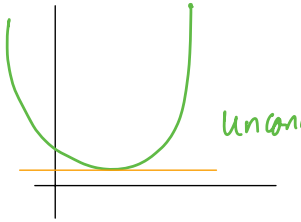
The Gradient Descent (GD) Algorithm

Outline:

$C = \mathbb{R}^p$

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 setting:

- **Constrained** optimization: $C \subset \mathbb{R}^p$

Convex and bounded subset of $\mathbb{R}^d$

Required: Solve $\min_{w \in C} f(w)$

where f is a convex, ρ-Lipschitz function (for some $\rho > 0$)

In this case, $w^* = \arg\min_{w \in C} f(w)$ is called the constrained minimizer of f over C.

The Gradient Descent (GD) Algorithm

Outline:

$C = \mathbb{R}^p$

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 setting:

- **Constrained** optimization: $C \subset \mathbb{R}^p$

- The gradient descent (GD) algorithm is an iterative algorithm for solving the above optimization problem (for both the unconstrained and constrained settings).

- As the number of iterations increases, the output of the GD algorithm converges to the minimizer w* of f.

- Next, we will consider **Stochastic** GD (SGD) algorithm for minimizing the loss in bounded-convex-Lipschitz learning problems.

Basic GD algorithm (non-stochastic, unconstrained optimization)

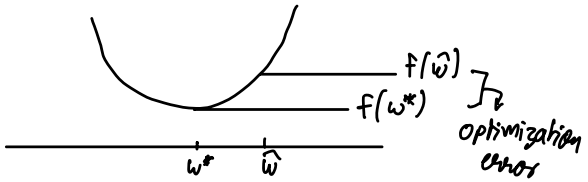
- **Inputs:**

➢ A **convex, ρ-Lipschitz** function: $f : \mathbb{R}^p \to \mathbb{R}$

we will assume that f is **differentiable**, i.e., the gradient $\nabla f(\mathbf{w})$ exists for all $\mathbf{w} \in \mathbb{R}^p$.

➢ Number of iterations: T

➢ Set of scalars: $\{\alpha_t : t = 1, \ldots, T-1\}$ (α_t called the ***step size*** or ***learning rate***).



- **Output:**

A vector $\hat{\mathbf{w}}$: an approximate version of $\mathbf{w}^*$ where $\mathbf{w}^* \in \arg\min_{\mathbf{w} \in \mathbb{R}^p} f(\mathbf{w})$

The quantity $f(\hat{w}) - f(w^*)$ is called the optimization error.

It measures the quality of the approximate solution $\hat{w}$.

NOT $\hat{w} - w^*$

Basic GD algorithm (non-stochastic, unconstrained optimization)

- **Inputs:** $f : \mathbb{R}^p \to \mathbb{R}$, T, $\{\alpha_t : t = 1, \ldots, T-1\}$

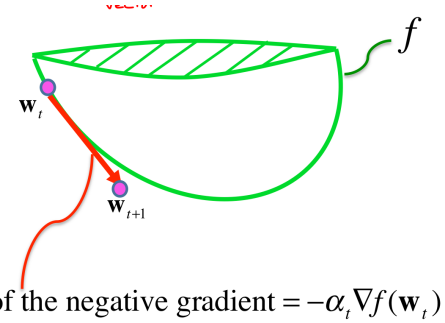
1. Initialization: Choose an initial point $\mathbf{w}_1 = \mathbf{0}$ (all-zero vector in $\mathbb{R}^p$)

2. **FOR** *t = 1, …, T – 1*

   Take a GD step: $\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha_t \nabla f(\mathbf{w}_t)$

   scalar; vector

   At iteration *t*



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

Averaging helps smooth out gradient because of generalness of our assumptions

gradient gives us dir of largest growth locally.

Theorem: (Convergence of Basic GD Algorithm)

In general, $\|\mathbf{w}^* - \mathbf{w}_1\| \le M$ → w_1 is initialized w. our case = 0.

