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

Lecture 5: Gradient Projection and Stochastic Gradient Descent-Part I

Instructor: Dimitrios Katselis *Scribe: Xingyu Bai and Tiancheng Zhao*

5.1 Introduction

Suppose we want to solve a problem of the form

$$\min_{x\in\mathcal{X}} \quad f(x), \tag{5.1}$$

where $\mathcal{X}$ is usually (but not always) a closed convex set[1]. This corresponds to a constrained optimization problem. Assuming that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, gradient-based methods can be applied. Nevertheless, the application of such methods poses a problem: if an iteration starts at a point x_k close to the boundary[2] $\partial\mathcal{X}$ of $\mathcal{X}$, one step in the opposite direction of the gradient might lead to a point not in $\mathcal{X}$. Denote the point obtained by the application of the gradient step by y_{k+1}. The most natural method to ensure the feasibility of the sequence of iterates $\{x_k\}$ is to obtain x_{k+1} by projecting y_{k+1} onto $\mathcal{X}$. A different approach to resolve this problem is to avoid stepping outside $\mathcal{X}$ by relying on the *conditional gradient method*, also known as the **Frank-Wolfe algorithm**.

Note: Often, the constraint set $\mathcal{X}$ has structure specified by equations and inequalities. If this structure is taken into account, then ideas from Lagrange multipliers and duality theory become useful. The aforementioned methods do not rely on any other structural characteristic of the constraint set $\mathcal{X}$ other than its convexity. These methods generate sequences of iterates $\{x_k\}$ by searching along descent directions. In this sense, they correspond to constrained versions of the unconstrained gradient methods that we encountered in previous lectures.

5.2 Some Background Material

Consider again the general setup of problem (5.1), where $\mathcal{X}$ can either be a general set or $\mathcal{X} = \{x \in \mathcal{U}_0 : h(x) \leq 0\} \subset \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathcal{U}_0$ is a nonempty open set in $\mathbb{R}^n$. We note here that if $\mathcal{X}$ contains in its description equality constraints of the form $\tilde{h}(x) = 0$, every such constraint can be always expressed as the pair of the inequality constraints $\tilde{h}(x) \leq 0$ and $-\tilde{h}(x) \leq 0$.

Definition 5.1. *$d \in \mathbb{R}^n$ is said to be a* feasible direction *of $\mathcal{X}$ at x if there exists a sufficiently small $\delta > 0$ such that $x + \varepsilon d \in \mathcal{X}, \forall \varepsilon \in [0, \delta]$.*

Two relevant concepts are:

Definition 5.2. *The **cone of feasible directions** of $\mathcal{X}$ at x is the set*

$$\mathcal{FD}(x) = \{d \in \mathbb{R}^n : x + \varepsilon d \in \mathcal{X},\ \forall \varepsilon \in [0, \delta] \ \ \textit{for some } \delta > 0\}.$$

[1] $\mathcal{X}$ is assumed closed when we deal with algorithms.

[2] $\mathcal{X} \subset \mathbb{R}^n$ is closed if and only if it contains all its boundary points.

Often the definitions of a feasible direction and the associated cone are given by assuming that $d \neq 0$ and $\varepsilon \in (0, \delta)$ for some $\delta > 0$. In the provided definition, $\mathcal{FD}(x)$ contains the origin. It does not need to be closed or convex. If $\mathcal{X}$ is convex, then $\mathcal{FD}(x)$ is a convex cone.

Definition 5.3. *The **cone of improving directions** at x is the set*

$$\mathcal{ID}(x) = \{d \in \mathbb{R}^n : f(x + \varepsilon d) < f(x) \ \ \textit{for suffciently small } \varepsilon > 0\}.$$

Let x^* be an extremum of f. If x^* is a local minimum and d is a feasible direction at x^*, then

$$f(x^* + \varepsilon d) \geq f(x^*), \quad \text{for suffciently small } \varepsilon > 0.$$

Since for a continuously differentiable f

$$f(x^* + \varepsilon d) = f(x^*) + \varepsilon \nabla f(x^*)^T d + o(\varepsilon),$$

then for all sufficiently small ε:

$$\varepsilon \nabla f(x^*)^T d + o(\varepsilon) \geq 0.$$

Dividing by ε and letting $\varepsilon \to 0$ we obtain:

$$\nabla f(x^*)^T d \geq 0. \tag{5.2}$$

Equation (5.2) corresponds to a **necessary condition** for x^* to be a local minimum of f over $\mathcal{X}$.

