Numerical Optimization

Lecture 8: Subgradient Descent Method

Hao Wang
Email: wanghao1@shanghaitech.edu.cn

Contents

1. Algorithm
2. Convergence
3. Example
4. Subgradient Projection Method

Contents

1. Algorithm
2. Convergence
3. Example
4. Subgradient Projection Method

Suppose $\min_x \ f(x)$

Construct local approximation at x^k:

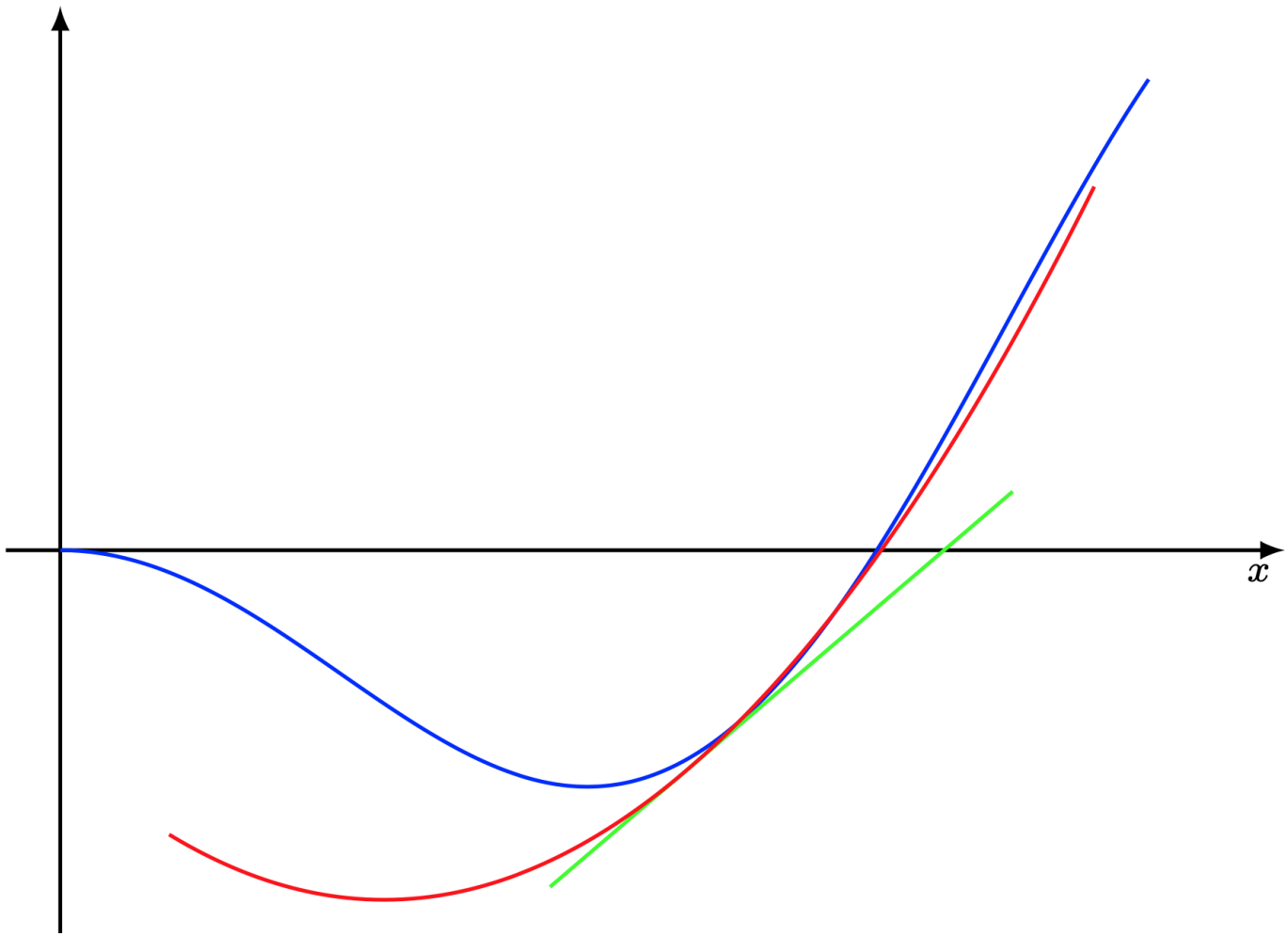
$$x^{k+1} \leftarrow \min_x \ f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2t} \|x - x^k\|_2^2$$

$$\implies x^{k+1} \leftarrow x^k - t \nabla f(x^k)$$

Interpretation:

