**Advanced Topics in Machine Learning and Algorithmic Game Theory**

**Fall semester, 2011**

# Lecture 2: November 7, 2011

*Lecturer: Yishay Mansour* *Scribe: Elad Liebman*[1]

## 2.1 Convex Learning Problems

In this lecture we continue to discuss convex learning problems. We should first begin by formally defining what "convexity" means. There are several essentially equivalent ways to define this term:

- A function $f$ is convex if and only if the following holds:

  $$f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha) f(v)$$

  for every $\alpha \in [0,1]$ and $u, v \in X$.

- A function $f$ is convex if and only if its *epigraph* is a convex set. An *epigraph* of a function $f : R^n \rightarrow R$ is the set of points lying on or above its graph:

  $$epigraph(f) = \{(x, \alpha) : x \in R^n, \alpha \in R, f(x) \leq \alpha\} \in R^{n+1}$$

  (*see Figure 2.1*).

  Informally, a set of points is convex if one could draw a line between every two points in the set, it would also be fully contained within it (*see Figure 2.2*).

- a function $f$ is convex if and only if its Hessian (matrix of second derivatives) is positive semidefinite, i.e., if the following holds:

  $$\bigtriangledown^2 f = H, \forall x . x^T H x \geq 0$$

  under the condition that $f$ is twice-differentiable.

The *gradient* of a function $f$ at point $w$ is the vector of partial derivatives of $f$ w.r.t. its elements, namely $\bigtriangledown f = (\frac{\partial f(w)}{\partial w_1}, \ldots, \frac{\partial f(w)}{\partial w_n})$. Let us observe the Taylor approximation of $f$, defined as $f(u) = f(w) + (u-w) \bigtriangledown f(w) + (u-w)^T \bigtriangledown^2 f(z)(u-v) = f(w) + (u-w) \bigtriangledown f(w) + (u-w)^T H_f(z)(u-v)$, where $H_f(z) = \bigtriangledown^2 f(z)$, for some $z$ between $u$ and $v$. If $f$ is convex it follows that $H_f$ is P.S.D., and therefore, $f(u) \geq f(w) + (u-w) \bigtriangledown f(w)$.

What do we do if $f$ is not differentiable? Let us a define a *sub-gradient*.

[1] Based on lecture notes by Shai Shalev-Shwartz (November 15, 2010), and by Stephen Boyd, Lin Xiao and Almir Mutapcic (October 1, 2003)

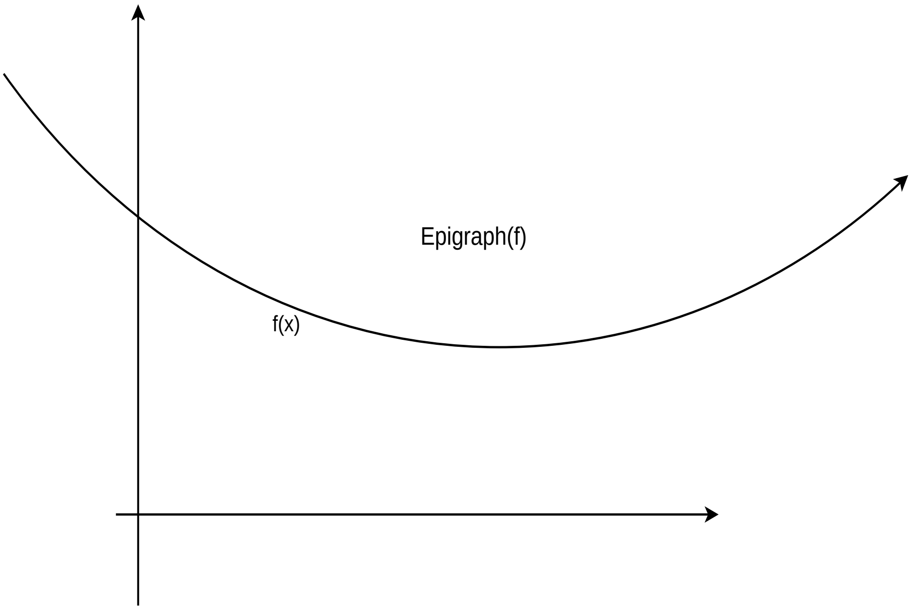

Figure 2.1: Epigraph Example

**Definition 1 Sub-gradient:** *a vector $\lambda$ is a* sub-gradient *of a function $f$ at point $w$ if for every $u$ the following holds:*

$$f(u) - f(w) \geq (u - w) \cdot \lambda$$

.

