Stochastic convex optimization

Vincent Divol

Recall the risk minimization paradigm. Let P be a probability distribution on $\mathcal{X} \times \mathcal{Y}$ (where $\mathcal{X}$ is the set of inputs and $\mathcal{Y}$ is the set of outputs). Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y}$, we defined the P-risk of a predictor $f : \mathcal{X} \times \mathcal{Y}$ as the quantity $\mathcal{R}_P(f) = \mathbb{E}_P[\ell(f(\mathbf{x}), \mathbf{y})]$. The Bayes predictor is defined as the predictor $f_P^\star$ minimizing the P-risk, with $\mathcal{R}_P^\star = \mathcal{R}_P(f_P^\star)$. To approach the Bayes predictor, we introduce a class of predictors $\mathcal{F} = \{f_\theta : \mathcal{X} \to \mathcal{Y},\ \theta \in \Theta\}$ indexed by some convex subset $\Theta \subset \mathbb{R}^d$ and consider the minimizer $\theta^\star$ of the function $\theta \in \Theta \mapsto \mathcal{R}_P(f_\theta)$. Note that in practice, we do not have access to the function $\mathcal{R}_P(f_\theta)$ (as P is unknown), so that computing $\theta^\star$ is not an option.

Previously, we proposed to approximate $\theta^\star$ by computing the minimum of the empirical risk $\mathcal{R}_n$ on the class of predictors $\mathcal{F}$. We here propose a slightly different perspective to achieve this same goal. What if we tried to apply the gradient descent algorithm to the function $F : \theta \mapsto \mathcal{R}_P(f_\theta)$? To do so, we only need to have access to the gradient of F, which is given at $\theta \in \Theta$ by

$$\nabla F(\theta) = \mathbb{E}_P[\nabla_\theta \ell(f_\theta(\mathbf{x}), \mathbf{y})]. \tag{1}$$

Of course, as we do not have access to P, we cannot compute this gradient. However, an unbiased estimator of this gradient is given by $\nabla_\theta \ell(f_\theta(\mathbf{x_i}), \mathbf{y_i})$, where $(\mathbf{x_i}, \mathbf{y_i})$ is an observation with distribution P. Running gradient descent with those approximations of the gradient is referred to as **stochastic gradient descent**. We explore in this chapter the performance of this method, and compare it with gradient descent applied to the empirical risk $\mathcal{R}_n$.

1 Stochastic gradient descent

We consider the more general setting where the goal is to minimize a function $F : \Theta \subset \mathbb{R}^d \to \mathbb{R}$. Stochastic gradient descent will iteratively compute a sequence of (random) iterates θ^t, based on (random) approximations of the gradients $\nabla F(\theta^t)$. More precisely, we assume that for every $t \geq 0$, we have access to a random vector $\mathbf{v^t}$ that satisfies $\mathbb{E}[\mathbf{v^t}|\theta^t] = \nabla F(\theta^t)$.

Algorithm 1: Stochastic gradient descent

Initialization: list of step sizes $(s_t)_{t=1,\dots,T-1}$, $\theta^1 \in \Theta$;
for $t = 1, \dots, T-1$ **do**
  draw $\mathbf{v^t}$ such that $\mathbb{E}[\mathbf{v^t}|\theta^t] = \nabla F(\theta^t)$;
  let $\theta^{t+1} = \theta^t - s_t \mathbf{v^t}$;
end
Output: $\overline{\theta} = \frac{1}{T} = \sum_{t=1}^{T} \theta^t$;

This general framework might appear quite abstract, so let us give directly the application we have in mind in this chapter. Let $F(\theta) = \mathbb{E}_P[\ell(f_\theta(\mathbf{x}), \mathbf{y})]$, so that $\nabla F(\theta) = \mathbb{E}_P[\nabla_\theta \ell(f_\theta(\mathbf{x}), \mathbf{y})]$. Assume that we have access to a sample of T i.i.d. samples $(\mathbf{x_1}, \mathbf{y_1}), \dots, (\mathbf{x_{T-1}}, \mathbf{y_{T-1}})$ from the distribution P. In this case, we define $\mathbf{v^t} = \nabla_\theta \ell(f_{\theta^t}(\mathbf{x_t}), \mathbf{y_t})$. As the vector θ^t only depends on the observation $(\mathbf{x_j}, \mathbf{y_j})$ for $j < t$, $\mathbf{v^t}$ is independent from θ^t, ensuring that

$$\mathbb{E}[\mathbf{v^t}|\theta^t] = \mathbb{E}_P[\nabla_{\theta^t} \ell(f_\theta(\mathbf{x}), \mathbf{y})] = \nabla F(\theta^t). \tag{2}$$

