epsilon > 0$, $\boldsymbol{a} + \epsilon\boldsymbol{v}$ is in a neighborhood of $\boldsymbol{a}$ in S. Then we have

$$\begin{aligned} f(\boldsymbol{a} + \epsilon\boldsymbol{v}) &\geq f(\boldsymbol{a}) + \epsilon\boldsymbol{w} \cdot \boldsymbol{v}, \\ f(\boldsymbol{a} + \epsilon\boldsymbol{v}) &= f(\boldsymbol{a}) + \epsilon\nabla f(\boldsymbol{a}) \cdot \boldsymbol{v} + o(\|\epsilon\boldsymbol{v}\|). \end{aligned}$$

By subtraction, we have

$$0 \geq \epsilon(\boldsymbol{w} - \nabla f(\boldsymbol{a})) \cdot \boldsymbol{v} + o(\|\epsilon\boldsymbol{v}\|).$$

Dividing by ϵ and letting $\epsilon \to 0$, we have

$$0 \geq (\boldsymbol{w} - \nabla f(\boldsymbol{a})) \cdot \boldsymbol{v}.$$

Suppose that $\boldsymbol{w} \neq \nabla f(\boldsymbol{a})$. By letting $\boldsymbol{v} = \boldsymbol{w} - \nabla f(\boldsymbol{a})$, we have

$$0 \geq \|\boldsymbol{w} - \nabla f(\boldsymbol{a})\|^2 > 0,$$

a contradiction. We conclude that $\boldsymbol{w} = \nabla f(\boldsymbol{a})$. □

- For $f(x) = |x|, x \in \mathbb{R}$, we have

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

- Let $f(\boldsymbol{x}) = \max_{1 \le i \le k} f_i(\boldsymbol{x})$, where f_i's are k convex functions on a convex subset S of $\mathbb{R}^n$. If there is an interior point $\boldsymbol{a}$ of S such that $j \in \arg\max_{1 \le i \le k} f_i(\boldsymbol{a})$ and f_j is differentiable at $\boldsymbol{a}$, then

$$\nabla f_j(\boldsymbol{a}) \in \partial f(\boldsymbol{a}).$$

Proof. Since f_j is a convex function on a convex set S and differentiable at $\boldsymbol{a}$, we have

$$f_j(\boldsymbol{x}) \ge f_j(\boldsymbol{a}) + \nabla f_j(\boldsymbol{a}) \cdot (\boldsymbol{x} - \boldsymbol{a}) \ \forall \ \boldsymbol{x} \in S.$$

Since $f(\boldsymbol{a}) = f_j(\boldsymbol{a})$, we have

$$f(\boldsymbol{x}) \geq f_j(\boldsymbol{x}) \geq f(\boldsymbol{a}) + \nabla f_j(\boldsymbol{a}) \cdot (\boldsymbol{x} - \boldsymbol{a}) \; \forall \; \boldsymbol{x} \in S,$$

which implies that $\nabla f_j(\boldsymbol{a})$ is a subgradient of f at $\boldsymbol{a}$. □

- $f(\boldsymbol{x}) = \max(0, 1 - \eta \boldsymbol{w} \cdot \boldsymbol{x}), \; \forall \; \boldsymbol{x} \in \mathbb{R}^n \; :$ the hinge loss function for some vector $\boldsymbol{w}$ and scalar η. For all points $\boldsymbol{x}$ such that $1 - \eta \boldsymbol{w} \cdot \boldsymbol{x} \leq 0$, we have

$$\mathbf{0} \in \partial f(\boldsymbol{x}).$$

For all points $\boldsymbol{x}$ such that $1 - \eta \boldsymbol{w} \cdot \boldsymbol{x} > 0$, we have

$$-\eta \boldsymbol{w} \in \partial f(\boldsymbol{x}).$$

Lipschitz Functions

A real-valued function $f : S \to \mathbb{R}$ defined on a subset S of $\mathbb{R}^n$ is call ρ-Lipschitz, $\rho > 0$, if

$$|f(\boldsymbol{y}) - f(\boldsymbol{x})| \leq \rho \|\boldsymbol{y} - \boldsymbol{x}\| \; \forall \; \boldsymbol{x}, \boldsymbol{y} \in S.$$

Comment

- Mean-value theorem: Let $\boldsymbol{f} : S \to \mathbb{R}^m$ be a differentiable mapping defined on an open subset S of $\mathbb{R}^n$. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two points in S such that $[\boldsymbol{x}, \boldsymbol{y}] \subseteq S$. Then for every vector $\boldsymbol{a}$ in $\mathbb{R}^m$, there is a point $\boldsymbol{z} \in [\boldsymbol{x}, \boldsymbol{y}]$ such that

$$\boldsymbol{a} \cdot (\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{a} \cdot (\boldsymbol{f}'(\boldsymbol{z})(\boldsymbol{y} - \boldsymbol{x})).$$

  - $\boldsymbol{z}$ depends on $\boldsymbol{a}$.

