MILA Optimisation Crash Course **Winter 2024**

Lecture 5: March 13th

Lecturer: Quentin Bertrand *Scribe(s): Danilo Vucetic*

These are the scribe notes for the fifth lecture of the optimisation crash course at MILA organised by Quentin Bertrand, Damien Scieur, Lucas Maes, and Danilo Vucetic. The purpose of this course is to provide proofs for standard optimisation techniques in order to help practitioners better understand why their algorithms learn. Here is the course web page.

5.1 Recapitulation

Last week we saw Nesterov acceleration, derived through the combination of dual averaging and gradient descent. The driving principle of the derivation of Nesterov acceleration was that the combination of smoothness upper bounds and (averaged) convexity lower bounds would allow an optimisation algorithm to leverage more information about the function, thereby converging more quickly. We saw that the convergence rate for the resulting algorithm was leagues faster than what we had seen previously for gradient descent in any of the cases: smooth, smooth + convex, smooth + strongly convex. Table 5.1 summarises the convergence guarantees we have derived thus far in the crash course.

Table 5.1: Comparison of the convergence rates we have derived for smooth and convex functions. GD = gradient descent. S = smoothness. C = convexity. SC = strong convexity. NA = Nesterov acceleration.

GD+S	$\min_{k\in[0,N]} \left\|\nabla f(x^k)\right\|^2 \leq \frac{2L}{N}(f(x^0) - f(x^\star))$
GD+S+C	$f(x^k) - f(x^*) \leq \frac{L}{2k} \left\|x_0 - x^*\right\|$
GD+S+SC	$f(x^k) - f(x^*) \leq (1 - \frac{\mu}{L})^k (f(x^0) - f(x^*))$
NA	$f(y^{k+1}) - f(y^*) \leq \frac{L}{2A_k} \left\|x_0 - x^*\right\|^2$

5.2 Introduction

This week, we will examine stochastic gradient descent (SGD). It is of interest to many in the field of machine learning to optimise functions that are sums of other functions. We may be more or less familiar with this in terms of empirical risk minimisation, where the empirical risk is defined in Equation 5.1 and the task at hand is defined by Equation 5.2.

$$R(h) = \mathop{\mathbb{E}}_{(x_i,y_i)\sim p_d} [L(h(x_i), y_i)] \tag{5.1}$$

$$h^\star = \arg\min_{h\in\mathcal{H}} R(h) \tag{5.2}$$

In other words, we will optimise over the expected loss of the observed data, where we assume the data is uniformly distributed. Thus, the expectation can be rewritten as follows:

$$\mathop{\mathbb{E}}_{(x_i,y_i)\sim p_d}[L(h(x_i),y_i)] = \frac{1}{n}\sum_{i=0}^{n-1} L(h(x_i),y_i)$$

In machine learning contexts, the hypothesis function, h, is usually parameterised and we are trying to select the best parameters, θ, for a given function, $f(\theta)$. For our purposes, we will represent the empirical risk minimisation problem as follows, where $f_i(\theta) = L(x_i, y_i, \theta)$ is the loss for a data point:

$$\arg\min_\theta f(\theta) = \arg\min_\theta \frac{1}{n}\sum_{i=0}^{n-1} f_i(\theta)$$

Notice that taking the gradient of the function $f(\theta)$ is significantly more expensive than taking the gradient of a single data point: $\nabla f(\theta) \in \mathcal{O}(np)$ versus $\nabla f_i(\theta) \in \mathcal{O}(p)$. Where n is the number of samples and p is the number of features. As such, we would prefer to optimise a function by estimating the gradient of $f(\theta)$ through the gradient of $f_i(\theta)$. This is the motivation for stochastic gradient descent.

5.3 Preliminaries

Definition 5.1 *A continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ with $x, y \in \mathbb{R}^p$ is L-smooth if*

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|. \tag{5.3}$$

From [1] and [2].

