Online Convex Programming and Generalized Gradient Ascent (Martin Zinkevich)

Definition 1: (Convex set) A set of vectors \( S \subseteq \mathbb{R}^n \) is convex if for all \( x, y \in S \) and all \( \lambda \in [0, 1] \) \( \lambda x + (1 - \lambda) y \in S \).

Definition 2: (Convex function) For a convex set \( F \), a function \( f : F \to \mathbb{R} \) is convex if for all \( x, y \in F \) and for all \( \lambda \in [0, 1] \) \( \lambda f(x) + (1 - \lambda) f(y) \leq f(\lambda x + (1 - \lambda) y) \).

The chord between two points on the graph lies above the graph.

Definition 3: (Convex programming problem) A convex programming problem consists of a convex feasible set \( F \) and a convex cost function \( c : F \to \mathbb{R} \). The optimal solution is the solution that minimizes the cost.

Definition 4: (Online convex programming problem) An online convex programming problem consists of a feasible set \( F \subseteq \mathbb{R}^n \) and infinite sequence \( \{c^1, c^2, \ldots\} \) where each \( c^t : F \to \mathbb{R} \) is a convex function.

Let \( \|x\| = \sqrt{x \cdot x} \) and \( d(x, y) = \|x - y\| \)

Assumptions:

1. The feasible set \( F \) is bounded. \( \forall t \in \mathbb{N} \) s.t. \( \forall x \in F \) \( d(x, y) \leq N \|x\| \)
2. The feasible set \( F \) is closed. \( \forall \) sequences \( \{x^t, x^{t+1}, \ldots\} \) where \( x^t \in F \)
For all \( t \), if there exists a \( \exists x^* \in \mathbb{R}^n \) such that \( x^* = \lim_{t \to \infty} x^t \) then \( x^* \in F \).
3. The feasible set \( F \) is nonempty. \( \exists x \in F \)
4. For all \( t \), \( c^t \) is differentiable.
5. \( \exists t \in \mathbb{N} \) s.t. \( \forall t \forall x \in F \) \( \|C^t(x)\| \leq N \|x\| \|\nabla c^t\| \)

6. \( \exists \) algorithm which produces \( \forall \) \( c^t(x) \forall x \in F \), \( \forall t \).
7. \( \exists \) algorithm which can produce \( \text{argmin}_{x \in F} d(x, y) \forall y \in \mathbb{R} \).

We define the projection \( P(y) = \text{argmin}_{x \in F} d(x, y) \forall y \in \mathbb{R} \).
Definition 5 (Cost) Given an algorithm $A$, and a convex programming problem $(F, \{c_1, c_2, \ldots, c_T\})$. If $\{x^1, x^2, \ldots, x^T\}$ are the vectors selected by $A$, then the cost of $A$ until time $T$ is

$$C_A(T) = \sum_{t=1}^{T} c_t(x^t)$$

Then cost of static feasible solution $x \in F$ until time $T$ is

$$C_x(T) = \sum_{t=1}^{T} c_t(x)$$

The regret of algorithm $A$ until time $T$ is

$$R_A(T) = C_A(T) - \min_{x \in F} C_x(T)$$

Algorithm 1 (Greedy Projection) Select an arbitrary $x^0 \in F$ and a sequence of learning rates $\eta_t \in \mathbb{R}^+$. In time $t$, after receiving the cost function, select the next vector $x^{t+1}$ according to:

$$x^{t+1} = P(x^t - \eta_t \nabla c(x^t))$$

Basic principle: If we consider the case where the cost functions are all the same, then the algorithm operating on unchanging valley, the boundary of the feasible set is the edge of the valley. By proceeding along the direction opposite the gradient, we walk down into the valley. By projection back into the convex set, we skirt the edge of the valley.
Analyzing the Performance of the Algorithm

The linear case is the worst case

- Begin with arbitrary $\{c^t, c^0, \ldots\}$, run the algorithm and compute $\{x^t, x^0, \ldots\}$. Let $g^t = \nabla c^t(x^t)$. 