- Furthermore, if S is convex and all the partial derivatives $\partial f_i / \partial x_j$ are bounded on S, then there is a constant A such that

$$\| \boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x}) \| \leq A \| \boldsymbol{y} - \boldsymbol{x} \| \;\forall\; \boldsymbol{x}, \boldsymbol{y} \in S,$$

which says that the mapping $\boldsymbol{f}$ is Lipschitz.

Proof. Let $\boldsymbol{a} = \boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x})$. Then we have

$$\begin{aligned} \|\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x})\|^2 &= (\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x})) \cdot (\boldsymbol{f}'(\boldsymbol{z})(\boldsymbol{y} - \boldsymbol{x})) \\ &\leq \|\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x})\| \|\boldsymbol{f}'(\boldsymbol{z})(\boldsymbol{y} - \boldsymbol{x})\|, \end{aligned}$$

which says that $\|\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x})\| \leq \|\boldsymbol{f}'(\boldsymbol{z})(\boldsymbol{y} - \boldsymbol{x})\|$. Note that $\boldsymbol{f} = [f_1, \ldots, f_m]^T$ and $\boldsymbol{f}'(\boldsymbol{z}) = [\nabla f_1(\boldsymbol{z}), \ldots, \nabla f_m(\boldsymbol{z})]^T$. But

$$\begin{aligned} \|\boldsymbol{f}'(\boldsymbol{z})(\boldsymbol{y} - \boldsymbol{x})\| &= \|\boldsymbol{f}'(\boldsymbol{z})(\boldsymbol{y} - \boldsymbol{x})\| = \|\sum_{j=1}^{m} \nabla f_j(\boldsymbol{z})(\boldsymbol{y} - \boldsymbol{x})\boldsymbol{e}_j\| \\ &\leq \sum_{j=1}^{m} |\nabla f_j(\boldsymbol{z})(\boldsymbol{y} - \boldsymbol{x})| \leq \sum_{j=1}^{m} \|\nabla f_j(\boldsymbol{z})\| \|\boldsymbol{y} - \boldsymbol{x}\| \\ &\leq \left(\sum_{j=1}^{m} \|\nabla f_j(\boldsymbol{z})\| \right) \|\boldsymbol{y} - \boldsymbol{x}\|. \end{aligned}$$

Since all the partial derivatives $\partial f_i / \partial x_j$ are bounded on S, there is an A such that $\sum_{j=1}^{m} \|\nabla f_j(\boldsymbol{x})\| \leq A$ for all $\boldsymbol{x} \in S$. □

Subgradients of a Convex Lipschitz Function

Theorem 5: Let $f : S \to \mathbb{R}$ be a convex function defined on a nonempty open convex subset S of $\mathbb{R}^n$. Then f is ρ-Lipschitz over S if and only if for all $\boldsymbol{a} \in S$ and all $\boldsymbol{w} \in \partial f(\boldsymbol{a})$, we have $\|\boldsymbol{w}\| \leq \rho$.

Proof. "$\Rightarrow$" Let $\boldsymbol{a} \in S$ and $\boldsymbol{w} \in \partial f(\boldsymbol{a})$. Since S is open, $\boldsymbol{a} + \epsilon \boldsymbol{w}$ is in a neighborhood of $\boldsymbol{a}$ in S for sufficiently small $\epsilon > 0$. Since $\boldsymbol{w}$ is a subgradient of f at $\boldsymbol{a}$, we have

$$f(\boldsymbol{a} + \epsilon \boldsymbol{w}) \geq f(\boldsymbol{a}) + \epsilon \boldsymbol{w} \cdot \boldsymbol{w}.$$

Since f is ρ-Lipschitz, we have

$$|f(\boldsymbol{a} + \epsilon \boldsymbol{w}) - f(\boldsymbol{a})| \leq \rho \|\epsilon \boldsymbol{w}\|$$

and then

$$\epsilon \|\boldsymbol{w}\|^2 \leq |f(\boldsymbol{a} + \epsilon \boldsymbol{w}) - f(\boldsymbol{a})| \leq \rho \epsilon \|\boldsymbol{w}\|,$$

which shows that $\|\boldsymbol{w}\| \leq \rho$.

