Recall: 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

$$n_{\mathcal{H}}(\epsilon, \delta) = O\left(\frac{M^2 \rho^2}{\epsilon^2 \delta^2}\right)$$
**Stochastic Gradient Descent**

**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.
Euclidean projection

Definition: (Euclidean Projection)

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

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

That is, $\Pi_C(v)$ is the “closest” point in $C$ (w.r.t. the Euclidean distance) to $v$. 
Euclidean projection

**Fact:**
\[ \forall v \in \mathbb{R}^d, \forall w \in C, \text{ we have } \| v - w \| \geq \| \Pi_C(v) - w \| \]

**Proof:**
\[
\| v - w \|^2 = \| v - \Pi_C(v) + \Pi_C(v) - w \|^2 \\
= \| \Pi_C(v) - w \|^2 + 2 \langle v - \Pi_C(v), \Pi_C(v) - w \rangle \\
\quad + \| v - \Pi_C(v) \|^2 \\
\geq \| \Pi_C(v) - w \|^2 + 2 \langle v - \Pi_C(v), \Pi_C(v) - w \rangle \\
\]

The following claim completes the proof

**Claim:** \( \langle v - \Pi_C(v), \Pi_C(v) - w \rangle \geq 0 \)

\[ A \geq B \text{ for any } w \in C \]
Euclidean projection

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

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

\[
(1 - \lambda) \Pi_C(v) + \lambda w \in C
\]

By the definition of \( \Pi_C(v) \)

\[
\| \Pi_C(v) - v \|^2 \leq \| u_\lambda - v \|^2
\]

\[
= \| \Pi_C(v) - v \|^2 + 2\lambda \langle \Pi_C(v) - v, w - \Pi_C(v) \rangle + \lambda^2 \| \Pi_C(v) - w \|^2
\]

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

Taking the limit as \( \lambda \to 0 \), we reach the desired result.
GD for constrained optimization: Projection step

- **Inputs:**
  - A convex $\rho$-Lipschitz function: $f : \mathbb{R}^d \rightarrow \mathbb{R}$
  - A convex constraint set: $C \subset \mathbb{R}^d$
  - Number of iterations: $T$
  - Set of scalars: $\{\alpha_t : t = 1, \ldots, T - 1\}$ (the 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: Projection step**

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

1. Initialization: \( w_1 = 0 \) (assume all-zero vector is in \( C \))

2. **FOR** \( t = 1, \ldots, T - 1 \)

   Take a projected GD step: \( w_{t+1} = \Pi_C \left( w_t - \alpha_t \nabla f(w_t) \right) \)

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

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

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

Then, \( f(\hat{w}) - f(w^*) \leq M \rho / \sqrt{T} \)
**GD for constrained optimization: Projection step**

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

1. Initialization: (assume all-zero vector is in $C$)

2. For $t = 1, \ldots, T - 1$
   - Take a projected GD step:
     $$w_{t+1} = \Pi_C \left( w_t - \alpha_t \nabla f(w_t) \right)$$

3. Return $w_1 = 0$

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

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

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

Note that if $C$ is $M$-bounded, then this directly implies that

$$\|w^* - w_1\| \leq M$$

In general,

$$\|w^* - w_1\| \leq M$$
**GD for constrained optimization: Projection step**

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

Let \( f : \mathbb{R}^d \to \mathbb{R} \) be a convex, \( \rho \)-Lipschitz function. Let \( C \subset \mathbb{R}^d \) 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, \( f(\hat{\mathbf{w}}) - f(\mathbf{w}^*) \leq M \rho / \sqrt{T} \)

As before, define \( \psi_t = \| \mathbf{w}_t - \mathbf{w}^* \|^2, \forall t \)

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

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

**Claim 2:** \( \forall t, \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle \geq f(\mathbf{w}_t) - f(\mathbf{w}^*) \)
Projected GD algorithm: small modification in the proof

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:

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

Follows from the Fact: the property of Euclidean projections we just proved (since \( w^* \in C \))

Continue the proof exactly as before.
**Stochastic Gradient Descent**

- **Stochastic Gradients:**
  - Does not require that the direction taken in each iteration to be exactly equal to the (negative of the) gradient
  - Instead the direction can be a random vector whose expected is equal to the (negative of the) gradient.