Lemma 5.2 *A smooth function f is upper-bounded by a parabola, i.e.,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2. \tag{5.4}$$

Definition 5.3 *A function $f : \mathbb{R}^p \to \mathbb{R}$ with $x, y \in \mathbb{R}^p$ is convex if and only if its domain is a convex set (see [1], to be defined) and if for all $\alpha \in [0, 1]$ it satisfies*

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y). \tag{5.5}$$

Lemma 5.4 *A convex function f is lower-bounded by a line, i.e.,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle. \tag{5.6}$$

Definition 5.5 *A function $f : \mathbb{R}^p \to \mathbb{R}$ with $x, y \in \mathbb{R}^p$ is strongly convex if and only if its domain is a convex set (see [1], to be defined) and if for $\alpha \in [0, 1]$ it satisfies*

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\mu}{2}\alpha(1-\alpha)\|x - y\|^2 \tag{5.7}$$

for some positive μ.

Proposition 5.6 *A (differentiable) strongly convex function is lower-bounded by a parabola* $\forall x, y$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \tag{5.8}$$

Definition 5.7 *Gradient descent is characterized by the following equation.*

$$x^{k+1} = x^k - \alpha \nabla_x f(x^k) \tag{5.9}$$

Lemma 5.8 *The law of total expectation*[1] *states that an expectation can be restated as the expectation of conditional expectations.*

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

5.4 Stochastic Gradient Descent

The update rule for stochastic gradient descent is similar to that of gradient descent, with one notable change: instead of computing the full, expensive, gradient of the function $f(\theta)$, we will compute the cheaper gradient of $f_i(\theta)$, using this to update the parameters. Remember, the index i simply denotes some data point in a dataset. As such, we can choose any sampling distribution to select i since we ultimately control this factor. For simplicity, we select a uniform distribution and stochastically sample the index on each iteration. The update rule of SGD is stated in Equation 5.10.

$$\theta^{k+1} = \theta^k - \gamma \nabla f_i(\theta^k) \tag{5.10}$$

Note that i is a random variable, and the iterates θ^k are also random variables since they depend on i. It is important to note that $\nabla f_i(\theta)$ is an unbiased estimator of the full gradient $\nabla f(\theta)$. This can be seen easily by taking the expectation of $\nabla f_i(\theta)$ over i (which is uniform), with a constant θ:

$$\mathbb{E}[\nabla f_i(\theta)] = \sum_{n=0}^{N} \nabla f_n(\theta) \frac{1}{N} = \nabla f(\theta) \tag{5.11}$$

If we instead take θ to be a random variable $\theta \leftarrow \theta^k$, then we must take the expectation over both i and θ^k. This again produces an unbiased estimator, but requires the use of the law of total expectation (see Lemma 5.8). Note that $\mathbb{E}[\nabla f_i(\theta)] = \mathbb{E}\left[\nabla f_i(\theta^k)|\theta^k\right]$, *i.e.*, we see θ^k as a constant rather than a random variable.

$$\mathbb{E}\left[\nabla f_i(\theta^k)\right] = \mathop{\mathbb{E}}_{\theta^k}\left[\mathop{\mathbb{E}}_{i}\left[\nabla f_i(\theta^k)|\theta^k\right]\right] = \mathop{\mathbb{E}}_{\theta^k}\left[\nabla f(\theta^k)\right] \tag{5.12}$$

Theorem 5.9 *The convergence rate of stochastic gradient descent. Assuming the component functions, $f_i(\theta)$, are μ-strongly convex and L-smooth, with a constant $c \in \mathbb{R}^{\geq 0}$, SGD has a general rate of*

$$\mathbb{E}\left[\left\|\theta^k - \theta^\star\right\|^2\right] \leq (1 - \gamma\mu)^k \left\|\theta_0 - \theta^\star\right\|^2 + c \tag{5.13}$$

This implies that the iterates of SGD do not converge in general unless the constant c is zero.[2]

We now provide two proofs for the convergence rate of SGD, starting with a strong assumption, then using a weaker assumption.

[1] See here.

[2] Note: Quentin implied that it might very well be the case for deep learning that the constant is zero, but no sources have been provided, so treat this as conjecture.

5.4.1 Proof with a strong assumption

We know that f is strongly convex, and that the domain of f is a convex set Q. For the first proof, we make an assumption on the gradients of f_i.[3]

1. The gradients must be bounded. This means we are dealing with finite gradients, or equivalently, that $\nabla f_i(x)$ is Lipshitz. $G \in \mathbb{R}$.
$$\max_{x \in Q} \|\nabla f_i(x)\| \leq G$$

Note, this is the same assumption we made for the convergence rate of subgradient descent. It is a "strong" assumption because it relies on the largest gradient of the function, which may be huge, thereby giving a terrible rate.

