Machine Learning: Boosting

Mahbub Majumdar

BRAC University
66 Mohakhali
Dhaka, Bangladesh

December 10, 2019

Convexity sets

- Convex sets:

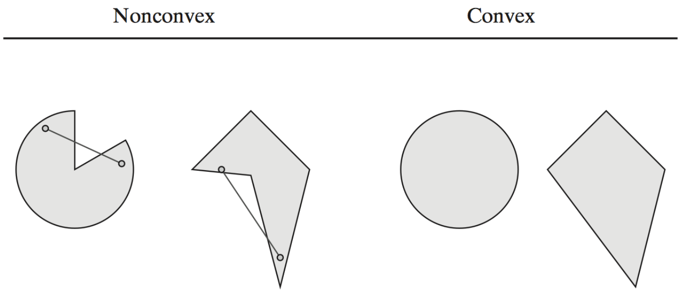

Figure: Convex and non-convex sets

- Convex: Straight line between two points is contained in the set.
- Non-convex: straight lines pass through the exterior

Convex functions

- Convex function:

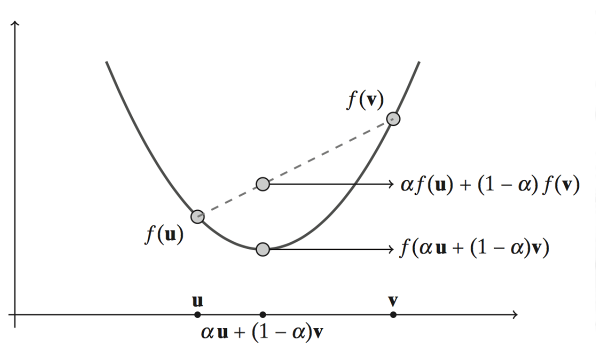

Figure: Convex function

- Straight line between two points lies above the function

$$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \leq \alpha f(\mathbf{u}) + (1-\alpha) f(\mathbf{v}) \tag{1}$$

Global Minima

- ERM $\Rightarrow$ minimize the error.
- Local minima of convex functions are global minima
- Draw a straight line between any two points. Minimum will be below the straight line $\Rightarrow$ Global Minimum
- Loss functions that are convex are convenient.

Derivative of a convex function

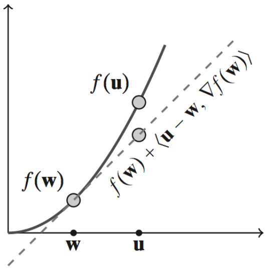

Figure: Tangent lies below f

- The tangent line to f lies below f,

$$\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle \tag{2}$$

Conditions for convexity

Lemma 12.3

Let $f : \mathbb{R} \to \mathbb{R}$ be a scalar twice differential function, and let f', f'' be its first and second derivatives, respectively. Then, the following are equivalent

1. f is convex
2. f is monotonically nondecreasing
3. f'' is nonnegative

- $f(x) = x^2$ is convex
- $f(x) = \log(1 + e^x)$ is convex

More examples

- Suppose $\{f_1, \ldots f_i, \ldots f_n\}$ are convex then linear combinations and the following are convex,
- $g(x) = \max\{f_i\}$, is convex
- $g(x) = \sum_i \alpha_i f_i(x)$, is convex
- Also, compositions of convex functions and linear functions are convex: $f(\langle w, x \rangle)$ is convex.

Lipschitzness

Denition 12.6 (Lipschitzness)

Let $C \subset \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is ρ -Lipschitz over C if for every $\mathbf{w}_1, \mathbf{w}_2 \in C$ we have that $\|f(\mathbf{w}_1) - f(\mathbf{w}_2)\| \leq \rho \|\mathbf{w}_1 - \mathbf{w}_2\|$

- Lipschitz functions don't change very fast (MVT):

$$f(w_1) - f(w_2) = f'(u)(w_1 - w_2) \tag{3}$$

- $|x|$ is Lipschitz
- $\log(1 + e^x)$ is Lipschitz
- $f(x) = x^2$ is not ρ-Lipschitz over the set $C = \mathbb{R}$ for any ρ. But it is ρ-Lipschitz over the set $C = \{x | x \leq \rho/2\}$.
- $f : \mathbb{R}^d \to \mathbb{R}$ is $\mathbf{v}$-Lipschitz for $\mathbf{v} \in \mathbb{R}^d$, where $f(\mathbf{w}) = \langle \mathbf{w}, \mathbf{v} \rangle + b$.
- Let $f(x) = g_1(g_2(x))$, where g_1 is ρ_1-Lipschitz and g_2 is ρ_2-Lipschitz. Then, f is $(\rho_1 \rho_2)$-Lipschitz.

Smoothness

Denition 12.8 (Smoothness)

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if its gradient is β-Lipschitz; namely, for all v,w we have

$$\|\nabla f(\mathbf{v}) - \nabla f(\mathbf{w})\| \leq \beta \|\mathbf{v} - \mathbf{w}\|$$

.