Interior Case: Let $\mathcal{X}$ be a subset of $\mathbb{R}^n$ and let f be a $\mathcal{C}^1$ (i.e., continuously differentiable) real-valued function on $\mathcal{X}$. If x^* is a local minimizer of f over $\mathcal{X}$ and $x^* \in \text{int}(\mathcal{X})$ (interior of $\mathcal{X}$), then necessarily $\nabla f(x^*) = 0$.

For a (closed) convex set $\mathcal{X}$, $d = x - x^*$ is a feasible direction $\forall x \in \mathcal{X}$ at $x^* \in \mathcal{X}$ (in fact, for a convex set $\mathcal{X}$ the feasible directions or *feasible variations* are of the form $d = x - x^*$ for $x \in \mathcal{X}$). This is because

$$x^* + \varepsilon d = x^* + \varepsilon(x - x^*) = (1 - \varepsilon)x^* + \varepsilon x \in \mathcal{X}, \forall \varepsilon \in [0, 1].$$

Therefore, a **necessary condition** for x^* to be a local minimum of f over a **convex** constraint set $\mathcal{X}$ is

$$\nabla f(x^*)^T (x - x^*) \geq 0, \ \ \forall x \in \mathcal{X}. \tag{5.3}$$

Note: (**First-Order Sufficient Optimality Condition**) If f is convex, (5.3) is also **sufficient** for x^* to minimize f over $\mathcal{X}$.

Proof. Let f be convex. Suppose that $x^* \in \mathcal{X}$ and that (5.3) holds for x^*. Then, due to the convexity of f we have that:

$$f(x) \geq f(x^*) + \nabla f(x^*)^T (x - x^*), \quad \forall x \in \mathcal{X}.$$

Therefore, $f(x) \geq f(x^*), \ \ \forall x \in \mathcal{X}$.

□

A point x^* satisfying (5.3) is called **stationary**.

Remarks:

1. In summary, for a convex set $\mathcal{X}$ the necessity of (5.3) is independent of the convexity of f. The convexity of f is only used in establishing sufficiency.

2. If $\mathcal{X} = \mathbb{R}^n$, then (5.3) reduces to the first-order unconstrained optimality condition $\nabla f(x^*) = 0$.

3. If f is convex and $\nabla f(x^*) \neq 0$, then $-\nabla f(x^*)$ defines a **supporting hyperplane** to $\mathcal{X}$ at x^*.

4. If $x^* \in \text{int}(\mathcal{X})$, then we must have $\nabla f(x^*) = 0$.

5. A more general result is the following: Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex. A vector x^* minimizes f over a convex set $\mathcal{X} \subset \mathbb{R}^n$ if and only if there exists a **subgradient** $g \in \partial f(x^*)$ such that

$$g^T(x - x^*) \geq 0, \ \ \forall x \in \mathcal{X}.$$

   Here, $\partial f(x^*)$ is the *subdifferential* of f at x^*, i.e., the set of all subgradients of f at x^*.

   For completeness, recall that $g \in \mathbb{R}^n$ is a subgradient of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at x if it satisfies:

$$f(y) \geq f(x) + g^T(y - x), \ \ \forall y \in \mathbb{R}^n. \tag{5.4}$$

   Therefore, a subgradient defines an affine global underestimator of f at x and exists always. If f is differentiable, then $\partial f(x) = \{\nabla f(x)\}$. Moreover, $\partial f(x)$ is a closed convex set, since by (5.4) it is the intersection of a collection of closed halfspaces, one for each $y \in \mathbb{R}^n$. Finally, the same subgradient definition applies also to any nonconvex f, although a subgradient may not exist in this case.