- Go along the steepest descent direction with a stepsize t
- Linear model but not too far away from x^k
- Local quadratic approximation model at x^k
- Newton method, quasi-Newton method also use local models.

$$x^{k+1} \leftarrow \min_x \; f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2t} \|x - x^k\|_2^2$$



Now suppose $f(x)$ is convex but not necessarily differentiable.

$$x^{k+1} \leftarrow \min_x \; f(x^k) + \langle g^k, x - x^k \rangle + \frac{1}{2t}\|x - x^k\|_2^2, \quad g^k \in \partial f(x^k)$$

$$\Longrightarrow \; x^{k+1} \leftarrow x^k - tg^k$$

Does this still work? — the answer is NO!

Subgradient method is not descent

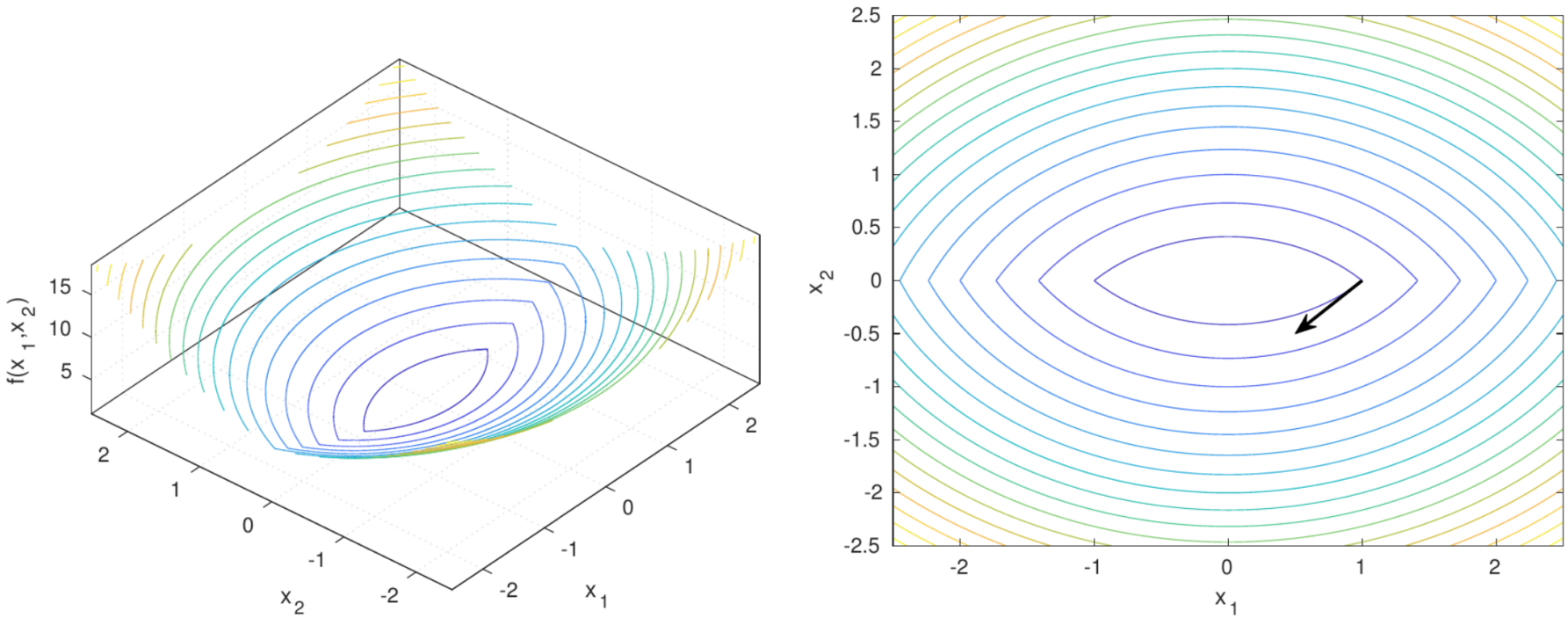


Figure 2. 3D plot (left) and level sets (right) of $f(\boldsymbol{x}) = \max[x_1^2 + (x_2+1)^2, x_1^2 + (x_2-1)^2]$. The negative subgradient is indicated by the black arrow.