Example 1. Another set of examples where stochastic gradient descent can be utilized is when we are looking for the minimum of a function F of the form $\theta \mapsto \frac{1}{n} \sum_{i=1}^{n} F_i(\theta)$, where the functions $F_i : \Theta \to \mathbb{R}$ are arbitrary functions. There is nothing random in the definition of the function F. We may however define a uniform random variable $\mathbf{i}$ on the set of indexes $\{1, \dots, n\}$, and remark that one can express F as

$$F(\theta) = \mathbb{E}[F_{\mathbf{i}}(\theta)]. \tag{3}$$

In this situation, one can obtain unbiased estimates of the gradients, by letting $\mathbf{i_1}, \dots, \mathbf{i_{T-1}}$ be T i.i.d. uniform random indexes and then by defining $\mathbf{v^t} = \nabla F_{\mathbf{i_t}}(\theta^t)$.

We are now ready to state our first theorem: stochastic gradient descent converges as long as the function F is convex, Lipschitz continuous, and that the estimates $\mathbf{v^t}$ of the gradients are bounded.

Theorem 2. *Let $B(0;R)$ be the open ball centered at 0 in $\mathbb{R}^d$. Let $F : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function, with minimizer $\theta^\star \in B(0;R)$. Let $T \geq 1$ be an integer and assume that for every $t = 1, \dots, T-1$, it holds that $\mathbb{E}[\|\mathbf{v^t}\|^2] \leq \rho^2$ for some constant $\rho > 0$. Consider stochastic gradient descent with constant step size $s = \sqrt{\frac{4R^2}{\rho^2 T}}$ and initialization $\theta^0 \in B(0;R)$. Then, the output $\overline{\theta}$ of stochastic gradient descent after T steps satisfies*

$$\mathbb{E}[F(\overline{\theta})] - F(\theta^\star) \leq 2\frac{R\rho}{\sqrt{T}}. \tag{4}$$

First proof of Theorem 2: without randomness. We first give a proof of Theorem 2 in the case where we always have $v^t = \nabla F(\theta^t)$. Note that in this case, the algorithm boils down to classical gradient descent, where we use the average of the iterates as our final output. To insist on the non-randomness of the method, we write v^t instead of $\mathbf{v^t}$. We can first apply Jensen's inequality: it holds that

$$F(\overline{\theta}) \leq \frac{1}{T}\sum_{t=1}^{T} F(\theta^t). \tag{5}$$

Also, by convexity of F, we have

$$F(\theta^t) - F(\theta^\star) \leq \langle v^t, \theta^t - \theta^\star \rangle \tag{6}$$

(remember that $v^t = \nabla F(\theta^t)$ by assumption in this simplified setting). Putting those two equation together yields that

$$F(\overline{\theta}) - F(\theta^\star) \leq \frac{1}{T}\sum_{t=1}^{T} \langle v^t, \theta^t - \theta^\star \rangle. \tag{7}$$

To bound the sum in (7), we are going to make a telescopic sum appear. To do so, we use the general identity

$$\langle a, b \rangle = \frac{1}{2}(\|a+b\|^2 - \|a\|^2 - \|b\|^2) \tag{8}$$

with $a = \theta^\star - \theta^t$, $b = sv^t$. This yields

$$\begin{aligned}\langle v^t, \theta^t - \theta^\star \rangle &= -\frac{1}{s}\langle sv^t, \theta^\star - \theta^t \rangle \\ &= -\frac{1}{2s}(\|\theta^\star - \theta^t + sv^t\|^2 - \|\theta^t - \theta^\star\|^2 - s^2\|v^t\|^2) \\ &= -\frac{1}{2s}(\|\theta^{t+1} - \theta^\star\|^2 - \|\theta^t - \theta^\star\|^2 - s^2\|v^t\|^2) \\ &= \frac{1}{2s}(\|\theta^t - \theta^\star\|^2 - \|\theta^{t+1} - \theta^\star\|^2) + \frac{s}{2}\|v^t\|^2. \end{aligned} \tag{9}$$

