Convex analysis: preliminaries

Convex set:
A subset $C$ of a vector space is convex if for any $u, v \in C$, $\forall \lambda \in [0,1]$, $\lambda u + (1 - \lambda)v \in C$

Convex function

Let $C$ be a convex set. A fn. $f : C \rightarrow \mathbb{R}$ is convex if $\forall u, v \in C$, $\forall \lambda \in [0,1]$, $f(\lambda u + (1 - \lambda)v) \leq \lambda f(u) + (1 - \lambda)f(v)$

$f$ is called strictly convex if the above inequality is strict for all $\lambda \in (0,1)$. 
Convex function (equivalent definition)

Let $C$ be a convex set. A fn. $f : C \to \mathbb{R}$ is convex if

$\forall \ u \in C, \ \exists$ a vector $\partial f(u)$ such that $\forall \ v \in C$, we have

$$f(v) \geq f(u) + \left\langle \partial f(u), v - u \right\rangle$$

$f$ is called strictly convex if the above inequality is strict for all $v \neq u$. 
Convex functions – cont’d

- The definition says that \( f \) is convex if at any point there is a hyperplane tangential to and lies below the surface of \( f \).
- \( \partial f(u) \) is called a subgradient of \( f \) at \( u \).
- The subgradient of \( f \) at any given point is not necessarily unique. However, if \( f \) is differentiable over \( C \), then at every point \( u \in C \) there is a unique subgradient for \( f \) denoted as \( \nabla f(u) \).

**Ex:** \( f(x) = |x|, x \in \mathbb{R} \) has infinitely many subgradients at \( x = 0 \).

- When \( f \) is differentiable over \( C \subseteq \mathbb{R}^d \), then \( \forall \ w = (w_1, \ldots, w_d) \in C \), \( \nabla f(w) = \begin{bmatrix} \frac{\partial f(w)}{\partial w_1} \\ \vdots \\ \frac{\partial f(w)}{\partial w_d} \end{bmatrix} \)
Convex functions – cont’d

Minimum/minimizer of a convex function:

Let $C$ be a convex set. If $f : C \rightarrow \mathbb{R}$ is a convex function, then $f$ has a unique minimum over $C$ (but not necessarily unique minimizer). If $f$ is strictly convex over $C$, then the minimizer of $f$ over $C$ is also unique.
Convex functions – cont’d

Composition of convex functions:
Let $f_1 : \mathbb{R}^d \rightarrow \mathbb{R}$ and $f_2 : \mathbb{R} \rightarrow \mathbb{R}$ be convex functions. Then, the composition $f_2 \circ f_1$ defined as $f_2 \circ f_1 (w) = f_2 \left( f_1 (w) \right)$ is convex if

**either** one of the following conditions holds:

- $f_1$ is affine (i.e., linear with a constant term, e.g., $f(w) = \langle w, x \rangle + b$)
- $f_2$ is non-decreasing.

One can easily generalize this to more than two functions:

Let $f_1 : \mathbb{R}^d \rightarrow \mathbb{R}$, $f_i : \mathbb{R} \rightarrow \mathbb{R}$, $i = 2, \ldots, r$

$f_r \circ f_{r-1} \circ \cdots \circ f_j \circ f_{j-1} \circ \cdots \circ f_1$ is convex if each individual function is convex, and

**either** one of the following conditions holds:

- $f_{r-1}, \ldots, f_1$ are affine
- $f_r, f_{r-1}, \ldots, f_2$ are non-decreasing
- there is a $j \in \{2, \ldots, r-1\}$ s.t. $f_{j-1}, \ldots, f_1$ are affine, and $f_r, f_{r-1}, \ldots, f_{j+1}$ are non-decreasing.
**Lipschitz functions**

**Definition:** \((\rho\text{-Lipschitz function})\)

A function \(f : \mathbb{R}^d \rightarrow \mathbb{R}\) is \(\rho\text{-Lipschitz}\) (w.r.t. \(L_2\) norm) if for all \(v, w \in \mathbb{R}^d\),

\[
|f(v) - f(w)| \leq \rho \|v - w\|
\]

(From now on, we will use \(\|\cdot\|\) to denote the \(L_2\) norm.)