Constant stepsize does not work

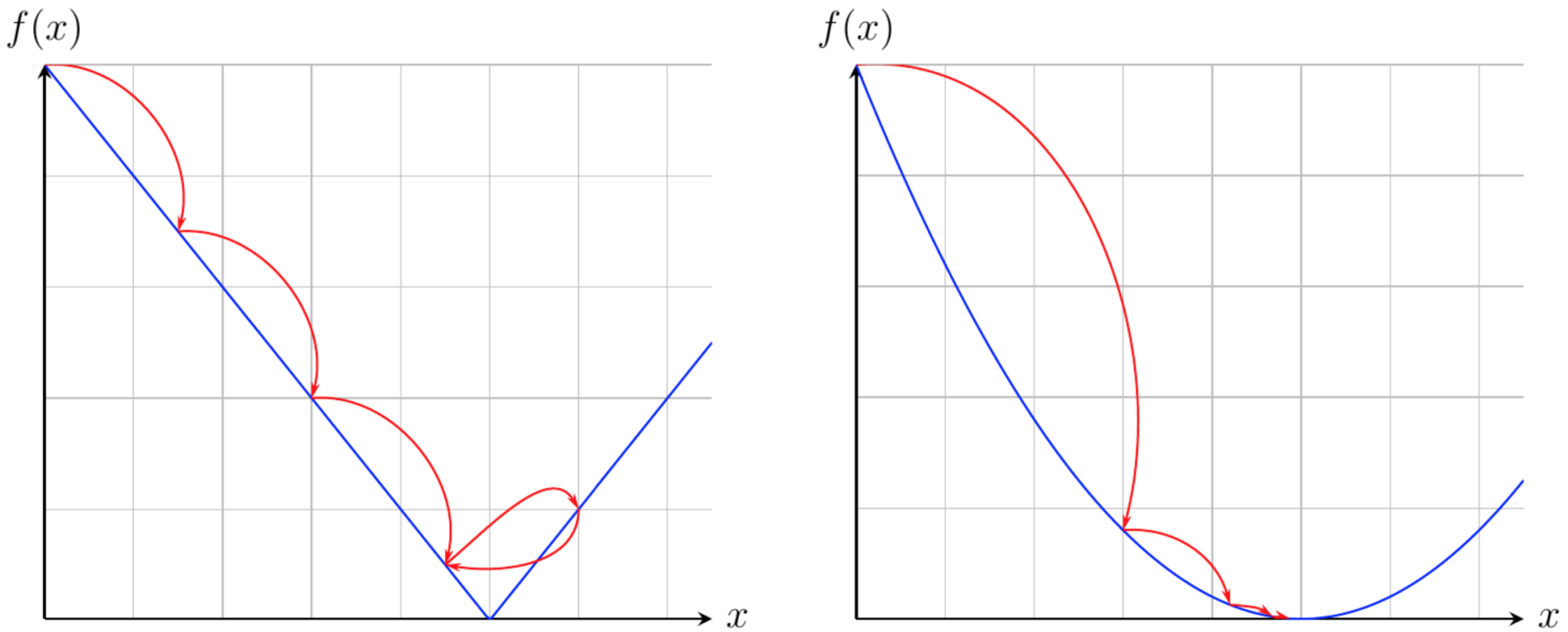


Figure 3. The effect of a constant stepsize on non-differentiable (left) and smooth (right) functions.

Now suppose $f(x)$ is convex but not necessarily differentiable.

$$x^{k+1} \leftarrow \min_x \; f(x^k) + \langle g^k, x - x^k \rangle + \frac{1}{2\alpha_k}\|x - x^k\|_2^2, \quad g^k \in \partial f(x^k)$$

$$\Longrightarrow \; x^{k+1} \leftarrow x^k - \alpha_k g^k$$

Stepsize:

- Fixed stepsize: $\alpha_k = \alpha$;
- Fixed $\|x^{k+1} - x^k\|$, i.e., $\alpha_k \|g^k\|$ is constant;
- Diminishing $\alpha_k \to 0$ and $\sum_{k=0}^{\infty} \alpha_k = +\infty$.

