Chapter 6

Stochastic approximation

This chapter is dedicated to stochastic approximation for large sums. Stochastic approximation has a long history starting with Robbins-Monro algorithm [53] with the ODE method [38] from Ljung and latter extensions, see [11] for a complete exposition and [16]. The idea of using stochastic approximation in large scale setting gained significance interest in the machine learning litterature see for example [17]. We provide example of non asymptotic convergence rate analyses for stochastic subgradient and stochastic proximal gradient for finite sums.

6.1 Motivation, large n

6.1.1 Lasso estimator

The Lasso estimator is given as follows:

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2n}\|\mathbb{X}\theta - Y\|^2 + \lambda\|\theta\|_1.$$

We have seen that the optimization problem has a favorable structure which allow to devise efficient algorithms. Another way to write the same optimization problem is to consider

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(x_i^T\theta - y_i)^2 + \lambda\|\theta\|_1,$$

which actually exhibits an additional sum structure. In this chapter we will be considering optimization problems of the form

$$\min_{x\in\mathbb{R}^d} \quad F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + g(x). \tag{6.1}$$

where f_i and g are convex lower semicontinuous convex functions.

6.1.2 Stochastic approximation

To solve problem (6.1), one may use first order methods such as the ones described in the previous chapter. Computing a subgradient in this case require to compute subgradient of f_i, $i = 1, \dots, n$ and average them. The computational cost is of the order of n subgradient computation and n vector operations. When n is very large, or even infinite, this could be prohibitive. Intuitively, if there is redundancy in the elements of the sum, one should be able to take advantage of it. For example, suppose that $f_i = f$, for all i, then blindly computing the gradient of F has a cost of the order $n \times d$, while only d operations (computing one gradient would suffice).

More generally, one could rewrite the objective function in (6.1) in the following form:

$$F\colon x \mapsto \mathbb{E}\left[f_I(x)\right] + g(x),$$

where I denotes a uniform random variable over $\{1, \dots, n\}$. Stochastic approximation, or stochastic optimization algorithm allow to handle such objectives. The main algorithmic step is as follows:

- For any $x \in \mathbb{R}^d$,
- Sample i uniformly at random in $\{1, \dots, n\}$.
- Perform an algorithmic step using only the value of $f_i(x)$ and $\nabla f_i(x)$ or eventually $v \in \partial f_i(x)$

The simple example can be extended to more general random variables I and under proper integrability and domination conditions, one can invert gradient (or subgradient) and expectation, assuming $g = 0$ for simplicity

- If for each value of I, f_I is continuously differentiable, we then have for any $x \in \mathbb{R}^d$,

$$\mathbb{E}\left[\nabla f_I(x)\right] = \nabla \mathbb{E}\left[f_I(x)\right] = \nabla F(x)$$

- Assume that f_I is convex for all realizations of I. Assume that we have access to a random variable v_I such that $v_I \in \partial f_I(x)$ almost surely, then the expectation is convex and

$$\mathbb{E}\left[v_I\right] \in \partial \mathbb{E}\left[f_I(x)\right] = \partial F(x).$$

Hence the process of using a single element of the sum in an algorithm can be seen as performing optimization based on noisy unbiased estimates of the gradient, or subgradient, of the objective. This intuition is described more formaly in the coming section.

6.2 Prototype stochastic approximation algorithm

This section describes Robbins-Monro algorithm for stochastic approximation. Consider a Lipschitz map $h\colon \mathbb{R}^p \mapsto \mathbb{R}^p$, the goal is to find a zero of h. The operator only has access to unbiased noisy estimates of h. The Robins-Monro algorithm is described as follows, $(X_k)_{k\in\mathbb{N}}$ is a sequence of random variables such that for any $k \in \mathbb{N}$

$$X_{k+1} = X_k + \alpha_k \left(h(X_k) + M_{k+1}\right) \tag{6.2}$$

where

- $(\alpha_k)_{k\in\mathbb{N}}$ is a sequence of positive step sizes satisfying

$$\sum_{i=1}^{n} \alpha_k = +\infty$$
$$\sum_{i=1}^{n} \alpha_k^2 < +\infty$$

- $(M_k)_{k\in\mathbb{N}}$ is a martingale difference sequence with respect to the increasing family of σ-fields

$$\mathcal{F}_k = \sigma(X_m, M_m, m \leq k) = \sigma(X_0, M_1, \dots, M_k).$$

  This means that $\mathbb{E}\left[M_{k+1}|\mathcal{F}_k\right] = 0$, for all $k \in \mathbb{N}$.