By summing over t, we obtain from (7) that

$$F(\overline{\theta}) - F(\theta^\star) \leq \frac{1}{2Ts}(\|\theta^0 - \theta^\star\|^2 - \|\theta^T - \theta^\star\|^2) + \frac{s}{2T}\sum_{t=1}^{T}\|v^t\|^2. \tag{10}$$

To conclude, we use that both θ^0 and $\theta^\star$ belong to R, and that all the gradients v^t have a norm smaller than ρ:

$$F(\overline{\theta}) - F(\theta^\star) \leq \frac{(2R)^2}{2Ts} + \frac{s\rho^2}{2}. \tag{11}$$

One obtains the conclusion by plugging in the value $s = \sqrt{\frac{4R^2}{\rho^2 T}}$. □

Second proof of Theorem 2: general case. In the general case, we adopt the same proof technique, but have to be careful when taking expectations. First, note that as before, by Jensen inequality,

$$\begin{aligned} F(\overline{\theta}) - F(\theta^\star) &\leq \frac{1}{T}\sum_{t=1}^{T}\langle \nabla F(\theta^t), \theta^t - \theta^\star \rangle \\ &= \frac{1}{T}\sum_{t=1}^{T}\langle \mathbb{E}[\mathbf{v^t}|\theta^t], \theta^t - \theta^\star \rangle \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\langle \mathbf{v^t}, \theta^t - \theta^\star \rangle | \theta^t]. \end{aligned} \tag{12}$$

Therefore, by using the law of total expectation,

$$\begin{aligned}\mathbb{E}[F(\overline{\theta})] - F(\theta^\star) &\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\langle \mathbf{v^t}, \theta^t - \theta^\star\rangle | \theta^t]\right] \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\langle \mathbf{v^t}, \theta^t - \theta^\star\rangle] \\ &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{v^t}, \theta^t - \theta^\star\rangle\right].\end{aligned} \tag{13}$$

As before, it holds that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{v^t}, \theta^t - \theta^\star\rangle &= \frac{1}{2Ts}(\|\theta^0 - \theta^\star\|^2 - \|\theta^T - \theta^\star\|^2) + \frac{s}{2T}\sum_{t=1}^{T}\|\mathbf{v^t}\|^2 \\ &\leq \frac{(2R)^2}{2Ts} + \frac{s}{2T}\sum_{t=1}^{T}\|\mathbf{v^t}\|^2.\end{aligned} \tag{14}$$

By putting (13) and (14) together, we obtain that

$$\begin{aligned}\mathbb{E}[F(\overline{\theta})] - F(\theta^\star) &\leq \frac{(2R)^2}{2Ts} + \frac{s}{2T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{v^t}\|^2] \\ &\leq \frac{(2R)^2}{2Ts} + \frac{s}{2T}\sum_{t=1}^{T}\rho^2 \\ &\leq \frac{(2R)^2}{2Ts} + \frac{s\rho^2}{2}.\end{aligned}$$

The conclusion is obtained as before by choosing $s = \sqrt{\frac{4R^2}{\rho^2 T}}$. □

Remark 3.

1. The randomness in (4) comes from the randomness in the estimates $\mathbf{v^t}$ of the gradients. In particular, we want to insist on the fact that the function F is not random here.

2. As a particular case of this theorem, we may consider the case where $\mathbf{v^t} = \nabla F(\theta^t)$ (exact gradients). This exactly corresponds to classical gradient descent with the final output being equal to the average of the iterates θ_t.

3. In practice, it is common to discard the first iterates when computing the average $\overline{\theta}$. This allows one to "forget" about the initialization θ^1 (that has no reason to be relevant).

Example 4 (Toy example). Consider the function $F : \mathbb{R}^2 \to \mathbb{R}$ defined by $F(\theta) = \frac{a_1}{2}\theta_1^2 + \frac{a_2}{2}\theta_2^2$. If we have access to the gradients of F, then iterates of the gradient descent will linearly converge to 0. If, on the opposites, we do not have access to $\nabla F(\theta)$, but to a corrupted version $\mathbf{v} = \nabla F(\theta) + \mathbf{u}$ where $\mathbf{u}$ is a bounded random variable, then we can implement stochastic gradient descent. The iterates of the stochastic gradient descent are displayed in Figure 1.