Contents

1 Algorithm

2 Convergence

3 Example

4 Subgradient Projection Method

Assumption 1

Suppose $f(x)$ satisfies

1. *f is convex.*
2. *f has at least one minimizer x^*, and $f(x^*) > -\infty$.*
3. *f is Lipschitz continuous,*

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

where $G > 0$ is Lipschitz constant.

Lemma 2

Suppose $f(x)$ is convex. $f(x)$ is G-Lipschitz continuous if and only if f has bounded subgradients, i.e.,

$$\|g\| \leq G, \quad \forall g \in \partial f(x), \ x \in \mathbb{R}^n.$$

Proof.

($\Rightarrow$) Suppose $\|g\| \leq G$ for any subgradient. Now choose $g_y \in \partial f(y), g_x \in \partial f(x)$, it follows from convexity that

$$g_x^T(x-y) \geq f(x) - f(y) \geq g_y^T(x-y).$$

Cauchy-Schwartz inequality yields

$$g_x^T(x-y) \leq \|g_x\|\|x-y\| \leq G\|x-y\|,$$
$$g_y^T(x-y) \geq -\|g_y\|\|x-y\| \geq -G\|x-y\|.$$

This leads

$$|f(x) - f(y)| \leq G\|x-y\|.$$

($\Leftarrow$) Suppose $f(x)$ is Lipschitz continuous. Suppose by contradiction that there exists x and $g \in \partial f(x)$ such that $\|g\| > G$. Now choose $y = x + \frac{g}{\|g\|}$, it holds that

$$f(y) \geq f(x) + g^T(y-x) = f(x) + \|g\| > f(x) + G,$$

contradicting with the Lipschitz continuity. $\square$

Theorem 3

Let x^* be a minimizer of f, $f^* = f(x^*)$, and $\{x^k\}$ be generated by SG, for any $k \geq 0$,

$$2\left(\sum_{i=0}^{k}\alpha_i\right)\left(\min_{0\leq i\leq k} f(x^i) - f^*\right) \leq \|x^0 - x^*\|^2 + \sum_{i=0}^{k}\alpha_i^2 G^2.$$

Proof.

The SG iteration yields

$$\begin{aligned}\|x^{i+1} - x^*\|^2 =& \|x^i - \alpha g^i - x^*\|^2 \\ =& \|x^i - x^*\|^2 - 2\alpha_i\langle g^i, x^i - x^*\rangle + \alpha_i^2\|g^i\|^2 \\ \leq& \|x^i - x^*\|^2 - 2\alpha_i(f(x^i) - f^*) + \alpha_i^2 G^2,\end{aligned}$$

The last inequality follows from the definition of subgradient and $\|g^i\| \leq G$.

□

Proof.

It follows that

$$2\alpha_i(f(x^i) - f^*) \leq \|x^i - x^*\|^2 - \|x^{i+1} - x^*\|^2 + \alpha_i^2 G^2.$$

Summing up both sides,

$$\begin{aligned} 2\sum_{i=0}^{k} \alpha_i(f(x^i) - f^*) &\leq \|x^0 - x^*\|^2 - \|x^{k+1} - x^*\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2 \\ &\leq \|x^0 - x^*\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2. \end{aligned}$$

On the other hand,

$$\sum_{i=0}^{k} \alpha_i(f(x^i) - f^*) \geq \left(\sum_{i=0}^{k} \alpha_i\right) (\min_{0 \leq i \leq k} f(x^i) - f^*).$$

Combining above two inequalities yields the result.

$$2\left(\sum_{i=0}^{k}\alpha_i\right)\left(\min_{0\le i\le k} f(x^i)-f^*\right)\le \|x^0-x^*\|^2+\sum_{i=0}^{k}\alpha_i^2G^2.$$