- In addition, we assume that there exists a positive constant C such that

$$\sup_{k\in\mathbb{N}} \mathbb{E}\left[\|M_{k+1}\|_2^2 | \mathcal{F}_k\right] \leq C.$$

The intuition here is that our hypotheses on the step size ensure that the quantity $\sum_{k=0}^{+\infty} \mathbb{E}\left[\alpha_k^2 \|M_{k+1}\|^2 | \mathcal{F}_k\right]$ is finite and hence the zero mean martingale $\sum_{k=0}^{K} \alpha_k M_{k+1}$ has square summable increments and converges to a square integrable random varible M in $\mathbb{R}^p$ both almost surely and in L^2 (see for example [29, Section 5.4]).

The long term behaviour of such recursions is at the heart of the field of stochastic approximation. The fact that the step sizes tends to 0 and that the sum of perturbation stabilizes suggests that in the limit one obtains trajectories of a continuous time differential equation. This is formalized in the next section.

6.3 The ODE approach

For optimization we may choose $h = -\nabla F(x)$ assuming that F has Lispchitz gradient. We consider Robbins-Monro algorithm in this setting. This idea dates back to Ljung [38], see also [11] for an advanced presentation. An accessible exposition of the following result is found in [16],

Theorem 6.3.1. *Conditioning on boundedness of $\{X_k\}_{k\in\mathbb{N}}$, almost surely, the (random) set of accumulation point of the sequence is compact connected and invariant by the flow generated by the continuous time limit:*

$$\dot{x} = h(x).$$

This theorems means that for any $\bar{x}$ accumulation point of the algorithm, the unique solution $x\colon t \mapsto \mathbb{R}^p$ of the continuous time ODE satisfying $x(0) = \bar{x}$ remains bounded for all $t \in \mathbb{R}$. This allows to conclude in the convex case.

Corollary 6.3.1. *If F is convex, differentiable and attains its minimum, setting $h = -\nabla F$, conditioning on the event that $\sup_{k\in\mathbb{N}} \|X_k\|$ is finite, almost surely, all the accumulation points of X_k are critical points of F.*

Proof. Fix $\bar{x} \in \mathbb{R}^p$ such that $\nabla F(\bar{x}) \neq 0$, this means that $F(\bar{x}) - F^* > 0$. Consider the solution to

$$\dot{x} = \nabla F(x),$$

starting at $\bar{x}$, we have

$$\begin{aligned} \frac{\partial}{\partial t} F(x(t)) &= \|\nabla F(x(t))\|_2^2 \geq 0 \\ \frac{\partial}{\partial t} \|x(t) - x^*\|_2^2 &= \langle \nabla F(x(t)), x(t) - x^* \rangle \geq F(x(t)) - F^* \geq F(\bar{x}) - F^* > 0. \end{aligned}$$

We deduce that F is increasing along the trajectory and diverges, hence the solution escapes any compact set which means that $\bar{x}$ does not belong to a compact invariant set. □

The power of the ODE approach lies in the fact that it allows to treat much more complicated situations beyond convexity and differentiability.

6.4 Rates for convex optimization

In the context of convex optimization problems of the form described in the introduction of this chapter, one can obtain precise covergence rate estimates using elementary arguments.

6.4.1 Stochastic subgradient descent

Proposition 6.4.1. *Consider the problem*

$$\min_{x\in\mathbb{R}^d} F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x),$$

where each f_i is convex and L-Lipschitz. Choose $x_0 \in \mathbb{R}$ and a sequence of random variables $(i_k)_{k\in\mathbb{N}}$ independently identically distributed uniformly on $\{1,\dots,n\}$ and a sequence of positive step sizes $(\alpha_k)_{k\in\mathbb{N}}$. Consider the recursion

$$x_{k+1} = x_k - \alpha_k v_k \tag{6.3}$$
$$v_k \in \partial f_{i_k}(x_k) \tag{6.4}$$

Then for all $K \in \mathbb{N}$, $K \geq 1$

$$\mathbb{E}\left[F(\bar{x}_K) - F^*\right] \leq \frac{L\|x_0 - x^*\|_2^2 + \frac{2G^2}{L}\sum_{k=0}^{K}\alpha_k^2}{2\sum_{k=0}^{K}\alpha_k}$$

where $\bar{x}_K = \frac{\sum_{k=0}^{K}\alpha_k x_k}{\sum_{k=0}^{K}\alpha_k}$.