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a convex, ρ-Lipschitz function. Let $\mathbf{w}^* \in \arg\min_{\mathbf{w} \in \mathbb{R}^p} f(\mathbf{w})$

and suppose that $\|\mathbf{w}^*\| \le M$. If we run the GD algorithm above for ***T*** steps

with $\alpha_t = \alpha = \frac{M}{\rho\sqrt{T}}, \forall t$. Then, $f(\hat{\mathbf{w}}) - f(\mathbf{w}^*) \le \frac{M\rho}{\sqrt{T}}$

constant step size

optimization error

This is optimal convergence rate for this type of objective functions

error bound scales as $\frac{1}{\sqrt{T}}$ (goes to 0 as $T \uparrow \infty$)

OPTIMAL WRT first-order methods (using gradients).

Basic GD algorithm: Proof of the main theorem

Theorem: (Convergence of Basic GD Algorithm)

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

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

We prove the theorem using the following claims:

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

The claims involve upper and lower bounds on the same quantity $\langle \nabla f(w_t), w_t - w^* \rangle$

Claim 2: $\forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \ge f(\mathbf{w}_t) - f(\mathbf{w}^*)$

Claim 1 gives an upper bound on $\langle \nabla f(w_t), w_t - w^* \rangle$ in terms of the difference of the potential functions at consecutive iterations $\psi_t - \psi_{t+1}$, step size α, and the Lipschitz constant ρ.

Basic GD algorithm: Proof of the main theorem

Theorem: (Convergence of Basic GD Algorithm)

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

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

We prove the theorem using the following claims:

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

inner product = projection

Claim 2: $\forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \ge f(\mathbf{w}_t) - f(\mathbf{w}^*)$

Claim 2 gives a lower bound on $\langle \nabla f(w_t), w_t - w^* \rangle$ in terms of the optimization error at iteration t: $f(w_t) - f(w^*)$.

Basic GD algorithm: Proof of the main theorem

Theorem: (Convergence of Basic GD Algorithm)

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

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

We prove the theorem using the following claims:

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

Claim 2: $\forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \ge f(\mathbf{w}_t) - f(\mathbf{w}^*)$

The quantity $\langle \nabla f(w_t), w_t - w^* \rangle$ can be viewed as a "proxy" quantity: combining the two claims we can upper bound the optimization error at iteration t in terms of the difference of potential functions, step size, and Lipschitz constant: $f(w_t) - f(w^*) \le \frac{\psi_t - \psi_{t+1}}{2\alpha} + \frac{\alpha\rho^2}{2}$

Basic GD algorithm: Proof of the main theorem

Theorem: (Convergence of Basic GD Algorithm)

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

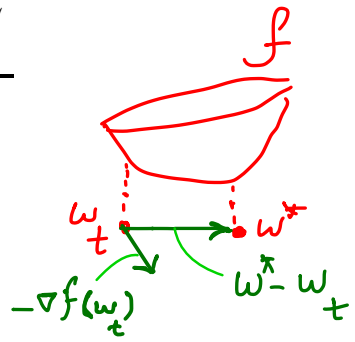
First, let's define $\psi_t = \left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2, \forall t$ (ψ_t usually called the potential).

We prove the theorem using the following claims:

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

Claim 2: $\forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \ge f(\mathbf{w}_t) - f(\mathbf{w}^*)$

$= \langle -\nabla f(w_t), w^* - w_t \rangle$



For now, suppose the two claims are true.

(We will first show how to use the two claims to prove the theorem, then we will prove the two claims.)

