COM S 578X: Optimization for Machine Learning

Lecture Note 8: Subgradient Method

Jia (Kevin) Liu

Assistant Professor
Department of Computer Science
Iowa State University, Ames, Iowa, USA

Fall 2019

Outline

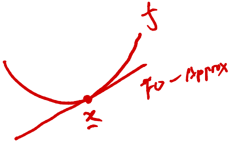
In this lecture:

- Subgradients
- Subgradient method and step-size rule
- Convergence rate analysis and proofs
- Optimal step-size and alternating projections
- Speeding up subgradient methods

Basic Inequality

Recall that the basic inequality for convex differentiable f:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$



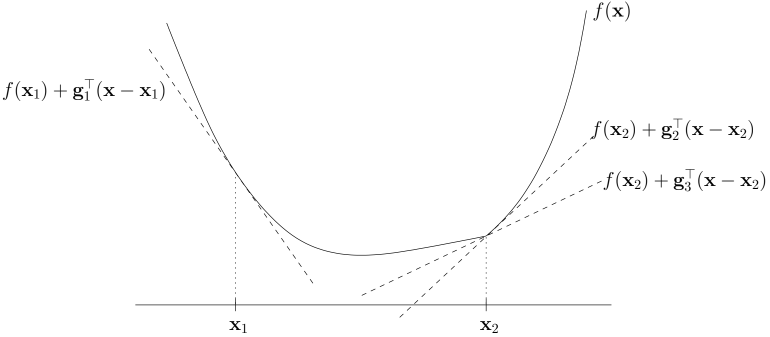
- First-order approximation of f at $\mathbf{x}$ is a global underestimator
- But what if f is not differentiable?

Subgradient of a Function

Definition 1

$\mathbf{g}$ is a subgradient of f (not necessarily convex) at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) \quad \text{for all } \mathbf{y}$$



- The collection of subgradients of f at $\mathbf{x}$ is called subdifferential of f at $\mathbf{x}$
- $\mathbf{g}$ is called a supergradient if $f(\mathbf{y}) \leq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x})$ for all $\mathbf{y}$

Subgradients and Convex Sets/Functions

Theorem 2

Let $\mathcal{S}$ be a nonempty convex set in $\mathbb{R}^n$ and let $f : \mathcal{S} \to \mathbb{R}$ be convex. Then for $\bar{\mathbf{x}} \in \mathrm{int}\{\mathcal{S}\}$, there exists a vector $\mathbf{g}$ such that $f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \mathbf{g}^\top(\mathbf{x} - \bar{\mathbf{x}})$, i.e., $\mathbf{g}$ is a subgradient at $\bar{\mathbf{x}}$.

Theorem 3

Let $\mathcal{S}$ be a nonempty convex set in $\mathbb{R}^n$ and let $f : \mathcal{S} \to \mathbb{R}$. Suppose that for each point $\bar{\mathbf{x}} \in \mathrm{int}\{\mathcal{S}\}$ there exists a subgradient vector $\mathbf{g}$ such that $f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \mathbf{g}^\top(\mathbf{x} - \bar{\mathbf{x}})$ for each $\mathbf{x} \in \mathcal{S}$. Then, f is convex on $\mathrm{int}\{\mathcal{S}\}$.

Generalized Optimality Conditions (Unconstrained)

Recall: For f convex & differentiable, the unconstrained optimal solution satisfies:

$$f(\mathbf{x}^*) = \inf_{\mathbf{x}} f(\mathbf{x}) \iff \nabla f(\mathbf{x}^*) = \mathbf{0}$$

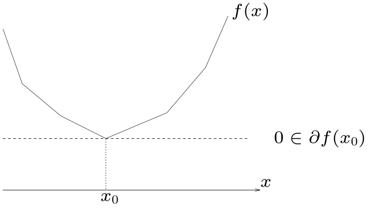
Generalize to nondifferentiable convex f:

$$f(\mathbf{x}^*) = \inf_{\mathbf{x}} f(\mathbf{x}) \iff \partial f(\mathbf{x}^*) \ni \mathbf{0}$$

