ECE586 MDPs and Reinforcement Learning **University of Illinois at Urbana-Champaign**
Spring 2019 **Acknowledgment: R. Srikant's Notes and Discussions**

Lecture 6: Stochastic Gradient Descent-Part II

Instructor: Dimitrios Katselis *Scribe: Kunhao Li, Zhikai Guo*

6.1 Introduction

Recall the following stochastic optimization problem

$$\min_x F(x) = \mathbb{E}_\xi[f(x,\xi)], \tag{6.1}$$

where $\xi \in \mathbb{R}^p$ is a random vector such that $\xi \sim P$ for some distribution P and $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ is a measurable function. Stochastic Gradient Descent (SGD) can be used to solve this problem by assuming that an i.i.d. sequence $\xi_0, \xi_1, \ldots$ is observed or that for any x a noisy gradient $\hat{\nabla} F(x)$ is available by querying an oracle. In the context of machine learning, P is unknown and a finite data record $\{\xi_0, \xi_1, \ldots, \xi_{N-1}\}$ is assumed available. In this case, (6.1) is replaced by the empirical risk minimization problem[1]:

$$\min_x F(x) = \frac{1}{N}\sum_{i=0}^{N-1} f(x;\xi_i). \tag{6.2}$$

Here, $F(x)$ corresponds to an expected value with respect to the empirical measure $N^{-1}\sum_{i=0}^{N-1}\delta_{\xi_i}$. In the context of machine learning, $\nabla F(x) = N^{-1}\sum_{i=0}^{N-1}\nabla_x f(x;\xi_i)$ is called *full* or *batch* gradient or simply gradient. If N is very large, the gradient $\nabla F(x) = N^{-1}\sum_{i=0}^{N-1}\nabla_x f(x;\xi_i)$ is difficult to compute. Again, this motivates SGD as we described in the previous lecture by sampling at every time step from the fixed training set with replacement.

Relying on SGD, we assume that at (every) step k we have evaluations $g(x_k,\xi_k)$ of $g(x,\xi)$, which is an unbiased estimator of the gradient $\nabla F(x_k)$ given $x_0, \xi_0, \ldots, \xi_{k-1}$ and therefore a surrogate for $\nabla F(x_k)$. We then perform steps of gradient descent with $g(x_k,\xi_k)$ replacing $\nabla F(x_k)$:

Intput: Initial vector $x_0 \in \mathbb{R}^n$ and learning rates ε_k
while *termination condition is not met* **do**
  Update $x_{k+1} = x_k - \varepsilon_k g(x_k,\xi_k)$
end

Algorithm 1: SGD

Some comments:

1. In the context of stochastic programming, often the expectation in (6.1) cannot be computed explicitly, particularly when f does not have a closed form. This motivates the use of sequences of observations $\xi_0, \xi_1, \ldots$ to solve this problem.
2. An important conclusion is that the SGD in solving the stochastic program (6.1) passes over the data only once, while in the machine learning context, the SGD in solving (6.2) relies on sampling with replacement from the

[1]The same symbol $F(x)$ is deliberately used here for unification purposes.

training set. Therefore, the two procedures aim to minimize different objectives. The first procedure is called *stochastic approximation* (SA) and goes after the expected risk, while the second procedure is called *sample average approximation* (SAA) and goes after the empirical risk. This conclusion motivates the question of how far the two objectives and their corresponding solutions are from each other. Some answers are provided in some of the subsequent comments.

3. From a statistical perspective, if one allows N to increase, then $\{N^{-1}\sum_{i=0}^{N-1} f(x;\xi_i)\}$ corresponds to a sequence of random approximations of $F(x)$ in (6.1). Moreover, for any N, $N^{-1}\sum_{i=0}^{N-1} f(x;\xi_i)$ corresponds to a *(standard) Monte Carlo* estimator of $F(x)$ in (6.1) when $\{\xi_0, \xi_1, \ldots, \xi_{N-1}\}$ is an i.i.d. sample. Such estimators are unbiased estimators of the objective function $F(x)$ in (6.1).