Basic GD algorithm: Proof of the main theorem

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: $\forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \geq f(\mathbf{w}_t) - f(\mathbf{w}^*)$

Combining the two claims: $\forall t,\ f(\mathbf{w}_t) - f(\mathbf{w}^*) \leq \frac{\psi_t - \psi_{t+1}}{2\alpha} + \frac{\alpha\rho^2}{2}$ **(#)**

optimization error at

Note that our goal is to show → Goal is to bound $\hat{w}$ & not w_t.

$$f(\hat{w}) - f(w^*) \leq \frac{M\rho}{\sqrt{T}}$$

$\hat{w}$ is the output of the algorithm $= \frac{1}{T}\sum_{t=1}^{T} w_t$

Let's see how can we show that given #

Basic GD algorithm: Proof of the main theorem

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: $\forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \geq f(\mathbf{w}_t) - f(\mathbf{w}^*)$

Combining the two claims: $\forall t,\ f(\mathbf{w}_t) - f(\mathbf{w}^*) \leq \frac{\psi_t - \psi_{t+1}}{2\alpha} + \frac{\alpha\rho^2}{2}$ **(#)**

Given those claims are true, now observe

$$f(\hat{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{w}_t) - f(\mathbf{w}^*)\right)$$

By convexity of f and the definition of $\hat{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$

$$f\left(\frac{1}{T}\sum_{t=1}^{T} w_t\right) \leq \frac{1}{T}\sum_{t=1}^{T} f(w_t)$$

Basic GD algorithm: Proof of the main theorem

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: $\forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \geq f(\mathbf{w}_t) - f(\mathbf{w}^*)$

Combining the two claims: $\forall t,\ f(\mathbf{w}_t) - f(\mathbf{w}^*) \leq \frac{\psi_t - \psi_{t+1}}{2\alpha} + \frac{\alpha\rho^2}{2}$ **(#)**

Given those claims are true, now observe

$$f(\hat{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{w}_t) - f(\mathbf{w}^*)\right)$$

By convexity of f and the definition of $\hat{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$

Recall: A fn $f : \mathbb{R}^p \to \mathbb{R}$ is convex if $\forall\, \lambda \in (0,1)$
$\forall\, u, v \in \mathbb{R}^p$ we have $f(\lambda u + (1-\lambda) v) \leq \lambda f(u) + (1-\lambda) f(v)$

In fact, this definition of convexity is equivalent to the following definition:

A fn $f : \mathbb{R}^p \to \mathbb{R}$ is convex if for any integer $K \geq 2$, any non-negative $\lambda_1, \ldots, \lambda_K$ s.t. $\sum_{j=1}^{K} \lambda_j = 1$, and any $v_1, \ldots, v_K \in \mathbb{R}^p$, we have $f\left(\sum_{j=1}^{K} \lambda_j v_j\right) \leq \sum_{j=1}^{K} \lambda_j f(v_j)$

$w \in \mathbb{R}^d : \|w\| \leq 1$

$\rho = 2(20)(10)$

$M = 2$

$(\langle w, x \rangle - y)^2$

$\nabla_w(\quad) = 2(\langle w, x \rangle - y)x$

$|\langle w, x \rangle - y| \leq 20$

$\| \quad \| = 2|\langle w, x \rangle - y|\|x\|$

≤ 10

$\langle w, x \rangle \geq 10$

Basic GD algorithm: Proof of the main theorem

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: $\forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \geq f(\mathbf{w}_t) - f(\mathbf{w}^*)$

Combining the two claims: $\forall t,\ f(\mathbf{w}_t) - f(\mathbf{w}^*) \leq \frac{\psi_t - \psi_{t+1}}{2\alpha} + \frac{\alpha\rho^2}{2}$ **(#)**

Given those claims are true, now observe

$$f(\hat{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{w}_t) - f(\mathbf{w}^*)\right)$$

By convexity of f and the definition of $\hat{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$

Let $k = T$, and let $\lambda_1 = \lambda_2 = \cdots = \lambda_T = \frac{1}{T}$ (Note $\sum_{j=1}^{T}\lambda_j = 1$).

Consider the parameter vectors $w_1, \ldots, w_T \in \mathbb{R}^p$.

