E0 270 Machine Learning **Jan 23, 2015**

Optimization Tutorial 3

Projected Gradient Descent, Duality

Lecture by Harikrishna Narasimhan

Note: This tutorial shall assume background in elementary calculus and linear algebra.

In the last lecture, we started with constrained optimization, and explained the KKT conditions for optimality of a solution to a constrained optimization problem. Based on these conditions, we also derived the method of Lagrange multipliers for solving a problem with equality constraints. We now describe a more general technique for finding the minimizer of a constrained (convex) optimization problem, namely projected gradient descent. We will then discuss the notion of a dual problem associated with any constrained optimization problem, which in many settings will be useful in designing efficient solvers for the problem.

1 Projected Gradient Descent

We wish to design a method to find the optimal solution to a constrained optimization problem. We shall be interested here in the special case of minimization of a (bounded) convex function $f : \mathbb{R}^d \to \mathbb{R}$ in C^1 over a convex constraint set $\mathcal{C} \subseteq \mathbb{R}^d$, where as mentioned earlier, all local minimizers are also global minimizers. Note that the gradient descent method described in the first lecture cannot be directly used here, as applying the prescribed gradient-based update to a point in the constraint set $\mathcal{C}$ can take us outside $\mathcal{C}$. Indeed, one can consider performing some form of post-processing on the point obtained from such an update to get a new point that lies in $\mathcal{C}$. For instance, having obtained $\mathbf{x}'_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$ from the current point $\mathbf{x}_t \in \mathcal{C}$, the next iterate in the optimization method can simply be set to the point closest to $\mathbf{x}'_{t+1}$ in $\mathcal{C}$ (under the ℓ_2-norm), i.e., to $\mathbf{x}_{t+1} \in \text{argmax}_{\mathbf{x}\in\mathcal{C}} \|\mathbf{x}'_{t+1} - \mathbf{x}\|_2$. This projection operation will always yield a unique point for the convex set $\mathcal{C}$. Moreover, convexity of $\mathcal{C}$ also gives us that a projection of $\mathbf{x}'_{t+1}$ onto $\mathcal{C}$ will always result in a point that is closer than $\mathbf{x}'_{t+1}$ to the minimizer in $\mathcal{C}$. The resulting technique known as the projected gradient method is thus guaranteed to converge to the (global) minimizer of the given problem (for appropriate choices of step size η_t) [1, 2]. An outline of the method is provided below.

```
Projected Gradient Descent Method
Input: $f : \mathbb{R}^d \to \mathbb{R}$, $\mathcal{C} \subseteq \mathbb{R}^d$
Initialize: $\mathbf{x}_1 \in \mathbb{R}^d$
Parameter: $T$
for $t = 1$ to $T$
    – Select step-size $\eta_t > 0$
    – $\mathbf{x}'_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$
    – $\mathbf{x}_{t+1} = \Pi_{\mathcal{C}}[\mathbf{x}'_{t+1}]$, where $\Pi_{\mathcal{C}}$ denotes projection onto $\mathcal{C}$
Output: $\mathbf{x}_{T+1}$
```

Example 1. *Consider the task of maximizing a convex function $f : \mathbb{R}^d \to \mathbb{R}$ subject to the constraint $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is a non-singular matrix and $\mathbf{b} \in \mathbb{R}^d$. It can be verified that the projection of any point $\mathbf{x}'$ onto the given convex constraint set $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^d \,|\, \mathbf{A}\mathbf{x} = \mathbf{b}\}$ is the unique point $\mathbf{x}' - \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}(\mathbf{A}\mathbf{x}' - \mathbf{b})$ (see for e.g. [3]). The above projected gradient update for this problem then becomes:*

$$\begin{aligned} \mathbf{x}_{t+1} &= \Pi_{\mathcal{C}}[\mathbf{x}'_{t+1}] = \mathbf{x}'_{t+1} - \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}(\mathbf{A}\mathbf{x}'_{t+1} - \mathbf{b}) \\ &= \mathbf{x}_t - \eta_t(\mathbf{I} - \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A})\nabla f(\mathbf{x}_t), \end{aligned}$$

which follows from the expansion of $\mathbf{x}'_{t+1}$ and from $\mathbf{A}\mathbf{x}_t = \mathbf{b}$ (as a result of $\mathbf{x}_t$ belonging to $\mathcal{C}$).