The rate of convergence in this theorem is of order $1/\sqrt{T}$. Previously, we showed that if F is β-smooth (that is the gradient of F is β-Lipschitz), then we can have a rate of convergence of order $1/T$. It is natural to wonder if this faster rate also holds for stochastic gradient descent. It turns out that assuming smoothness does not improve the $1/\sqrt{T}$ rate of convergence here. However, assuming that the function is α-strongly convex is enough to obtain this $1/T$-rate of convergence. To do so, we use a variant of the previous stochastic gradient algorithm where we use an additional projection step to ensure that the different iterates θ^t do not blow up. For $R > 0$, we let

$$\mathrm{proj}_R(\theta) = \begin{cases} \theta & \text{if } \|\theta\| \leq R \\ R\frac{\theta}{\|\theta\|} & \text{otherwise,} \end{cases} \tag{15}$$

see also Figure 2.

Algorithm 2: Stochastic gradient descent with projection step

Initialization: list of step sizes $(s_t)_{t=1,\dots,T-1}$, $\theta^1 \in \Theta$, radius $R > 0$;
for $t = 1, \dots, T-1$ **do**
  draw $\mathbf{v^t}$ such that $\mathbb{E}[\mathbf{v^t}|\theta^t] = G(\theta^t)$;
  let $\theta^{t+1} = \mathrm{proj}_R(\theta^t - s_t\mathbf{v^t})$;
end
Output: $\overline{\theta} = \frac{1}{T} = \sum_{t=1}^{T} \theta^t$;

Theorem 5. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a α-strongly convex differentiable function, with minimizer $\theta^\star \in B(0; R)$. Let $T \geq 1$ be an integer and assume that for every $t = 1, \dots, T-1$, it holds that $\mathbb{E}[\|\mathbf{v^t}\|^2] \leq \rho^2$ for some constant $\rho >$*

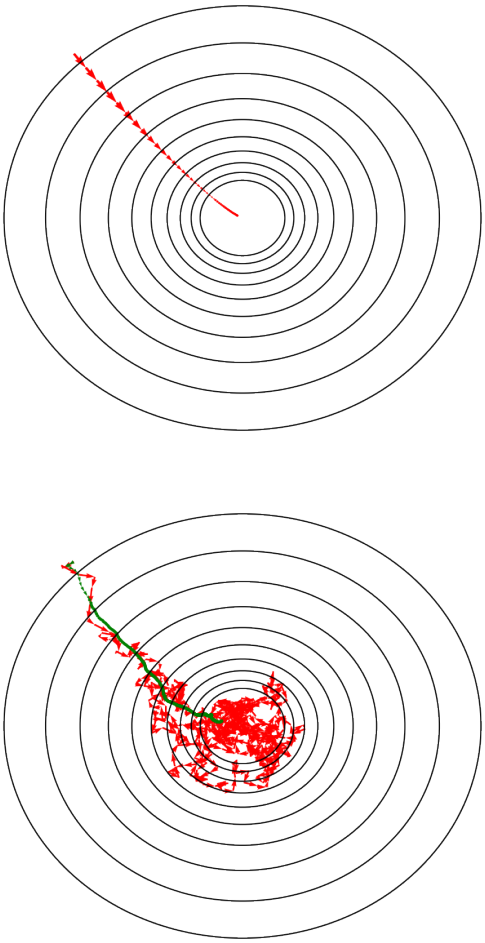

Figure 1: Top: iterates of the gradient descent. Bottom: iterates of the stochastic gradient descent (red), and average of the first iterates (green).

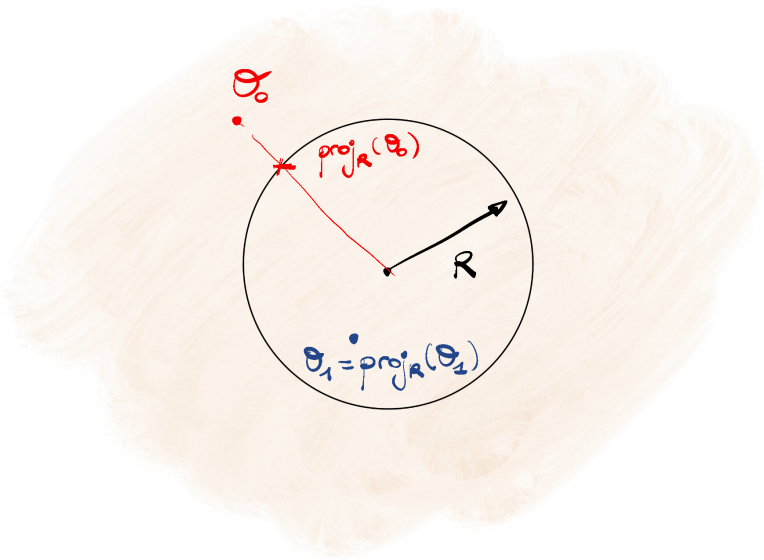

Figure 2: The proj_R function.