In each iteration $t = 1, \ldots, T$, given that the current parameter value is $\mathbf{w}_t$, a step will be taken in a random direction $-\mathbf{G}_t$ where the random vector $\mathbf{G}_t$ satisfies: $\mathbf{E}\left[ \mathbf{G}_t | \mathbf{w}_t \right] = \nabla f(\mathbf{w}_t)$ where $\nabla f(\mathbf{w}_t)$ is the true gradient at $\mathbf{w}_t$.

- Instead of having $f$ given to the algorithm as an input, the algorithm will be given access to an “oracle” that given an input $\mathbf{w} \in \mathbb{R}^d$, it outputs a random vector $\mathbf{G}_t$ with the property: $\mathbf{E}\left[ \mathbf{G}_t | \mathbf{w}_t \right] = \nabla f(\mathbf{w}_t)$

- The oracle knows $f$, but the algorithm does not need to. The oracle is a “black-box” (the algorithm does not need to know how it runs).
**Stochastic Gradient Descent Algorithm**

- **Inputs:** A gradient oracle (denoted as **G-Oracle**),
  - Constraint set $C$,
  - Number of iterations $T$,
  - Learning rate \( \{\alpha_t : t = 1, \ldots, T - 1\} \)

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

2. **FOR** \( t = 1, \ldots, T - 1 \)
   
   a. Given \( \mathbf{w}_t \), **G-Oracle**($\mathbf{w}_t$) generates a random vector $G_t$ such that
      \[
      \mathbb{E}[G_t | \mathbf{w}_t] = \nabla f(\mathbf{w}_t)
      \]
   
   b. Take a projected step in direction of $G_t$: \( \mathbf{w}_{t+1} = \Pi_C (\mathbf{w}_t - \alpha_t \ G_t) \)

3. Return \( \hat{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t \)
**Theorem: (Convergence of SGD Algorithm)**

Let \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) be a convex function. Let \( C \subset \mathbb{R}^d \) be a closed convex set. Let \( w^* \in \arg\min_{w \in C} f(w) \) where \( \|w^*\| \leq M \). Suppose we run SGD algorithm for \( T \) steps with \( \alpha_t = \alpha = \frac{M}{\rho \sqrt{T}} \), \( \forall t \). Suppose that \( \forall t, \|G_t\| \leq \rho \) with probability 1. Then, \( \mathbb{E}[ f(\hat{w}) ] - f(w^*) \leq \frac{M \rho}{\sqrt{T}} \)

Before we prove this theorem, we will first show an important implication. If \( C \) is \( M \)-bounded, then this will immediately follow. This implies that \( f \) is \( \rho \)-Lipschitz.
Learning with SGD

Given a parameter set $C$, and a loss function $\ell : C \times Z \rightarrow \mathbb{R}_+$, define

$$L(w ; D) \triangleq \mathbb{E}_{z \sim D}[\ell(w, z)]$$

The true risk. Analogous to $err(h_w ; D)$ in the standard PAC model (with 0-1 loss)

The excess risk incurred by an algorithm that uses a training set $S = \{z_1, \ldots, z_n\} \sim D^n$ to output a hypothesis $\hat{w}_S \in C$ is given by

$$L(\hat{w}_S ; D) - \min_{w \in C} L(w ; D)$$

Recall that learnability is tied to this quantity

Will show that SGD aims at directly minimizing the expected excess risk

$$\mathbb{E}_{S \sim D^n}[L(\hat{w}_S ; D)] - \min_{w \in C} L(w ; D)$$
**SGD Learner**

- **Inputs:** A loss function \( \ell : C \times Z \rightarrow \mathbb{R}_+ \),
  A sample oracle: returns a fresh sample \( z \sim D \)
  Parameter set \( C \),
  Number of iterations \( T \),
  Learning rate \( \{\alpha_t : t = 1, \ldots, T - 1\} \)