Proof. We fix $k \in \mathbb{N}$ and condition on $i_1,\dots,i_k$ so that x_k and x_{k+1} are fixed. We have for any $k \in \mathbb{N}$

$$\begin{aligned}\frac{1}{2}\|x_{k+1} - x^*\|_2^2 &= \frac{1}{2}\|x_k - \alpha_k v_k - x^*\|_2^2 \\ &= \frac{1}{2}\|x_k - x^*\|_2^2 + \alpha_k v_k^T(x^* - x_k) + \frac{\alpha_k^2}{2}\|v_k\|_2^2 \\ &\leq \frac{1}{2}\|x_k - x^*\|_2^2 + \alpha_k(f_{i_k}(x^*) - f_{i_k}(x_k)) + \frac{\alpha_k^2}{2}L^2.\end{aligned}$$

Conditioning on x_k and taking expectation with respect to i_k,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{2}\|x_{k+1} - x^*\|_2^2 | x_k\right] &\leq \mathbb{E}\left[\frac{1}{2}\|x_k - x^*\|_2^2 | x_k\right] + \frac{\alpha_k^2 L^2}{2} + \mathbb{E}\left[\alpha_k(f_{i_k}(x^*) - f_{i_k}(x_k)) | x_k\right] \\ &= \frac{1}{2}\|x_k - x^*\|_2^2 + \frac{\alpha_k^2 L^2}{2} + \alpha_k(F(x^*) - F(x_k)).\end{aligned}$$

Taking expectation with respect to x_k, using tower property of conditional expectation, we have

$$\mathbb{E}\left[\frac{1}{2}\|x_{k+1} - x^*\|_2^2\right] \leq \mathbb{E}\left[\frac{1}{2}\|x_k - x^*\|_2^2\right] + \frac{\alpha_k^2 L^2}{2} + \alpha_k\mathbb{E}\left[(F(x^*) - F(x_k))\right].$$

By summing up, we obtain, for all $K \in \mathbb{N}$, $K \geq 1$

$$\frac{\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[F(x_k) - F^*\right]}{\sum_{i=0}^{k}\alpha_i} \leq \frac{\|x_0 - x_*\|^2 + L^2\sum_{k=0}^{K}\alpha_k^2}{2\sum_{k=0}^{K}\alpha_i}$$

and the result follows from convexity of f. □

Corollary 6.4.1. *Under the hypotheses of Proposition 6.4.1, we have the following*

- *If $\alpha_k = \alpha$ is constant, we have*

$$\mathbb{E}\left[F(\bar{x}_k) - F^*\right] \leq \frac{\|x_0 - x^*\|^2}{2(k+1)\alpha} + \frac{L^2\alpha}{2}.$$

- *In particular, choosing* $\alpha_i = \frac{\|x_0 - x^*\|/L}{\sqrt{k+1}}$*, we have*

$$\mathbb{E}\left[F(\bar{x}_k) - F^*\right] \leq \frac{\|x_0 - x^*\|L}{\sqrt{k+1}}.$$

- *Choosing* $\alpha_k = \|x_0 - x^*\|/(L\sqrt{k})$ *for all* k*, we obtain for all* k

$$\mathbb{E}\left[F(\bar{x}_k) - F^*\right] = O\left(\frac{\|x_0 - x^*\|_2 L(1+\log(k))}{\sqrt{k}}\right).$$

6.4.2 Stochastic proximal gradient descent

This method is sometimes called FOBOS in the litterature. I could not find a reference for the following result.

Proposition 6.4.2. *Consider the problem*

$$\min_{x\in\mathbb{R}^d} F(x) \;:=\; \frac{1}{n}\sum_{i=1}^{n} f_i(x) + g(x)$$

where each f_i *is convex with* L*-Lipschitz gradient and* g *is convex. Choose* $x_0 \in \mathbb{R}$ *and a sequence of random variables* $(i_k)_{k\in\mathbb{N}}$ *independently identically distributed uniformly on* $\{1,\dots,n\}$ *and a sequence of positive step sizes* $(\alpha_k)_{k\in\mathbb{N}}$*. Consider the recursion*

$$x_{k+1} = \operatorname{prox}_{\alpha_k g/L}\left(x_k - \alpha_k/L\nabla f_{i_k}(x_k)\right). \tag{6.5}$$

Assume the following

- $0 < \alpha_k \leq 1$*, for all* $k \in \mathbb{N}$*.*
- f_i *and* g *are* G*-Lipschitz for all* $i = 1,\dots,n$*;*