**Remark:**

When \(f : \mathbb{R}^d \rightarrow \mathbb{R}\) is differentiable over its domain, then \(\rho\text{-Lipschitzness of } f\) is equivalent to: \(\forall w \in \mathbb{R}^d, \|\nabla f(w)\| \leq \rho\)

**Fact:** Let \(f(w) = g_2(g_1(w)) \quad \forall w\). If \(g_1\) is \(\rho_1\)-Lipschitz and \(g_2\) is \(\rho_2\)-Lipschitz, then \(f\) is \(\rho_1\rho_2\)-Lipschitz.
Learnability of **convex-Lipschitz-bounded** problems

**Theorem:**
In the convex learning model, a hypothesis class $\mathcal{H}$ (described by a parameter set $C \subset \mathbb{R}^d$) is learnable (in the sense of the generalized agnostic-PAC learnability) if **all** the following conditions hold:

- The parameter set $C$ is convex and $M$-bounded for some $M > 0$, (i.e., $\exists M > 0$ s.t. $\forall v, w \in C$, $\|v - w\| \leq M$).
- The loss function $\ell : C \times \mathcal{Z} \rightarrow \mathbb{R}_+$ satisfies the following: $\exists \rho > 0$ s.t. $\forall z \in \mathcal{Z}$, $\ell(\cdot, z)$ is convex and $\rho$-Lipschitz.

Under these conditions, $\mathcal{H}$ is learnable via an **efficient convex optimization** algorithm that requires a training set of size

$$n_{\mathcal{H}}(\epsilon, \delta) = O\left(\frac{M^2 \rho^2}{\epsilon^2 \delta^2}\right)$$
Examples of Convex-Lipschitz loss

- Hinge loss:

\[ \ell(w, (x, y)) = \max\left( 0, 1 - y \langle w, \tilde{x} \rangle \right) \]

where \( x \in \mathbb{R}^m \), \( \tilde{x} \triangleq (x, 1) \), \( y \in \{-1, +1\} \)

\( w \in \mathbb{R}^{m+1} \)

Let's define

\[ g_1(w) \triangleq y \langle w, \tilde{x} \rangle \], \( g_2(a) = \max(0, 1 - a) \)

Note \( \ell(\cdot, (x,y)) = g_2 \circ g_1 \)

\( g_1 \) is linear and \( \|\tilde{x}\|-\text{Lipschitz} \).

\( g_2 \) is convex and 1-Lipschitz.

Hence, \( \ell(\cdot, (x,y)) \) is convex and \( \|\tilde{x}\|-\text{Lipschitz} \).

- Soft-threshold loss for linear classification (as opposed to hard-threshold 0-1 loss).
- Used in practice in implementation of Support Vector Machines (SVMs).
Examples of Convex-Lipschitz loss

• Hinge loss:

\[ \ell(w, (x, y)) = \max(0, 1 - y \langle w, \tilde{x} \rangle) \]

where \( x \in \mathbb{R}^m, \tilde{x} \triangleq (x, 1), y \in \{-1, +1\} \)

\[ w \in \mathbb{R}^{m+1} \]

Let's define

\[ g_1(w) \triangleq y \langle w, \tilde{x} \rangle, \quad g_2(a) = \max(0, 1 - a) \]

Note \( \ell(\cdot, (x, y)) = g_2 \circ g_1 \)

\( g_1 \) is linear and \( \| \tilde{x} \| \)-Lipschitz.

\( g_2 \) is convex and 1-Lipschitz.

Hence, \( \ell(\cdot, (x, y)) \) is convex and \( \| \tilde{x} \| \)-Lipschitz.

Does this hinge-loss model fit directly into the convex-Lipschitz-bounded learning model, or are there extra conditions that must be satisfied?
Examples of Convex-Lipschitz loss

- Logistic loss:

\[ \ell(w, (x, y)) = \log(1 + \exp(-y \langle w, \tilde{x} \rangle)) \]

where \( x \in \mathbb{R}^m \), \( \tilde{x} \triangleq (x, 1) \), \( y \in \{-1, +1\} \)

\[ w \in \mathbb{R}^{m+1} \]

Let’s define

\[ g_1(w) \triangleq y \langle w, \tilde{x} \rangle, \quad g_2(a) = \log(1 + \exp(-a)) \]

Note \( \ell(\cdot, (x, y)) = g_2 \circ g_1 \)

\( g_1 \) is linear and \( \| \tilde{x} \| \)-Lipschitz.