We shall denote the group of sub-gradients for a function $f$ in $u$ as $\partial f(u)$.

**Observation 1** *If $f$ is differentiable at point $u$, the only possible sub-gradient for $f(u)$ is the gradient of f itself. Otherwise, many sub-gradients are possible (*see Figure 2.3*).*

*an example:* Consider $f(w) = max_i(g_i(w)), i \in [0 \dots n]$. We are assured that all $g_i$ are both differetiable and convex.

- First, we show that $f(w)$ is convex.

  **Proof:** $f(\alpha w + (1-\alpha)v) = max_i(g_i(\alpha w + (1-\alpha)v))$, and since all $g_i$ are convex, it follows that $f(\alpha w + (1-\alpha)v) \leq max_i(\alpha g_i(w) + (1-\alpha)g_i(u) \leq \alpha max_i(g_i(w)) + (1-\alpha)max_i(g_i(u)) \leq \alpha f(w) + (1-\alpha)f(u)$ ■

- Second, we show that if $argmax_i(g_i(w)) = j$, then the gradient of $g_j$ at $\mathbf{u}$ is a subgradient of $f$ at $\mathbf{u}$.

  **Proof:** $f(u) \geq g_j(u)$, and because $g_j$ is convex, it follows that, $f(u) \geq g_j(w) + (u-w) \bigtriangledown g_j(w) = f(w) + (u-w) \bigtriangledown g_j(w)$. Meaning that $\bigtriangledown g_j$ is a sub-gradient of $f$, as required. ■