4. Denoting by $\hat{x}_N$ and $\hat{F}_N$ an optimal solution and the optimal value of problem (6.2), $\hat{x}_N$ and $\hat{F}_N$ provide approximations to an optimal solution x^* and the optimal value F^* of problem (6.1). When x^* is the unique optimal solution of (6.1), then $\hat{x}_N \to x^*$ and $\hat{F}_N \to F^*$ under fairly general conditions. More explicitly, under proper conditions $P(|F(\hat{x}_N) - F(x^*)| \leq \epsilon)$ for $F(x)$ in (6.1) and $P(\|\hat{x}_N - x^*\| \leq \epsilon)$ converge exponentially fast to one in the sample size N for any given $\epsilon > 0$. Under further conditions, more can be established, namely that $P(\hat{x}_N = x^*)$ converges to one exponentially fast in the sample size N.

5. The sequence of optimal values $\{\hat{F}_N\}$ satisfies a *Central Limit Theorem* (CLT): $\sqrt{N}(\hat{F}_N - F^*) \xrightarrow{d} \mathcal{N}(0, \sigma_*^2)$, where $\sigma_*^2 = \mathrm{Var}(f(x^*, \xi))$. Here, $\xrightarrow{d}$ denotes convergence in distribution. An immediate conclusion is that $\hat{F}_N = F^* + O_P\left(1/\sqrt{N}\right)$, which gives the corresponding rate of convergence[2]. This rate of convergence is not surprising since it holds for pointwise estimators: $\sqrt{N}\left[N^{-1}\sum_{i=0}^{N-1} \nabla_x f(x;\xi_i) - F(x)\right] / \sqrt{\mathrm{Var}(f(x,\xi))} \xrightarrow{d} \mathcal{N}(0,1)$ for any fixed x by the CLT, which leads to the error $N^{-1}\sum_{i=0}^{N-1} \nabla_x f(x;\xi_i) - F(x)$ converging to zero at the same rate as before. Here, again $F(x)$ is the objective in (6.1).

6. Often in practice, N has to be very large in order to obtain a reasonable approximation in the described setup. This is crucial, especially when the evaluation of $f(x,\xi)$ for a given ξ is computationally expensive. This motivates *variance reduction techniques*, which lead to estimators with smaller variance than the ones obtained with standard sampling. Therefore, the same error can be obtained with less computational effort, which is a critical step for the use of sampling-based methods in large-scale problems.

Finally, some remarks on the implementation of SGD and its convergence properties are:

- *Biased but consistent* gradient estimators can be also motivated in some setups when implementing SGD.
- **Convergence analysis for a strongly convex objective with a Lipschitz gradient (previous lecture):** In the previous lecture we proved that the convergence rate of SGD for a strongly convex F with a Lipschitz gradient is $O\left(\frac{1}{k}\right)$.

  Nemirovski: "When minimizing strongly convex functions, no algorithm performing k queries to noisy first-order oracles, can achieve better accuracy than $O\left(\frac{1}{k}\right)$."

  This means that if we only use noisy gradients in minimizing F, the best achievable accuracy by any algorithm is $O\left(\frac{1}{k}\right)$. Sometimes SGD is implemented to return $\tilde{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$. This variation improves robustness, but still the rate is $O\left(\frac{1}{k}\right)$.

[2]We write $X_N = O_P(Y_N)$ if for every $\epsilon > 0$, $\exists M, N$ such that

$$P\left(\left|\frac{X_N}{Y_N}\right| < M\right) > 1 - \epsilon \quad \forall n > N.$$

Any $O_P(1)$ sequence is referred to as a *bounded in probability sequence*. In the provided CLT, the normal distribution to the right of $\xrightarrow{d}$ is $O_P(1)$. This justifies the $O_P(1/\sqrt{N})$ term in the convergence of $\{\hat{F}_N\}$.

- Depending on if the SGD is applied in solving (6.1) or (6.2), the expectation in the associated convergence rates (in the previous lecture and in subsequent sections here) is with respect to the corresponding underlying distributions, which correspond to either sampling directly from P in the case of SA and passing over the data once or sampling with replacement from the dataset in the case of SAA.