0. *Consider stochastic gradient descent with projection step, with step size* $s_t = 1/(\alpha t)$ *and initialization* θ^1*. Then, the output* $\overline{\theta}$ *of stochastic gradient descent after* T *steps satisfies*

$$\mathbb{E}[F(\overline{\theta})] - F(\theta^\star) \leq \frac{\rho^2}{2\alpha T}(1 + \log(T)). \tag{16}$$

Proof. For sake of simplicity, we prove the result in the case $R = \infty$ (that is without the projection step). Our starting point is once again the identity (9), that states that

$$\begin{aligned} \langle \mathbf{v^t}, \theta^t - \theta^\star \rangle &= \frac{1}{2s_t}(\|\theta^t - \theta^\star\|^2 - \|\theta^{t+1} - \theta^\star\|^2) + \frac{s_t}{2}\|\mathbf{v^t}\|^2 \\ &= \frac{\alpha t}{2}(\|\theta^t - \theta^\star\|^2 - \|\theta^{t+1} - \theta^\star\|^2) + \frac{1}{2\alpha t}\|\mathbf{v^t}\|^2. \end{aligned} \tag{17}$$

We then use strong convexity, which implies that for every $t \geq 1$,

$$\begin{aligned} F(\theta^t) - F(\theta^\star) &\leq \langle \nabla F(\theta^t), \theta^t - \theta^\star \rangle - \frac{\alpha}{2}\|\theta^t - \theta^\star\|^2 \\ &= \langle \mathbb{E}[\mathbf{v^t}|\theta^t], \theta^t - \theta^\star \rangle - \frac{\alpha}{2}\|\theta^t - \theta^\star\|^2 \\ &= \mathbb{E}\Big[\langle \mathbf{v^t}, \theta^t - \theta^\star \rangle - \frac{\alpha}{2}\|\theta^t - \theta^\star\|^2 \Big| \theta^t\Big] \\ &= \mathbb{E}\Big[\frac{\alpha(t-1)}{2}\|\theta^t - \theta^\star\|^2 - \frac{\alpha t}{2}\|\theta^{t+1} - \theta^\star\|^2 + \frac{1}{2\alpha t}\|\mathbf{v^t}\|^2 \Big| \theta^t\Big]. \end{aligned} \tag{18}$$

The next steps are now similar to the previous proof. We first apply Jensen's inequality to obtain that $F(\overline{\theta}) - F(\theta^\star) \leq \frac{1}{T}\sum_{t=1}^T (F(\theta^t) - F(\theta^\star))$. Summing the inequality (18) for $t = 1, \ldots, T-1$, we obtain that

$$\begin{aligned} \mathbb{E}[F(\overline{\theta}) - F(\theta^\star)] &\leq \frac{1}{T}\sum_{t=1}^T \mathbb{E}[F(\theta^t) - F(\theta^\star)] \\ &\leq \frac{1}{T}\sum_{t=1}^T \mathbb{E}[F(\theta^t) - F(\theta^\star)] \\ &\leq \frac{1}{T}\sum_{t=1}^T \mathbb{E}\Big[\mathbb{E}\Big[\frac{\alpha(t-1)}{2}\|\theta^t - \theta^\star\|^2 - \frac{\alpha t}{2}\|\theta^{t+1} - \theta^\star\|^2 + \frac{1}{2\alpha t}\|\mathbf{v^t}\|^2 \Big| \theta^t\Big]\Big] \\ &\leq \frac{1}{T}\mathbb{E}\Big[\sum_{t=1}^T \frac{\alpha(t-1)}{2}\|\theta^t - \theta^\star\|^2 - \frac{\alpha t}{2}\|\theta^{t+1} - \theta^\star\|^2 + \frac{1}{2\alpha t}\|\mathbf{v^t}\|^2\Big] \\ &\leq \mathbb{E}\Big[-\frac{\alpha T}{2T}\|\theta^T - \theta^\star\|^2 + \frac{1}{T}\sum_{t=1}^T \frac{1}{2\alpha t}\|\mathbf{v^t}\|^2\Big] \\ &\leq \frac{1}{T}\sum_{t=1}^T \frac{\rho^2}{2\alpha t} \leq \frac{\rho^2}{2\alpha T}(1 + \log(T)), \end{aligned}$$