Proof. By definition:

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + \mathbf{0}^\top(\mathbf{y} - \mathbf{x}^*), \quad \forall \mathbf{y}$$
$$\iff \mathbf{0} \in \partial f(\mathbf{x}^*)$$

□



Generalized Optimality Conditions (Constrained)

$$\text{Minimize } f(\mathbf{x})$$
$$\text{Subject to } g_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m$$

where:

- $f_i : \mathbb{R}^n \to \mathbb{R}$ is convex (hence subdifferentiable)
- Strict feasibility (Slater's condition)

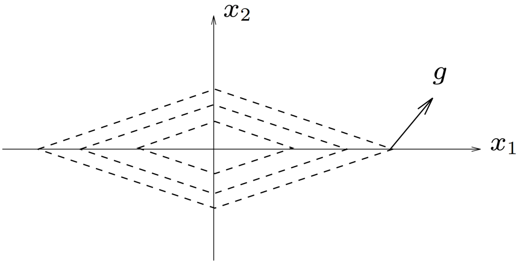
Then $\mathbf{x}^*$ is primal optimal ($\mathbf{u}^*$ is dual optimal) iff

$$\text{(ST): } 0 \in \partial f(\mathbf{x}^*) + \sum_{i=1}^{m} u_i^* \partial g_i(\mathbf{x}^*)$$
$$\text{(PF): } g_i(\mathbf{x}^*) \leq 0, \quad \forall i = 1, \ldots, m$$
$$\text{(DF): } u_i^* \geq 0, \quad \forall i = 1, \ldots, m$$
$$\text{(CS): } u_i^* g_i(\mathbf{x}^*) = 0, \quad \forall i = 1, \ldots, m$$

Generalizes KKT for nondifferentiable f and g_i

Subgradient Needs Not Be A Descent Direction

- For differentiable f, $-\nabla f(\mathbf{x})$ is always a descent direction (except when it's zero)
- But for nondifferentiable (convex) functions, a negative subgradient $-\mathbf{g}$, where $\mathbf{g} \in \partial f(\mathbf{x})$, needs not be a descent direction
- Example: $f(\mathbf{x}) = |x_1| + 2|x_2|$



But It Does Bring Us Closer to An Optimal Solution!

- If f convex and $f(\mathbf{z}) < f(\mathbf{x})$, $\mathbf{g} \in \partial f(\mathbf{x})$, then for $s > 0$ small enough,

$$\|\mathbf{x} - s\mathbf{g} - \mathbf{z}\|_2 < \|\mathbf{x} - \mathbf{z}\|_2$$

  Proof.

$$\begin{aligned}\|\mathbf{x} - s\mathbf{g} - \mathbf{z}\|_2^2 &= \|\mathbf{x} - \mathbf{z}\|_2^2 - 2s\mathbf{g}^\top(\mathbf{x} - \mathbf{z}) + s^2\|\mathbf{g}\|_2^2 \\ &\leq \|\mathbf{x} - \mathbf{z}\|_2^2 - 2s(f(\mathbf{x}) - f(\mathbf{z})) + s^2\|\mathbf{g}\|_2^2\end{aligned}$$

  □

- Thus $-\mathbf{g}$ is a descent direction for $\|\mathbf{x} - \mathbf{z}\|_2$ for any $f(\mathbf{z}) < f(\mathbf{x})$ (e.g., $\mathbf{x}^*$)
- Hence, negative subgradient is descent direction for distance to optimal point

Subgradient Method

Subgradient method is a simple algorithm to minimize nondifferentiable convex function f:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \mathbf{g}_k$$

where:

- $\mathbf{x}_k$ is the kth iterate
- $\mathbf{g}_k$ is any subgradient of f at $\mathbf{x}_k$
- s_k is the kth step-size.