6.2 Convergence Analysis for a Convex Objective

Continuing the convergence analysis of SGD initiated in the previous lecture, we now assume that:

- F is a convex function,
- $E[\|g(x,\xi)\|^2] \leq C^2, \forall x,$
- $\{\xi_0, \xi_1, \ldots, \xi_{N-1}\}$ is an i.i.d. random sample, independent of x_0,
- $g(x_k, \xi_k)$ is an unbiased estimator of $\nabla F(x_k)$ given $x_0, \xi_0, \ldots, \xi_{k-1}$,
- $\tilde{x}_k = \sum_{t=0}^{k} \frac{\varepsilon_t}{\sum_{j=0}^{k} \varepsilon_j} x_t$ is returned instead of $x_k = x_{k-1} - \varepsilon_{k-1} g(x_{k-1}, \xi_{k-1})$.

Theorem 6.1. *Under the above assumptions*

$$\mathbb{E}[F(\tilde{x}_k) - F(x^*)] \leq \frac{\frac{1}{2}E\left[\|x_0 - x^*]\,\| + \frac{1}{2}C^2 \sum_{t=0}^{k} \varepsilon_t^2\right.}{\sum_{t=0}^{k} \varepsilon_t}.$$

If $\varepsilon_k \asymp \frac{1}{\sqrt{k+1}}$*, then:*

$$E[F(\tilde{x}_k) - F(x^*)] \lesssim \frac{\log(k+1)}{\sqrt{k+1}}.$$

Proof. By the convexity of F:

$$\begin{aligned} & F(x^*) \geq F(x_k) + \nabla F(x_k)^T (x^* - x_k) \\ \text{or} \quad & \nabla F(x_k)^T (x^* - x_k) \leq F(x^*) - F(x_k) \\ \text{or} \quad & \nabla F(x_k)^T (x_k - x^*) \geq F(x_k) - F(x^*) \end{aligned}$$

and therefore:

$$E[\nabla F(x_k)^T (x_k - x^*)] \geq E[F(x_k) - F(x^*)]. \tag{6.3}$$

We now expand $\|x_{k+1} - x^*\|^2$ as in the case of a strongly convex F in the previous lecture:

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 + \varepsilon_k^2 \|g(x_k, \xi_k)\|^2 - 2\varepsilon_k (x_k - x^*)^T g(x_k, \xi_k). \tag{6.4}$$

Using the fact that $E[(x_k - x^*)^T g(x_k, \xi_k)] = E[(x_k - x^*) \nabla F(x_k)]$ and taking the expectation in (6.4) by employing (6.3) and the assumption $E[\|g(x,\xi)\|^2] \leq C^2, \forall x$, we obtain:

$$2\varepsilon_k E\left[F(x_k) - F(x^*)\right] \leq E\left[\|x_k - x^*\|^2\right] - E\left[\|x_{k+1} - x^*\|^2\right] + \varepsilon_k^2 C^2. \tag{6.5}$$

Summing (6.5) over $t = 0, \dots, k$ we get:

$$\sum_{t=0}^{k} 2\varepsilon_t E[F(x_t) - F(x^*)] \leq E[\|x_0 - x^*\|^2] - E[\|x_{k+1} - x^*\|^2] + C^2 \sum_{t=0}^{k} \varepsilon_t^2$$
$$\leq E[\|x_0 - x^*\|^2] + C^2 \sum_{t=0}^{k} \varepsilon_t^2. \tag{6.6}$$

We now divide both sides by $2\sum_{j=0}^{k} \varepsilon_j$ to obtain:

$$\sum_{t=0}^{k} \frac{\varepsilon_t}{\sum_{j=0}^{k} \varepsilon_j} E[F(x_t) - F(x^*)] \leq \frac{\frac{1}{2}E[\|x_0 - x^*\|^2] + \frac{1}{2}C^2 \sum_{t=0}^{k} \varepsilon_t^2}{\sum_{t=0}^{k} \varepsilon_t} \tag{6.7}$$

or

$$E\left[\sum_{t=0}^{k} \frac{\varepsilon_t}{\sum_{j=0}^{k} \varepsilon_j} (F(x_t) - F(x^*))\right] \leq \frac{\frac{1}{2}E[\|x_0 - x^*\|^2] + \frac{1}{2}C^2 \sum_{t=0}^{k} \varepsilon_t^2}{\sum_{t=0}^{k} \varepsilon_t}. \tag{6.8}$$