6. A general result in terms of the aforementioned cones (necessary condition for optimality of x^*): Consider the problem

$$\min_{x \in \mathcal{X}} f(x), \tag{5.5}$$

   where $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathcal{X} \subseteq \mathbb{R}^n$ with $\mathcal{X} \neq \emptyset$. Let $x^* \in \mathcal{X}$ be a local minimum of f over $\mathcal{X}$. Then,

$$\mathcal{ID}(x^*) \cap \mathcal{FD}(x^*) = \emptyset. \tag{5.6}$$

   Intuition: At the local minimum x^*, no improving direction is also a feasible direction.

7. **Using the gradient of the loss function f to characterize the set of descent directions**: Suppose that $f \in \mathcal{C}^1$. For any $x \in \mathcal{X}$ we define the set

$$\mathcal{ID}_0(x) = \{d \in \mathbb{R}^n : \nabla f(x)^T d < 0\}.$$

   Recall that $\nabla f(x)^T d < 0 \Rightarrow f(x + \varepsilon d) < f(x)$ for sufficiently small $\varepsilon > 0$. Therefore, d is a *descent direction*, i.e., $d \in \mathcal{ID}(x)$. This implies that $\mathcal{ID}_0(x) \subseteq \mathcal{ID}(x)$. Then, for the local minimum $x^* \in \mathcal{X}$ of f over $\mathcal{X}$, condition (5.6) is often replaced by $\mathcal{ID}_0(x^*) \cap \mathcal{FD}(x^*) = \emptyset$.

5.3 Gradient Projection Algorithms

Suppose that we want to solve the problem

$$\min_{x \in \mathcal{X}} \quad f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $\mathcal{X}$ is a closed convex set. Solving this problem or equivalently solving (5.3) may be difficult for complex problems. Therefore, we need some numerical methods.

Consider the iteration: pick $x_0 \in \mathcal{X}$ and $\varepsilon > 0$. Define the iterative scheme:

$$x_{k+1} = [x_k - \varepsilon \nabla f(x_k)]^+, \tag{5.7}$$

where $[\cdot]^+$ is the projection operator of x onto $\mathcal{X}$, that is,

$$[x]^+ = \arg\min_{y \in \mathcal{X}} \|y - x\|. \tag{5.8}$$

Clearly, $[x]^+$ corresponds to the closest point to x in $\mathcal{X}$. The projection $[x]^+$ is also **unique**: Assume that $[x]_1^+ \neq [x]_2^+ \in \mathcal{X}$ are two different projections of x onto $\mathcal{X}$. Then, their midpoint $\frac{1}{2}([x]_1^+ + [x]_2^+) \in \mathcal{X}$ is strictly closer to x than at least one of $[x]_1^+, [x]_2^+$, a contradiction!

The constant stepsize projected steepest descent given by (5.7) can be shown to converge to a minimizer of f over $\mathcal{X}$ under some conditions. The following theorem is used to analyze gradient projection methods:

Theorem 5.4. ***(Projection Theorem)*** *Consider the problem (5.8), where $\mathcal{X} \subset \mathbb{R}^n$ is closed and convex.*

1. *The projection $[x]^+$ of any $x \in \mathbb{R}^n$ onto $\mathcal{X}$ is unique.*
2. *For any given $x \in \mathbb{R}^n$, $x^* = [x]^+$ if and only if*

$$(x - x^*)^T (y - x^*) \leq 0, \quad \forall y \in \mathcal{X}.$$

3. *The mapping $\phi : \mathbb{R}^n \to \mathcal{X}$ defined by $\phi(x) = [x]^+$ is continuous and nonexpansive, i.e.,*

$$\left\| [x]^+ - [y]^+ \right\| \leq \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

4. *(**Orthogonality Principle**) If $\mathcal{X}$ is a subspace, then $x^* = [x]^+$ if and only if*

$$(x - x^*)^T y = 0, \quad \forall y \in \mathcal{X}.$$

The next fact says that the gradient projection method terminates if and only if it encounters a stationary point:

Fact: $x^* = [x^* - \varepsilon \nabla f(x^*)]^+$ for any $\varepsilon > 0$ if and only if x^* is stationary.