Subgradient method is not a descent method, so we keep track of the best solution so far:

$$f_{\text{best}}^{(k)} = \min_{i=1,\ldots,k} f(\mathbf{x}_i)$$

Step Size Rules

Several commonly used step-size strategies for subgradient method:

- Constant step-size: $s_k = s$ (constant), $\forall k$
- Constant step-length: $s_k = \gamma / \|\mathbf{g}_k\|_2$ (so $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 = \gamma$)
- Square summable but not summable: Step size satisfy

$$\sum_{k=1}^{\infty} s_k^2 < \infty, \qquad \sum_{k=1}^{\infty} s_k = \infty$$

- Nonsummable diminishing: Step size satisfy

$$\lim_{k \to \infty} s_k = 0, \qquad \sum_{k=1}^{\infty} s_k = \infty$$

Some Assumptions for Convergence Proof

- $f^* = \inf_{\mathbf{x}} f(\mathbf{x}) > -\infty$, with $f(\mathbf{x}^*) = f^*$
- $\|\mathbf{g}\|_2 \leq G$ for all $\mathbf{g} \in \partial f$ (similar to Lipschitz continuity condition on f)
- $\|\mathbf{x}_1 - \mathbf{x}^*\|_2 \leq R$

Convergence Results of Subgradient Method

Theorem 4

Let $\bar{f} \triangleq \lim_{k\to\infty} f_{\text{best}}^{(k)}$. The the subgradient method achieves the following convergence results:

- *Constant step-size:* $\bar{f} - f^* \leq G^2 s/2$, *i.e., subgradient method converges to a* $G^2\alpha/2$*-neighborhood around* $\mathbf{x}^*$
- *Constant step-length:* $\bar{f} - f^* \leq G\gamma/2$, *i.e., subgradient method converges to a* $G\gamma/2$*-neighborhood around* $\mathbf{x}^*$
- *Diminishing step-size rule:* $\bar{f} = f^*$, *i.e., subgradient method converges to* $\mathbf{x}^*$

Convergence Proof for Subgradient Method

Proof Sketch:

- Consider the distance to the optimal solution set (rather than the function value): Let $\mathbf{x}^*$ be any minimizer of f. We can show that

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 \leq R^2 - 2\sum_{i=1}^{k} s_i(f(\mathbf{x}_i) - f^*) + G^2 \sum_{i=1}^{k} s_i^2$$

- This implies that:

$$f_{\text{best}}^{(k)} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^{k} s_i^2}{2\sum_{i=1}^{k} s_i}$$

- The result for each step-size selection strategy follows from plugging in the respective step-size definition. □

Stopping Criterion

- Terminating when $\dfrac{R^2 + G^2 \sum_{i=1}^k s_i^2}{2 \sum_{i=1}^k s_i} \le \epsilon$ is very slow

- Optimal choice of s_i to achieve $\dfrac{R^2 + G^2 \sum_{i=1}^k s_i^2}{2 \sum_{i=1}^k s_i} \le \epsilon$ for smallest k:

$$s_i = \frac{R}{G\sqrt{k}}, \quad i = 1, \ldots, k$$

  The number of steps required: $k = (RG/\epsilon)^2$

- The reality: There really isn't a good stopping criterion for the subgradient method

Piecewise Linear Minimization

$$\text{Minimize } f(\mathbf{x}) = \max_{i=1,\ldots,m} (\mathbf{a}_i^\top \mathbf{x} + b_i)$$