We further note that

$$\frac{\varepsilon_t}{\sum_{j=0}^{k} \varepsilon_j} > 0 \quad \text{and} \quad \sum_{t=0}^{k} \frac{\varepsilon_t}{\sum_{j=0}^{k} \varepsilon_j} = 1.$$

Employing the fact that F is convex, we finally have:

$$\mathbb{E}[F(\tilde{x}_k) - F(x^*)] \leq \frac{\frac{1}{2}E\left[\|x_0 - x^*]\right\| + \frac{1}{2}C^2 \sum_{t=0}^{k} \varepsilon_t^2}{\sum_{t=0}^{k} \varepsilon_t}.$$

For the second part, let $\varepsilon_k \asymp \frac{1}{\sqrt{k+1}}$. Then, by employing standard results for the harmonic sum we have:

$$\sum_{t=0}^{k} \varepsilon_t^2 \asymp \sum_{t=0}^{k} \frac{1}{t+1} = \sum_{t=1}^{k+1} \frac{1}{t} \asymp \log(k+1).$$

Moreover,

$$\sum_{t=0}^{k} \varepsilon_t \asymp \sum_{t=0}^{k} \frac{1}{\sqrt{t+1}} = \sum_{t=1}^{k+1} \frac{1}{\sqrt{t}},$$

which satisfies

$$2\sqrt{k+2} - 2 < \sum_{t=1}^{k+1} \frac{1}{\sqrt{t}} < 2\sqrt{k+1}.$$

Combining these results, we obtain:

$$E[F(\tilde{x}_k) - F(x^*)] \lesssim \frac{\log(k+1)}{\sqrt{k+1}}.$$

□

6.3 Almost Sure Convergence and Relevant Properties

Consider again the SGD iteration $x_{k+1} = x_k - \varepsilon_k g(x_k, \xi_k)$. Assume that the learning rate schedule satisfies:

$$\sum_{k=0}^{\infty} \varepsilon_k = \infty \quad \text{and} \quad \sum_{k=0}^{\infty} \varepsilon_k^2 < \infty.$$

Then, for any initialization point $x_0 \in \mathbb{R}^n$ and under some additional mild conditions,

$$x_k \to x^* \quad \text{almost surely,} \tag{6.9}$$

where x^* is either a global minimum of F when F is *convex* or *pseudoconvex* or a local minimum of F otherwise.

6.3.1 A bit of background material

For completeness purposes, we provide some background material on pseudoconvexity and related concepts in optimization. *Not all* this material is relevant to our discussion here. Nevertheless, it is provided in a similar spirit as in previous lectures for general background purposes.

Definition 6.2. *Let $\mathcal{X} \subset \mathbb{R}^n$ be a nonempty open set and let $f : \mathcal{X} \to \mathbb{R}$ be a differentiable function on $\mathcal{X}$. Then, f is pseudoconvex on $\mathcal{X}$ if*

$$f(x) < f(y) \Rightarrow \nabla f(y)^T (x - y) < 0, \ \ \forall x, y \in \mathcal{X}$$

or equivalently if

$$\nabla f(y)^T (x - y) \geq 0 \Rightarrow f(x) \geq f(y), \ \forall x, y \in \mathcal{X}.$$

Remark: Often in the definition of a pseudoconvex function, $\mathcal{X}$ is also assumed to be convex. Nevertheless, convexity of $\mathcal{X}$ is not necessary.

This definition states that if the directional derivative of a pseudoconvex function at any point y in the direction $x - y$ is nonnegative, then the function values are nondecreasing in that direction. The definition also implies that if f is pseudoconvex and $\nabla f(y) = 0$, then y is a global minimum of f over $\mathcal{X}$. Pseudoconvexity leads to sufficient optimality conditions in nonlinear optimization, since if a differentiable objective function is pseudoconvex, then the usual first-order stationarity conditions produce a global minimum.

Note: In relevance to prior lecture notes:

$$\begin{aligned} \text{convexity} &\Rightarrow \text{pseudoconvexity}, \\ \text{pseudoconvexity} &\Rightarrow \text{strict quasiconvexity}. \end{aligned} \tag{6.10}$$

The converses are not true. For both implications, a *convex domain* $\mathcal{X}$ is assumed.