Since f is convex, then

$f\left(\sum_{j=1}^{T}\lambda_j w_j\right) \leq \sum_{j=1}^{T}\lambda_j f(w_j)$ i.e., $f\left(\frac{1}{T}\sum_{j=1}^{T} w_j\right) \leq \sum_{j=1}^{T}\frac{1}{T} f(w_j) = \frac{1}{T}\sum_{j=1}^{T} f(w_j)$

Basic GD algorithm: Proof of the main theorem

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: $\forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \geq f(\mathbf{w}_t) - f(\mathbf{w}^*)$

Combining the two claims: $\forall t,\ f(\mathbf{w}_t) - f(\mathbf{w}^*) \leq \frac{\psi_t - \psi_{t+1}}{2\alpha} + \frac{\alpha\rho^2}{2}$ **(#)**

Given those claims are true, now observe

$$f(\hat{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{w}_t) - f(\mathbf{w}^*)\right)$$

$$\leq \frac{1}{2\alpha T}\sum_{t=1}^{T}\left(\psi_t - \psi_{t+1}\right) + \frac{\alpha\rho^2}{2}$$

$$= \frac{1}{2\alpha T}\left(\psi_1 - \psi_{T+1}\right) + \frac{\alpha\rho^2}{2}$$

$$\leq \frac{\psi_1}{2\alpha T} + \frac{\alpha\rho^2}{2} \leq \frac{M^2}{2\alpha T} + \frac{\alpha\rho^2}{2}$$

By convexity of f and the definition of $\hat{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$

Follows directly from **(#)**

Telescopic sum: terms cancel.

Since $\psi_1 = \left\|\mathbf{w}^*\right\|^2 \leq M^2$

note $\psi_1 = \|w_1 - w^*\|^2 = \|w^*\|^2$ since $w_1 = 0$

Basic GD algorithm: Proof of the main theorem

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: $\forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \geq f(\mathbf{w}_t) - f(\mathbf{w}^*)$

Combining the two claims: $\forall t,\ f(\mathbf{w}_t) - f(\mathbf{w}^*) \leq \frac{\psi_t - \psi_{t+1}}{2\alpha} + \frac{\alpha\rho^2}{2}$ **(#)**

Given those claims are true, now observe

$$f(\hat{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{w}_t) - f(\mathbf{w}^*)\right)$$

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

Telescopic Sum, will cancel

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

drop always -ve term

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

bounded by M^2.

minimizing α gives this bound

By convexity of f and the definition of $\hat{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$

Follows directly from **(#)**

Telescopic sum: terms cancel.

Since $\psi_1 = \left\|\mathbf{w}^*\right\|^2 \leq M^2$

Last equality: by substitution with $\alpha = M/(\rho\sqrt{T})$

Basic GD algorithm: Proof of the main theorem

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

Proof :

$$\psi_{t+1} = \left\|\mathbf{w}_{t+1} - \mathbf{w}^*\right\|^2 = \left\|\mathbf{w}_t - \alpha\,\nabla f(\mathbf{w}_t) - \mathbf{w}^*\right\|^2$$

$$= \left\langle \mathbf{w}_t - \mathbf{w}^* - \alpha\,\nabla f(\mathbf{w}_t)\,,\, \mathbf{w}_t - \mathbf{w}^* - \alpha\,\nabla f(\mathbf{w}_t) \right\rangle \quad (A-B)^2$$

$$= \left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2 - 2\alpha \left\langle \nabla f(\mathbf{w}_t)\,,\, \mathbf{w}_t - \mathbf{w}^* \right\rangle + \alpha^2 \left\|\nabla f(\mathbf{w}_t)\right\|^2$$

$$\leq \psi_t - 2\alpha \left\langle \nabla f(\mathbf{w}_t)\,,\, \mathbf{w}_t - \mathbf{w}^* \right\rangle + \alpha^2\rho^2$$

since $\left\|\nabla f(\mathbf{w}_t)\right\| \leq \rho$ by ρ-Lipschitzness of f

By rearranging the terms, the proof is complete.