Proof: To begin the proof, we take the expectation of the norm of the difference of the k-th iterate and the optimum.

$$\begin{aligned}
\mathbb{E}\left[\left\|\theta^{k+1} - \theta^\star\right\|^2\right] &= \mathbb{E}\left[\left\|\theta^k - \gamma \nabla f_i(\theta^k) - \theta^\star\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\theta^k - \theta^\star\right\|^2 + \gamma^2 \left\|\nabla f_i(\theta^k)\right\|^2 - 2\gamma \langle \nabla f_i(\theta^k), \theta^k - \theta^\star \rangle\right] \\
&= \mathbb{E}\left[\left\|\theta^k - \theta^\star\right\|^2\right] + \gamma^2 \mathbb{E}\left[\left\|\nabla f_i(\theta^k)\right\|^2\right] - 2\gamma \mathbb{E}\left[\langle \nabla f_i(\theta^k), \theta^k - \theta^\star \rangle\right]
\end{aligned}$$

We want to get rid of the $\nabla f_i(\theta^k)$ term in the inner-product. To do this, we will again use the law of total expectation (Lemma 5.8), this time noting that the inner product is a sum over the products of each element of the vectors $\nabla f_i(\theta^k), \theta^k - \theta^\star \in \mathbb{R}^{\mathbb{D}}$. This allows us to, eventually, bring the expectation into the sum, thereby producing a result similar to Equation 5.12.

$$\begin{aligned}
-2\gamma \mathbb{E}\left[\langle \nabla f_i(\theta^k), \theta^k - \theta^\star \rangle\right] &= -2\gamma \mathop{\mathbb{E}}_{\theta^k}\left[\mathop{\mathbb{E}}_{i}\left[\langle \nabla f_i(\theta^k), \theta^k - \theta^\star \rangle | \theta^k\right]\right] \\
&= -2\gamma \mathop{\mathbb{E}}_{\theta^k}\left[\mathop{\mathbb{E}}_{i}\left[\sum_{j=1}^{D} \nabla f_{i,j}(\theta^k)(\theta^k - \theta^\star)_j \middle| \theta^k\right]\right] \\
&= -2\gamma \mathop{\mathbb{E}}_{\theta^k}\left[\sum_{j=1}^{D} \mathop{\mathbb{E}}_{i}\left[\nabla f_{i,j}(\theta^k)(\theta^k - \theta^\star)_j \middle| \theta^k\right]\right] \\
&= -2\gamma \mathop{\mathbb{E}}_{\theta^k}\left[\sum_{j=1}^{D} \mathop{\mathbb{E}}_{i}\left[\nabla f_{i,j}(\theta^k) \middle| \theta^k\right](\theta^k - \theta^\star)_j\right] \\
&= -2\gamma \mathop{\mathbb{E}}_{\theta^k}\left[\left\langle \mathop{\mathbb{E}}_{i}\left[\nabla f_i(\theta^k) \middle| \theta^k\right], \theta^k - \theta^\star \right\rangle\right] \\
&= -2\gamma \mathop{\mathbb{E}}_{\theta^k}\left[\langle \nabla f(\theta^k), \theta^k - \theta^\star \rangle\right]
\end{aligned}$$

Plugging this into the previous equation and using strong convexity (Proposition 5.6) to get rid of the inner product

$$\begin{aligned}
\mathbb{E}\left[\left\|\theta^{k+1} - \theta^\star\right\|^2\right] &= \mathbb{E}\left[\left\|\theta^k - \theta^\star\right\|^2\right] + \gamma^2 \mathbb{E}\left[\left\|\nabla f_i(\theta^k)\right\|^2\right] - 2\gamma \mathop{\mathbb{E}}_{\theta^k}\left[\langle \nabla f(\theta^k), \theta^k - \theta^\star \rangle\right] \\
&\leq \mathbb{E}\left[\left\|\theta^k - \theta^\star\right\|^2\right] + \gamma^2 \mathbb{E}\left[\left\|\nabla f_i(\theta^k)\right\|^2\right] + 2\gamma \mathop{\mathbb{E}}_{\theta^k}\left[f(\theta^\star) - f(\theta^k) - \frac{\mu}{2}\left\|\theta^k - \theta^\star\right\|^2\right] \\
&= (1 - \gamma\mu)\,\mathbb{E}\left[\left\|\theta^k - \theta^\star\right\|^2\right] + \gamma^2 \mathbb{E}\left[\left\|\nabla f_i(\theta^k)\right\|^2\right] + 2\gamma \mathop{\mathbb{E}}_{\theta^k}\left[f(\theta^\star) - f(\theta^k)\right]
\end{aligned}$$