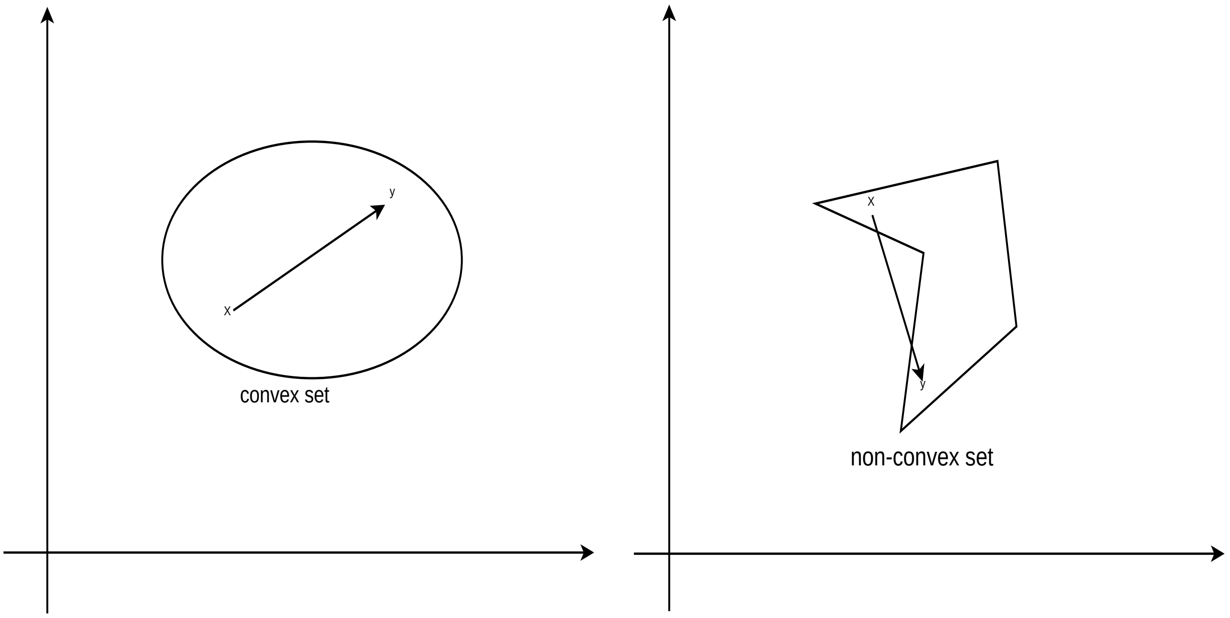

Figure 2.2: convex vs. non-convex sets

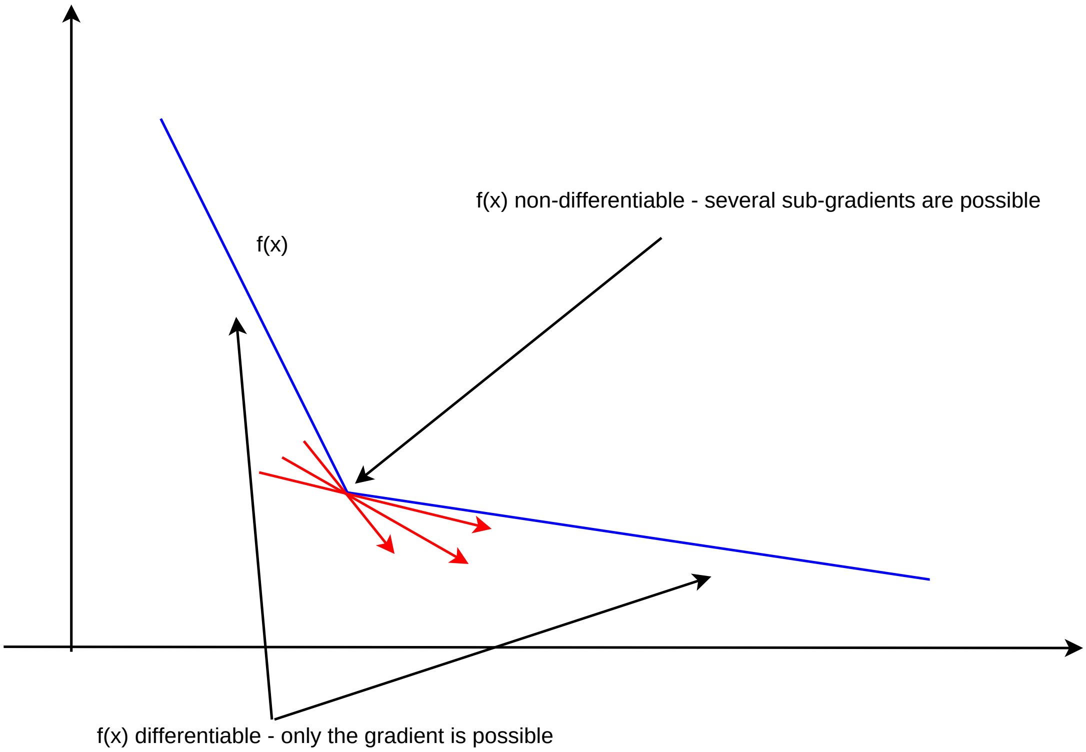
f(x) non-differentiable - several sub-gradients are possible
f(x)
f(x) differentiable - only the gradient is possible

Figure 2.3: Sub-Gradients - degrees of freedom

### 2.1.1 Hinge Loss

Quite often, we face a non-convex loss function. For the class of halfspaces, for instance, we deal with the non-convex $0-1$ loss, which is noncontinous and cannot be efficiently used in an optimization program. To circumvent the hardness implied by this, we would like to upper-bound the $0-1$ loss function using a *hinge loss* function. *Hinge loss* is defined as follows: $l_H(w, x, y) = max\{0, 1 - y\langle w, x\rangle\}$ (*see Figure 2.4*), where $w$ represents a separating hyperplane (essentially, an hypothesis), $x \in X$ and $y \in [-1, 1]$. The hinge loss is a maximum of linear functions and therefore convex. However, when $1 - y\langle w, x\rangle = 0$, it is not differentiable. It is easy to verify that

$$\partial l_H(w,x,y) = \begin{cases} -yx & \text{if } 1 - y\langle w,x\rangle & > 0 \\ 0 & \text{if } 1 - y\langle w,x\rangle & < 0 \\ -\alpha yx & \text{if } 1 - y\langle w,x\rangle & = 0; \quad \alpha \in [0,1] \end{cases}$$