- To find a subgradient of f: Find index j for which

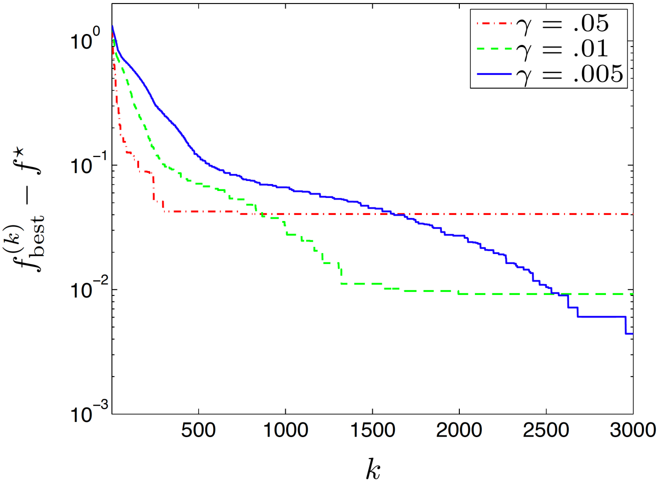
  $$\mathbf{a}_j^\top \mathbf{x} + \mathbf{b}_j = \max_{i=1,\ldots,m} (\mathbf{a}_i^\top \mathbf{x} + b_i)$$

  and take $\mathbf{g} = \mathbf{a}_j$

- Then the subgradient method is: $\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \mathbf{a}_j$

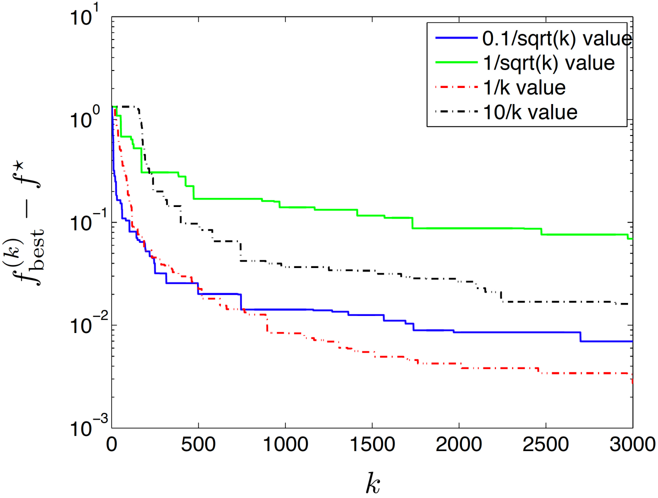
Numerical Example: Constant Step Size

Problem instance with $n = 20$ variables, $m = 100$ terms, $f^* \approx 1.1$, $f_{\text{best}}^{(k)} - f^*$, constant step-size $\gamma = 0.05, 0.01, 0.005$



Numerical Example: Diminishing Step Size

Same problem with diminishing step size rules: $s_k = 0.1/\sqrt{k}$ and $s_k = 1/\sqrt{k}$, square summable step size rules $s_k = 1/k$ and $s_k = 10/k$



Optimal Step Size When f^* Is Known

- Polyak's step-size choice:

$$s_k = \frac{f(\mathbf{x}_k) - f^*}{\|\mathbf{g}_k\|_2^2}$$

Note: f^* can also be replaced by an estimated optimal value

- Rationale: Start with basic inequality

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - 2s_k(f(\mathbf{x}_k) - f^*) + s_k^2\|\mathbf{g}_k\|_2^2$$

and choose s_k to minimize the RHS

Optimal Step Size When f^* Is Known

- Noting the RHS is a quadratic function of s_k, we have

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 \le \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \frac{(f(\mathbf{x}_k) - f^*)^2}{\|\mathbf{g}_k\|_2^2}$$

Observation: $\|\mathbf{x}_k - \mathbf{x}^*\|_2$ decreases each step

- After telescoping, we have

$$\sum_{i=1}^{k} \frac{(f(\mathbf{x}_k) - f^*)^2}{\|\mathbf{g}_k\|_2^2} \le R^2$$