1. Initialization: \( w_1 = 0 \) (assume all-zero vector is in \( C \))

2. **FOR** \( t = 1, \ldots, T - 1 \)
   a. Draw a fresh example \( z_t \sim D \)
   b. Compute \( G_t = \nabla \ell(w_t, z_t) \)
   c. Update: \( w_{t+1} = \Pi_C \left( w_t - \alpha_t G_t \right) \)

3. Return \( \hat{w}_S = \frac{1}{T} \sum_{t=1}^{T} w_t \) (where \( S = \{z_1, \ldots, z_{T-1}\} \). The notation \( \hat{w}_S \) is to remind us that the output depends on \( S \).)
**SGD Learner**

- **Inputs:** A loss function \( \ell : C \times Z \rightarrow \mathbb{R}_+ \),
  A sample oracle: returns a fresh sample \( z \sim D \)
  Parameter set \( C \),
  Number of iterations \( T \),
  Learning rate \( \{ \alpha_t : t = 1, \ldots, T - 1 \} \)

1. **Initialization:** \( w_1 = 0 \)

2. **FOR** \( t = 1, \ldots, T - 1 \)
   
   a. Draw a fresh example \( z_t \sim D \)
   
   b. Compute \( G_t = \nabla \ell(w_t, z_t) \)
   
   c. Update: \( w_{t+1} = \Pi_C \left( w_t - \alpha_t G_t \right) \)

3. Return \( \hat{W}_S = \frac{1}{T} \sum_{t=1}^{T} w_t \)

**This plays the role of the G-Oracle:** Randomness here is due to \( z_t \sim D \)

**Note that** at step \( t \), given that the current point is \( w_t \), the random vector \( G_t \) is given by \( \nabla \ell(w_t, z_t) \).
**SGD Learner**

- **Inputs:** A loss function \( \ell : C \times \mathbb{Z} \to \mathbb{R}^+ \), a sample oracle that returns a fresh sample \( z_t \), a parameter set \( \mathbf{w} \), number of iterations \( T \), learning rate \( \alpha_t \). \( \mathbf{w}_t = 1 \) to \( T - 1 \setminus \{ \ell \} \)

1. **Initialization:**
   - (assume all-zero vector is in \( D \))

2. **FOR** \( t = 1, \ldots, T \)
   a. Draw a fresh example \( z_t \sim D \)
   b. **Compute** \( G_t = \nabla \ell (\mathbf{w}_t, z_t) \)
   c. **Update:** \( \mathbf{w}_{t+1} = \Pi_C \left( \mathbf{w}_t - \alpha_t G_t \right) \)

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

\[ \mathbb{E}\left[ G_t \mid \mathbf{w}_t \right] = \mathbb{E}_{z_t \sim D} \left[ \nabla \ell (\mathbf{w}_t, z_t) \right] \]

Conditioned on \( \mathbf{w}_t \), the remaining randomness comes from \( z_t \). Note that \( \mathbf{w}_t \) is independent of \( z_t \).
**SGD Learner**

- **Inputs:**
  - A loss function \( \ell \)
  - A sample oracle: returns a fresh sample \( z_t \)
  - Parameter set \( \theta \)
  - Number of iterations \( T \)
  - Learning rate \( \alpha_t \);

\[
\theta_t = 1, \ldots, T - 1
\]

\[
\ell : \mathbb{C} \times \mathbb{Z} \rightarrow \mathbb{R}^+ 
\]

1. **Initialization:** (assume all-zero vector is in \( \mathbb{D} \))

2. **FOR** \( t = 1, \ldots, T - 1 \)
   a. Draw a fresh example \( z_t \sim \mathbb{D} \)
   b. **Compute** \( G_t = \nabla \ell(\theta_t, z_t) \)
   c. Update: \( \theta_{t+1} = \pi_C(\theta_t - \alpha_t G_t) \)

3. **Return** \( \hat{\theta}_S = \frac{1}{T} \sum_{t=1}^{T} \theta_t \)

\[
\mathbb{E}[G_t | \theta_t] = \mathbb{E}_{z_t \sim \mathbb{D}}[\nabla \ell(\theta_t, z_t)] \\
= \nabla \mathbb{E}_{z_t \sim \mathbb{D}}[\ell(\theta_t, z_t)] \\
\]