- A property of convex functions (reminder: $c^t$ is convex function)
  $$\forall x \in E \quad c^t(x) \geq \langle \nabla c^t(x^t), x - x^t \rangle + c^t(x^t)$$

- Let $x^* \in E$ be a statically optimal vector.
  $$c^t(x^t) - c^t(x^*) \leq c^t(x^t) - \langle g^t(x^t), x^* - x^t \rangle$$

  $$= g^t \cdot x^t - g^t \cdot x^*.$$  

  **Bound the regret at time $t$**

- $y^{t+1} = x^t - \eta_g^t g^t \quad \Rightarrow \quad x^{t+1} = \rho(y^{t+1})$

- $(y^{t+1} - x^*)^2 = (x^t - x^* - \eta_g^t g^t)^2 = (x^t - x^*)^2 - 2 \eta_g^t g^t \cdot (x^t - x^*) + \eta_g^t \|g^t\|^2$

- $\left( \frac{1}{\lambda} \right)^{\frac{1}{2}} \leq \left( \frac{1}{\lambda} \right)^{\frac{1}{2}} \cdot \frac{(x^t - x^*)^2}{\|\nabla c\|^2}$

- $$(x^t - x^*) g^t \leq \frac{1}{2 \lambda t} \left( (x^t - x^*)^2 - (x^{t-1} - x^*)^2 \right) + \frac{\eta_g^2}{2} \|\nabla c\|^2$$

Proof in the appendix (lemma 5)
Bound the regret

\[ R_0(T) = \sum_{t=1}^{T} \left( C^t(x^t) - C^t(x^*) \right) \leq \sum_{t=1}^{T} \| \nabla C_t \|^2 \]

\[ \leq \sum_{t=1}^{T} \left( \frac{1}{2} \nabla_t^2 (x^t - x^*)^2 - \frac{1}{2} \nabla_t^2 (x^{t-1} - x^*)^2 \right) + \frac{\eta_t}{2} \| \nabla C_t \|^2 \]

\[ = \frac{1}{2} \sum_{t=1}^{T} \nabla_t^2 (x^t - x^*)^2 - \frac{1}{2} \sum_{t=1}^{T} \nabla_t^2 (x^{t-1} - x^*)^2 + \frac{\| \nabla C_t \|^2}{2} \leq \frac{\| \nabla C_t \|^2}{2} \leq \eta_t \]

\[ \leq \frac{1}{2} \eta_t (x^t - x^*)^2 + \frac{1}{2} \sum_{t=1}^{T} \left( \frac{\eta_t}{\eta_{t-1}} - \frac{\eta_{t-1}}{\eta_t} \right)(x^{t-1} - x^*)^2 \leq \| \nabla C_t \|^2 \]

\[ \leq \frac{1}{2} \eta_t (x^t - x^*)^2 + \frac{1}{2} \sum_{t=1}^{T} \eta_t (x^t - x^*)^2 \leq \| \nabla C_t \|^2 \]

\[ R_0(T) \leq \| \nabla C_t \|^2 \frac{T}{2} + \| \nabla C_t \|^2 \frac{T}{2} \leq \| \nabla C_t \|^2 \frac{T}{2} + \| \nabla C_t \|^2 \frac{T}{2} \]

Optimizing

\[ \eta_t = \frac{\alpha}{t} \quad 0 < \alpha < 1 \]

\[ \sum_{t=1}^{T} \eta_t = \sum_{t=1}^{T} \frac{1}{t^\alpha} \leq \sum_{t=1}^{T} \frac{1}{t} \leq T \]

\[ \sum_{t=1}^{T} t^{-\alpha} dt = \frac{1}{1-\alpha} \left[ t^{-\alpha} \right]_{t=1}^{T} = \frac{T^{-1+\alpha} - 1}{1-\alpha} \]

\[ T^{-1+\alpha} = T^{1-\alpha} \]

\[ \alpha = \frac{1}{2} \]

\[ R_0(T) \leq \| \nabla C_t \|^2 \frac{T}{2} + \| \nabla C_t \|^2 \frac{T}{2} \]
Algorithm 2: (Lazy Projection) Select an arbitrary $x \in F$ and a sequence of learning rates $\eta_1, \eta_2, \ldots \in \mathbb{R}^+$. Define $y^0 = x$.