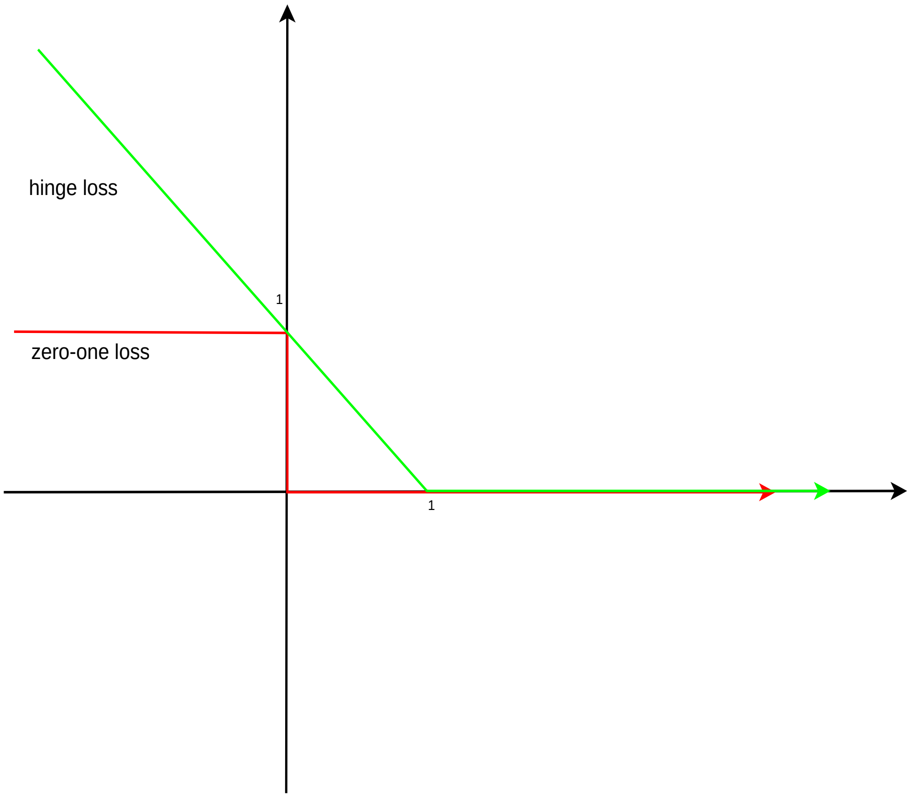

Figure 2.4: Hinge Loss

**Proof:** The first two cases are trivial. We are interested in the third case. $l_H(w, x, y) = 0$, which means that $l_H(u, x, y) - l_H(w, x, y) = max\{0, 1 - y\langle u, x\rangle\}$. We would like to bound this result using a sub-gradient: $\langle u - w, -\alpha yx\rangle = -\alpha y\langle u, x\rangle + \alpha y\langle w, x\rangle$. We know that $y\langle x, w\rangle = 1$, therefore: $\langle u - w, -\alpha yx\rangle = \alpha(1 - y\langle u, w\rangle) \leq max\{0, 1 - y\langle u, x\rangle\} = l_H(u, x, y) - l_H(w, x, y)$, as required. ■

## 2.2 Optimization With Sub-Gradients

***A reminder:*** let $f(w)$ be a convex, differentiable function which we would like to minimzie. *Gradient descent* (also occasionally known as "steepest descent") is one of the simplest methods to achieve this - we begin with an initial vector $w_1$ and on each iteration we follow the update rule $w_{t+1} = w_t - \alpha_t \bigtriangledown f(w_t)$. That is, we move from $w_t$ in the direction *opposite* to the gradient, whereas $\alpha_t$ is the *step size.* Because we're following the gradient, we are guaranteed to improve on every step (hence the "descent"), and since there's only a single minimum to $f(w)$, we will eventually converge to it.

Let us define a *sub-gradient method* for the same optimization problem, using the following iteration rule:

$$w_{t+1} = w_t - \alpha_t g_t$$

Where $g_t$ is *any* sub-gradient of $f$ in $w_t$.

When working with *sub-gradients*, we are not assured to consistently improve on each step, so proving convergence is more difficult. We must keep track of the best result observed thus far. Let us define a new term, $f_t^{best}$, which "remembers" the best result at step $t$:

**Definition 2** $f_t^{best} = min\{f_{t-1}^{best}, f(w_t)\}$

Let $f^*$ be the minimum value of $f$, and let $w^*$ be the $w$ value which obtains it: $f^* = f(w^*)$. Let us assume also that the value of the sub-gradient is bound by some constant $G$: $g_t \leq G$ (this is true for Lipschitz functions).

