(Sub)-gradient method

Axel Böhm

October 5, 2021

Smooth vs. nonsmooth

$$\min_x \; f(x)$$

$$x_{k+1} = x_k - \alpha g_k$$

- If f is smooth we take $g_k = \nabla f(x_k) \to$ **Gradient Descent**.
- stepsize can be constant $1/L$ (smoothness constant)
- convergence rate $f(x_k) - f^* = \mathcal{O}(1/k)$

- If not we take g_k a *subgradient* $\to$ **Subgradient method**.
- stepsize has to be chosen *small or decreasing* $\approx 1/\sqrt{k}$
- convergence rate is *worse* $f(x_k) - f^* = \mathcal{O}(1/\sqrt{k})$

Intuition behind GD

- derivative (gradient) points in the direction of steepest ascent
  $\rightarrow$ GD is also called **steepest descent**
- GD update is equivalent to

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} \left\{ \underbrace{f(x_k) + \langle \nabla f(x_k), x - x_k \rangle}_{\text{linearization of } f} + \frac{1}{2\alpha} \|x - x_k\| \right\}$$

  - solves a linear model of f
  - minimizing unconstrained linear models is no good
  - so we add a "proximity term"

Subgradients

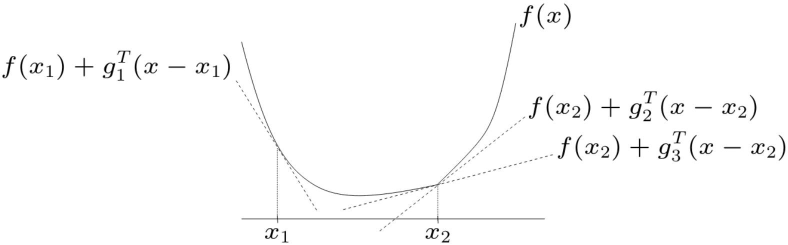
What if f is not differentiable?

Definition

$g \in \mathbb{R}^d$ is a subgradient of f at x if

$$f(y) \geq f(x) + g^T(y - x)$$

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

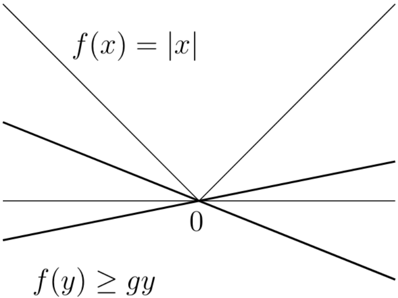


Subgradients II

Definition

The subdifferential $\partial f(x)$ is the *set of all subgradients* of f at x.

Example



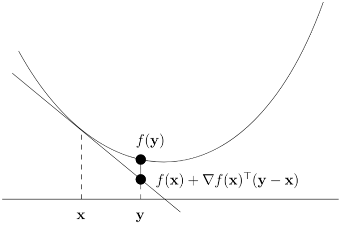
Subgradient condition at $x = 0$ is $f(y) \geq f(0) + g(y - 0) = gy$.
What is $\partial f(0)$?

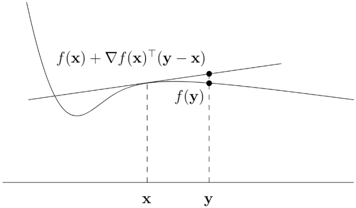
Subgradients III

Lemma

If f is differentiable at x then $\partial f(x) \subset \{\nabla f(x)\}$

So either one subgradient or none.





Subgradient characterization of convexity

Lemma

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if $\partial f(x)$ is not empty for all x.

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

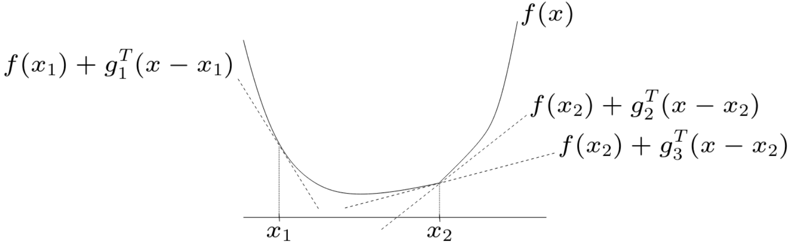


Figure: Subgradients at every point.