”$\Leftarrow$” Let $\boldsymbol{a}, \boldsymbol{b}$ be in S with subgradients $\boldsymbol{w}$ and $\boldsymbol{u}$ respectively. Then we have

$$\begin{aligned} f(\boldsymbol{b}) &\geq f(\boldsymbol{a}) + \boldsymbol{w} \cdot (\boldsymbol{b} - \boldsymbol{a}), \\ f(\boldsymbol{a}) &\geq f(\boldsymbol{b}) + \boldsymbol{u} \cdot (\boldsymbol{a} - \boldsymbol{b}) \end{aligned}$$

which implies that

$$\begin{aligned} f(\boldsymbol{a}) - f(\boldsymbol{b}) &\leq \boldsymbol{w} \cdot (\boldsymbol{a} - \boldsymbol{b}) \leq \|\boldsymbol{w}\| \|\boldsymbol{a} - \boldsymbol{b}\| \leq \rho \|\boldsymbol{a} - \boldsymbol{b}\|, \\ f(\boldsymbol{b}) - f(\boldsymbol{a}) &\leq \boldsymbol{u} \cdot (\boldsymbol{b} - \boldsymbol{a}) \leq \|\boldsymbol{u}\| \|\boldsymbol{a} - \boldsymbol{b}\| \leq \rho \|\boldsymbol{a} - \boldsymbol{b}\| \end{aligned}$$

so that $|f(\boldsymbol{a}) - f(\boldsymbol{b})| \leq \rho \|\boldsymbol{a} - \boldsymbol{b}\|$. □

Subgradient Descent Algorithm

- $f : S \to \mathbb{R}$: a real-valued convex function on a convex subset S of $\mathbb{R}^n$.

With an initial interior point $\boldsymbol{x}^{(1)}$ of S, the recursive update rule is

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \eta \boldsymbol{v}_t, \ \forall \ t \geq 1,$$

where $\boldsymbol{v}_t$ is a subgradient of f at $\boldsymbol{x}^{(t)}$.

- Assume that $\eta < \frac{r^{(t)}}{\|\boldsymbol{v}_t\|}$ for all $t \geq 1$, where $B(\boldsymbol{x}^{(t)}; r^{(t)})$ is a neighborhood of $\boldsymbol{x}^{(t)}$ in S.
- Output of the subgradient descent (SD) algorithm : After T iterations, the algorithm outputs the averaged vector,

$$\bar{\boldsymbol{x}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}^{(t)}.$$

A Lemma

Lemma 1: Let

- $\boldsymbol{x}^*$: an arbitrary vector in $\mathbb{R}^n$.
- $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_T$: an arbitrary sequence of vectors in $\mathbb{R}^n$.
- $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(T)}$: a sequence of vectors in $\mathbb{R}^n$ generated by a recursive formula

  $$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \eta \boldsymbol{v}_t, \ \forall\, 1 \leq t \leq T-1,$$

  with an arbitrary initial vector $\boldsymbol{x}^{(1)}$ in $\mathbb{R}^n$, where $\eta > 0$ is a constant.

Then we have

$$\sum_{t=1}^{T} (\boldsymbol{x}^{(t)} - \boldsymbol{x}^*) \cdot \boldsymbol{v}_t \leq \frac{\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\boldsymbol{v}_t\|^2.$$

Furthermore, if $\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\| \leq B$ and $\|\boldsymbol{v}_t\| \leq \rho$ for all t for some constants $B, \rho > 0$, by setting $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, we have

$$\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{x}^{(t)} - \boldsymbol{x}^*) \cdot \boldsymbol{v}_t \leq \frac{B\rho}{\sqrt{T}}.$$

Proof. First note that

$$\begin{aligned}
(\boldsymbol{x}^{(t)} - \boldsymbol{x}^*) \cdot \boldsymbol{v}_t = \frac{1}{\eta}(\boldsymbol{x}^{(t)} - \boldsymbol{x}^*) \cdot (\eta \boldsymbol{v}_t) \\
= \quad & \frac{1}{2\eta}(-\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^* - \eta\boldsymbol{v}_t\|\|^2 + \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|^2 + \eta^2\|\boldsymbol{v}_t\|^2) \\
= \quad & \frac{1}{2\eta}(-\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\|\|^2 + \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|^2) + \frac{\eta}{2}\|\boldsymbol{v}_t\|^2.
\end{aligned}$$