We shall prove convergence relying on the distance between $w_t$ and $w^*$, rather than relying directly on the distance between $f_t$ and $f^*$.

**Proof:** Because $w^*$ yields the minimum for $f$, it follows that $||w_{t+1} - w^*||_2^2 = ||w_t - \alpha_t g_t - w^*||_2^2$. This leads to $||w_{t+1} - w^*||_2^2 = ||w_t - w^*||_2^2 - 2\alpha_t \langle g_t, w_t - w^* \rangle + \alpha_t^2 ||g_t||_2^2$. $\langle g_t, w_t - w^* \rangle$ is the sub-gradient, so $\langle g_t, w_t - w^* \rangle \leq f(w_t) - f(w^*)$ holds, and $||g_t||_2^2$ is bound by $G^2$, meaning that $||w_{t+1} - w^*||_2^2 \leq ||w_t - w^*||_2^2 - 2\alpha_t(f(w_t) - f^*) + \alpha_t^2 G^2$. We can repeat this step recursively to obtain:

$$||w_{t+1} - w^*||_2^2 \leq ||w_1 - w^*||_2^2 - 2\sum_{t=1}^{t} \alpha_t (f(w_t) - f^*) + G^2 \sum_{t=1}^{t} \alpha_t^2$$

We note that $2\sum_{t=1}^{t} \alpha_t (f(w_t) - f^*) \geq 0$ because $f^*$ is the minimum, and $||w_{t+1} - w^*||_2^2 \geq 0$ necessarily. Reorganizing the terms of the inequality yields:

$$2\sum_{t=1}^{t} \alpha_t (f(w_t) - f^*) \leq ||w_1 - w^*||_2^2 + G^2 \sum_{t=1}^{t} \alpha_t^2$$

Which $f(w_t)$ value brings $f(w_t) - f^*$ to minimum? Clearly, $f_t^{best}$, therefore:

$$2\sum_{t=1}^{t} \alpha_t (f_t^{best} - f^*) \leq ||w_1 - w^*||_2^2 + G^2 \sum_{t=1}^{t} \alpha_t^2$$

Which implies that:

$$f_t^{best} - f^* \leq \frac{||w_1 - w^*||_2^2 + G^2 \sum_{t=1}^{t} \alpha_t^2}{2 \sum_{t=1}^{t} \alpha_t}$$

Given that $||w_1 - w^*||_2^2$ is bound by some constant $D$, we now have all we need in order to determine sufficient conditions for convergence:

- For instance, if $\sum \alpha_t \to \infty$ and $\sum \alpha_t^2$ is bound, it's clear to see that $f_t^{best}$ converges to $f^*$. The most intuitive example for this is selecting $\alpha_t = \frac{1}{t}$.
- Another example: let $\alpha_t = h$, then

  $$f_t^{best} - f^* \leq \frac{D + G^2 h^2 t}{2th} = \frac{D}{2th} + G^2 h$$

  meaning we converge in a rate proportionate to the number of steps taken, $t$. Notice that when $f_t^{best}$ converges, $\frac{D}{2Th} = G^2 h$, and therefore $h \sim \frac{1}{\sqrt{T}}$.

■

### 2.2.1 The Restricted Case

Let us assume we're facing the following optimization problem: we need to find the minimum of $f(w)$ s.t. $w \in K$ for some domain $K$. The naive sub-gradient algorithm will not do, because we'll get constantly thrown out of the domain K. We must define a *projection function* $\Pi$, and use it in the update rule:

$$w_{t+1} = \Pi(w_t - \alpha_t g_t)$$

Let $y_t = w_t - \alpha_t g_t$, then

$$\Pi(w_t - \alpha_t g_t) = argmin_{x \in K} ||y_t - x||_2^2$$

Now, $||y_t - w^*||_2^2 = ||w_t - \alpha_t g_t - w^*||_2^2 \leq ||w_t - w^*||_2^2 - 2\alpha_t(f(w_t) - f^*) + \alpha_t^2 G^2$. For our algorithm to work in the current case, we must show that $||y_t - w^*||_2^2 \geq ||w_{t+1} - w^*||_2^2$, meaning to say that projecting $y_t$ can only reduce our distance from $w^*$.