Lipschitz = bounded subgradients

Definition

We call f L-Lipschitz (continuous) if

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

Lemma

Let f be convex. Then the following two are equivalent.

(i) *All **subgradients are uniformly bounded**.*

$$\|g\| \leq L \quad \forall x, \forall g \in \partial f(x)$$

(ii) *f is L-Lipschitz*

Subgradient optimality condition

Lemma

Let $0 \in \partial f(\bar{x})$, then $\bar{x}$ is a global minimum.

Proof.

By the definition of subgradients, $g = 0 \in \partial f(\bar{x})$ gives

$$f(y) \geq f(\bar{x}) + g^T(y - \bar{x}) = f(\bar{x}).$$

$\square$

Convergence statement

We assume there exists minimizer x^* and we write $f^* = f(x^*)$.

Theorem

f is convex, subgradients are bounded $\|g(x)\| \leq G$ for all $g(x) \in \partial f(x)$. Then,

$$f(\bar{x}_k) - f^* \leq \frac{\|x_1 - x^*\|^2 G}{\sqrt{k}}$$

*for the **averaged** iterates $\bar{x}_k = \frac{1}{k} \sum_{i=0}^{k-1} x_i$*

- Also holds for the "**best**" iterate.
- Dimension independent! (no d)

Proof

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq \|x_k - \alpha g_k - x^*\|^2 \\
&= \|x_k - x^*\|^2 + 2\alpha\langle g_k, x^* - x_k\rangle + \alpha^2\|g_k\|^2.
\end{aligned}
$$

Using the **subgradient inequality** $\langle g_k, x^* - x_k\rangle \leq f(x^*) - f(x_k)$

$$
\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + 2\alpha(f(x^*) - f(x_k)) + \alpha^2\|g_k\|^2.
$$

Summing up (telescoping) yields

$$
2\sum_{i=0}^{k-1} \alpha(f(x_i) - f(x^*)) + \|x_k - x^*\|^2 \leq \|x_0 - x^*\|^2 + \alpha^2 \sum_{i=0}^{k-1} \|g_k\|^2. \quad (1)
$$

Via the *bounded subgradient* assumption

$$
2\sum_{i=0}^{k-1} \alpha(f(x_i) - f(x^*)) \leq \|x_0 - x^*\|^2 + \alpha^2 k G^2.
$$

Proof [contd]

We divide by 2α and k

$$\frac{1}{k}\sum_{i=0}^{k-1} f(x_i) - f^* \leq \frac{1}{2\alpha k}\|x_0 - x^*\|^2 + \alpha G^2$$

Using Jensens inequality (convexity with more than 2 points)

$$\sum_{i=0}^{k-1}\frac{1}{k}f(x_i) \geq \sum_i f\left(\frac{1}{k}\sum_{i=0}^{k-1} x_i\right)$$

we obtain

$$f(\bar{x}_k) - f^* \leq \frac{1}{2\alpha k}\|x_0 - x^*\|^2 + \alpha G^2.$$

How to choose the stepsize?

We have

$$f(\bar{x}_k) - f^* \leq \frac{1}{2\alpha k}\|x_0 - x^*\|^2 + \alpha G^2.$$

Choose α such that *RHS is minimized*, i.e.