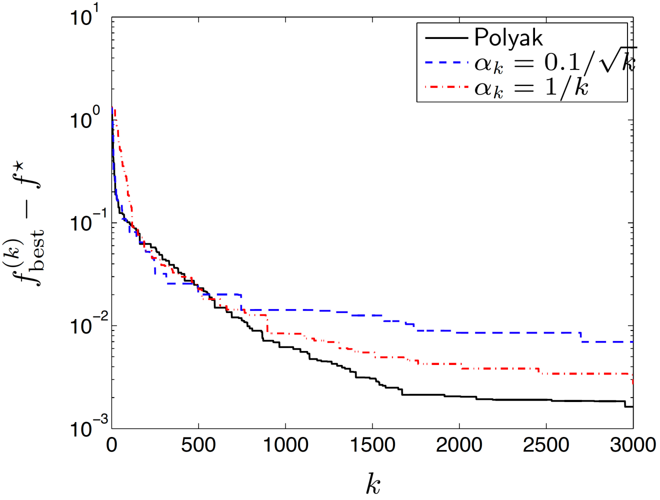
which implies

$$\sum_{i=1}^{k} (f(\mathbf{x}_k) - f^*)^2 \le R^2 G^2$$

which proves that $\bar{f} \to f^*$

Numerical Example: Polyak's Step Size

Piecewise linear maximization with Polyak's step-size: $s_k = 0.1/\sqrt{k}$ and $s_k = 1/k$



Polyak's Step Size When f^* Isn't Known

- Use step-size

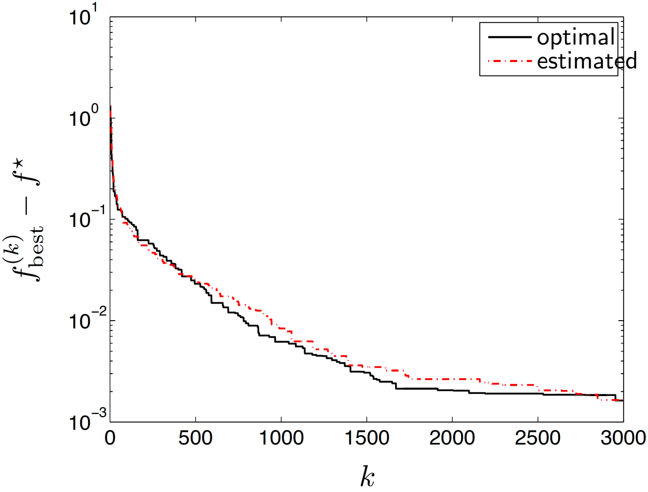
$$s_k = \frac{f(\mathbf{x}_k) - f_{\text{best}}^k + \gamma_k}{\|\mathbf{g}_k\|_2^2}$$

  with $\sum_{k=1}^{\infty} \gamma_k = \infty$, $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$

- $f(\mathbf{x}_k) - f_{\text{best}}^k$ serves as estimate of f^*
- γ_k is in scale of objective value
- Can show that $f_{\text{best}}^k \to f^*$

Numerical Example: Polyak's Step Size for Unknown f^*

Piecewise linear maximization with Polyak's step-size, using f^*, and estimated with $\gamma_k = 10/(10+k)$



Speeding up Subgradient Methods

- Subgradient methods are very slow
- Often convergence can be improved by keeping memory of past steps (heavy-ball)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \mathbf{g}^k + \beta(\mathbf{x}_k - \mathbf{x}_{k-1})$$

- **Other ideas**: Localization methods, conjugate directions, ...

Several Speedup Algorithms

$$\mathbf{x}_{k+1} = x_k - s_k \mathbf{d}_k, \qquad s_k = \frac{f(\mathbf{x}_k) - f^*}{\|\mathbf{d}_k\|_2^2},$$