Summing over t, we have

$$\begin{aligned}
& \sum_{t=1}^{T} (\boldsymbol{x}^{(t)} - \boldsymbol{x}^*) \cdot \boldsymbol{v}_t \\
= \quad & \frac{1}{2\eta}(\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^*\|\|^2) + \frac{\eta}{2}\sum_{t=1}^{T}\|\boldsymbol{v}_t\|^2 \\
\leq \quad & \frac{\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\boldsymbol{v}_t\|^2.
\end{aligned}$$

With $\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\| \leq B$ and $\|\boldsymbol{v}_t\| \leq \rho$ for all t, we have

$$\frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{x}^{(t)} - \boldsymbol{x}^*) \cdot \boldsymbol{v}_t \leq \frac{B^2}{2\eta T} + \frac{\eta\rho^2}{2}.$$

The upper bound is minimized if and only if $\eta = \sqrt{\frac{B^2}{T\rho^2}}$. The minimum value of the upper bound is $\frac{B\rho}{\sqrt{T}}$. □

Convergence Rate of Subgradient Descent Algorithm for Convex-Lipschitz Functions

Theorem 6: Let

- $f : S \to \mathbb{R}$: a convex ρ-Lipschitz function defined on a nonempty open convex subset S of $\mathbb{R}^n$;
- $\bar{B}(\boldsymbol{x}_0; B) \subseteq S$ for some point $\boldsymbol{x}_0$ in S and some $B > 0$;
- $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in \bar{B}(\boldsymbol{x}_0; B)} f(\boldsymbol{x})$.

If we run the subgradient descent algorithm on f for T steps with the initial point $\boldsymbol{x}^{(1)} = \boldsymbol{x}_0$ and the step size $\eta = \sqrt{\frac{B^2}{T\rho^2}}$, then the output vector $\bar{\boldsymbol{x}}$ has

$$f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq \frac{B\rho}{\sqrt{T}}.$$

Furthermore, for every $\epsilon > 0$, to achieve $f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq \epsilon$, it suffices to run the subgradient descent algorithm for a number of iterations that satisfies

$$T \geq \frac{B^2 \rho^2}{\epsilon^2}.$$

Proof.

$$\begin{aligned} f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}^*) &= f\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}^{(t)}\right) - f(\boldsymbol{x}^*) \\ &\leq \frac{1}{T}\sum_{t=1}^{T} f\left(\boldsymbol{x}^{(t)}\right) - f(\boldsymbol{x}^*) \text{ by the convexity of } f \\ &= \frac{1}{T}\sum_{t=1}^{T} \left(f\left(\boldsymbol{x}^{(t)}\right) - f(\boldsymbol{x}^*)\right) \leq \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{v}_t \cdot (\boldsymbol{x}^{(t)} - \boldsymbol{x}^*), \end{aligned}$$

since $\boldsymbol{v}_t$ is a subgradient of f at $\boldsymbol{x}^{(t)}$. Since f is a convex ρ-Lipschitz function on a nonempty open convex subset, all subgradients at any point of S has the norm $\leq \rho$ by Theorem 5. In particular, $\|\boldsymbol{v}_t\| \leq \rho$

for all $t \geq 1$. Since $\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\| \leq B$ and $\eta = \sqrt{\frac{B^2}{T\rho^2}}$, we have

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{v}_t \cdot \left(\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right) \leq \frac{B\rho}{\sqrt{T}}$$

by Lemma 1 and then $f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq \frac{B\rho}{\sqrt{T}}$. □

- We have implicitly assumed that $\boldsymbol{x}^{(t)} \in S$ for all $t \geq 1$.