Basic GD algorithm: Proof of the main theorem

Claim 2: $\forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \geq f(\mathbf{w}_t) - f(\mathbf{w}^*)$

Proof :

second definition

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

$$f(\mathbf{w}^*) \geq f(\mathbf{w}_t) + \langle \nabla f(\mathbf{w}_t), \mathbf{w}^* - \mathbf{w}_t \rangle$$

$$= f(\mathbf{w}_t) - \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle$$

Note the change in the sign in the second term inside the inner-product. (change sign)

By rearranging the terms, the proof is complete.

We just proved:

Theorem: (Convergence of Basic GD Algorithm)

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a convex, ρ-Lipschitz function. Let $\mathbf{w}^* \in \arg\min_{\mathbf{w} \in \mathbb{R}^p} f(\mathbf{w})$ 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, $f(\hat{\mathbf{w}}) - f(\mathbf{w}^*) \leq M\rho / \sqrt{T}$

The Gradient Descent (GD) 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 setting:

- **Constrained** optimization: $C \subset \mathbb{R}^p$ ← ***Now***

- Next, we will consider **Stochastic** GD (SGD) algorithm for minimizing the loss in bounded-convex-Lipschitz learning problems.

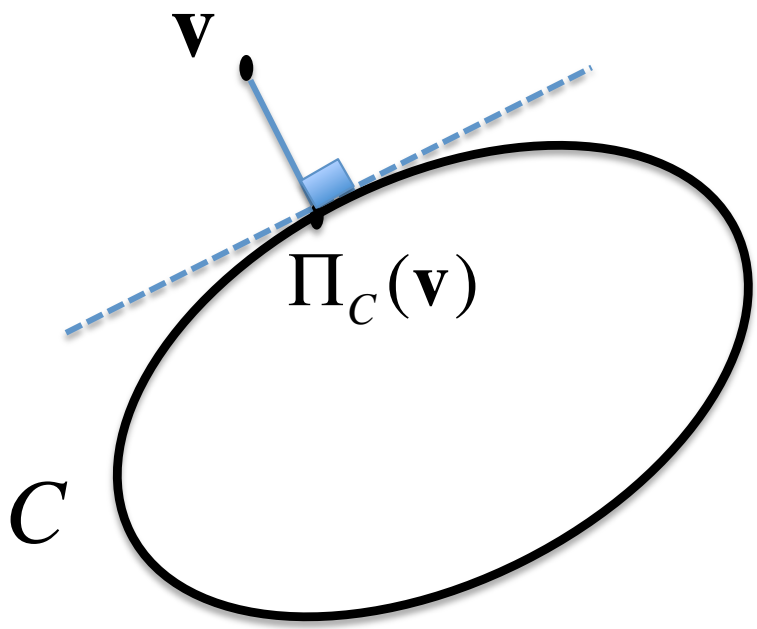
Euclidean projection

Definition: (Euclidean Projection)

Let $C \subseteq \mathbb{R}^p$ be a closed convex set. The Euclidean projection $\Pi_C : \mathbb{R}^p \to C$ is defined as:

$$\forall \mathbf{v} \in \mathbb{R}^p, \ \Pi_C(\mathbf{v}) = \arg\min_{\mathbf{w} \in C} \| \mathbf{v} - \mathbf{w} \|$$

That is, $\Pi_C(\mathbf{v})$ is the "closest" point in C (w.r.t. the Euclidean distance) to $\mathbf{v}$