Then for all $K \in \mathbb{N}$*,* $K \geq 1$

$$\mathbb{E}\left[F(\bar{x}_K) - F^*\right] \leq \frac{L\|x_0 - x^*\|_2^2 + \frac{2G^2}{L}\sum_{k=0}^{K}\alpha_k^2}{2\sum_{k=0}^{K}\alpha_k}$$

where $\bar{x}_K = \frac{\sum_{k=0}^{K}\alpha_k x_k}{\sum_{k=0}^{K}\alpha_k}$*.*

Proof. We fix $k \in \mathbb{N}$ and condition on $i_1,\dots,i_k$ so that x_k and x_{k+1} are deterministic. Note that the prox iteration gives

$$\frac{\alpha_k}{L}\partial g(x_{k+1}) + x_{k+1} = x_k - \frac{\alpha_k}{L}\nabla f_{i_k}(x_k)$$
$$\|x_{k+1} - x_k\|_2 \leq 2G\frac{\alpha_k}{L}$$

Fix $k \in \mathbb{N}$, applying Lemma 5.4.1 with $x = x_k$, $z = x^*$ and $y = x_{k+1}$, using the fact that f_{i_k} has L/α_k Lipschitz gradient,

$$\begin{aligned}
&f_{i_k}(x^*) + g(x^*) + \frac{L}{2\alpha_k}\|x^* - x_k\|_2^2 - \frac{L}{2\alpha_k}\|x_{k+1} - x^*\|_2^2 \\
&\geq f_{i_k}(x_{k+1}) + g(x_{k+1}) \\
&\geq f_{i_k}(x_k) + g(x_k) - 2G\|x_{k+1} - x_k\|_2 \\
&\geq f_{i_k}(x_k) + g(x_k) - 4G^2\frac{\alpha_k}{L}
\end{aligned}$$

And

$$\frac{\alpha_k}{L}(f_{i_k}(x_k) + g(x_k) - F^*) \leq \frac{1}{2}\|x^* - x_k\|_2^2 - \frac{1}{2}\|x_{k+1} - x^*\|_2^2 + 4G^2\frac{\alpha_k^2}{L^2}$$

We have, considering tower expectation, with respect to i_k first and the remaining randomness in a second step

$$\begin{aligned}
\mathbb{E}\left[\frac{\alpha_k}{L}(f_{i_k}(x_k) + g(x_k) - F^*)\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\alpha_k}{L}(f_{i_k}(x_k) + g(x_k) - F^*)|x_k\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\alpha_k}{L}(F(x_k) - F^*)|x_k\right]\right] \\
&= \mathbb{E}\left[\frac{\alpha_k}{L}(F(x_k) - F^*)\right] \\
&\leq \mathbb{E}\left[\frac{1}{2}\|x^* - x_k\|_2^2\right] - \mathbb{E}\left[\frac{1}{2}\|x_{k+1} - x^*\|_2^2\right] + 4G^2\frac{\alpha_k^2}{L^2}
\end{aligned}$$

By summing, we obtain, for any $K \in \mathbb{N}$

$$\frac{\mathbb{E}\left[\sum_{k=0}^K \alpha_k(F(x_k) - F^*)\right]}{\sum_{k=0}^K \alpha_k} \leq \frac{L\|x_0 - x^*\|_2^2 + \frac{2G^2}{L}\sum_{k=0}^K \alpha_k^2}{2\sum_{k=0}^K \alpha_k}$$

and the result follows by Jensen's inequality. □

Corollary 6.4.2. *Under the hypotheses of Proposition 7.4.1.*

- *If $\alpha_k = \alpha$ is constant, we have for all $k \geq 1$*

$$F(\bar{x}_k) - F^* \leq \frac{L\|x_0 - x^*\|^2}{2(k+1)\alpha} + \frac{G^2\alpha}{L}.$$

- *In particular, choosing $\alpha_i = \frac{1}{\sqrt{2k+2}}$, for $i = 1 \ldots, k$, for some $k \in \mathbb{N}$, we have*

$$F(\bar{x}_k) - F^* \leq \frac{L\|x_0 - x^*\|_2^2 + \frac{G^2}{L}}{\sqrt{2k+2}}.$$

- *Choosing $\alpha_k = 1/\sqrt{2k+2}$ for all k, we obtain for all k*

$$F(x_k) - F^* = O\left(\frac{L\|x_0 - x^*\|_2^2 + \frac{G^2}{L}\log(k)}{\sqrt{2k+2}}\right).$$