Finally, observe that the implication

$$\text{strict quasiconvexity} \Rightarrow \text{quasiconvexity} \tag{6.11}$$

that one would expect to appear in (6.10) has not been added in the second line. Assuming that the reader is familiar with the role of quasiconvexity in optimization, we note that the implication (6.11) does *not* hold in general. To see this, let $\mathcal{X} = [-1, 1]$ and consider the function

$$f(x) = \begin{cases} 1, & x = 0 \\ 0, & -1 \leq x < 0, \ \ 0 < x \leq 1. \end{cases} \tag{6.12}$$

This function is strictly quasiconvex but not quasiconvex.

Note: This conclusion appears to be due to the definition of strict quasiconvexity used here. A small survey showed to us that there exist definitions of strict quasiconvexity in the optimization literature, which are *not* exactly equivalent.

We first start by giving the definition of a quasiconvex function:

Definition 6.3. *Let $\mathcal{X}$ be a convex subset of $\mathbb{R}^n$. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be quasiconvex on $\mathcal{X}$ if all its sublevel sets $S_c = \{x \in \mathcal{X} : f(x) \leq c\}$ for $c \in \mathbb{R}$ are convex.*

Theorem 6.4. *Let $\mathcal{X}$ be a convex subset of $\mathbb{R}^n$. A function $f : \mathcal{X} \to \mathbb{R}$ is quasiconvex on $\mathcal{X}$ if and only if for every $x, y \in \mathcal{X}$:*

$$f(x) \leq f(y) \Rightarrow f(\lambda x + (1-\lambda)y) \leq f(y), \quad \forall \lambda \in [0,1]$$

or alternatively

$$f(\lambda x + (1-\lambda)y) \leq \max\{f(x), f(y)\}, \quad \forall \lambda \in [0,1].$$

This theorem motivates different definitions or notions of strict quasiconvexity existing in the literature. The definition of strict quasiconvexity used here and for the remaining of this subsection is the following:

Definition 6.5. *Let $\mathcal{X}$ be a convex subset of $\mathbb{R}^n$. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be strictly quasiconvex on $\mathcal{X}$ if for every $x, y \in \mathcal{X}$ such that $x \neq y$:*

$$f(x) < f(y) \Rightarrow f(\lambda x + (1-\lambda)y) < f(y), \quad \forall \lambda \in (0,1).$$

The following notion of strict quasiconvexity is also very common in the literature:

Definition 6.6. *Let $\mathcal{X}$ be a convex subset of $\mathbb{R}^n$. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be strictly quasiconvex on $\mathcal{X}$ if f is quasiconvex and if for every $x, y \in \mathcal{X}$ such that $x \neq y$:*

$$f(\lambda x + (1-\lambda)y) < \max\{f(x), f(y)\}, \quad \forall \lambda \in (0,1).$$

Remark: According to the second definition, $f(x)$ in (6.12) is not strictly quasiconvex.

To connect strict quasiconvexity and quasiconvexity according to Definition 6.5, the concept of lower semi-continuity is required:

Definition 6.7. *Let $\mathcal{X}$ be a metric space and let $f : \mathcal{X} \to \mathbb{R}$ be a real-valued function. We say that f is lower semi-continuous at $x_0 \in \mathcal{X}$ if for every $\epsilon > 0$ there exists a neighborhood $\mathcal{U}$ of x_0 such that $f(x) > f(x_0) - \epsilon$ for all $x \in \mathcal{U}$. Equivalently, this can be expressed as:*

$$\liminf_{x \to x_0} f(x) \geq f(x_0).$$

Intuition: Roughly speaking, $f(x)$ for values of x near x_0 is either close to or greater than $f(x_0)$.

Remark: A function may be lower semi-continuous at a point x_0 without being either *left* or *right continuous* at x_0.

Relevance to optimization: Lower and upper semi-continuity correspond to weaker concepts than continuity. Every lower semi-continuous function on a compact space $\mathcal{X}$ has a minimum on $\mathcal{X}$.

Theorem 6.8. *Let $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^n$ is a convex set. If f is strictly quasiconvex and lower semi-continuous on $\mathcal{X}$, then it is quasiconvex on $\mathcal{X}$.*

6.3.2 Back to our Discussion

Example: (Mean Estimation via SGD) Suppose that we want to solve

$$\min_{x \in \mathbb{R}} F(x) = \frac{1}{2} E\left[(x - \xi)^2\right]$$

by employing the SGD algorithm. Observe that $(x-\xi)^2$ is convex in x for any $\xi \in \mathbb{R}$ and therefore $F(x)$ is convex.