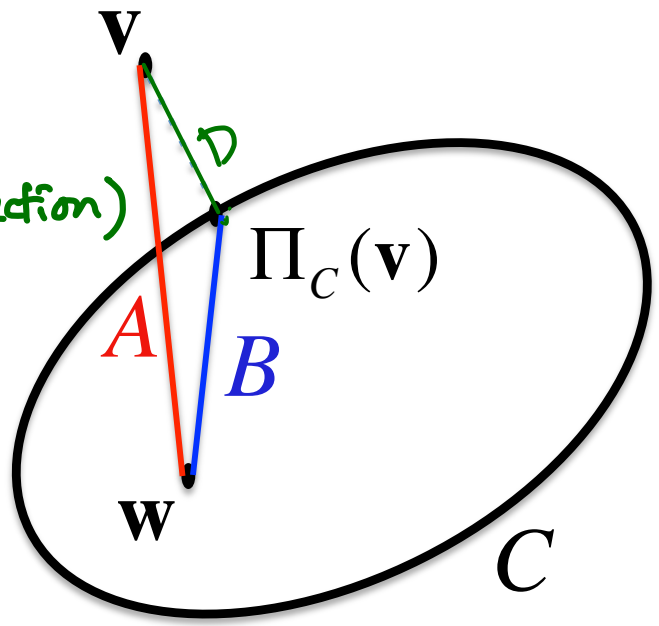
Euclidean projection

Fact:

$\forall \mathbf{v} \in \mathbb{R}^p$, $\forall \mathbf{w} \in C$, we have $\|\mathbf{v}-\mathbf{w}\| \geq \|\Pi_C(\mathbf{v})-\mathbf{w}\|$

Note that we are not aiming to prove that $A \geq D$ (which directly follows from the definition of Euclidean projection)

Before proving this fact, we will first see its useful application in the Gradient Descent algorithm.



$A \geq B$ for any $\mathbf{w} \in C$

GD for constrained optimization: Projected GD

- **Inputs:**
  - A convex ρ-**Lipschitz** function: $f : \mathbb{R}^p \rightarrow \mathbb{R}$
  - A convex constraint set: $C \subset \mathbb{R}^p$
  - Number of iterations: T
  - Set of scalars: $\{\alpha_t : t = 1, \ldots, T-1\}$ (***step size*** or ***learning rate***).

- **Output:**

An estimate $\hat{\mathbf{w}}$ of $\mathbf{w}^*$ where $\mathbf{w}^* \in \arg\min_{\mathbf{w} \in C} f(\mathbf{w})$

GD for constrained optimization: Projected GD

- **Inputs:** $f,\ C,\ T, \{\alpha_t : t = 1, \ldots, T-1\}$

1. Initialization: $\mathbf{w}_1 = \mathbf{0}$ (assume all-zero vector is in C)

Note that if C is M-bounded, then this directly implies that $\| \mathbf{w}^* - \mathbf{w}_1 \| \leq M$

2. **FOR** *t = 1, …, T – 1*

   Take a projected GD step: $\mathbf{w}_{t+1} = \Pi_C \left(\mathbf{w}_t - \alpha_t \nabla f(\mathbf{w}_t) \right)$

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

In general, $\| \mathbf{w}^* - \mathbf{w}_1 \| \leq M$

Theorem: (Convergence of Projected GD Algorithm)

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a convex, ρ-Lipschitz function. Let $C \subset \mathbb{R}^p$ be a closed convex set. Let $\mathbf{w}^* \in \arg\min_{\mathbf{w} \in C} f(\mathbf{w})$ where $\| \mathbf{w}^* \| \leq M$. If we run the projected GD algorithm above for ***T*** steps with $\alpha_t = \alpha = \frac{M}{\rho \sqrt{T}}, \forall t.$

Then, $f(\hat{\mathbf{w}}) - f(\mathbf{w}^*) \leq M \rho / \sqrt{T}$

GD for constrained optimization: Projected GD

Theorem: (Convergence of Projected GD Algorithm)

Let $f:\mathbb{R}^p \to \mathbb{R}$ be a convex, ρ-Lipschitz function. Let $C \subset \mathbb{R}^p$ be a closed convex set. Let $\mathbf{w}^* \in \arg\min_{\mathbf{w}\in C} f(\mathbf{w})$ where $\|\mathbf{w}^*\| \le M$. If we run the projected GD algorithm above for T steps with $\alpha_t = \alpha = \frac{M}{\rho\sqrt{T}}, \forall t.$