where f^* can be estimated

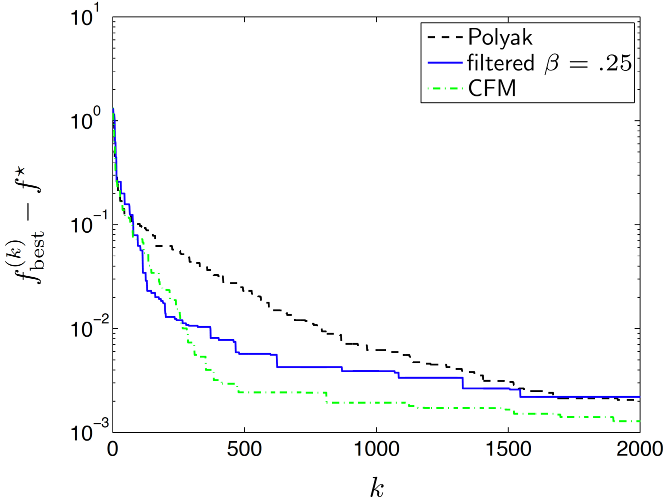
- "Filtered subgradient": $\mathbf{d}_k = (1-\beta)\mathbf{g}_k + \beta \mathbf{d}_{k-1}$, where $\beta \in [0, 1)$
- Camerini, Fratta, and Maffioli (1975)

$$\mathbf{d}_k = \mathbf{g}_k + \beta_k \mathbf{d}_{k-1}, \qquad \beta_k = \max\{0, -\gamma_k \mathbf{d}_{k-1}^\top \mathbf{g}_k / \|\mathbf{d}_{k-1}\|_2^2\},$$

  where $\gamma_k \in [0, 2)$ ($\gamma_k = 1.5$ is recommended)

Numerical Example: Subgradient Method Speedup

Piecewise linear maximization: Polyak's step, filtered subgradient, CFM step



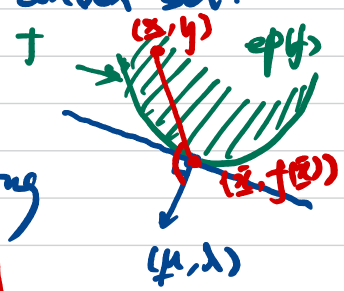
Next Class

Stochastic Gradient Descent

Thm 2: Let S be convex set in $\mathbb{R}^n$ and $f: S \to \mathbb{R}$ be convex. Then, for any $\bar{x} \in int(S)$, $\exists\, g \in \mathbb{R}^n$ s.t. $f(x) \geq f(\bar{x}) + g^T(x - \bar{x})$, $\forall x$, i.e., g is a subgradient.

Proof. Recall that f is convex iff the epigraph is convex set.

$$ep(f) \triangleq \{(x, y) \in S \times \mathbb{R} : f(x) \leq y\}.$$



Note $(\bar{x}, f(\bar{x}))$ at bndry of $ep(f)$. (by supporting hyperplane prop. of convex sets, $\exists$ a non-zero vector $(\mu, \lambda) \in \mathbb{R}^n \times \mathbb{R}$, s.t.

$$\mu^T(x - \bar{x}) + \lambda(y - f(\bar{x})) \leq 0, \quad \forall (x, y) \in ep(f). \qquad (1).$$

Suppose $\lambda < 0$. Dividing both sides of (1) by $|\lambda|$, and letting $g = \frac{\mu}{|\lambda|}$, we have $g^T(x - \bar{x}) - y + f(\bar{x}) \leq 0$, $\forall (x, y) \in ep(f)$.

Let's define $H \triangleq \{(x, y) : y = f(\bar{x}) + g^T(x - \bar{x})\}$. $\quad (\Delta)$.

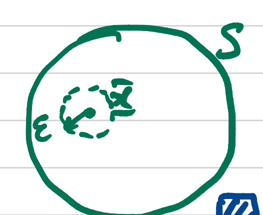
supports $ep(f)$ at $(\bar{x}, f(\bar{x}))$. By letting $y = f(x)$ in (Δ), then we have $f(x) \geq f(\bar{x}) + g^T(x - \bar{x})$, $\forall x \in S$. then we're done.

Claim 1: λ can't be positive, b/c if otherwise, by choosing y suff. large, (1) will be violated.