**Proof:** In fact, the inequality $\forall x \in K. ||y_t - x||_2^2 \geq ||w_{t+1} - x||_2^2$ is true, for geometric reasons - when we project a point onto $K$, we move closer to every point in $K$ (*See Figure 2.5*). $||y_t - x||_2^2 = ||(y_t - w_{t+1}) + (w_{t+1} - x)||_2^2 = ||y_t - w_{t+1}||_2^2 + ||w_{t+1} - x||_2^2 - 2\langle y - w_{t+1}, w_{t+1} - x \rangle$, and $2\langle y - w_{t+1}, w_{t+1} - x \rangle < 0$, which proves the required. ■

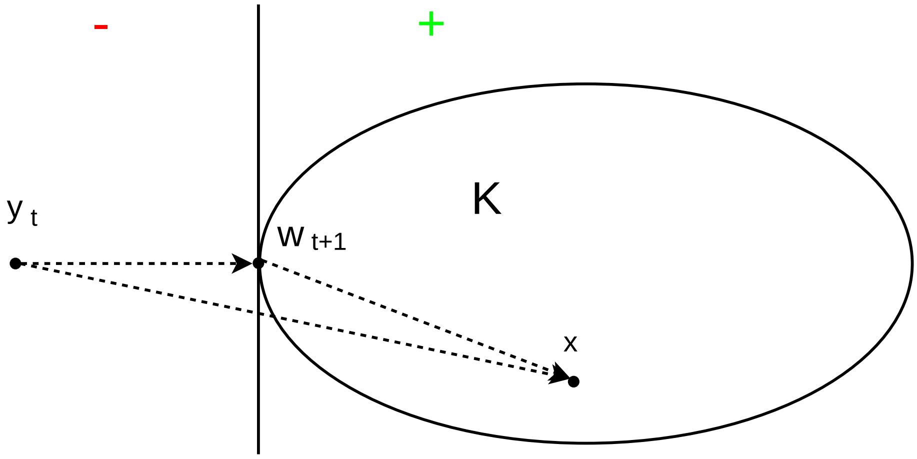

Figure 2.5: geometric explanation why the inequality $\forall x \in K.||y_t - x||_2^2 \geq ||x_{t+1} - x||_2^2$ holds

## 2.3 Stochastic Sub-Gradient Descent

The main difference between optimization problems and online learning is that the function $f$ is unknown - only fragments of it are revealed in each step. For this reason we cannot base our algorithm on the sub-gradient (or the gradient, for that matter) of the function $f$ we would like to minimize, but rather use the update rule $w_{t+1} = w_t - \alpha_t r_t$, where $r_t$ is a random vector whose expected value is a sub-gradient of $f$ at $w_t$ (so we hope).

Let us consider a loss function $l(w, z)$ denoting our loss when using $w$ at point $z$. What really interests us is the average over a given distribution: $L_D(w) = E_{z\sim D}[l(w, z)]$. Our approach would be to sample $z$ (meaning to say, select examples randomly) and take the observed sub-gradient of our incurred losses. The following lemma formalizes this approach.

**Lemma 2.1** *For all $w$ let $\lambda = \partial l(w, z)$ where $z$ is sampled: $z \sim D$. Then*

$$E_{z\sim D}[\lambda] \in \partial L_D(w)$$

**Proof:** By the definition of sub-gradients: $l(u, z) - l(w, z) \geq \; < u - w, \lambda >$. Let us take the expectations from both sides (the expectation is linear so the inequality holds): $E[l(u, z)] - E[l(w, z)] \geq E[< u - w, \lambda >]$, which means that:

$$L_D(u) - L_D(w) \geq \; < u - w, E_{z\sim D}[\lambda] >$$

and therefore $E_{z\sim D}[\lambda]$ is a sub-gradient of $L_D$. ■

We now define a stochastic sub-gradient algorithm:

**Algorithm 1** Stochastic Gradient Descent Algorithm

```
$w_1 = 0$
init $\mu_1 \in R$ (this is our "pace" parameter - the learning rate).
for $t \in [0, \dots, T]$ do
  choose random vector $r_t$ s.t. $E[r_t] \in \partial f(w_t)$
  $\mu_t = \frac{\mu_1}{\sqrt{t}}$
  $y_t = w_t - \mu_t r_t$
  $w_{t+1} = argmin_{w \in K} ||y_t - w||_2^2$
end for
return $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$
```

We note that we've seen how to produce $r_t$ s.t. $E[r_t] \in \partial f(w_t)$ through sampling.