\( g_2 \) is convex and 1-Lipschitz.

Hence, \( \ell(\cdot, (x, y)) \) is convex and \( \| \tilde{x} \| \)-Lipschitz.

- Soft-threshold loss for linear classification (Smöother than hinge loss).
- Used in logistic regression (a powerful model for linear classification).
The Gradient Descent Algorithm

Outline:

We will first consider the general form of the basic (non-stochastic) GD algorithm for unconstrained convex optimization of Lipschitz functions (not specifically in the learning context).

Then, we will consider a more advanced algorithm:
- **Constrained** optimization: \( C \subset \mathbb{R}^d \)
- **Stochastic** GD

By a direct instantiation of the SGD algorithm, we obtain an efficient learner in the convex-Lipschitz-bounded model.
Basic GD algorithm (non-stochastic, unconstrained optimization)

- **Inputs:**
  - A convex $\rho$-Lipschitz function: $f: \mathbb{R}^d \rightarrow \mathbb{R}$

  (for simplicity, we will assume that $f$ is differentiable, i.e., the gradient $\nabla f(w)$ exists for all $w \in \mathbb{R}^d$. Note that this is not necessary: we can use a sub-gradient $\partial f(w)$ instead of $\nabla f(w)$ whenever $f$ is non-differentiable.)

  - Number of iterations: $T$

  - Set of scalars: $\{\alpha_t : t = 1, \ldots, T - 1\}$ ($\alpha_t$ called the learning rate).

- **Output:**

  An estimate $\hat{w}$ of $w^*$ where $w^* \in \arg\min_{w \in \mathbb{R}^d} f(w)$
Basic GD algorithm (non-stochastic, unconstrained optimization)

- **Inputs:** \( f : \mathbb{R}^d \to \mathbb{R}, \ T, \ \{\alpha_t : t = 1, \ldots, T - 1\} \)

1. **Initialization:** Choose an initial point \( \mathbf{w}_1 = \mathbf{0} \) (all-zero vector in \( \mathbb{R}^d \))

2. **FOR** \( t = 1, \ldots, T - 1 \)
   
   Take a GD step: \( \mathbf{w}_{t+1} = \mathbf{w}_t - \alpha_t \nabla f(\mathbf{w}_t) \)

3. **Return** \( \hat{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t \)

At iteration \( t \)

Direction of the negative gradient = \( -\alpha_t \nabla f(\mathbf{w}_t) \)
Basic GD algorithm (non-stochastic, unconstrained optimization)

- **Inputs:** \( f : \mathbb{R}^d \rightarrow \mathbb{R}, \ T, \ \{\alpha_t : t = 1, \ldots, T - 1\} \)

1. Initialization: Choose an initial point \( \mathbf{w}_1 = \mathbf{0} \) (all-zero vector in \( \mathbb{R}^d \))

2. FOR \( t = 1, \ldots, T - 1 \)
   
   Take a GD step: \( \mathbf{w}_{t+1} = \mathbf{w}_t - \alpha_t \nabla f(\mathbf{w}_t) \)

3. Return \( \hat{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t \)

**Theorem:** (Convergence of Basic GD Algorithm)

Let \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) be a convex, \( \rho \)-Lipschitz function. Let \( \mathbf{w}^* \in \text{arg min}_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}) \) and suppose that \( \| \mathbf{w}^* \| \leq M \). If we run the GD algorithm above for \( T \) steps with \( \alpha_t = \alpha = \frac{M}{\rho \sqrt{T}}, \forall t \). Then, \( f(\hat{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{M \rho}{\sqrt{T}} \)
Basic GD algorithm: Proof of the main theorem

Theorem: (Convergence of Basic GD Algorithm)
Let \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) be a convex, \( \rho \)-Lipschitz function. Let \( \mathbf{w}^* \in \arg \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}) \) and suppose that \( \| \mathbf{w}^* \| \leq M \). If we run the GD algorithm above for \( T \) steps with \( \alpha_t = \alpha = M / (\rho \sqrt{T}) \). Then, \( f(\hat{\mathbf{w}}) - f(\mathbf{w}^*) \leq M \rho / \sqrt{T} \)

First, let’s define \( \psi_t = \left\| \mathbf{w}_t - \mathbf{w}^* \right\|^2 \), \( \forall t \) (\( \psi_t \) usually called the potential).