Lemma 6.9. *For any random variable X,*

$$E[(X - E[X])^2] \leq E[(X - c)^2] \quad \textit{for any } c \in \mathbb{R}.$$

This lemma shows that the optimal solution in our problem is

$$x^* = \arg\min_{x\in\mathbb{R}} F(x) = E[\xi] = \mu$$

and is unique. Therefore, for any initialization x_0 of the SGD, (6.9) implies that

$$x_k \to E[\xi] \quad \text{almost surely.}$$

To see this, note that

$$x_{k+1} = x_k - \varepsilon_k g(x_k, \xi_k) = x_k - \varepsilon_k(x_k - \xi_k) = (1-\varepsilon_k)x_k + \varepsilon_k\xi_k. \tag{6.13}$$

We now choose $\varepsilon_k = \frac{1}{k+1}$. Then, for any $x_0 \in \mathbb{R}$:

$$x_{k+1} = \frac{k}{k+1}x_k + \frac{1}{k+1}\xi_k = \frac{1}{k+1}\sum_{t=0}^{k}\xi_t \to E[\xi] \quad \text{almost surely}$$

by the Strong Law of Large Numbers[3]. Since by *Fatou's Lemma*, almost sure convergence implies convergence in probability, the SGD estimator is also *consistent*[4].

Expected Evolution: We observe further that each iterate of the sequence $\{x_k\}_{k\geq 1}$ is an unbiased estimator of μ:

$$E[x_{k+1}] = \frac{1}{k+1}\sum_{t=0}^{k} E[\xi_t] = \mu.$$

Therefore, trivially:

$$\lim_{k\to\infty} E[x_k] = \mu.$$

Convergence in the Mean Square Sense: Let $\mathrm{Var}(\xi_k) = \sigma^2 < \infty$. We can easily establish that

$$E\left[(x_k - \mu)^2\right] = \frac{\sigma^2}{k} = O(\varepsilon_k) \xrightarrow[k\to\infty]{} 0.$$

Almost sure convergence to μ holds also for more general learning schedules satisfying the conditions $\sum_{k=0}^{\infty}\varepsilon_k = \infty$ and $\sum_{k=0}^{\infty}\varepsilon_k^2 < \infty$. An example of using these conditions can be obtained by analyzing the asymptotic mean of the SGD estimator:

$$\begin{aligned} E[x_{k+1} - \mu] &= (1-\varepsilon_k)E[x_k] + \varepsilon_k E[\xi_k] - \mu \\ &= (1-\varepsilon_k)E[x_k - \mu] + \varepsilon_k E[\xi_k - \mu] \\ &= (1-\varepsilon_k)E[x_k - \mu]. \end{aligned}$$

By iterating this equation, we obtain:

$$E[x_{k+1} - \mu] = \prod_{t=0}^{k}(1-\varepsilon_t)E[x_0 - \mu].$$

By taking the absolute value to both sides and by assuming that $\varepsilon_k \in (0,1), \forall k$, the following bound follows:

$$|E[x_{k+1} - \mu]| = \prod_{t=0}^{k}(1-\varepsilon_t)\,|E[x_0 - \mu]| \leq \prod_{t=0}^{k} e^{-\varepsilon_t}\,|E[x_0 - \mu]| = e^{-\sum_{t=0}^{k}\varepsilon_t}\,|E[x_0 - \mu]|\,.$$

[3]Clearly, we implicitly assume that $|E[\xi]| < \infty$.

[4]This is also a demonstration of the Weak Law of Large Numbers.