[3] For more on convex sets see [2].

Now, we note that $f(\theta^\star) \leq f(\theta^k)$, so the rightmost expectation can be set to zero. In addition, we have bounded the gradients of the individual data points, hence the middle expectation will be replaced by G^2. All that remains is to solve the recurrence.

$$\begin{aligned}
\mathbb{E}\left[\left\|\theta^{k+1} - \theta^\star\right\|^2\right] &\leq (1-\gamma\mu)\,\mathbb{E}\left[\left\|\theta^k - \theta^\star\right\|^2\right] + \gamma^2 G^2 \\
&\leq (1-\gamma\mu)\left((1-\gamma\mu)\,\mathbb{E}\left[\left\|\theta^{k-1} - \theta^\star\right\|^2\right] + \gamma^2 G^2\right) + \gamma^2 G^2 \\
&= (1-\gamma\mu)^2\,\mathbb{E}\left[\left\|\theta^{k-1} - \theta^\star\right\|^2\right] + \gamma^2 G^2\,(1+1-\gamma\mu) \\
&\cdots \\
&\leq (1-\gamma\mu)^{k+1}\,\mathbb{E}\left[\left\|\theta^0 - \theta^\star\right\|^2\right] + \gamma^2 G^2 \sum_{j=0}^{k} (1-\gamma\mu)^j
\end{aligned}$$

The expectation can be removed since θ^0 is not a random variable. Additionally, since μ and γ are less than one, we can rewrite this as a series: $\sum_{j=0}^{k}(1-\gamma\mu)^j = \frac{1}{1-1+\mu\gamma} = \frac{1}{\mu\gamma}$. The desired result is finally obtained:

$$\mathbb{E}\left[\left\|\theta^{k+1} - \theta^\star\right\|^2\right] \leq (1-\gamma\mu)^{k+1}\left\|\theta^0 - \theta^\star\right\|^2 + \frac{\gamma}{\mu}G^2 \tag{5.14}$$

■

5.4.2 Proof with a weaker assumption

In practice, $\mathbb{E}\left[\left\|\nabla f_i(\theta^k)\right\|^2\right]$ is hard to bound: we don't actually know what value G takes. Further, we know that at the optimum $\nabla f(\theta^\star) = 0$, but this does not imply that $\nabla f_i(\theta^\star) = 0$! Thus, we have no direct information on the gradients of the individual functions $f_i(\theta)$. Instead of assuming that our gradients are bounded by some G, we will now assume that the norm of the gradient is finite, *i.e.*, that $\|\nabla f_i(\theta^\star)\|^2 = \sigma_f < \infty$. Starting from $\mathbb{E}\left[\left\|\nabla f_i(\theta^k)\right\|^2\right]$, we will now prove a convergence bound but with the weaker assumption.

Proof:

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla f_i(\theta^k)\right\|^2\right] &= \mathbb{E}\left[\left\|\nabla f_i(\theta^k) - \nabla f_i(\theta^\star) + \nabla f_i(\theta^\star)\right\|^2\right] \\
&\leq 2\,\mathbb{E}\left[\left\|\nabla f_i(\theta^k) - \nabla f_i(\theta^\star)\right\|^2 + \left\|\nabla f_i(\theta^\star)\right\|^2\right] \\
&= 2\,\mathbb{E}\left[\left\|\nabla f_i(\theta^k) - \nabla f_i(\theta^\star)\right\|^2\right] + 2\sigma_f
\end{aligned}$$

We used the fact that $(a+b)^2 \leq 2a^2 + 2b^2$. We still need to linearise the term inside the expectation so that we can simplify the inequality. We will use the assumption of smoothness on the functions $f_i(\cdot)$ to accomplish this goal.