$$\alpha = \frac{\|x_0 - x^*\|}{G\sqrt{k}},$$

which gives

$$f(\bar{x}_k) - f^* \leq \frac{\|x_0 - x^*\| G}{2\sqrt{k}}. \quad \square$$

When ignoring constants (and focusing on the rate) we sometimes write

$$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right).$$

Complexity

For convex Lipschitz functions we require $\mathcal{O}(\epsilon^{-2})$ iterations. For $D := \|x_0 - x^*\|$

$$f(\bar{x}_k) - f^* \leq \frac{DG}{\sqrt{k}}$$

Q: How many iterations to get

$$f(\bar{x}_k) - f^* \leq \epsilon?$$

A: We get this if

$$\frac{DG}{\sqrt{k}} \leq \epsilon$$

Equivalently

$$k \geq \frac{D^2 G^2}{\epsilon^2}.$$

Projected subgradient method

$$\text{(constrained setting)} \quad \min_{x \in C} f(x)$$

Algorithm Projected subgradient method

1: **for** $k = 0, 1, \ldots$ **do**
2: Pick $g_k \in \partial f(x_k)$
3: $x_{k+1} = P_C(x_k - \alpha g_k)$

By using the fact that the projection is a contraction

$$\|P_C(x) - P_C(y)\| \leq \|x - y\|$$

Projected subgradient method II

Proof.

We can deduce the exact same inequality as before

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|P_C(x_k - \alpha g_k) - x^*\|^2 \\ &\leq \|x_k - \alpha g_k - x^*\|^2 \\ &= \|x_k - x^*\|^2 + 2\alpha\langle g_k, x^* - x_k\rangle + \alpha^2\|g_k\|^2 \\ &\leq \|x_k - x^*\|^2 + 2\alpha(f^* - f(x_k)) + \alpha^2\|g_k\|^2.\end{aligned}$$

□

Polyak stepsize

Let's revisit the convergence proof of the subgradient method

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq \|x_k - \alpha_k g_k - x^*\|^2 \\
&= \|x_k - x^*\|^2 + 2\alpha\langle g_k, x^* - x_k\rangle + \alpha^2\|g_k\|^2 \\
&\leq \|x_k - x^*\|^2 + 2\alpha(f^* - f(x_k)) + \alpha^2\|g_k\|^2.
\end{aligned}
$$

Can we pick α such that the RHS is minimized?

$$\min_{\alpha} \alpha^2\|g_k\|^2 + 2\alpha_k(f^* - f(x_k))$$

gives

$$\alpha^* = \frac{f(x_k) - f^*}{\|g_k\|^2}$$

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - \left(\frac{f(x_k) - f^*}{\|g_k\|}\right)^2$$

Polyak stepsize [contd]

- Requires us to know the optimal objective function value
- can be the case in certain setting: separable data, feasibility problems
- modern deep learning interpolation setting

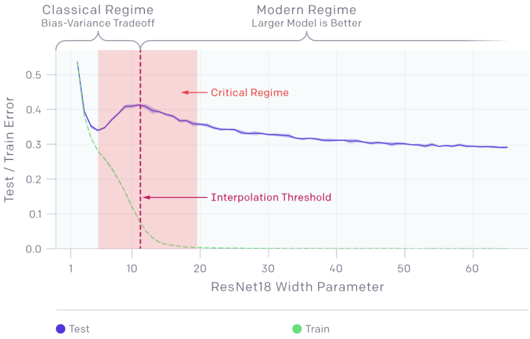


Figure: from openai.com

Can we do better?

If f is in addition *strongly convex* the rate improves to

$$f(\bar{x}_k) - f(x^*) \leq \frac{L\|x_1 - x^*\|^2}{\mu T}$$

by choosing the stepsize $\alpha_k \approx \frac{1}{T}$.

Can we do better if the function is smooth?

Definition

We call a function L-smooth if

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2.$$

Can be upper bounded by a quadratic.