Note: Gradient projection has the same rate as gradient descent (unconstrained case).

Strongly convex and L-smooth f: If f is strongly convex and L-smooth ($mI \leq \nabla^2 f(x) \leq LI$), the projected steepest descent converges geometrically fast to x^*. Also, the dependence on the upper bound $\frac{L}{m}$ of the corresponding condition number is similar to the unconstrained case.

Remarks:

1. Computing the operator $[\cdot]^+$ can be a challenging problem itself, but often less challenging than minimizing the loss function f directly. A first reason is that the projection objective $\|y - x\|$ or equivalently its squared version $\|y - x\|^2$ is smooth and strongly convex with condition number 1, which is more or less as well-behaved as a convex objective function could possibly be. Also, the constraint set $\mathcal{X}$ might have a shape which permits an easy computation of $[x]^+$, e.g., when $\mathcal{X}$ is a sphere or a rectangular box. However, often in practice computing $[x]^+$ in the gradient projection iteration is the most computationally demanding step.
2. Some examples of easy sets to project onto are:
   - Box constraints of the form $\{x \in \mathbb{R}^n : l \leq x \leq u\}$ with an elementwise interpretation of the inequalities.
   - Solutions of linear systems $\{x \in \mathbb{R}^n : Ax = b\}$ or more generally *simple* polyhedral sets $\{x \in \mathbb{R}^n : Ax = b, Bx \leq c\}$ for a small number of equality and inequality constraints, again with an elementwise interpretation of the inequalities. Recall that a *polyhedron* is the solution set of a finite number of linear equality and inequality constraints, i.e., the intersection of a finite number of halfspaces and hyperplanes. Often

in the literature, a polyhedron is defined as the solution set of a finite[3] number of inequality constraints (with no explicit reference to equality constraints), i.e., as the intersection of finitely many halfspaces. We clarify here that the definition can include equality constraints since $Ax = b$ is equivalent to $Ax \leq b$ and $-Ax \leq -b$.

- Norm balls $\{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$ for $p = 1, 2, \infty$ and norm cones, i.e., sets of the form $\{(x, t) \in \mathbb{R}^n \times \mathbb{R} : \|x\| \leq t\}$ for some norm $\|\cdot\|$. Recall that norm balls and norm cones are convex sets.
- Probability simplex $\{x \geq 0 : \sum_{i=1}^n x_i = 1\}$.
- Nonnegative orthant $\mathbb{R}^n_+$.

Note: Projection onto seemingly simple sets $\mathcal{X}$ can be very hard. For example, it is typically hard to project onto the solution set of arbitrary linear inequalities, i.e., onto an arbitrary polyhedron.

Conditional Gradient Method: Introduced by Marguerite Frank and Philip Wolfe in 1956. At every iteration, it obtains $\bar{x}_k$ as the solution of the problem:

$$\bar{x}_k = \arg\min_{x \in \mathcal{X}} \nabla f(x_k)^T (x - x_k), \tag{5.9}$$

where $\mathcal{X}$ is assumed to be compact so that (5.9) has a solution and implements the iterative scheme

$$x_{k+1} = x_k + \varepsilon_k(\bar{x}_k - x_k), \quad \varepsilon_k \in (0, 1],$$

using $d_k = \bar{x}_k - x_k$ as a descent direction.