Swapping order of \( \mathbb{E} \) (integration) w.r.t. \( z_t \) and \( \nabla \) (differentiation) w.r.t. \( \theta \) at \( \theta_t \) (Note again that \( \theta_t \) does not depend on \( z_t \))
**SGD Learner**

- **Inputs:** A loss function, a sample oracle: returns a fresh sample, parameter set, number of iterations, learning rate $T\alpha_t$: $t = 1, \ldots, T - 1$

1. **Initialization:** (assume all-zero vector is in $C$)

2. **FOR** $t = 1, \ldots, T - 1$
   a. Draw a fresh example $z_t \sim D$
   b. **Compute** $G_t = \nabla \ell(\mathbf{w}_t, z_t)$
   c. **Update:** $\mathbf{w}_{t+1} = \Pi_C(\mathbf{w}_t - \alpha_t G_t)$

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

\[\mathbb{E}[G_t | \mathbf{w}_t] = \mathbb{E}_{z_t \sim D}[\nabla \ell(\mathbf{w}_t, z_t)] \]
\[= \nabla \mathbb{E}_{z_t \sim D}[\ell(\mathbf{w}_t, z_t)] \]
\[= \nabla \mathbb{E}_{z \sim D}[\ell(\mathbf{w}_t, z)] \]

Since $z_t$ is a fresh sample that is independent of $\mathbf{w}_t$, the expectation will remain the same if we replace $z_t$ by a fresh sample $z \sim D$. 
SGD Learner

- **Inputs:**
  - A loss function
  - A sample oracle: returns a fresh sample
  - Parameter set
  - Number of iterations
  - Learning rate $T_\alpha = 1, \ldots, T-1$

1. **Initialization:**
   Assume all-zero vector is in

2. **FOR** $t = 1, \ldots, T - 1$
   a. Draw a fresh example $z_t \sim D$
   b. **Compute** $G_t = \nabla \ell(w_t, z_t)$
   c. **Update**: $w_{t+1} = \Pi_C \left( w_t - \alpha_t G_t \right)$

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

\[
\mathbb{E}\left[ G_t \mid w_t \right] = \mathbb{E}_{z_t \sim D} \left[ \nabla \ell(w_t, z_t) \right]
\]
\[
= \nabla \mathbb{E}_{z_t \sim D} \left[ \ell(w_t, z_t) \right]
\]
\[
= \nabla \mathbb{E}_{z \sim D} \left[ \ell(w_t, z) \right]
\]
\[
= \nabla L(w_t \mid D)
\]

By definition of the true risk $L(w_t \mid D)$
**SGD Learner**

- **Inputs:**
  - A loss function \( \ell \)
  - A sample oracle: returns a fresh sample \( z_t \)
  - Parameter set \( \theta \)
  - Number of iterations \( T \)
  - Learning rate \( \alpha_t \):
    \[ t = 1, \ldots, T - 1 \]

1. **Initialization:**
   Assume all-zero vector is in \( C \).

2. **FOR** \( t = 1, \ldots, T - 1 \)
   a. Draw a fresh example \( z_t \sim D \).
   b. Compute \( G_t = \nabla \ell(w_t, z_t) \)
   c. Update: \( w_{t+1} = \Pi_C(w_t - \alpha_t G_t) \)

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

Hence, we have

\[
\mathbb{E}[G_t | w_t] = \nabla L(w_t ; D)
\]

That is, the objective function here is actually \( L(\cdot ; D) \)

i.e., here \( L(\cdot ; D) \) is an instantiation of \( f \) in the

generic SGD algorithm.
Learning with SGD

**Theorem: (Convergence of the Generic SGD Algorithm)**

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex function. Let $C \subset \mathbb{R}^d$ be a closed convex set.

Let $w^* \in \arg\min_{w \in C} f(w)$ where $\|w^*\| \leq M$. Suppose we run SGD algorithm for $T$ steps with $\alpha_t = M / (\rho \sqrt{T})$, $\forall t$. Suppose that $\forall t$, $\|G_t\| \leq \rho$ with probability 1.