6.5 Minimizing the population risk

The methods which we have seen can be used to minimize functions of the form

$$x \mapsto \mathbb{E}_Z\left[f(x, Z)\right]$$

where x denotes some model parameters and Z denotes a random variable describing our population. In this case, Z could the input output pair (X, Y) of a regression problem, for which we try to minimize the exepected prediction error over a certain parametric regression function class $\mathcal{F}$.

$$R(f) = \mathbb{E}\left[(f(X) - Y)^2\right] = \int_{\mathcal{X}\times\mathcal{Y}} (f(x) - y))^2 P(dx, dy).$$

This can be done by replacing the finite sum by an expectation and sampling of independant indices by iis samples of the random variable Z. The results are exactly the same.

Such a procedure are usually called "single pass" procedure: given a dataset $(x_i, y_i)_{i=1}^n$ for a regression problem, performing one pass of a stochastic algorithm, looking at each data point only once amount to perform n step of the same stochastic algorithm on the population risk.

This illustrates a strong relation between stochastic optimization and statistics. We have seen that in the linear regression setting, there is no hope to obtain estimators with statistical rates much faster than $1/n$ in terms of mean squared error. Similarly, the rates which we obtained for stochastic algorithms are of the order of $1/\sqrt{k}$. This is also optimal in a precise sense. These algorithms provide estimator with statistical efficiency of the order of $1/\sqrt{n}$.

The gap stands because we considered regression problems with squared loss, a very special structure, while here the convex functions which we considered are arbitrary. For strongly convex functions, stochastic optimization algorithms may show faster convergence rate of the order $1/k$.

Chapter 7

Block coordinate methods

Block decomposition methods appeared as alternatives to solve optimization problems involving large number of dimensions. The idea is to reduce the complexity of a single iteration by updating only a few coordinates at a time. The use of such methods was advocated by Nesterov [44], extensions such as [50] appeared in the continuity of these works. The survey [64] is a good entry point to the litterature.

7.1 Motivation, large d

The Lasso estimator is given as follows:

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2n}\|\mathbb{X}\theta - Y\|^2 + \lambda\|\theta\|_1.$$

We have seen that the optimization problem has a favorable structure which allow to devise efficient algorithms. The cost of each iteration is depends on the dimension (here d^2) which for some problems may be limiting. A possible alternative is to update coordinates independantly, reducing the cost of each iteration.

In general, this approach is not convergent for nonsmooth functions (can you see why?), however, the Lasso problem, despite being nonsmooth, fits coordinate descent methods because the nonsmooth part is separable. We shall see two variations of such algorithms, deterministic and random, with convergence rate estimates in both cases. A good introduction to the topic cand be found in [64] and a pioneering work in optimization is described in [44]. The litterature on the subject has completely exploded in the past years.

7.2 Description of the algorithm

We consider optimization problems of the form

$$\min_{x\in\mathbb{R}^p} F(x) = f(x) + \sum_{i=1}^{p} g_i(x_i),$$

where $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ has L-Lipschitz gradient and $g_i\colon \mathbb{R} \mapsto \mathbb{R}$ are convex lower semicontinuous univariate functions. We denote by $e_1, \dots, e_p$ the elements of the canonical basis. Block coordinate descent algorithms are given a sequence of coordinate indices $(i_k)_{k\in\mathbb{N}}$, and, starting at $x_0 \in \mathbb{R}^p$ updates coordinates one by one at each iteration. For example

$$\begin{aligned} x_{k+1} &= \arg\min_{y=x_k+te_{i_k}} f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{L}{2}\|y - x_k\|_2^2 + g_{i_k}(y) \\ x_{k+1} &= \arg\min_{y=x_k+te_{i_k}} f(y) + g_{i_k}(y). \end{aligned}$$

The first option corresponds to a block proximal gradient algorithm, the second option corresponds to exact block minimization. Block coordinate descent algorithms are usualy analysed under coercvity assumptions:

Assumption 7.2.1. *The sublevelset $\{y \in \mathbb{R}^p,\ F(y) \leq F(x_0)\}$ is compact, for any $y \in \mathbb{R}^p$ such that $F(y) \leq F(x_0)$, $\|y - x^*\|_2 \leq R$.*

7.3 Convergence rate analysis using random blocks

7.3.1 Smooth setting

The following technical Lemma is classical.

Lemma 7.3.1. *Let $(A_k)_{k\in\mathbb{N}}$ be a sequence of positive real numbers and $\gamma > 0$ be such that*

$$A_k - A_{k+1} \geq \gamma A_k^2$$

then for all $k \in \mathbb{N}$, $k \geq 1$, $A_k \leq (\gamma k)^{-1}$.