- For all $\mathbf{v}, \mathbf{w}$, smoothness implies

$$f(\mathbf{v}) \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{\beta}{2} \|\mathbf{v} - \mathbf{w}\|^2 \tag{4}$$

- Convexity implies: $f(\mathbf{v}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle$
- Thus smoothness and convexity give an upper and lower bound on f.

Properties of Smoothness

- Setting $\mathbf{v} = \mathbf{w} - \frac{1}{\beta}\nabla f(\mathbf{w})$, we can show that

$$\frac{1}{2\beta}\|\nabla f(\mathbf{w})\|^2 \leq f(\mathbf{w}) - f(\mathbf{v}) \tag{5}$$

- If $f(\mathbf{v}) \geq 0$ for all $\mathbf{v}$ then, f is *self-bounded*,

$$\|\nabla f(\mathbf{w})\|^2 \leq 2\beta f(\mathbf{w}) \tag{6}$$

- $f(x) = x^2$ is 2-smooth

- $f(x) = \log(1 + e^x)$ is $\frac{1}{4}$-smooth. $f'(x)$ is $\frac{1}{4}$-Lipschitz. Thus f is also self-bounded.

Examples of smooth functions

- Let $f(\mathbf{w}) = g(\langle \mathbf{w}, \mathbf{x} \rangle + b)$, where $g : \mathbb{R} \to \mathbb{R}$ is a β-smooth function, $\mathbf{x} \in \mathbb{R}^d$, and $b \in \mathbb{R}$. Then, f is $(\beta \|x\|^2)$-smooth.

  Proof: Using the chain rule, $\nabla f(\mathbf{w}) = g'(\langle \mathbf{w}, \mathbf{x} \rangle + b)\mathbf{x}$,

$$
\begin{aligned}
f(\mathbf{v}) &= g(\langle \mathbf{v}, \mathbf{x} \rangle + b) \\
&\leq g(\langle \mathbf{w}, \mathbf{x} \rangle + b) + g'(\langle \mathbf{w}, \mathbf{x} \rangle + b)\langle \mathbf{v} - \mathbf{w}, \mathbf{x} \rangle + \frac{\beta}{2}(\langle \mathbf{v} - \mathbf{w}, \mathbf{x} \rangle)^2 \\
&\leq g(\langle \mathbf{w}, \mathbf{x} \rangle + b) + g'(\langle \mathbf{w}, \mathbf{x} \rangle + b)\langle \mathbf{v} - \mathbf{w}, \mathbf{x} \rangle + \frac{\beta}{2}(\|\mathbf{v} - \mathbf{w}\| \|\mathbf{x}\|)^2 \\
&= f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{\beta \|\mathbf{x}\|^2}{2} \|\mathbf{v} - \mathbf{w}\|^2
\end{aligned}
$$

- Then $f(\mathbf{w}) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$ is $2\|x\|^2$-smooth, and $f(\mathbf{w}) = \log(1 + \exp(-y\langle \mathbf{w}, \mathbf{x} \rangle))$ is $\frac{1}{4}\|x\|^2$-smooth.

Convex Learning Problems

- $\mathcal{H} \subset C$.
- For example $C = \mathbb{R}^d$.
- Then for every $\mathbf{w} \in \mathbb{R}^d$, there is a h in $\mathcal{H}$.
- Regression problems $\rightarrow$ find the weights $\mathbf{w}$.

Definition 12.10 (Convex Learning Problem)

A learning problem, $(H, \ell, Z,)$, is called convex if the hypothesis class $\mathcal{H}$ is a convex set and for all $z \in Z$, the loss function, $\ell(\cdot, z)$, is a convex function (where, for any $z, \ell(\cdot, z)$ denotes the function $f : \mathcal{H} \rightarrow R$ defined by $f(\mathbf{w}) = \ell(w, z)$).

Convex Learning Problems

Lemma 12.11

If ℓ is a convex loss function and the class $\mathcal{H}$ is convex, then the $\text{ERM}_{\mathcal{H}}$ problem, of minimizing the empirical loss over $\mathcal{H}$, is a convex optimization problem (that is, a problem of minimizing a convex function over a convex set).

- **Proof:**

$$\text{ERM}_{\mathcal{H}}(S) = \underset{\mathbf{w}\in\mathcal{H}}{\text{argmin}} L_S(\mathbf{w}) \tag{7}$$

- $S = \{z_1, \ldots, z_n\}$
- $L_S(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m} \ell(\mathbf{w}, z_i)$ is a convex function.
- Therefore, the ERM rule is a problem of minimizing a convex function subject to the constraint that the solution should be in a convex set.

Convex-Lipschitz/Smooth-Bounded Learning Problems

Denition 12.12 (Convex-Lipschitz-Bounded Learning Problem)

The hypothesis class $\mathcal{H}$ is a convex set and for all $\mathbf{w} \in \mathcal{H}$ we have $\|\mathbf{w}\| \leq B$. For all $z \in Z$, the loss function, $\ell(\cdot, z)$, is a convex and ρ -Lipschitz function.

-