concluding the proof. □

Note that one can actually choose $R = +\infty$ in the previous theorem, so that the same result holds without the projection step. However, it is most of the time delicate to ensure that the expected norm of the gradients $\mathbb{E}[\|\mathbf{v^t}\|^2]$ stay bounded without this projection step. For example, for a quadratic function of the form $F(\theta) = \frac{\alpha}{2}\|\theta\|^2$, the norm of the gradient will blow up if $\|\theta\|$ diverges. An alternative method to ensure that the iterates do not blow

up consists in adding a regularization term to the objective function, that is we minimize $F(\theta) + \lambda\|\theta\|^2$ for some $\lambda > 0$ instead. See Theorem 5.5 in [Bach, 2022] for details.

2 Application to risk minimization

We now review how those theorems translate in the setting of risk minimization. Let $F(\theta) = \mathcal{R}_P(f_\theta) = \mathbb{E}_P[\ell(f_\theta(\mathbf{x}), \mathbf{y})]$ be the P-risk of some predictor f_θ indexed by $\theta \in \mathbb{R}^k$, with minimizer $\theta^\star$. Assume that we have access to n i.i.d. samples $(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})$ from distribution P. We compare two methods:

- (SGD) Let $\hat{\theta}_{\mathrm{SGD}}$ be the output of stochastic gradient descent (with projection) with n steps using the gradient estimates $\nabla_\theta \ell(f_\theta(\mathbf{x_i}), \mathbf{y_i})$.
- (GD-ER) Let $\hat{\theta}_T$ be the output of gradient descent applied for T steps on the empirical risk

$$\theta \mapsto \mathcal{R}_n(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(\mathbf{x_i}), \mathbf{y_i}). \tag{19}$$

Note that for any of those predictors $\hat{\theta}$, it holds that the expected excess of risk can be decomposed into

$$\mathbb{E}[\mathcal{R}_P(f_{\hat{\theta}}) - \mathcal{R}_P^\star] = \underbrace{\mathbb{E}[\mathcal{R}_P(f_{\hat{\theta}}) - \mathcal{R}_P(f_{\theta^\star})]}_{\text{optimization error}} + \underbrace{\mathcal{R}_P(f_{\theta^\star}) - \mathcal{R}_P^\star}_{\text{approximation error}}. \tag{20}$$

The approximation error will depend only the "size" of the set of predictors $\mathcal{F} = \{f_\theta,\ \theta \in \mathbb{R}^k\}$. On the contrary, the optimization error will depend on our method to find the minimum of F. We consider two questions.

- (Q1) What is the minimal number n of samples required to get an optimization error smaller than ε using (SGD) or (GD-ER)?
- (Q2) What is the associated time complexity of the algorithm?

For sake of conciseness, we will only answer (Q1) and (Q2) in the "favorable" case where, for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the function $\theta \in \mathbb{R}^k \mapsto \ell(f_\theta(x), y)$ is α-strongly convex and β-smooth. We further assume that the minimizer

$\theta^\star$ belong to $B(0;R)$ with R of the form $c/\sqrt{\beta}$. One can try as an exercise to answer those questions by removing for instance the α-strongly convex assumption. We let $\kappa = \beta/\alpha \geq 1$ be the condition number. Also, we use the notation $\tilde{O}(\varepsilon^a)$ to denote a quantity of the form $\varepsilon^a(\log(\varepsilon)^{-1})^b$. This allows us to hide logarithmic factors that are almost constant in practice.