Here, we have used the inequality $1 - x \leq e^{-x}, \ \forall x \in \mathbb{R}$. Letting $k \to \infty$ and using the fact that $\sum_{k=0}^{\infty} \varepsilon_k = \infty$ we establish the **asymptotic unbiasedness** of $\{x_k\}$.

$$\lim_{k \to \infty} E[x_k] = \mu.$$

Example: (M-estimation) Suppose that we have i.i.d. observations $x_1, x_2, \ldots, x_N \sim P \in \mathcal{P}$ and a parametric family of distributions $\mathcal{P}_\theta = \{P_\theta : \theta \in \Theta\} \subset \mathcal{P}$ for the purpose of approximating P. Our goal is to choose θ or equivalently P_θ based on the available data record to approximate P in some sense. Let $f_\theta(x)$ be the associated likelihood and assume that it is strictly positive for all admissible θ, x. The Maximum-Likelihood (ML) estimator is a statistic of the form $\hat{\theta}_N = T_N(x_1, x_2, \ldots, x_N)$ for some measurable function T_N which maximizes $\prod_{i=1}^{N} f_\theta(x_i)$ or equivalently minimizes $\sum_{i=1}^{N} [-\log(f_\theta(x_i))]$. In 1964, Peter J. Huber proposed generalizing this approach to *M-estimators* ("M" stands for "Maximum Likelihood-type") which are defined as the statistics $\hat{\theta}_N$ minimizing

$$\sum_{i=1}^{N} \rho(x_i; \theta),$$

where ρ is some real-valued function, often with certain properties. Assuming differentiability of ρ with respect to θ, clearly

$$\sum_{i=1}^{N} \nabla_\theta \rho(x_i; \hat{\theta}_N) = 0$$

is an optimality condition. Choosing $\rho(x; \theta) = -\log f_\theta(x)$, we recover the ML estimators, which are M-estimators. Suppose further that our observation record is of the form $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where (x_i, y_i) is an input-output pair assumed to be linearly related. Let $x_i, \theta \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ for any i. Then, Ordinary Least Squares (OLS) estimation for the linear regression model also provides an M-estimator:

$$\sum_{i=1}^{N} \left(y_i - x_i^T \theta\right)^2, \qquad \rho((x, y); \theta) = (y - x^T \theta)^2.$$

In general,

$$C_N(\theta) = \phi\left(\frac{1}{N} \sum_{i=1}^{N} \rho(x_i; \theta)\right)$$

is called *criterion function*, where $\phi(\cdot)$ is a continuous function and $\hat{\theta}_N = \arg\min_{\theta \in \Theta} C_N(\theta)$. In the previous definition, M-estimators were defined with $\phi(x) = x$. M-estimators enjoy similar consistency and asymptotic normality properties as the ML estimators, occasionally with higher asymptotic variance. There are several reasons for studying M-estimators:

- They may be more computationally efficient than ML estimators.
- They are often used in robust statistics, because they are more resistant to deviations from the underlying assumptions than ML estimators.
- They can be analyzed without assuming that the true model $P \in \mathcal{P}_\theta$.

The following theorem establishes the *consistency* of M-estimators under some regularity conditions:

Theorem 6.10. *Suppose that the parameter space $\Theta \subset \mathbb{R}^p$ is compact and that the true parameter θ_o is an interior point of Θ, i.e., $\theta_o \in \text{int}(\Theta)$ and therefore $P \in \mathcal{P}_\theta$. Moreover, assume that*

$$C_N(\theta) \xrightarrow{p} \bar{C}(\theta) \quad \textit{uniformly in } \Theta,$$

i.e., $\sup_{\theta\in\Theta}|C_N(\theta)-\bar{C}(\theta)|\xrightarrow{p}0$*, where* $\bar{C}(\theta)$ *is a deterministic function of* θ *and that* θ_o *is the unique minimum of* $\bar{C}(\theta)$*. Then,*

$$\hat{\theta}_N\xrightarrow{p}\theta_o.$$

Example of SGD application: Let

$$\hat{\theta}_N=\arg\min_{\theta\in\mathbb{R}^p}\sum_{i=1}^{N}\varrho(y_i-x_i^T\theta),$$

where $\varrho:\mathbb{R}\to\mathbb{R}^+$ is a convex function and $y_i=x_i^T\theta_o+w_i,\forall i$, where $\{w_i\}$ corresponds to an i.i.d. zero-mean noise sequence. We can then apply

$$\theta_{k+1}=\theta_k+\varepsilon_k\varrho'(y_k-x_k^T\theta_k)x_k$$
$$\tilde{\theta}_{k+1}=\frac{1}{k+1}\sum_{r=1}^{k+1}\theta_r$$

to approximate M-estimators.