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

**Corollary: (Excess Risk of the SGD Learner)**

Consider a convex-Lipschitz-bounded model where the parameter set $C \subset \mathbb{R}^d$ is $M$-bounded and there is $\rho > 0$ such that for all $z \in \mathcal{Z}$, the loss function $\ell(\cdot; z)$ is convex and $\rho$-Lipschitz. Suppose we run the SGD learner with number of iterations (i.e., number of examples) $T \geq \frac{M^2 \rho^2}{\epsilon^2}$, and with $\alpha_t = M / (\rho \sqrt{T})$, $\forall t$.

Then, $\mathbb{E}_{S \sim \mathcal{D}^T} \left[ L(\hat{w}_S; D) \right] - \min_{w \in C} L(w; D) \leq \epsilon$
Learning with SGD: Remarks

• SGD takes a rather direct approach towards learning.
• Instead of minimizing the empirical error \( \hat{L}(w ; S) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(w , z_i) \) over \( w \in C \) w.r.t. a training set \( S = \{z_1, \ldots, z_n\} \), it seeks to minimize the true risk directly.
• This is done by taking a random update step at each iteration. In each iteration \( t \), given a current parameter estimate \( w_t \), the update step is taken along \( -\nabla \ell(w_t , z_t) \) where \( z_t \) is a fresh sample from the unknown distribution.
• So, this SGD learner is not an ERM algorithm.
• Note that the SGD learner is an online algorithm where training samples can arrive one by one (or in small batches) rather than requiring that the entire training set to be available in advance.
**Theorem: (Convergence of SGD Algorithm)**

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Let $C \subset \mathbb{R}^d$ be a closed convex set.

Let $w^* \in \arg \min_{w \in C} f(w)$ where $\|w^*\| \leq M$. Suppose we run SGD algorithm for $T$ steps with $\alpha_t = \alpha = \frac{M}{\rho \sqrt{T}}$, $\forall t$. Suppose that $\forall t$, $\|G_t\| \leq \rho$ with probability 1. Then, $E[f(\hat{w})] - f(w^*) \leq \frac{M \rho}{\sqrt{T}}$

Define $\tilde{\psi}_t = E[\|w_t - w^*\|^2], \forall t$

**Claim 1**: $\forall t$, $E[\langle G_t, w_t - w^* \rangle] \leq \frac{\tilde{\psi}_t - \tilde{\psi}_{t+1}}{2\alpha} + \frac{\alpha \rho^2}{2}$

**Claim 2**: $\forall t$, $E[\langle G_t, w_t - w^* \rangle] \geq E[f(w_t)] - f(w^*)$

Given these claims, the rest of the proof follows similar steps as in the proof of the analogous theorem for the basic GD algorithm.
**SGD algorithm: Proof of convergence guarantees**

**Claim 1**: \( \forall t, \mathbb{E}[\langle G_t, w_t - w^* \rangle] \leq \frac{\tilde{\psi}_t - \tilde{\psi}_{t+1}}{2\alpha} + \frac{\alpha \rho^2}{2} \)

**Claim 2**: \( \forall t, \mathbb{E}[\langle G_t, w_t - w^* \rangle] \geq \mathbb{E}[f(w_t)] - f(w^*) \)

The key to prove these claims is the following fact:

\[
\mathbb{E}[\langle G_t, w_t - w^* \rangle] = \mathbb{E}[\langle \nabla f(w_t), w_t - w^* \rangle]
\]
SGD algorithm: Proof of convergence guarantees

Let’s first prove this fact. Let $G^{t-1}_1$ denote the sequence of random vectors $(G_1, \ldots, G_{t-1})$. Observe that by applying the rule of iterated expectation, we have

$$
\mathbb{E}\left[\langle G_t, w_t - w^* \rangle \right] = \mathbb{E}_{G^{t-1}_1} \left[ \mathbb{E}_{G_t|G^{t-1}_1} \left[ \langle G_t, w_t - w^* \rangle \big| G^{t-1}_1 \right] \right]
$$

Now, consider the inner conditional expectation. Since we condition on $G^{t-1}_1$ (i.e., we fix $G^{t-1}_1$) in the inner expectation, $w_t$ is also fixed.