Stochastic Subgradient Descent Algorithm

```
StochasticSubgradientDescent($\boldsymbol{x}_0, T, \eta$)
  1. $\boldsymbol{x}^{(1)} \leftarrow \boldsymbol{x}_0$          $\triangleright$ $\boldsymbol{x}_0$ initial point
  2. for $t \leftarrow 1$ to $T$ do
  3.       $\boldsymbol{v}_t \leftarrow$ RandomSubgradient($\boldsymbol{x}^{(t)}$)   $\triangleright$ $E[\boldsymbol{v}_t \mid \boldsymbol{x}^{(t)}] \in \partial f(\boldsymbol{x}^{(t)})$
  4.       $\boldsymbol{x}^{(t+1)} \leftarrow \boldsymbol{x}^{(t)} - \eta \boldsymbol{v}_t$
  5. return $\bar{\boldsymbol{x}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}^{(t)}$
```

Convergence Rate of Stochastic Subgradient Descent Algorithm for Convex Functions

Theorem 7: Let

- $f : S \to \mathbb{R}$: a convex function defined on a nonempty open convex subset S of $\mathbb{R}^n$;
- $\bar{B}(\boldsymbol{x}_0; B) \subseteq S$ for some point $\boldsymbol{x}_0$ in S and some $B > 0$;
- $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in \bar{B}(\boldsymbol{x}_0; B)} f(\boldsymbol{x})$.

Assume that when we run the stochastic subgradient descent algorithm,

- the random subgradient $\boldsymbol{v}_t$ generated by RandomSubgradient at the tth iteration has $E[\boldsymbol{v}_t \mid \boldsymbol{x}^{(t)}] \in \partial f(\boldsymbol{x}^{(t)})$;
- $E[\|\boldsymbol{v}_t\|^2] \le \rho^2$ for all $t \ge 1$.

If we run the stochastic subgradient descent algorithm on f for T steps with the initial point $\boldsymbol{x}^{(1)} = \boldsymbol{x}_0$ and the step size $\eta = \sqrt{\frac{B^2}{T\rho^2}}$, then the output vector $\bar{\boldsymbol{x}}$ has

$$E[f(\bar{\boldsymbol{x}})] - f(\boldsymbol{x}^*) \leq \frac{B\rho}{\sqrt{T}}.$$

Furthermore, for every $\epsilon > 0$, to achieve $E[f(\bar{\boldsymbol{x}})] - f(\boldsymbol{x}^*) \leq \epsilon$, it suffices to run the stochastic subgradient descent algorithm for a number of iterations that satisfies

$$T \geq \frac{B^2\rho^2}{\epsilon^2}.$$

Proof. As in the proof of Theorem 6, we have

$$E[f(\bar{\boldsymbol{x}})] - f(\boldsymbol{x}^*) = E\left[f\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}^{(t)}\right)\right] - f(\boldsymbol{x}^*)$$

$$\leq \quad E\left[\frac{1}{T}\sum_{t=1}^{T} f\left(\boldsymbol{x}^{(t)}\right)\right] - f(\boldsymbol{x}^*) \text{ by the convexity of } f$$

$$= \quad \frac{1}{T}\sum_{t=1}^{T} E\left[\left(f\left(\boldsymbol{x}^{(t)}\right) - f(\boldsymbol{x}^*)\right)\right].$$

Since $E[\boldsymbol{v}_t \mid \boldsymbol{x}^{(t)}] \in \partial f(\boldsymbol{x}^{(t)})$, we have

$$f\left(\boldsymbol{x}^{(t)}\right) - f(\boldsymbol{x}^*) \leq E[\boldsymbol{v}_t \mid \boldsymbol{x}^{(t)}] \cdot \left(\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right)$$

so that

$$\begin{aligned}
& E\left[\left(f\left(\boldsymbol{x}^{(t)}\right)-f(\boldsymbol{x}^*)\right)\right] \\
\leq\ & E\left[E\left[\boldsymbol{v}_t \mid \boldsymbol{x}^{(t)}\right]\cdot\left(\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right)\right] \\
=\ & E\left[E\left[\boldsymbol{v}_t\cdot\left(\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right)\ \mid \boldsymbol{x}^{(t)}\right]\right]=E\left[\boldsymbol{v}_t\cdot\left(\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right)\right].
\end{aligned}$$

Now we have

$$\begin{aligned}
E[f(\bar{\boldsymbol{x}})]-f(\boldsymbol{x}^*) &\leq \frac{1}{T}\sum_{t=1}^{T}E\left[\boldsymbol{v}_t\cdot(\boldsymbol{x}^{(t)}-\boldsymbol{x}^*)\right] \\
&= E\left[\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{v}_t\cdot(\boldsymbol{x}^{(t)}-\boldsymbol{x}^*)\right] \\
&\leq \frac{\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^*\|^2}{2\eta T}+\frac{\eta}{2T}\sum_{t=1}^{T}E\left[\|\boldsymbol{v}_t\|^2\right]\ \text{ by Lemma 1.}
\end{aligned}$$