Claim 2: $\lambda \neq 0$. By contradiction, if $\lambda = 0$, then $\mu^T(x - \bar{x}) \leq 0$, $\forall x \in S$. $(*)$

Since $\bar{x} \in int(S)$, there $\exists\, \varepsilon > 0$ s.t. $\bar{x} + \mu\varepsilon \in S$. Then, from $(*)$, $\mu^T(\bar{x} + \mu\varepsilon - \bar{x}) \leq 0$

$\Rightarrow \varepsilon\mu^T\mu \leq 0 \Rightarrow \mu = 0 \Rightarrow (\mu, \lambda) = (0, 0)$, $\rightarrow\!\leftarrow$ $(\mu, \lambda) \neq 0$.



☒

Thm 3: Let S be non-empty convex set in $\mathbb{R}^n$. Let $f: S \to \mathbb{R}$. Suppose also $\bar{\underline{x}} \in \text{int}\{S\}$. $\exists$ a subgradient $g \in \mathbb{R}^n$ s.t. $f(\underline{x}) \geq f(\bar{\underline{x}}) + g^T(\underline{x} - \bar{\underline{x}})$, $\forall \underline{x} \in S$. Then, f is convex in $\text{int}\{S\}$.

Proof. Pick $\underline{x}_1, \underline{x}_2 \in \text{int}\{S\}$, & pick some $\lambda \in (0,1)$.

Since S is convex $\Rightarrow$ $\text{int}\{S\}$ convex. Then,

$\lambda \underline{x}_1 + (1-\lambda)\underline{x}_2 \in \text{int}\{S\}$.

Since $\exists$ subgrad g, $\forall$ pt. in $\text{int}\{S\}$, then:

$$f(\underline{x}_1) \geq f(\lambda \underline{x}_1 + (1-\lambda)\underline{x}_2) + g^T[\underline{x}_1 - (\lambda \underline{x}_1 + (1-\lambda)\underline{x}_2)] \leftarrow (1).$$

$$f(\underline{x}_2) \geq f(\lambda \underline{x}_1 + (1-\lambda)\underline{x}_2) + g^T[\underline{x}_2 - (\lambda \underline{x}_1 + (1-\lambda)\underline{x}_2)] \leftarrow (2).$$

$(1) \times \lambda + (2) \times (1-\lambda) \Rightarrow$

$$\lambda f(\underline{x}_1) + (1-\lambda) f(\underline{x}_2) \geq f(\lambda \underline{x}_1 + (1-\lambda)\underline{x}_2).$$

∎

Thm: Let $\bar{f} = \lim_{k\to\infty} f_{best}^{(k)}$. The subgrad method achieves:

(1) Const. step-size: $\bar{f} - f^* \le \frac{G^2 s}{2}$

(2) const. step-length: $\bar{f} - f^* \le \frac{G\gamma}{2}$

(3) Diminishing step-size: $\bar{f} = f^*$.

Proof. $\|\underline{x}_{k+1} - \underline{x}^*\|_2^2 = \|\underline{x}_k - s_k g_k - \underline{x}^*\|_2^2$

$$= \|\underline{x}_k - \underline{x}^*\|_2^2 + s_k^2 \|g_k\|_2^2 - 2 s_k g_k^T(\underline{x}_k - \underline{x}^*) . \qquad (\Delta).$$

Now, note that $f^* = f(\underline{x}^*) \ge f(\underline{x}_k) + g_k^T(\underline{x}^* - \underline{x}_k)$.

$\Rightarrow -g_k^T(\underline{x}_k - \underline{x}^*) \le -(f(\underline{x}_k) - f(\underline{x}^*))$. (*)

Thus, $(\Delta) \Rightarrow \|\underline{x}_{k+1} - \underline{x}^*\|_2^2 \le \|\underline{x}_k - \underline{x}^*\|_2^2 - 2 s_k (f(\underline{x}_k) - f(\underline{x}^*)) + s_k^2 \|g_k\|_2^2$

Repeat the process recursively, we have:

$$\|\underline{x}_{k+1} - \underline{x}^*\|_2^2 \le \|\underline{x}_1 - \underline{x}^*\|_2^2 - 2\sum_{i=1}^k s_i (f(\underline{x}_i) - f(\underline{x}^*)) + \sum_{i=1}^k s_i^2 \|g_i\|_2^2$$

$$\le R^2 - 2\sum_{i=1}^k s_i (f(\underline{x}_i) - f(\underline{x}^*)) + G^2 \sum_{i=1}^k s_i^2 \qquad (\Delta\Delta).$$

best in mem.