The use of conditional gradient descent is justified if the solution of (5.9) is simpler than solving the original problem. If f is nonlinear and $\mathcal{X}$ is specified by linear equality and inequality constraints, then (5.9) becomes a linear program, which is generally easy to solve. Additionally, minimizing a linear function over $\mathcal{X}$ is generally easier than computing $[x]^+$ in every iteration, which corresponds to solving a quadratic program. Finally, when moving towards $\bar{x}_k$ rather than in the direction of $-\nabla f(x_k)$, we can take larger steps without intersecting $\partial\mathcal{X}$. Larger steps tend to reduce the number of iterations required to find a near-optimal point of the original problem.

5.4 Stochastic Gradient Descent

Optimization methods for machine learning fall into two broad categories often called *stochastic* and *batch*. The main representative of the first category is stochastic gradient descent (SGD). Each iteration of this method is very cheap, involving only the computation of the gradient $\nabla f(x_k)$ based on one sample (the resulting gradient is often denoted by $\hat{\nabla} f(x_k)$). On the other hand, the simplest batch method is steepest descent, which is also known as *batch gradient* or *full gradient* method. The distinction between these methods has been enhanced due to the emergence of large-scale machine learning problems. Large-scale machine learning corresponds to a very particular setup in which the stochastic gradient method has traditionally played a core role. This is in contrast to conventional gradient-based nonlinear optimization methods, which are typically effective for solving small-scale learning problems.

Consider the general optimization problem:

$$\min_{x \in \mathbb{R}^n} F(x) = \mathbb{E}_\xi[f(x; \xi)],$$

where $F(x) = \mathbb{E}_\xi[f(x; \xi)]$ is called *expected* or *population risk*. If $f(\cdot; \xi)$ is convex for each possible value of ξ, then $F(x)$ is also convex. In the machine learning literature, the problem often appears in the following form:

$$\min_\theta F(\theta) = \mathbb{E}_{XY}[\ell((X, Y); \theta)].$$

[3]The solution set of the *infinite* set of linear inequalities $a^T x \leq 1$ for all a such that $\|a\| = 1$ corresponds to the unit ball $\{x : \|x\| \leq 1\}$, which is not a polyhedron.

Here, $\ell((X,Y);\theta)$ is the loss function, $\xi=(X,Y)$ is the input-output pair and θ is a parameterization of the functional relationship between X,Y.

For the cases where the gradients are easy to compute, we can easily apply traditional gradient-based methods introduced in previous lectures. However, in many cases computing the gradient of the objective function $F(x)$ is computationally expensive, thus the idea of stochastic gradient descent arises. In practice, the distribution of ξ is often unknown to us, but we can observe a set of i.i.d. samples $\{\xi_1,\xi_2,\ldots,\xi_N\}$ drawn from this distribution. We therefore replace the expected risk with the *empirical risk*:

$$\frac{1}{N}\sum_{i=1}^{N} f(x;\xi_i).$$

General or batch gradient-based methods need to compute or estimate

$$\nabla F(x)=\nabla E_{\xi}\left[f(x;\xi)\right]\approx\frac{1}{N}\sum_{i=1}^{N}\nabla_x f(x;\xi_i),$$

which is inefficient for large-scale optimization.

SGD idea: Estimate the required gradient based on a single observation:

$$g(x_k,\xi_k)=\nabla_x f(x_k;\xi_k),$$

where ξ_k is chosen at random from the observation record $\{\xi_1,\xi_2,...,\xi_N\}$ and $E_{\xi_k}[g(x_k,\xi_k)]=\nabla F(x_k)$ (unbiased gradient estimator). Then, perform the following iteration:

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

We note here that due to stochasticity, SGD is not necessarily a real descent, i.e. $F(x_{k+1})<F(x_k)$ may not hold for each iteration.