Since $\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|^2 \leq B^2$ and $E\left[\|\boldsymbol{v}_t\|^2\right] \leq \rho^2$ for all $t \geq 1$, we have

$$E[f(\bar{\boldsymbol{x}})] - f(\boldsymbol{x}^*) \leq \frac{B^2}{2\eta T} + \frac{\eta\rho^2}{2}.$$

The upper bound is minimized when setting $\eta = \sqrt{\frac{B^2}{T\rho^2}}$. Thus we have $f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq \frac{B\rho}{\sqrt{T}}$. □

- We have implicitly assumed that $\boldsymbol{x}^{(t)} \in S$ for all $t \geq 1$ w.p.1.

Stochastic Subgradient Descent with Projection Algorithm

```
StochasticSubgradientDescentProjection($\boldsymbol{x}_0, T, \eta$)
$\boldsymbol{x}^{(1)} \leftarrow \boldsymbol{x}_0$          ▷ $\boldsymbol{x}_0$ initial point
for $t \leftarrow 1$ to $T$ do
    $\boldsymbol{v}_t \leftarrow$ RandomSubgradient($\boldsymbol{x}^{(t)}$)   ▷ $E[\boldsymbol{v}_t \mid \boldsymbol{x}^{(t)}] \in \partial f(\boldsymbol{x}^{(t)})$
    $\boldsymbol{x}^{(t+\frac{1}{2})} \leftarrow \boldsymbol{x}^{(t)} - \eta \boldsymbol{v}_t$
    $\boldsymbol{x}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{x} \in \bar{B}(\boldsymbol{x}_0; B)} \|\boldsymbol{x} - \boldsymbol{x}^{(t+\frac{1}{2})}\|$  ▷ projection to $\bar{B}(\boldsymbol{x}_0; B)$
return $\bar{\boldsymbol{x}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}^{(t)}$
```

Projection Lemma

Lemma 2: Let

- $\mathcal{H}$: a closed convex subset of $\mathbb{R}^n$;
- $\boldsymbol{w}$: an arbitrary vector in $\mathbb{R}^n$;
- $\boldsymbol{h}^*$: the projection of $\boldsymbol{w}$ onto $\mathcal{H}$, i.e.,

$$\boldsymbol{h}^* = \arg\min_{\boldsymbol{h}\in\mathcal{H}} \|\boldsymbol{h} - \boldsymbol{w}\|.$$

Then we have

$$\|\boldsymbol{w} - \boldsymbol{h}\| \geq \|\boldsymbol{h}^* - \boldsymbol{h}\| \;\forall\; \boldsymbol{h} \in \mathcal{H}.$$

Proof. By the convexity of $\mathcal{H}$, $\boldsymbol{h}^* + \lambda(\boldsymbol{h} - \boldsymbol{h}^*)$ is in $\mathcal{H}$ for all $\lambda \in (0,1)$. From the optimality of $\boldsymbol{h}^*$, we have

$$\begin{aligned} \|\boldsymbol{h}^* - \boldsymbol{w}\|^2 &\leq \|\boldsymbol{h}^* + \lambda(\boldsymbol{h} - \boldsymbol{h}^*) - \boldsymbol{w}\|^2 \\ &= \|\boldsymbol{h}^* - \boldsymbol{w}\|^2 + 2\lambda(\boldsymbol{h}^* - \boldsymbol{w}) \cdot (\boldsymbol{h} - \boldsymbol{h}^*) + \lambda^2\|\boldsymbol{h} - \boldsymbol{h}^*\|^2, \end{aligned}$$

which implies that

$$(\boldsymbol{h}^* - \boldsymbol{w}) \cdot (\boldsymbol{h} - \boldsymbol{h}^*) \geq -\frac{\lambda}{2}\|\boldsymbol{h} - \boldsymbol{h}^*\|^2.$$

By letting $\lambda \to 0$, we have

$$(\boldsymbol{h}^* - \boldsymbol{w}) \cdot (\boldsymbol{h} - \boldsymbol{h}^*) \geq 0.$$