Proof. We have for all $k \in \mathbb{N}$,

$$\frac{1}{A_{k+1}} - \frac{1}{A_k} = \frac{A_k - A_{k+1}}{A_k A_{k+1}} \geq \gamma \frac{A_k^2}{A_{k+1}A_k} = \gamma \frac{A_k}{A_{k+1}} \geq \gamma.$$

Hence for all $k \in \mathbb{N}$,

$$\frac{1}{A_k} \geq \frac{1}{A_0} + \gamma k \geq k\gamma.$$

□

Proposition 7.3.1. *Consider the problem*

$$\min_{x\in\mathbb{R}^p} f(x)$$

where $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ is convex differentiable with L-Lipschitz gradient. Choose $x_0 \in \mathbb{R}$ and a sequence of random variables $(i_k)_{k\in\mathbb{N}}$ independently identically distributed uniformly on $\{1, \dots, p\}$ and a sequence of positive step sizes. Consider the recursion

$$x_{k+1} = x_k - \frac{1}{L}\nabla_{i_k} f(x_k) \tag{7.1}$$

Then for all $k \in \mathbb{N}$, $k \geq 1$

$$\mathbb{E}\left[f(x_k) - f^*\right] \leq \frac{2pLR^2}{k}.$$

Proof. Fix, $k \in \mathbb{N}$, and condition on x_k and $i_0, \dots, i_k$ so that x_{k+1} is deterministic. We remark that $t \mapsto f(x_k + te_{i_k})$ is convex with L-Lipschitz gradient. Applying Lemma 7.3.1 with $x = z = x_k$, $y = x_{k+1}$,

$$f(x_k) \geq f(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 = f(x_{k+1}) + \frac{1}{2L}\|\nabla_i f(x_k)\|_2^2,$$

and in particular f is decreasing along the sequence. Taking expectation with respect to i_k, we obtain

$$\mathbb{E}\left[f(x_{k+1})|x_k\right] \leq f(x_k) - \frac{1}{2pL}\|\nabla f(x_k)\|_2^2 \tag{7.2}$$

From convexity, we have $f^* \geq f(x) - \|\nabla f(x)\|\|x - x^*\|$ and using the fact that f is decreasing along the sequence, $\|x_k - x^*\|$ remains bounded. We have

$$\|\nabla f(x)\|_2^2 \geq \frac{(f(x_k) - f^*)^2}{R^2}$$

and

$$\mathbb{E}\left[f(x_{k+1})|x_k\right] - f^* \leq f(x_k) - f^* - \frac{1}{2pL}\|\nabla f(x_k)\|_2^2$$
$$f(x_k) - f^* - \frac{(f(x_k) - f^*)^2}{2pLR^2}$$

Taking expectation with respect to x_k and using the fact that $\mathbb{E}\left[Z^2\right] \geq \mathbb{E}\left[Z\right]^2$, we obtain

$$\mathbb{E}\left[f(x_{k+1}) - f^*\right] \leq \mathbb{E}\left[f(x_k) - f^*\right] - \frac{\mathbb{E}\left[f(x_k) - f^*\right]^2}{2pLR^2}.$$

Applying Lemma 7.3.1, we obtain for all $k \in \mathbb{N}$, $k \geq 1$,

$$\mathbb{E}\left[f(x_k) - f^*\right] \leq \frac{2pLR^2}{k}.$$

□

7.3.2 Extension to the nonsmooth setting

The following is a simplification of the arguments given in [50].

Proposition 7.3.2. *Consider the problem*

$$\min_{x \in \mathbb{R}^d} F(x) \;:=\; f(x) + \sum_{i=1}^{p} g_i(x)$$

where $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ is convex differentiable with L-Lipschitz gradient, each $g_i : \mathbb{R}^p \mapsto \mathbb{R}$ is convex and lower semicontinuous and only depends on coordinate i. Choose $x_0 \in \mathbb{R}$ and a sequence of random variables $(i_k)_{k\in\mathbb{N}}$ independently identically distributed uniformly on $\{1, \ldots, p\}$ and a sequence of positive step sizes. Consider the recursion

$$x_{k+1} = \arg\min_y f(x_k) + \langle \nabla_{i_k} f(x_k), y - x_k\rangle + \frac{L}{2}\|y - x_k\|_2^2 + g_{i_k}(y) \tag{7.3}$$
$$= \operatorname{prox}_{g_{i_k}/L}\left(x_k - \frac{1}{L}\nabla_{i_k} f(x_k)\right). \tag{7.4}$$