**Theorem 2.2** *Given that $\rho^2 \geq E[||r_t||^2]$, and $D \geq sup||w - u||$ for $u, v \in K$, then:*

$$E[f(\bar{w})] - f(w^*) \leq \frac{1}{\sqrt{T}}(\frac{D^2}{2\mu_1} + \rho^2 \mu_1)$$

*Which means that for $\mu = \frac{D}{\rho\sqrt{2}}$ we'll get:*

$$E[f(\bar{w})] - f(w^*) \leq D\rho\sqrt{\frac{2}{T}}$$

**Proof:** According to Jensen's inequality, $f(\bar{w}) \leq \frac{1}{T}\sum_{t=1}^{T} f(w_t)$. Taking expectations from both sides yields $E[f(\bar{w})] \leq E[\frac{1}{T}\sum_{t=1}^{T} f(w_t)] = \frac{1}{T}\sum_{t=1}^{T} E[f(w_t)]$.

As before, we will prove convergence using the distance from $w^*$ and not from $f^*$.

**Claim 2.3**

$$\forall t. E[f(w_t)] - f(w^*) \leq E[\frac{||w_t - w^*||_2^2 - ||w_{t+1} - w^*||_2^2}{2\mu_t} + \mu_t \frac{\rho^2}{2}$$

**Proof:** The proof relies on the norm $||x_t - x^*||_2^2$. $||y_t - w^*|| = ||w_t - \mu_t r_t - w^*|| = ||w_t - w^*||_2^2 + \mu_t^2 ||r_t||_2^2 - 2\mu_t < w_t - w^*, r_t >$. Taking expectations from both sides, and using the bound $\rho^2 \geq E[||r_t||^2]$, we get: $E[||y_t - w^*||_2^2] \leq E[||w_t - w^*||_2^2] + \mu_t^2 \rho^2 - 2\mu_t < w_t - w^*, E[r_t] >$. Since $< w_t - w^*, E[r_t] > \ \geq E[f(w_t) - f(w^*)]$, we get:

$$E[||w_t - w^*||_2^2] - E[||y_t - w^*||_2^2] \geq 2\mu_t(E[f(w_t)] - f(w^*)) - \mu_t^2 \rho^2$$

If we replace $y_t$ with $w_{t+1}$ we can only increase the left side expression of this inequality, so it will still hold. All that's left is to re-arrange the terms, getting:

$$E[f(w_t)] - f(w^*) \leq E[\frac{||w_t - w^*||_2^2 - ||w_{t+1} - w^*||_2^2}{2\mu_t}] + \frac{\rho^2 \mu_t}{2}$$

Thus concluding the proof of our claim. ■

(*Note:* this implies that convergence is dependent on the step size and on the rate of learning.)

We now continue, using our claim, to prove that the stochastic gradient descent algorithm converges to the desired value. Let us consider the sum over $t$ of the inequality proven in the previous claim:

$$\sum_{t=1}^{T} E[f(w_t)] - f(w^*) \leq E[||w_1 - w^*||_2^2]\frac{1}{2\mu_1} + \sum_{t=2}^{T} E[||w_t - w^*||_2^2](\frac{1}{2\mu_t} - \frac{1}{2\mu_{t-1}}) -$$

$$-E[||w_{T+1} - w^*||_2^2]\frac{1}{2\mu_T} + \frac{\rho^2}{2}\sum_{t=1}^{T}\mu_t$$

For all $t$, $||w_t - w^*|| \leq D^2$. Also, we note that $E[||w_{T+1} - w^*||_2^2]\frac{1}{2\mu_T}$ is the last term of the summation, it is negative so we can erase it without harming the inequality. Thus we obtain:

$$\sum_{t=1}^{T} E[f(w_t)] - f(w^*) \leq D^2(\frac{1}{2\mu_1} + \sum_{t=2}^{T}(\frac{1}{2\mu_t} - \frac{1}{2\mu_{t-1}})) + \frac{\rho^2}{2}\sum_{t=1}^{T}\mu_t$$