Let us first consider (SGD). In this setting, every gradient $\mathbf{v^t}$ is of the form $\nabla_\theta \ell_{\theta^t}(f_\theta(\mathbf{x_i}), \mathbf{y_i})$. By definition of β-smoothness, the norm of this gradient is smaller than

$$\beta\|\theta^t - \theta^\star\| \leq 2\beta R = \rho.$$

According to Theorem 5 with $n = T$, we need $n = \tilde{O}(\beta^2 R^2/(\alpha\varepsilon))$ samples to reach a precision ε. Evaluating the gradient $\nabla\ell(f_\theta(\mathbf{x_i}), \mathbf{y_i})$ requires $O(k)$ operations, so that the time complexity of SGD is $\tilde{O}(k\beta^2 R^2/(\alpha\varepsilon))$. Recalling that we assume that $R^2 \leq c^2/\beta$, we obtain a time complexity of order

$$\tilde{O}(k\kappa/\varepsilon).$$

The analysis of (GD-ER) is slightly more complicated. Let $\hat{\theta}_\infty$ be the minimizer of $\theta \mapsto \mathcal{R}_n(\theta)$ (that is the actual empirical risk minimizer). One can further bound the optimization error:

$$\begin{aligned}
\mathcal{R}_P(f_{\hat{\theta}_T}) - \mathcal{R}_P(f_{\theta^\star}) &\leq \mathcal{R}_P(f_{\hat{\theta}_T}) - \mathcal{R}_n(f_{\hat{\theta}_T}) \\
&\quad + \mathcal{R}_n(f_{\hat{\theta}_T}) - \mathcal{R}_n(f_{\hat{\theta}_\infty}) \\
&\quad + \mathcal{R}_n(f_{\hat{\theta}_\infty}) - \mathcal{R}_n(f_{\theta^\star}) \\
&\quad + \mathcal{R}_n(f_{\theta^\star}) - \mathcal{R}_P(f_{\theta^\star}) \\
&\leq 2 \cdot \sup_\theta |\mathcal{R}_n(f_\theta) - \mathcal{R}_P(f_\theta)| + \mathcal{R}_n(f_{\hat{\theta}_T}) - \mathcal{R}_n(f_{\hat{\theta}_\infty})
\end{aligned}$$

where at the last line we use that $\mathcal{R}_n(f_{\hat{\theta}_\infty}) \leq \mathcal{R}_n(f_{\theta^\star})$ by definition of the empirical risk minimizer. The first term in this last inequality can be shown to be at least of order $1/\sqrt{n}$ (this follows from $\mathcal{R}_n(f_\theta)$ being the average of n i.i.d. random variables). In particular, to reach an optimization error $\mathbb{E}[\mathcal{R}_P(f_{\hat{\theta}_T}) - \mathcal{R}_P(f_{\theta^\star})]$ of order ε, we need at least $n = O(\varepsilon^{-2})$ samples. After T steps of gradient descent, we have a control of the form

$$\begin{aligned}
\mathcal{R}_n(\theta_T) - \mathcal{R}_n(\theta_\infty) &\leq \exp(-T/\kappa)(\mathcal{R}_n(\theta_0) - \mathcal{R}_n(\theta_\infty)) \\
&\leq \exp(-T/\kappa)\frac{\beta}{2}\|\theta_0 - \theta_\infty\|^2 \\
&\leq \exp(-T/\kappa)2R^2\beta \leq \exp(-T/\kappa)2c^2,
\end{aligned} \tag{21}$$

where we also use the definition of β-smoothness and the fact that both $\|\theta_0\|$ and $\|\hat{\theta}_\infty\|$ are smaller than R. Therefore, to make this quantity smaller than ε, a number $T = \tilde{O}(\kappa)$ of steps of gradient descent are required. As computing a single gradient $\nabla_\theta \mathcal{R}_n(f_\theta)$ requires to compute n gradients $\nabla_\theta \ell(f_\theta(\mathbf{x_i}), \mathbf{y_i})$, the final complexity of gradient descent is in this situation

$$\tilde{O}(knT) = \tilde{O}(kn\kappa) = \tilde{O}(k/\varepsilon^2\kappa).$$

We summarize the different results in the following table. **Stochastic gradient descent requires less samples and a smaller time complexity to attain a given accuracy.**

Algo.	Num. of samples	Complexity
SGD	$n = \tilde{O}(\kappa/\varepsilon)$	$\tilde{O}(k\kappa/\varepsilon)$
GD	$n = O(1/\varepsilon^2)$	$\tilde{O}(k\kappa/\varepsilon^2)$

Table 1: Summary of the convergence rates in the α-strongly convex and β-smooth case.

References

[Bach, 2022] Bach, F. (2022). *Learning theory from first principles.*