Lemma

If the gradient of f is L-Lipschitz

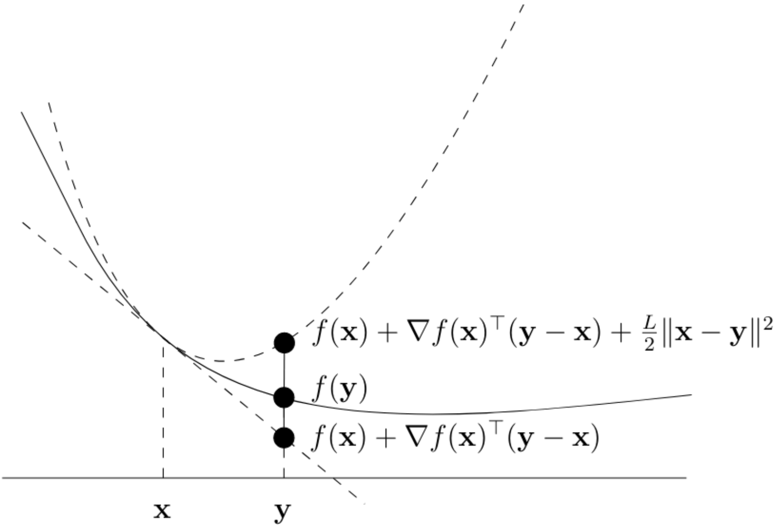
$$\|\nabla f(x) - f(y)\| \leq L\|x - y\|.$$

then it is also L-smooth.

Note: Definition does not require convexity.

Smoothness

If f is convex we get upper and lower bound:



Smooth vs. Lipschitz

- Bounded (sub)gradients $\Leftrightarrow$ Lipschitz continuity of f
- Smoothness $\Leftrightarrow$ Lipschitz continuity of ∇f (if convex)

Lemma

Let f be convex and differentiable, then the following are equivalent

(i) *f is smooth with parameter L*

(ii) *∇f is L-Lipschitz*

Sufficient decrease

Lemma

If f is L-smooth with stepsize $\alpha = 1/L$, then gradient descent satisfies

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2$$

Proof.

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \gamma\|\nabla f(x_k)\|^2 + \frac{L}{2\gamma^2}\|\nabla f(x_k)\|^2 \\ &= f(x_k) - \left(\frac{1}{L} - \frac{1}{2L}\right)\|\nabla f(x_k)\|^2 \end{aligned}$$

$\square$

Smooth convex functions

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth and the stepsize $\alpha = 1/L$, then gradient descent yields

$$f(x_k) - f^* \leq \frac{L}{2k}\|x_0 - x^*\|^2.$$

- holds for last iterate
- independet of dimension d

Complexity of gradient method

Denote $D^2 := \|x_1 - x^*\|^2$

$$\text{iteration: } k \geq \frac{D^2 L}{2\epsilon} \Rightarrow \text{error} \leq \frac{LD^2}{2k} \leq \epsilon$$

Given error $\epsilon = 0.01$ results in

- $50 \cdot D^2 L$ iterations for *smooth* case
- $10000 \cdot D^2 G^2$ for nonsmooth but Lipschitz

What if we don't know L?

Proof of $\mathcal{O}(\epsilon^{-1})$ for smooth functions

Subgradient analysis gave us

$$\sum_{i=0}^{k-1}(f(x_i) - f^*) \leq \frac{1}{2\alpha}\|x_0 - x^*\|^2 + \frac{\alpha}{2}\sum_{i=0}^{k-1}\|g_k\|^2,$$

see (1). This time we use sufficient decrease to bound gradient norm

$$\frac{1}{2L}\sum_{i=0}^{k-1}\|\nabla f(x_k)\|^2 \leq \sum_{i=0}^{k-1}(f(x_i) - f(x_{i+1})) = f(x_0) - f(x_k)$$

Combining the above two (with $\alpha = 1/L$)

$$\begin{aligned}\sum_{i=0}^{k-1}(f(x_i) - f^*) &\leq \frac{L}{2}\|x_0 - x^*\|^2 + \frac{1}{2L}\sum_{i=0}^{k-1}\|g_k\|^2 \\ &\leq \frac{L}{2}\|x_0 - x^*\|^2 + f(x_0) - f(x_k)\end{aligned}$$

Proof II

By rewriting:

$$\sum_{i=1}^{k}(f(x_i) - f^*) \leq \frac{L}{2}\|x_0 - x^*\|^2$$

As last iterate is the best (sufficient decrease):

$$f(x_k) - f^* \leq \frac{1}{k}\sum_{i=1}^{k} f(x_i) - f^* \leq \frac{L}{2k}\|x_0 - x^*\|^2 \quad \square$$