Lemma 5.10 *An arbitrary smooth function $g(\cdot)$ obeys the bound*

$$\frac{1}{2L}\left\|\nabla g(x)\right\|^2 \leq g(x) - g(y)$$

Proof: To prove this lemma, we begin with Lemma 5.2, to get the smoothness upper bound and find the minimiser with respect to y.

$$\begin{aligned} g(y) &\leq g(x) + \langle \nabla g(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 \\ \min_y g(y) &\leq \min_y g(x) + \langle \nabla g(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 \\ g^\star &\leq \min_y g(x) + \langle \nabla g(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 \end{aligned}$$

Now, taking the gradient with respect to y on the right hand side will provide the equation for y.

$$\begin{aligned} 0 &= \nabla_y \left(g(x) + \langle \nabla_x g(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 \right) \\ &= \nabla_x g(x) + L(y - x) \\ y &= x - \frac{1}{L} \nabla_x g(x) \end{aligned}$$

Plugging this into the above, we get the minimiser of the right hand side.

$$\begin{aligned} g^\star &\leq g(x) + \left\langle \nabla g(x), x - \frac{1}{L} \nabla_x g(x) - x \right\rangle + \frac{L}{2} \left\| x - \frac{1}{L} \nabla_x g(x) - x \right\|^2 \\ &= g(x) - \frac{1}{2L} \|\nabla_x g(x)\|^2 \end{aligned}$$

Rearranging, we get the desired result:

$$\frac{1}{2L} \|\nabla_x g(x)\|^2 \leq g(x) - g^\star = g(x) - g(y)$$

■

We can use this lemma to linearise $\left\|\nabla f_i(\theta^k) - \nabla f_i(\theta^\star)\right\|^2$ by defining a smooth function $h_{\theta^\star}(\cdot)$:

$$h_{\theta^\star}(\theta) = f_i(\theta) - \langle \nabla f_i(\theta^\star), \theta \rangle$$

We can easily show that $h_{\theta^\star}(\cdot)$ is smooth from the fact the $f_i(\cdot)$ is smooth:

$$\|\nabla_{\theta_1} h_{\theta^\star}(\theta_1) - \nabla_{\theta_2} h_{\theta^\star}(\theta_2)\| = \|\nabla_{\theta_1} f_i(\theta_1) - \nabla_{\theta_2} f_i(\theta_2)\| \leq L \|\theta_1 - \theta_2\|$$

Since $h_{\theta^\star}(\cdot)$ is smooth, we can use the above lemma.

$$\begin{aligned} \frac{1}{2L} \left\|\nabla_{\theta^k} h_{\theta^\star}(\theta^k)\right\|^2 &\leq h_{\theta^\star}(\theta^k) - h_{\theta^\star}(\theta^\star) \\ \left\|\nabla_{\theta^k} f_i(\theta^k) - \nabla_{\theta^\star} f_i(\theta^\star)\right\| &\leq 2L \left(f_i(\theta^k) - f_i(\theta^\star) - \left\langle \nabla f_i(\theta^\star), \theta^k - \theta^\star \right\rangle \right) \end{aligned}$$

Using this result, we can complete the proof, again using the law of total expectation, Lemma 5.8.