Hence, using linearity of expectation, the inner expectation becomes:

$$
\mathbb{E}_{G_t|G^{t-1}_1} \left[ \langle G_t, w_t - w^* \rangle \big| G^{t-1}_1 \right] = \langle \mathbb{E}_{G_t|G^{t-1}_1} \left[ G_t \big| G^{t-1}_1 \right], w_t - w^* \rangle
$$

Since $w_t$ only depends on $G^{t-1}_1$, and $G_t$ only depends on $w_t$, we have

$$
\mathbb{E}_{G_t|G^{t-1}_1} \left[ G_t \big| G^{t-1}_1 \right] = \mathbb{E}_{G_t|G^{t-1}_1, w_t} \left[ G_t \big| G^{t-1}_1, w_t \right] = \mathbb{E}_{G_t|w_t} \left[ G_t \big| w_t \right] = \nabla f(w_t)
$$

We arrive at the desired fact by applying the outer expectation.
Namely, we get
\[
\mathbb{E}\left[ \langle G_t, w_t - w^* \rangle \right] = \mathbb{E}_{G_{t-1}} \left[ \langle \nabla f(w_t), w_t - w^* \rangle \right] \\
= \mathbb{E} \left[ \langle \nabla f(w_t), w_t - w^* \rangle \right]
\]

Given this fact, the proofs of the Claims 1* and 2* should be straightforward as they follow similar lines to their counterparts for the basic GD. However, for completeness, let’s briefly go over their proofs.
SGD algorithm: Proof of convergence guarantees

Claim 1*: \( \forall t, \mathbb{E}\left[ \langle G_t, w_t - w^* \rangle \right] \leq \frac{\tilde{\psi}_t - \tilde{\psi}_{t+1}}{2\alpha} + \frac{\alpha \rho^2}{2} \)

Proof:

\[ \tilde{\psi}_{t+1} = \mathbb{E}\left[ \|w_{t+1} - w^*\|^2 \right] = \mathbb{E}\left[ \|\Pi_C (w_t - \alpha G_t) - w^*\|^2 \right] \]

\[ \leq \mathbb{E}\left[ \|w_t - \alpha G_t - w^*\|^2 \right] \]

\[ = \mathbb{E}\left[ \langle w_t - w^* - \alpha G_t, w_t - w^* - \alpha G_t \rangle \right] \]

\[ = \mathbb{E}\left[ \|w_t - w^*\|^2 \right] - 2\alpha \mathbb{E}\left[ \langle G_t, w_t - w^* \rangle \right] + \alpha^2 \mathbb{E}\left[ \|G_t\|^2 \right] \]

\[ \leq \tilde{\psi}_t - 2\alpha \mathbb{E}\left[ \langle G_t, w_t - w^* \rangle \right] + \alpha^2 \rho^2 \]

By rearranging the terms, the proof is complete.

By the property of Euclidean projections that we proved

since \( \|G_t\| \leq \rho \) with prob. 1
SGD algorithm: Proof of convergence guarantees

Claim 2*:
\[ \forall t, \mathbb{E}\left[ \langle G_t, w_t - w^* \rangle \right] \geq \mathbb{E}\left[ f(w_t) - f(w^*) \right] \]

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, we have
\[
\langle \nabla f(w_t), w_t - w^* \rangle \geq f(w_t) - f(w^*)
\]
Taking expectation of both sides:
\[
\mathbb{E}\left[ \langle \nabla f(w_t), w_t - w^* \rangle \right] \geq \mathbb{E}\left[ f(w_t) - f(w^*) \right]
\]
Using the key fact we just proved the LHS is equal to \( \mathbb{E}\left[ \langle G_t, w_t - w^* \rangle \right] \) which completes the proof.
**SGD algorithm: Proof of convergence guarantees**

Combining those claims, we have

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

\[
= \frac{1}{2\alpha T} \sum_{t=1}^{T} (\tilde{\psi}_t - \tilde{\psi}_{t+1}) + \frac{\alpha \rho^2}{2}
\]

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

Continue exactly as the proof of the basic GD theorem.