A first SGD convergence result is provided by the following theorem:

Theorem 5.5. *Suppose that F is strongly convex with parameter m and has a Lipschitz gradient with Lipschitz constant L. Moreover, assume that $E[\|g(x,\xi)\|^2]\le C^2,\forall x$ and $\varepsilon_k=\frac{1}{m(k+1)}$. Then, for x_0 independent of $\{\xi_k\}$:*

$$E\left[\|x_k-x^*\|^2\right]\le\frac{1}{k+1}\max\left\{E\left[\|x_0-x^*\|^2\right],\frac{C^2}{m^2}\right\},$$

Additionally, we have

$$\mathbb{E}\left[F(x_k)-F(x^*)\right]\le\frac{L}{2}\mathbb{E}\left[\|x_k-x^*\|^2\right]=O\left(\frac{1}{k}\right).$$

Proof.

$$\begin{aligned}\|x_{k+1}-x^*\|^2&=\|x_k-\varepsilon_k g(x_k,\xi_k)-x^*\|^2\\&=\|x_k-x^*\|^2+\varepsilon_k^2\|g(x_k,\xi_k)\|^2-2\varepsilon_k g(x_k,\xi_k)^T(x_k-x^*)\end{aligned}\tag{5.10}$$

It is not hard to see that $x_k=h(x_0,\xi_0,\cdots,\xi_{k-1})$. Since ξ_k is independent of $\{x_0,\xi_0,\cdots,\xi_{k-1}\}$, then ξ_k is independent of x_k. By the smoothing property of conditional expectation,

$$\begin{aligned}E[(x_k-x^*)^T g(x_k,\xi_k)]&=E[E[(x_k-x^*)^T g(x_k,\xi_k)|x_0,\xi_0,\cdots,\xi_{k-1}]]\\&=E[(x_k-x^*)^T E[g(x_k,\xi_k)|x_0,\xi_0,\cdots,\xi_{k-1}]]\\&=E[(x_k-x^*)^T\nabla F(x_k)]\\&=E[(x_k-x^*)^T(\nabla F(x_k)-\nabla F(x^*))],\qquad(\nabla F(x^*))=0\\&\ge mE[\|x_k-x^*\|^2].\qquad\text{(strong convexity)}\end{aligned}\tag{5.11}$$

By taking the expectation in (5.10) and using (5.11) we obtain:

$$E[\|x_{k+1}-x^*\|^2] \leq E[\|x_k-x^*\|^2] + \varepsilon_k^2 C^2 - 2\varepsilon_k m E[\|x_k-x^*\|^2] = (1-2\varepsilon_k m)E[\|x_k-x^*\|^2] + \varepsilon_k^2 C^2. \quad (5.12)$$

We now apply induction. For $k=0$, it is clear that $E\left[\|x_0-x^*\|^2\right] \leq \max\left\{E\left[\|x_0-x^*\|^2\right], \frac{C^2}{m^2}\right\}$. Assume that the result holds for $k-1$; that is, $E[\|x_{k-1}-x^*\|^2] \leq \frac{1}{k}\max\left\{E\left[\|x_0-x^*\|^2\right], \frac{C^2}{m^2}\right\}$. Then for k and $\varepsilon_k = \frac{1}{m(k+1)}$ we have:

$$\begin{aligned}
E[\|x_k-x^*\|^2] &\leq \left(1-\frac{2}{k}\right)E[\|x_{k-1}-x^*\|^2] + \frac{C^2}{m^2k^2} \\
&\leq \left(1-\frac{2}{k}\right)\frac{1}{k}\max\left\{E\left[\|x_0-x^*\|^2\right], \frac{C^2}{m^2}\right\} + \frac{C^2}{m^2k^2} \\
&\leq \left[\left(1-\frac{2}{k}\right)\frac{1}{k} + \frac{1}{k^2}\right]\max\left\{E\left[\|x_0-x^*\|^2\right], \frac{C^2}{m^2}\right\} \\
&\leq \frac{1}{k+1}\max\left\{E\left[\|x_0-x^*\|^2\right], \frac{C^2}{m^2}\right\}.
\end{aligned}$$

Therefore, $E[\|x_k-x^*\|^2] \leq \frac{1}{k+1}\max\left\{E\left[\|x_0-x^*\|^2\right], \frac{C^2}{m^2}\right\}$ holds for all k.

Employing the Lipschitz gradient property and the fact that $\nabla F(x^*)=0$ we obtain:

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

□