In time step $t$, after receiving a cost function, define $y^{t+1}$:

$$y^{t+1} = y^t - \eta_t \nabla C_t(x^t)$$

and select a vector $x^{t+1} = P(y^{t+1})$

* Let $\eta_t = \frac{1}{t}$.

$$R_L(T) = \sum_{t=1}^{T} (C_t(x^t) - C_t(x^*)^T) = \sum_{t=1}^{T} C_t(y_t) - C_t(y^*)^T = \sum_{t=1}^{T} (C_t(y_t) - C_t(x^*))$$

1. **Projection Potential**
2. **Ident Potential**

(1) Plugging-in $y_t$ instead of $x_t$:

$$\sum_{t=1}^{T} C_t(y_t) - C_t(x^*) \leq \frac{\|F\|^2}{2\eta} + \frac{\|C\|_2^2}{2} \sum_{t=1}^{T} \eta_t^2 - \frac{1}{2\eta} \sum_{t=1}^{T} (y^{t+1} - x^*)^2 = \frac{\|F\|^2}{2\eta} + \frac{\|C\|_2^2}{2} \sum_{t=1}^{T} \eta_t^2 - \frac{1}{2\eta} \sum_{t=1}^{T} (y^{t+1} - x^*)^2$$

(2) Plugging-in $y_t$ instead of $x_t$:

$$\sum_{t=1}^{T} C_t(x^t) - C_t(y_t) \leq \eta \sum_{t=1}^{T} (y_t - y_{t+1}) (P(y_t) - y_t)$$

$$y_t = y_{t+1} = \eta_t$$

$$\sum_{t=1}^{T} (y_t - P(y_t)) (y^{t+1} - y_t) \leq \frac{\|F\|^2}{2\eta} + \frac{\|C\|_2^2}{2} \sum_{t=1}^{T} \eta_t^2$$

we will show that...

$$R_L(T) \leq \frac{\|F\|^2}{2\eta} + \frac{\|C\|_2^2}{2} \sum_{t=1}^{T} \eta_t^2$$
Let \( d(y, x) = \min_{x \in F} d(y, x) \) → it is enough to show

that

\[
\sum_{t=1}^{T} (y^{t+1} - y^t)(y^t - p(y^t)) \leq \frac{d(y^{1+1}, x)}{a^t}
\]

A motivating example

\( F = \{ x \in \mathbb{R} : x \leq a \} \). If \( y^t \geq a \), \( x^t = p(y^t) = a \).

\[
\sum_{t=1}^{T} (y^{t+1} - y^t)(y^t - p(y^t)) = \sum_{t=1}^{T} (y^{t+1} - y^t)(y^t - a)
\]

[Let \( z^t = b(y^t, x) = y^t - a \)]

\[
\sum_{t=1}^{T} (z^{t+1} - z^t) \geq b\left(\sum_{t=1}^{T} (z^{t+1})^2 - (z^t)^2\right) \leq \frac{1}{2} \sum_{t=1}^{T} (z^{t+1})^2 - (z^t)^2 = \frac{1}{2} \left((z^{T+1})^2 - (z^1)^2\right) \leq \frac{d(y^{1+1}, F)}{a^t}
\]

\( a, b, e \in \mathbb{R} \)

\[
(a - b)^2 > 0 \rightarrow a^2 - 2ab + b^2 > 0
\]

\[
a^2 - b^2 > 2ab - 2b^2
\]

\[
a^2 - b^2 \geq b(a - b)
\]

The problem is \( (y^{t+1} - y^t)(y^t - p(y^t)) \neq (z^{t+1} - z^t)(z^t) \) in the general case. However, it is always less.
Geometric Lemmas

Lemma 5: Given a convex set $F \subseteq \mathbb{R}^n$, let $y \in \mathbb{R}^n$ and $x \in F$ then

$$(y - P(y))(x - P(y)) \leq 0.$$ 

Moreover, if $x \neq P(y)$, then

$$\gamma y \cdot x > 90^\circ.$$ 

Proof: By contradiction, assume $x = P(y)$.

The set $F$ can be defined as $\{x \in \mathbb{R}^n : (y - x)(x - x') \geq 0\}$.

For $\lambda \in [0,1]$, define $z(\lambda) = (1-\lambda)x' + \lambda x = x' + \lambda(x - x')$

We will prove that for small values of $\lambda$, $z(\lambda)$ is closer to $y$ than $x'$.

$$(y - z(\lambda))^2 = (y - x' - \lambda(x - x'))^2 = \lambda^2 (x - x')^2 - 2\lambda(y - x')(x - x') + (y - x')^2$$

Observe that for $0 < \lambda < \frac{x' - (x - x')}{(x - x')^2}$, $(y - z(\lambda))^2 < (y - x')^2$

Hence $P(y) = x'$.