Corollary 4

- *Fix* $\alpha_i=t$. *Then*

$$\min_{0\le i\le k} f(x^i)-f^*\le \frac{\|x^0-x^*\|^2}{2kt}+\frac{G^2t}{2}\to\frac{G^2t}{2}$$

- *Fix* $\|x^{k+1}-x^k\|=\alpha_i\|g^i\|=s$. *Then*

$$\min_{0\le i\le k} f(x^i)-f^*\le \frac{G\|x^0-x^*\|^2}{2ks}+\frac{Gs}{2}\to\frac{Gs}{2}$$

- *For diminishing* α_i *satisfying* $\sum_{i=0}^{\infty}\alpha_i=\infty$ *and* $\sum_{i=0}^{\infty}\alpha_i^2<\infty$, *it holds*

$$\min_{0\le i\le k} f(x^i)-f^*\le \frac{\|x^0-x^*\|^2+G^2\sum_{i=0}^{k}\alpha_i^2}{2\sum_{i=0}^{k}\alpha_i}\to 0.$$

Suppose total budget k is given. Let $\|x^0 - x^*\| = R$, the first case yields

$$\min_{0\le i\le k} f(x^i) - f^* \le \frac{R^2}{2kt} + \frac{G^2 t}{2} \le \frac{R^2}{2kt} + \frac{G^2 t}{2}$$

If t satisfying $\frac{R^2}{2kt} = \frac{G^2 t}{2}$ and $t = \frac{R}{G\sqrt{k}}$, then

$$\min_{0\le i\le k} f(x^i) - f^* \le \frac{GR}{\sqrt{k}}.$$

To get $\min_{0\le i\le k} f(x^i) - f^* \le \epsilon$, need choose

$$k = O\left(1/\epsilon^2\right) \quad \text{and} \quad \alpha_k = O(1/\sqrt{k}).$$

The second case has similar result.

Contents

1. Algorithm
2. Convergence
3. Example
4. Subgradient Projection Method

LASSO regression:

$$\min_x f(x) = \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1$$

The subgradient of f is given

$$g = A^T(Ax - b) + \mu \text{sign}(x)$$

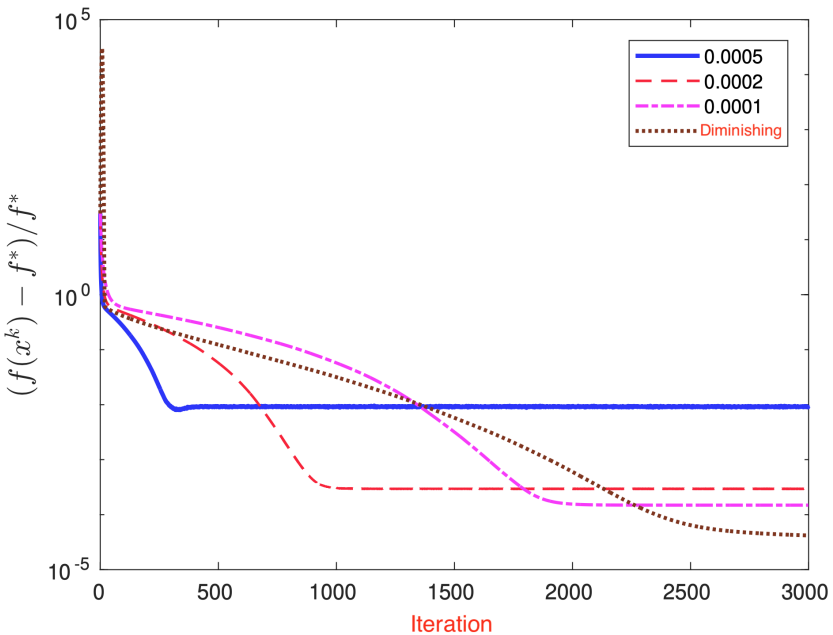
The algorithm is

$$x^{k+1} = x^k - \alpha_k(A^T(Ax^k - b) + \mu \text{sign}(x^k))$$

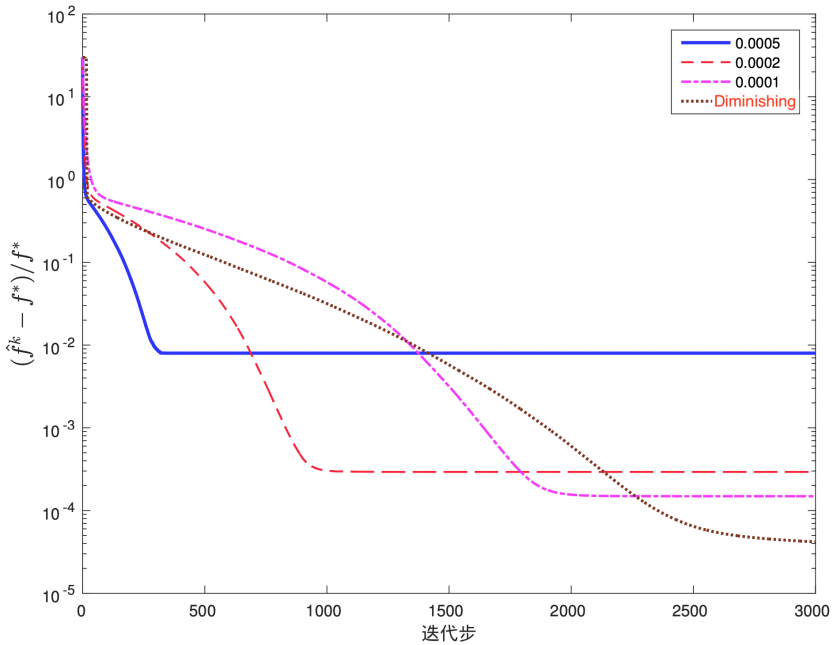
Stepsize choices:

$$\alpha_k = 0.0005, 0.0002, 0.0001, \quad \alpha_k = \frac{0.002}{\sqrt{k}}.$$

Example



(a) $f(x^k) - f^*$



(b) $\hat{f}^k - f^*$

Contents

1. Algorithm
2. Convergence
3. Example
4. Subgradient Projection Method

Constraint case:

$$\min \ f(x) \quad \text{s.t. } x \in \Omega$$

Subgradient Projection Method: choose $g^k \in \partial f(x^k)$,

$$\begin{aligned} x^{k+1} &\leftarrow \arg\min_{x\in\Omega} \langle g^k, x\rangle + \frac{1}{2\eta_k}\|x^k - x\|_2^2 \\ &= \arg\min_{x\in\Omega} \|\eta_k g^k\|_2^2 + 2\eta^k\langle g^k, x - x^k\rangle + \|x^k - x\|_2^2 \\ &= \arg\min_{x\in\Omega} \|x - x^k - \eta_k g^k\|_2^2 \\ &= P_\Omega(x^k - \eta_k g^k), \end{aligned}$$

Theorem 5

Let $\Omega \subset \mathbb{R}^d$ be a closed non-empty convex set with diameter D, i.e., $\max_{x,y\in\Omega} \|x-y\| \leq D$. Let f be a convex function from $\Omega \subset \mathbb{R}^d$ to $\mathbb{R}$. Set $x^1 \in \Omega$. Set $g^k \in \partial f(x^k)$. Then, $\forall u \in \Omega$, the following convergence bounds hold:

$$f\left(\frac{1}{T}\sum_{k=1}^{T} x^k\right) - f(x^*) \leq \frac{D^2}{\eta_T} + \frac{\sum_{k=1}^{T} \eta_k \|g^k\|_2^2}{2T},$$

$$f\left(\frac{\sum_{k=1}^{T} \eta_k x^k}{\sum_{k=1}^{T} \eta_k}\right) - f(x^*) \leq \frac{\|x^1-u\|_2^2 + \sum_{k=1}^{T} \eta_k^2 \|g^k\|_2^2}{2\sum_{k=1}^{T} \eta_k},$$

$$\min_{1\leq k\leq T} f(x^k) - f(x^*) \leq \min\left(\frac{D^2}{\eta_T} + \frac{\sum_{k=1}^{T} \eta_k \|g^k\|_2^2}{2T}, \frac{\|x^1-u\|_2^2 + \sum_{k=1}^{T} \eta_k^2 \|g^k\|_2^2}{2\sum_{k=1}^{T} \eta_k}\right).$$