Then, $f(\hat{\mathbf{w}}) - f(\mathbf{w}^*) \le M\rho/\sqrt{T}$

As before, define $\psi_t = \left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2, \ \forall t$

Recall the two claims used in the proof of the basic GD:

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

remains valid (with an extra step in the proof)

Claim 2: $\forall t, \ \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \ge f(\mathbf{w}_t) - f(\mathbf{w}^*)$

remains valid (same proof: convexity of f)

Projected GD algorithm: small modification in the proof

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

Proof :

$$\psi_{t+1} = \left\|\mathbf{w}_{t+1} - \mathbf{w}^*\right\|^2 = \left\|\Pi_C\left(\mathbf{w}_t - \alpha\,\nabla f(\mathbf{w}_t)\right) - \mathbf{w}^*\right\|^2$$

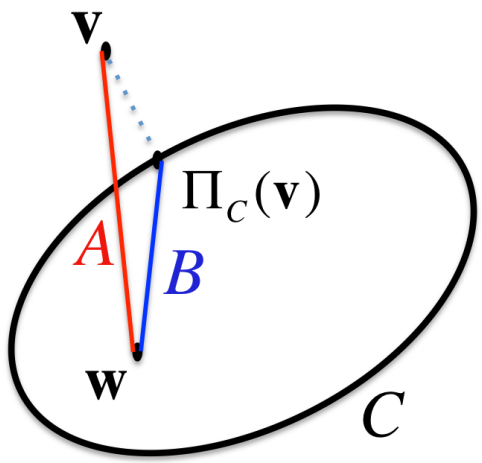
upper bound since projection on convex maps is a contractive mapping

$$\leq \left\|\mathbf{w}_t - \alpha\,\nabla f(\mathbf{w}_t) - \mathbf{w}^*\right\|^2$$

$$\vdots$$

Follows from the **Fact:** the property of Euclidean projections (since $\mathbf{w}^* \in C$)

Continue the proof exactly as before.



$A \geq B$ for any $\mathbf{w} \in C$

Euclidean projection

Fact:

$\forall \mathbf{v} \in \mathbb{R}^p$, $\forall \mathbf{w} \in C$, we have $\| \mathbf{v} - \mathbf{w} \| \geq \| \Pi_C(\mathbf{v}) - \mathbf{w} \|$

Proof:

$$\| \mathbf{v} - \mathbf{w} \|^2 = \| \mathbf{v} - \Pi_C(\mathbf{v}) + \Pi_C(\mathbf{v}) - \mathbf{w} \|^2$$

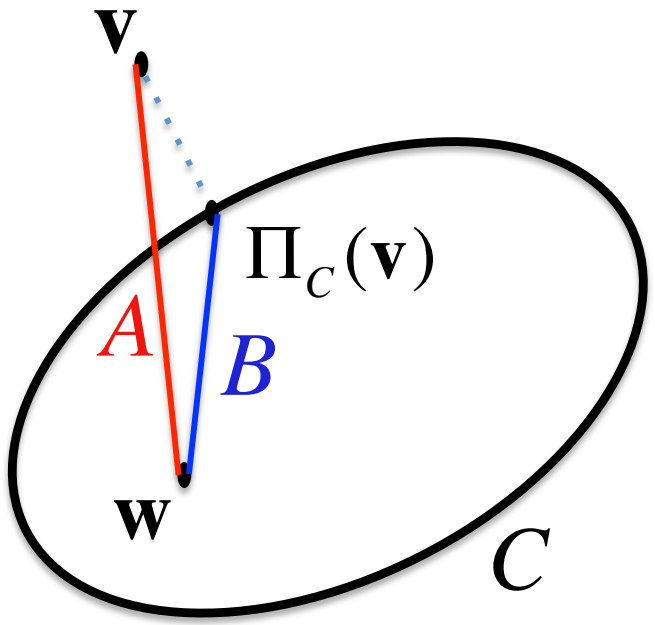
$$= \| \Pi_C(\mathbf{v}) - \mathbf{w} \|^2 + 2\langle \mathbf{v} - \Pi_C(\mathbf{v}),\ \Pi_C(\mathbf{v}) - \mathbf{w} \rangle$$