Set $C = \max\left\{R^2, F(x_0) - F^*\right\}$, where R is given in Assumption 7.2.1, we have, for all $k \geq 1$,

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \frac{2pC}{k}.$$

Proof. Fix, $k \in \mathbb{N}$, and condition on x_k and $i_0, \ldots, i_k$ so that x_{k+1} is deterministic. We remark that $t \mapsto f(x_k + te_{i_k})$ is convex with L-Lipschitz gradient. Noting that the iteration actually solves a univariate problem, applying Lemma 7.3.1 with $x = z = x_k$, $y = x_{k+1}$,

$$f(x_k) + g_{i_k}(x_k) \geq f(x_{k+1}) + g_{i_k}(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2.$$

By assumption, g_i depends only on coordinate i so that, $g_i(x_k) = g_i(x_{k+1})$ for $i \neq i_k$.

$$F(x_k) \geq F(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2,$$

So that F is non increasing along the sequence and for any $k \in \mathbb{N}$, $\|x_k - x^*\|_2^2 \leq C$, almost surely.

We write $g = \sum_{i=1}^p g_i$. From the definition of the proximity operator, we have

$$f(x_{k+1}) + g_{i_k}(x_{k+1}) \leq f(x_k) + \langle \nabla_{i_k} f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + g_{i_k}(x_{k+1})$$
$$F(x_{k+1}) \leq f(x_k) + \langle \nabla_{i_k} f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + g_{i_k}(x_{k+1}) + \sum_{i \neq i_k} g_i(x_k).$$

Since each g_i only depends on coordinate i, prox_g can be computed coordinate by coordinate. Taking expectation with respect to i_k and setting $z_k = \operatorname{prox}_{g/L}\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$, we obtain

$$\mathbb{E}\left[F(x_{k+1})|x_k\right] \leq \frac{1}{p}\left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_k)\right) + \frac{p-1}{p}F(x_k). \tag{7.5}$$

By definition of the proximity operator, we have for any $y \in \mathbb{R}^p$,

$$\begin{aligned} &f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_k) \\ &\leq f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|_2^2 + g(y) \\ &\leq F(y) + \frac{L}{2}\|y - x_k\|_2^2 \end{aligned}$$

In particular, for any $\alpha \in [0, 1]$,

$$\begin{aligned} &f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_k) \\ &\leq F(\alpha x^* + (1-\alpha)x_k) + \frac{\alpha^2 L}{2}\|x^* - x_k\|_2^2 \\ &\leq \alpha F(x^*) + (1-\alpha)F(x_k) + \frac{\alpha^2 L}{2}C \end{aligned}$$

The minimum is attained for $\alpha = (F(x_k) - F^*)/C \leq 1$ so that

$$\begin{aligned} &f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_k) - F^* \\ &\leq \left(1 - \frac{F(x_k) - F^*}{C}\right)(F(x_k) - F^*) \end{aligned}$$

Pluging this in (7.5), we obtain

$$\begin{aligned} \mathbb{E}\left[F(x_{k+1})|x_k\right] - F^* &\leq \frac{F(x_k) - F^*}{p}\left(1 - \frac{F(x_k) - F^*}{2C}\right) \\ &\quad + \frac{p-1}{p}(F(x_k) - F^*) \\ &= (F(x_k) - F^*)\left(1 - \frac{F(x_k) - F^*}{2pC}\right) \end{aligned} \tag{7.6}$$

Taking expectation with respect to x_k and using the fact that $\mathbb{E}[Z^2] \geq \mathbb{E}[Z]^2$, we have

$$\mathbb{E}\left[F(x_{k+1}) - F^*\right] \leq \mathbb{E}[F(x_k) - F^*] - \frac{1}{2pC}\mathbb{E}[F(x_k) - F^*]^2. \tag{7.7}$$

The result follows from Lemma 7.3.1. □

7.4 Convergence rates using deterministic blocks

Deterministic block selection may lead to similar theoretical guaranties, there is some computational overhead, but as we should see, this is affordable for Lasso instances. A broader discussion on this aspect is found in [48].