We prove the theorem using the following claims:

Claim 1: \( \forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \leq \frac{\psi_t - \psi_{t+1}}{2\alpha} + \frac{\alpha \rho^2}{2} \)

Claim 2: \( \forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \geq f(\mathbf{w}_t) - f(\mathbf{w}^*) \)
Basic GD algorithm: Proof of the main theorem

Claim 1: \( \forall t, \langle \nabla f(w_t), w_t - w^* \rangle \leq \frac{\psi_t - \psi_{t+1} + \alpha \rho^2}{2\alpha} \)

Claim 2: \( \forall t, \langle \nabla f(w_t), w_t - w^* \rangle \geq f(w_t) - f(w^*) \)

Combining the two claims:

\[
\forall t, \quad f(w_t) - f(w^*) \leq \frac{\psi_t - \psi_{t+1} + \alpha \rho^2}{2\alpha} \quad (#)
\]

Given those claims are true, now observe

\[
f(\hat{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^{T} (f(w_t) - f(w^*))
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} (\psi_t - \psi_{t+1}) + \frac{\alpha \rho^2}{2}
\]

\[
= \frac{1}{2\alpha T} (\psi_1 - \psi_{T+1}) + \frac{\alpha \rho^2}{2}
\]

\[
\leq \frac{\psi_1}{2\alpha T} + \frac{\alpha \rho^2}{2} \leq \frac{M^2}{2\alpha T} + \frac{\alpha \rho^2}{2} = \frac{M \rho}{\sqrt{T}}
\]

By convexity of \( f \) and the definition of \( \hat{w} = \frac{1}{T} \sum_{t=1}^{T} w_t \)

Follows directly from (#)

Telescopic sum: terms cancel.

2nd ineq.: \( \psi_1 = \|w^*\|^2 \leq M^2 \), Last equality: by substitution with \( \alpha = M / (\rho \sqrt{T}) \)
Basic GD algorithm: Proof of the main theorem

Claim 1:  \[ \forall t, \langle \nabla f(w_t), w_t - w^* \rangle \leq \frac{\psi_t - \psi_{t+1}}{2\alpha} + \frac{\alpha \rho^2}{2} \]

Proof:

\[ \psi_{t+1} = \left\| w_{t+1} - w^* \right\|^2 = \left\| w_t - \alpha \nabla f(w_t) - w^* \right\|^2 \]

\[ = \langle w_t - w^* - \alpha \nabla f(w_t), w_t - w^* - \alpha \nabla f(w_t) \rangle \]

\[ = \left\| w_t - w^* \right\|^2 - 2\alpha \left\langle \nabla f(w_t), w_t - w^* \right\rangle + \alpha^2 \left\| \nabla f(w_t) \right\|^2 \]

\[ \leq \psi_t - 2\alpha \left\langle \nabla f(w_t), w_t - w^* \right\rangle + \alpha^2 \rho^2 \quad \text{since} \quad \left\| \nabla f(w_t) \right\| \leq \rho \quad \text{by} \quad \rho\text{-Lipschitzness of } f \]

By rearranging the terms, the proof is complete.
Basic GD algorithm: Proof of the main theorem

Claim 2: \( \forall t, \langle \nabla f(w_t), w_t - w^* \rangle \geq f(w_t) - f(w^*) \)

Proof:

By convexity of \( f \) (recall the equivalent definition of convex functions):

\[
f(w^*) \geq f(w_t) + \langle \nabla f(w_t), w^* - w_t \rangle = f(w_t) - \langle \nabla f(w_t), w_t - w^* \rangle
\]

By rearranging the terms, the proof is complete.

We just proved:

Theorem: (Convergence of Basic GD Algorithm)

Let \( f : \mathbb{R}^d \to \mathbb{R} \) be a convex, \( \rho \)-Lipschitz function. Let \( w^* \in \arg \min_{w \in \mathbb{R}^d} f(w) \) and suppose that \( \|w^*\| \leq M \). If we run the GD algorithm above for \( T \) steps with \( \alpha_t = \alpha = M / (\rho \sqrt{T}) \). Then, \( f(\hat{w}) - f(w^*) \leq M \rho / \sqrt{T} \).