Note: $\sum_{i=1}^k s_i (f(\underline{x}_i) - f(\underline{x}^*)) \ge \sum_{i=1}^k s_i (\bar{f} - f(\underline{x}^*)) = (\bar{f} - f(\underline{x}^*)) \sum_{i=1}^k s_i$

$$(\Delta\Delta) = \bar{f} - f^* \le \frac{R^2 + G^2 \sum_{i=1}^k s_i^2 - \|\underline{x}_{k+1} - \underline{x}^*\|_2^2}{2\sum_{i=1}^k s_i} \le \frac{R^2 + G^2 \sum_{i=1}^k s_i^2}{2\sum_{i=1}^k s_i}$$

drop.

Case 1°: Const. step-size: $s_k = s$, $\forall k$,

$$\bar{f} - f^* \leq \frac{R^2 + G^2 k s^2}{2ks} = \frac{\frac{1}{k}R^2 + G^2 s^2}{2s} \to \frac{G^2 s}{2} \quad \text{as } k \to \infty.$$

Case 2°: Const. step-length: $s_k = \frac{\gamma}{\|g_k\|_2}$ $(\|g_k\|_2 \leq G)$

$$\bar{f} - f^* \leq \frac{R^2 + \sum_{i=1}^{k} s_i^2 \|g_i\|_2^2}{2\gamma \sum_{i=1}^{k} \frac{1}{\|g_i\|_2}} \leq \frac{R^2 + \gamma^2 k}{2\gamma k / G} \to \frac{G\gamma}{2} \quad \text{as } k \to \infty.$$

Case 3°: Diminishing step-size: $s_k \to 0$, $\sum_{k=1}^{\infty} s_k = \infty$, $\sum_{k=1}^{\infty} s_k^2 = B < \infty$.

$$\bar{f} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^{k} s_i^2}{2 \sum_{i=1}^{k} s_i} \to \frac{R^2 + G^2 B}{\infty} \to 0 \quad \text{as } k \to \infty.$$

($\sum_{i=1}^{k} s_i^2 = B$, $\sum_{i=1}^{k} s_i \to \infty$)

Case 4°: Diminishing step-size: $s_k \to \infty$, $\sum_{k=1}^{\infty} s_k \to \infty$.

When k suff. large; $k \geq K$, for some K. (A') becomes:

$$\|x_{k+1} - x^*\|_2^2 \leq \|x_k - x^*\|_2^2 - 2s_k (f(x_k) - f(x^*)) + o(s_k) \quad (*).$$

Summing $(*)$ for $k = K, \cdots, K+r$, yields:

$$\|x_{K+r} - x^*\|_2^2 - \|x_K - x^*\|_2^2 \leq -2 \sum_{i=K}^{K+r} (f(x_i) - f(x^*)) s_i$$

$$\Rightarrow 2 \sum_{i=K}^{K+r} s_i (f(x_i) - f(x^*)) \leq \|x_K - x^*\|_2^2 \underbrace{- \|x_{K+r} - x^*\|_2^2}_{\leq 0} \leq \|x_K - x^*\|_2^2, \forall r > 0.$$

$$\Rightarrow (\bar{f} - f^*) \left(2 \sum_{i=K}^{K+r} s_i \right) \leq \|x_K - x^*\|_2^2$$

Let $r \to \infty \Rightarrow$ $$\bar{f} - f^* \leq \frac{\|x_K - x^*\|^2}{2 \underbrace{\sum_{i=K}^{K+r} s_i}_{\to \infty}} \to 0.$$ ∎