Proposition 7.4.1. *Consider the problem*

$$\min_{x\in\mathbb{R}^d} f(x)$$

where $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ is convex differentiable with L-Lipschitz gradient. Choose $x_0 \in \mathbb{R}$, and consider the recursion

$$x_{k+1} = x_k - \frac{1}{L}\nabla_{i_k} f(x_k) \tag{7.8}$$

where i_k is the largest block of $\nabla f(x_k)$ in Euclidean norm. Then for all $k \in \mathbb{N}$, $k \geq 1$

$$f(x_k) - f^* \leq \frac{2pLR^2}{k}.$$

Proof. The proof is essentially the same as in Proposition 7.3.1, we have for any $k \in \mathbb{N}$,

$$\|\nabla f(x_k)\|_2^2 \leq p\|\nabla_i f(x_k)\|_2^2$$

and one obtain using the same arguments as in (7.2)

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2pL}\|\nabla f(x_k)\|_2^2 \tag{7.9}$$

and the rest of the analysis is the same. □

Using similar ideas, one obtains the same behaviour for deterministic block proximal gradient algorithm.

Proposition 7.4.2. *Consider the problem*

$$\min_{x\in\mathbb{R}^d} F(x) \;:=\; f(x) + \sum_{i=1}^{p} g_i(x)$$

where $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ is convex differentiable with L-Lipschitz gradient, each $g_i : \mathbb{R}^p \mapsto \mathbb{R}$ is convex and lower semicontinuous and only depends on coordinate i. Choose $x_0 \in \mathbb{R}$ and consider the recursion

$$x_{k+1} = \arg\min_y f(x_k) + \langle \nabla_{i_k} f(x_k), y - x_k\rangle + \frac{L}{2}\|y - x_k\|_2^2 + g_{i_k}(y) \tag{7.10}$$

$$= \operatorname{prox}_{g_{i_k}/L}\left(x_k - \frac{1}{L}\nabla_i f(x_k)\right). \tag{7.11}$$

where i_k is given by

$$\arg\min_i \left\{ \langle y - x_k, \nabla_i f(x_k)\rangle + \frac{L}{2}\|y - x_k\|_2^2 + g(y) - g_i(x_k),\; y = \operatorname{prox}_{g_i/L}\left(x_k - \frac{1}{L}\nabla_i f(x_k)\right)\right\}$$

Set $C = \max\left\{R^2, F(x_0) - F^*\right\}$, where R is given in Assumption 7.2.1, we have, for all $k \geq 1$,

$$F(x_k) - F^* \leq \frac{2pC}{k}.$$

Proof. Using the same arguments as in the proof of Proposition 7.3.2, we obtain for any $k \in \mathbb{N}$,

$$\begin{aligned} F(x_{k+1}) &\leq f(x_k) + \langle \nabla_{i_k} f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + g_{i_k}(x_{k+1}) + \sum_{i \neq i_k} g_i(x_k) \\ &= F(x_k) + \langle \nabla_{i_k} f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + g_{i_k}(x_{k+1}) - g_{i_k}(x_k). \end{aligned}$$

Since each g_i only depends on coordinate i, prox_g can be computed coordinate by coordinate. Setting $z_k = \operatorname{prox}_{g/L}\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$, and $g = \sum_i g_i$, we deduce from the definition of i_k,

$$\begin{aligned} F(x_{k+1}) &\leq F(x_k) + \frac{1}{p}\left(\langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_{k+1}) - g(x_k)\right) \\ &= F(x_k) + \frac{1}{p}\left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_k) - g(x_k) - f(x_k)\right) \\ &= \frac{p-1}{p}F(x_k) + \frac{1}{p}\left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_k)\right) \end{aligned}$$

This is similar to (7.5) and the result follows from the same arguments. □

7.5 Comments on complexity for quadratic problems

The Lasso problem is a special case for block descent methods since the objective is quadratic. This leads to the following remark

- Computing the gradient of the Lasso problem costs a matrix vector product which complexity is of the order of d^2.
- Given $\theta \in \mathbb{R}^d$ and $\beta = \mathbb{X}^T(\mathbb{X}\theta - Y)$, choosing $\tilde{\theta}$ differing from θ in at most one coordinate, computing $\mathbb{X}^T(\mathbb{X}\tilde{\theta} - Y)$ given β costs only of the order of d operations by only considering the corresponding column of $\mathbb{X}^T\mathbb{X}$.

As a consequence the cost of performing one iteration of full proximal gradient for the Lasso problem is roughly equivalent to the cost of performing d iterations of random block proximal gradient.

Given the value of the gradient, the added complexity of computing the deterministic block is of the order of d as it requires only one path through the coordinates of the gradient and the current estimate θ. Hence the deterministic rule has similar complexity per iteratation as the random block rules.