Cancelling out terms and using $\mu_t = \frac{\mu_1}{\sqrt{t}}$ we get

$= D^2\frac{1}{2\mu_T} + \frac{\rho^2\mu_1}{2}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}$, and since $\sum_{t=1}^{T}\frac{1}{\sqrt{t}}$ is bound by $2\sqrt{T}$, this leads to

$$\sum_{t=1}^{T} E[f(w_t)] - f(w^*) \leq D^2\frac{\sqrt{T}}{2\mu_1} + \rho^2\mu_1\sqrt{T}$$

Dividing both sides by T concludes our proof as required. ■

We note that the $\sqrt{T}$ factor originates in our learning rate. If we were to select $\mu_t = \frac{\mu_1}{t}$, for instance, our bound for the second term would have been $O(logT)$. The first term, however, whould have exploded. If in each step we could have "earned" a bit more, perhaps the first term could have been cancelled, allowing a better bound. In the next section we'll see a class of functions for which this is possible.

## 2.4 Strongly Convex Functions

**Definition 3** *A function $f$ is $\lambda$-*strongly convex *if $H = \bigtriangledown^2 f \succeq \lambda \cdot I$, meaning that $\bigtriangledown^2 f - \lambda \cdot I$ is positive semidefinite. I.e., $X^T H x \geq x^T(\lambda I)x = \lambda||x||^2$. This implies that for any $u, w$ and $v \in \partial f(w)$:*

$$\langle w - u, v\rangle \geq f(w) - f(u) + \frac{\lambda}{2}||u - w||_2^2$$

$\lambda$ is, in effect, the convexity factor of $f$, and the term added to the ineqaulity is what would enable us to "earn" more in each step, allowing for a more optimistic step size and a better bound.

Let us take the stochastic sub-gradient descent algorithm, and only modify it so that $\mu_t = \frac{1}{\lambda t}$. If $f$ is $\lambda$-strongly convex, the effect is rather dramatic, and is presented in the next theorem.

**Theorem 2.4** *Given that $\rho^2 \geq E[||r_t||^2]$, and $D \geq sup||w-u||$ for $u, v \in K$ (same as before), then:*

$$E[f(\bar{w})] - f(w^*) \leq \frac{\rho^2}{2\lambda T}(1 + lnT)$$

**Proof:** Let us denote $\bigtriangledown_t = E[r_t]$. Then $\bigtriangledown_t \in \partial f(w_t)$, and therefore:

$$\langle w_t - w^*, \bigtriangledown_t \rangle \geq f(w_t) - f(w^*) + \frac{\lambda}{2}||w_t - w^*||_2^2$$

We may observe that

$$\langle w_t - w^*, \bigtriangledown_t \rangle \leq \frac{E[||w_t - w^*||_2^2 - E[||w_{t+1} - w^*||_2^2}{2\mu_t} + \frac{\mu_t}{2}\rho^2$$

This observation is true because $w_{t+1}$ is the projection of $y_t$ onto $K$, and thus $||y_t - w^*||_2^2 \geq ||w_{t+1} - w^*||_2^2$, as we've seen in the previous section.

This means that $||w_t - w^*||_2^2 - ||w_{t+1} - w^*||_2^2 \geq ||w_t - w^*||_2^2 - ||y_t - w^*||_2^2 = 2\mu_2\langle w_t - w^*, r_t\rangle - \mu^2||r_t||_2^2$. Taking expectations and considering that $E[||r_t||_2^2] \leq \rho^2$ yields the desired inequality.

Combining these observations and summing over $t$ yields:

$$\sum_{t=1}^{T}(E[f(w_t)] - f(w^*)) \leq E[\sum_{t=1}^{T} \frac{||w_t - w^*||_2^2 - ||w_{t+1} - w^*||_2^2}{2\mu_t} - \frac{\lambda}{2}||w_t - w^*||_2^2$$

Using $\mu_t = \frac{1}{t\lambda}$, we notice that the first sum on the right hand side of the inequality collapses to $-\lambda(T+1)||w_{T+1} - f(w^*)||_2^2$, which is negative and thus can be taken off the inequality without changing its correctness. This yields

$$\sum_{t=1}^{T}(E[f(w_t)] - f(w^*)) \leq \frac{\rho^2}{2\lambda}\sum_{t=1}^{T}\frac{1}{t} \leq \frac{\rho^2}{2\lambda}(1 + lnT)$$

Applying Jensen's inequality leads directly to the theorem, as required. ■