Therefore,

$$\begin{aligned} & \|\boldsymbol{w} - \boldsymbol{h}\|^2 = \|\boldsymbol{w} - \boldsymbol{h}^* + \boldsymbol{h}^* - \boldsymbol{h}\|^2 \\ = \; & \|\boldsymbol{w} - \boldsymbol{h}^*\|^2 + 2(\boldsymbol{w} - \boldsymbol{h}^*) \cdot (\boldsymbol{h}^* - \boldsymbol{h}) + \|\boldsymbol{h}^* - \boldsymbol{h}\|^2 \geq \|\boldsymbol{h}^* - \boldsymbol{h}\|^2. \end{aligned}$$

□

Comments

- In the tth iteration of the stochastic subgradient descent with projection algorithm, we have

$$\begin{aligned} & -\|\boldsymbol{x}^{(t+\frac{1}{2})} - \boldsymbol{x}^*\| + \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\| \\ = \;& -\|\boldsymbol{x}^{(t+\frac{1}{2})} - \boldsymbol{x}^*\| + \|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\| - \|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\| + \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|. \end{aligned}$$

  Since $\boldsymbol{x}^{(t+1)}$ is the projection of $\boldsymbol{x}^{(t+\frac{1}{2})}$ onto the closed convex set $\bar{B}(\boldsymbol{x}_0; B)$ and $\boldsymbol{x}^* \in \bar{B}(\boldsymbol{x}_0; B)$, we have

$$\|\boldsymbol{x}^{(t+\frac{1}{2})} - \boldsymbol{x}^*\| \geq \|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\|$$

  so that

$$-\|\boldsymbol{x}^{(t+\frac{1}{2})} - \boldsymbol{x}^*\| + \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\| \leq -\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\| + \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|.$$

- This guarantees that Lemma 1 remains true when doing projection.

- Theorem 7 remains true for the stochastic subgradient descent with projection algorithm.

General Learning Problem

- $\mathscr{I}$: the input space of all possible items, associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$.
- $c : \mathscr{I} \to \mathscr{Y}$: a fixed unknown concept to learn, which is a function from the input space $\mathscr{I}$ to the label space $\mathscr{Y}$.
- $\mathcal{H} = \{h_{\boldsymbol{w}} \mid \boldsymbol{w} \in S\}$: the set of hypotheses, each of which is represented by a parameter vector $\boldsymbol{w}$ in a subset S of $\mathbb{R}^n$.
  - Each hypothesis $h_{\boldsymbol{w}}$ is a function from the input space $\mathscr{I}$ to the output space $\mathscr{Y}'$.
- $L : \mathscr{Y}' \times \mathscr{Y} \to \mathbb{R}$: the loss function, which is measurable.
- $R(\boldsymbol{w})$: the generalization error (or risk) or true error of the hypothesis $h_{\boldsymbol{w}}$ to the concept c,

$$R(\boldsymbol{w}) = \mathop{E}_{\omega \sim P}[L(h_{\boldsymbol{w}}(\omega), c(\omega))].$$

- Problem: find an $\boldsymbol{w}^*$ in S which minimizes the risk $R(\boldsymbol{w})$, i.e.,

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in S} R(\boldsymbol{w}).$$

S(S)GD for Risk Minimization

- Since we do not know the distribution P, we cannot simply calculate $R(\boldsymbol{w}^{(t)})$ and minimize it with the (S)GD method.
  - Here we assume that $R(\boldsymbol{x})$ is a convex function defined on a convex set S.
- With S(S)GD, however, all we need is to find an unbiased estimate of the gradient or a subgradient of $R(\boldsymbol{x})$ at $\boldsymbol{x} = \boldsymbol{w}^{(t)}$, that is, a random vector $\boldsymbol{v}$ whose conditional expected value $E[\boldsymbol{v} \mid \boldsymbol{w}^{(t)}]$ given $\boldsymbol{w}^{(t)}$ is $\nabla R(\boldsymbol{w}^{(t)})$ or in $\partial R(\boldsymbol{w}^{(t)})$ if a subgradient is used.
- We next see how such an estimate can be easily constructed.