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla f_i(\theta^k)\right\|^2\right] &\leq 2\,\mathbb{E}\left[\left\|\nabla f_i(\theta^k) - \nabla f_i(\theta^\star)\right\|^2\right] + 2\sigma_f \\
&\leq 4L\,\mathbb{E}\left[f_i(\theta^k) - f_i(\theta^\star) - \left\langle \nabla f_i(\theta^\star), \theta^k - \theta^\star \right\rangle\right] + 2\sigma_f \\
&= 4L \mathop{\mathbb{E}}_{\theta^k}\left[\mathop{\mathbb{E}}_{i}\left[f_i(\theta^k) - f_i(\theta^\star) - \left\langle \nabla f_i(\theta^\star), \theta^k - \theta^\star \right\rangle \mid \theta^k\right]\right] + 2\sigma_f \\
&= 4L \mathop{\mathbb{E}}_{\theta^k}\left[\mathop{\mathbb{E}}_{i}\left[f_i(\theta^k) \mid \theta^k\right] - \mathop{\mathbb{E}}_{i}\left[f_i(\theta^\star) \mid \theta^k\right] - \mathop{\mathbb{E}}_{i}\left[\left\langle \nabla f_i(\theta^\star), \theta^k - \theta^\star \right\rangle \mid \theta^k\right]\right] + 2\sigma_f \\
&= 4L \mathop{\mathbb{E}}_{\theta^k}\left[f(\theta^k) - f(\theta^\star) - \left\langle \mathop{\mathbb{E}}_{i}\left[\nabla f_i(\theta^\star) \mid \theta^k\right], \theta^k - \theta^\star \right\rangle\right] + 2\sigma_f \\
&= 4L \mathop{\mathbb{E}}_{\theta^k}\left[f(\theta^k) - f(\theta^\star)\right] + 2\sigma_f
\end{aligned}$$

Now, continuing from an intermediate result in Section 5.4.1.

$$\begin{aligned}
\mathbb{E}\left[\left\|\theta^{k+1} - \theta^\star\right\|^2\right] &\leq (1-\gamma\mu)\,\mathbb{E}\left[\left\|\theta^k - \theta^\star\right\|^2\right] + \gamma^2\,\mathbb{E}\left[\left\|\nabla f_i(\theta^k)\right\|^2\right] + 2\gamma \mathop{\mathbb{E}}_{\theta^k}\left[f(\theta^\star) - f(\theta^k)\right] \\
&\leq (1-\gamma\mu)\,\mathbb{E}\left[\left\|\theta^k - \theta^\star\right\|^2\right] + 4\gamma^2 L \mathop{\mathbb{E}}_{\theta^k}\left[f(\theta^k) - f(\theta^\star)\right] + 2\gamma^2\sigma_f + 2\gamma \mathop{\mathbb{E}}_{\theta^k}\left[f(\theta^\star) - f(\theta^k)\right] \\
&= (1-\gamma\mu)\,\mathbb{E}\left[\left\|\theta^k - \theta^\star\right\|^2\right] + 2\gamma^2\sigma_f + 2\gamma\,(1-2\gamma L) \mathop{\mathbb{E}}_{\theta^k}\left[f(\theta^\star) - f(\theta^k)\right]
\end{aligned}$$

By definition, $f(\theta^\star) - f(\theta^k) \leq 0$. Thus, we can remove the rightmost term from the inequality so long as $0 \leq 1 - 2\gamma L$. This implies that the learning rate should be set as follows: $\gamma \leq \frac{1}{2L}$. With this condition satisfied, we have a recurrence, whose solution yields the convergence rate of SGD under the milder assumption.

$$\begin{aligned}
\mathbb{E}\left[\left\|\theta^{k+1} - \theta^\star\right\|^2\right] &\leq (1-\gamma\mu)\,\mathbb{E}\left[\left\|\theta^k - \theta^\star\right\|^2\right] + 2\gamma^2\sigma_f \\
&\leq (1-\gamma\mu)\left((1-\gamma\mu)\,\mathbb{E}\left[\left\|\theta^{k-1} - \theta^\star\right\|^2\right] + 2\gamma^2\sigma_f\right) + 2\gamma^2\sigma_f \\
&= (1-\gamma\mu)^2\,\mathbb{E}\left[\left\|\theta^{k-1} - \theta^\star\right\|^2\right] + 2\gamma^2\sigma_f\,(1 + 1 - \gamma\mu) \\
&\leq \cdots \\
\mathbb{E}\left[\left\|\theta^{k+1} - \theta^\star\right\|^2\right] &\leq (1-\gamma\mu)^{k+1}\left\|\theta^0 - \theta^\star\right\|^2 + \frac{2\gamma\sigma_f}{\mu}
\end{aligned}$$

This completes the proof.

■

5.5 Concluding remarks

Notice how in both proofs, the convergence rate has a constant. This indicates that SGD does not generally converge.

References

[1] I. Mitliagkas *et al.*, "Gradients for smooth and for strongly convex functions." Course notes for IFT 6085 at UdeM.

[2] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 1–357, 2015.