$$+ \| \mathbf{v} - \Pi_C(\mathbf{v}) \|^2 \quad \text{← Drop +ve term}$$

$$\geq \| \Pi_C(\mathbf{v}) - \mathbf{w} \|^2 + 2\langle \mathbf{v} - \Pi_C(\mathbf{v}),\ \Pi_C(\mathbf{v}) - \mathbf{w} \rangle$$

The following claim completes the proof

Claim: $\langle \mathbf{v} - \Pi_C(\mathbf{v}),\ \Pi_C(\mathbf{v}) - \mathbf{w} \rangle \geq 0$



$A \geq B$ for any $\mathbf{w} \in C$

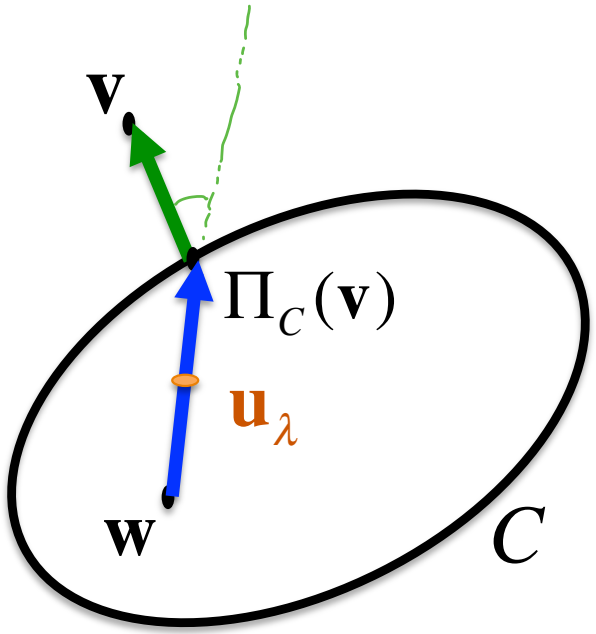
Euclidean projection

Proof of the claim: $\langle \mathbf{v} - \Pi_C(\mathbf{v}),\ \Pi_C(\mathbf{v}) - \mathbf{w} \rangle \geq 0$

∠ is always ≥ 90°

Since ***C*** is a convex set, then for any $\lambda \in (0,\ 1)$

$$\underbrace{(1-\lambda)\,\Pi_C(\mathbf{v}) + \lambda\,\mathbf{w}}_{\mathbf{u}_\lambda} \in C$$



By the definition of $\Pi_C(\mathbf{v})$

$$\|\Pi_C(\mathbf{v}) - \mathbf{v}\|^2 \leq \|\mathbf{v} - \mathbf{u}_\lambda\|^2$$

$$= \|\Pi_C(\mathbf{v}) - \mathbf{v}\|^2 + 2\lambda \langle \mathbf{v} - \Pi_C(\mathbf{v}),\ \Pi_C(\mathbf{v}) - \mathbf{w} \rangle + \lambda^2 \|\Pi_C(\mathbf{v}) - \mathbf{w}\|^2$$

Rearranging: $\langle \mathbf{v} - \Pi_C(\mathbf{v}),\ \Pi_C(\mathbf{v}) - \mathbf{w} \rangle \geq -\frac{\lambda}{2} \|\Pi_C(\mathbf{v}) - \mathbf{w}\|^2$

Taking the limit as $\lambda \to 0$, we reach the desired result.