Lemma 6: Given $y, x, y' \in \mathbb{R}^n$, then

$$(y - x)(y' - y) \leq \langle y, x \rangle \left(2\langle y', x \rangle - d(y, x)\right)$$

Proof: Start with simple example in $\mathbb{R}^2$, suppose that $x, y, y' \in \mathbb{R}^2$, $x = (0,0)$ and $y = (1,0)$.

$$(y - x)(y' - y) = (y', -1)$$

$$d(y, x) = \sqrt{(y')^2 + (y)^2} - 1 > (y', -1)$$
Now suppose that for some \( y, x, y' \), we have proven that the property holds.

Then it will hold for \( y+a, x+a, y'+a \) (a translation), because:

\[
(d(y+a) - (x+a)) \cdot (d(y'+a) - (y+a)) = (y-x) \cdot (y'-y) \leq d(y,x) \cdot d(y',x) - d(y,x)
\]

Also will hold for \( ky, kx, ky' \) (a scalar), because:

\[
(ky-x) \cdot (ky'-ky) = k^2(y-x) \cdot (y'-y) \leq k^2 d(ky,x) \cdot d(ky',x) - d(ky,x)
\]

Also the property is invariant under rotation. Suppose that \( R \) is an orthonormal matrix (where \( R^T R = I \)). Now for all \( a, b \in \mathbb{R}^n \):

1. \((Ra) \cdot (Rb) = (Ra)^T (Rb) = a^T R^T R b = a^T b = a \cdot b \)

2. \(d(Ra, Rb) = \sqrt{(R(a-b)) \cdot (R(a-b))^T} = (a-b) \cdot (a-b)^T = d(a,b) \)

so the property will hold for \( Ry, Rx, Ry' \), because:

\[
(Ry-Rx) \cdot (Ry'-Ry) = (R(y-x)) \cdot (R(y'-y)) = (y-x)(y'-y) \leq d(y,x) \cdot d(y',x) - d(y,x)
\]

\[
\leq d(Ry,Rx) \cdot d(Ry',Rx) - d(Ry,Rx)
\]

Observe that we can think of \( \mathbb{R}^2 \) as embedded in \( \mathbb{R}^n \) without changing distance or dot product. Any three vectors \( y, x, y' \) can be obtained from \( (0,1,0,0,\ldots), (0,0,\ldots,1) \), \( (y_1, y_2, 0, \ldots, 0) \) using translation, scaling, and rotation.
Lemma 73: Given \( y, x, x', y' \in \mathbb{R}^n \), where \((y - x) \cdot (x' - x) \leq 0\) then
\[
(y - x) \cdot (y' - y) \leq d(y, x) (d(y' | x') - d(y, x))
\]

Corollary 8: Given \( y, y' \in \mathbb{R}^n \), set \( z = t + c \cdot y, p(y) \) and \( z' = t(y', p(y')) \):
\[
(y - p(y)) \cdot (y' - y) \leq \varepsilon \left( \frac{\varepsilon' - \varepsilon}{2} \right). \]

Proof: If \( y = x \) then it is trivial. Thus, we begin with a simple example as before. We assume \( y, x, x', y' \in \mathbb{R}^3 \)
\[
y = (0, 0, 0), \quad x = (3, 3, 0) \text{ and } x' = (x_1, x_2, 0).
\]
Throughout the next part of the proof, we prove that the worst case occurs when \( x = x' \). We do this by defining \( y'' = y' - x + x \) and replacing \( y' \) with \( y'' \) and \( x' \) with \( x_0 \). Observe that \( d(y', x) = d(y', x') \). Since \( x = (0, 0, 0) \)

Observe that \((y - x) \cdot (y' - y) = (y, -1) \) and \((y - x) \cdot (y'' - y) = (y, -1) \).
Since \((y - x) \cdot (x' - x) = x_1, x_1 \leq 0\). Thus, \((y - x) \cdot (y' - y) \leq (y - x) (y'' - y), \) so the relationship only gets tighter as we force \( x = x' \). Hence, the property holds for these vectors by Lemma 6.

It is possible to complete the proof by scaling, translation, and rotation as in Lemma 6.