Conditionally Unbiased Estimate of the Gradient

- Assume that the loss function $L(h_{\boldsymbol{w}}(\omega), c(\omega))$ is a differentiable function of $\boldsymbol{w}$ in an open set S.
  - Then the risk $R(\boldsymbol{w})$ is also a differentiable function of $\boldsymbol{w}$ in the open set S.
- $S = \{\omega_1, \omega_2, \ldots, \omega_T\}$: a random sample of size T, drawn i.i.d. from the input space $\mathscr{I}$ under the unknown distribution P.
- At the tth iteration, define $\boldsymbol{v}_t(\omega_t)$ to be the gradient of the function $L(h_{\boldsymbol{w}}(\omega_t), c(\omega_t))$ with respect to $\boldsymbol{w}$, at the point $\boldsymbol{w}^{(t)}$, i.e.,

$$\boldsymbol{v}_t(\omega_t) \triangleq \nabla_{\boldsymbol{w}} L(h_{\boldsymbol{w}^{(t)}}(\omega_t), c(\omega_t)).$$

- $\boldsymbol{v}_t(\omega_t)$ is an unbiased estimate of the gradient of the risk $R(\boldsymbol{w})$ at $\boldsymbol{w}^{(t)}$:

$$\begin{aligned} E[\boldsymbol{v}_t(\omega_t) \mid \boldsymbol{w}^{(t)}] &= \underset{\omega_t \sim P}{E}[\nabla_{\boldsymbol{w}} L(h_{\boldsymbol{w}^{(t)}}(\omega_t), c(\omega_t))] \\ &= \nabla_{\boldsymbol{w}} \underset{\omega \sim P}{E}[L(h_{\boldsymbol{w}^{(t)}}(\omega), c(\omega))] \\ &= \nabla R(\boldsymbol{w}^{(t)}). \end{aligned}$$

Conditionally Unbiased Estimate of a Subgradient

- Assume that the loss function $L(h_{\boldsymbol{w}}(\omega), c(\omega))$ is a convex function of $\boldsymbol{w}$ in an open convex set S.
  - Then the risk $R(\boldsymbol{w})$ is also a convex function of $\boldsymbol{w}$ in the open convex set S.
- $S = \{\omega_1, \omega_2, \ldots, \omega_T\}$: a random sample of size T, drawn i.i.d. from the input space $\mathscr{I}$ under the unknown distribution P.
- At the tth iteration, define $\boldsymbol{v}_t(\omega_t)$ to be a subgradient of the function $L(h_{\boldsymbol{w}}(\omega_t), c(\omega_t))$ with respect to $\boldsymbol{w}$, at the point $\boldsymbol{w}^{(t)}$, i.e.,

$$L(h_{\boldsymbol{u}}(\omega_t), c(\omega_t)) - L(h_{\boldsymbol{w}^{(t)}}(\omega_t), c(\omega_t)) \geq \boldsymbol{v}_t(\omega_t) \cdot (\boldsymbol{u} - \boldsymbol{w}^{(t)}) \ \forall\, \boldsymbol{u} \in S.$$

- Taking expectation on both sides with respect to $\omega_t \sim P$ and conditioned on the value of $\boldsymbol{w}^{(t)}$, we have

$$\begin{array}{rcl} R(\boldsymbol{u}) - R(\boldsymbol{w}^{(t)}) & = & E[L(h_{\boldsymbol{u}}(\omega_t), c(\omega_t)) - L(h_{\boldsymbol{w}^{(t)}}(\omega_t), c(\omega_t)) \mid \boldsymbol{w}^{(t)}] \\ & \geq & E[\boldsymbol{v}_t(\omega_t) \cdot (\boldsymbol{u} - \boldsymbol{w}^{(t)}) \mid \boldsymbol{w}^{(t)}] \\ & = & E[\boldsymbol{v}_t(\omega_t) \mid \boldsymbol{w}^{(t)}] \cdot (\boldsymbol{u} - \boldsymbol{w}^{(t)}) \end{array}$$

which shows that $E[\boldsymbol{v}_t(\omega_t) \mid \boldsymbol{w}^{(t)}]$ is a subgradient of the risk R at the point $\boldsymbol{w}^{(t)}$.

Stochastic Subgradient Descent Algorithm for Risk Minimization

```
StochasticSubgradientDescentRiskMinimization($\boldsymbol{x}_0, T, \eta$)
$\boldsymbol{x}^{(1)} \leftarrow \boldsymbol{x}_0$        $\triangleright$ $\boldsymbol{x}_0$ initial point
for $t \leftarrow 1$ to $T$ do
    $\omega_t \leftarrow$ Sample($P$)
    $\boldsymbol{v}_t \leftarrow$ Subgradient($\boldsymbol{x}^{(t)}, \omega_t$)
    $\boldsymbol{x}^{(t+1)} \leftarrow \boldsymbol{x}^{(t)} - \eta \boldsymbol{v}_t$
return $\bar{\boldsymbol{x}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}^{(t)}$
```