2 Duality

We have so far seen conditions for optimality of a solution to a constrained optimization problems and algorithmic techniques for finding such an optimal solution. We now describe the concept of a dual problem associated with any constrained optimization problem, which often offers nice interpretations for the original problem and in many settings helps in developing efficient methods for solving the problem. Our explanation here shall be based on Chapter 5 of [3].

Consider an optimization problem with equality and inequality constraints:

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

$$\begin{aligned} \text{s.t.} \quad & h_i(\mathbf{x}) = 0, \quad i = 1,\dots,E \\ & g_j(\mathbf{x}) \le 0, \quad j = 1,\dots,I, \end{aligned}$$

where, as before, all functions f, h_i and g_j are in C^1, and f is bounded below within the constraint region. We shall refer to this problem as the primal problem and to all points $\mathbf{x} \in \mathbb{R}^d$ that satisfy the specified constraints as primal feasible points. Further, let $\mathbf{x}^*$ be a (global) minimizer of this function and $P^* = f(\mathbf{x}^*)$ be the corresponding optimal value. One can then introduce Lagrange multipliers $\mu_1,\dots,\mu_E \in \mathbb{R}$ and $\lambda_1,\dots,\lambda_I \in \mathbb{R}$ for the above constraints and define the associated Lagrangian as:

$$L(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}) \;=\; f(\mathbf{x}) + \sum_{i=1}^{E}\mu_i h_i(\mathbf{x}) + \sum_{j=1}^{I}\lambda_j g_j(\mathbf{x}).$$

Notice that by setting $\nabla L(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}) = \mathbf{0}$, we recover one of the KKT conditions for optimality of a solution to the above problem (see fifth condition in Proposition 1 in previous lecture). Next, we define the dual function for the problem as:

$$D(\boldsymbol{\mu},\boldsymbol{\lambda}) \;=\; \min_{\mathbf{x}\in\mathbb{R}^d} L(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}).$$

It can be shown that the dual function is always concave in $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ (i.e., a negative of a convex function in $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$) even if the primal problem is not convex (this follows from the dual function being a point-wise minimum of linear functions in $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$). It can further be shown that for any $\boldsymbol{\mu}$ and $\boldsymbol{\lambda} \ge \mathbf{0}$, the dual function serves as a lower bound on the optimal primal value P^*, as seen below:

$$D(\boldsymbol{\mu},\boldsymbol{\lambda}) \;\le\; L(\mathbf{x}^*,\boldsymbol{\mu},\boldsymbol{\lambda}) \;=\; f(\mathbf{x}^*) + \sum_{i=1}^{E}\mu_i h_i(\mathbf{x}^*) + \sum_{j=1}^{I}\lambda_j g_j(\mathbf{x}^*) \;\le\; f(\mathbf{x}^*) \;=\; P^*,$$

where we use the fact that the minimizer $\mathbf{x}^*$ of the primal problem is feasible (i.e., $h_i(\mathbf{x}^*) = 0$ for all i and $g_j(\mathbf{x}^*) \le 0$ for all j). Also, note that the above lower bound is useful only if $D(\boldsymbol{\mu},\boldsymbol{\lambda}) > -\infty$; let us refer to all $\boldsymbol{\mu},\boldsymbol{\lambda} \ge \mathbf{0}$ for which this is true as dual feasible and denote the set of all such points $(\boldsymbol{\mu},\boldsymbol{\lambda})$ as $\mathcal{D}$. If we now wish to find a dual feasible point that yields the tightest such lower bound, we have

$$\max_{(\boldsymbol{\mu},\boldsymbol{\lambda})\in\mathcal{D}} D(\boldsymbol{\mu},\boldsymbol{\lambda}) \;=\; \max_{(\boldsymbol{\mu},\boldsymbol{\lambda})\in\mathcal{D}} \min_{\mathbf{x}\in\mathbb{R}^d} L(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}). \tag{D}$$

This is called the (Lagrangian) dual problem associated with (P). As mentioned above this will always result in a convex optimization problem (i.e., in a maximization of a concave function) even if the primal problem is not convex, and hence can be solved efficiently. Let us use D^* to denote the optimal value of this dual problem. The following is then clearly seen.

Proposition 1 (Weak duality). $D^* \le P^*$.

If the optimal primal and dual values do indeed match for an optimization problem, then the problem is said to exhibit *strong duality.* This is the case with a large number of convex optimization problems such as those with linear constraints; in fact, these are precisely the problems where the KKT conditions described in the previous lecture are sufficient for optimality (see Slater's condition in [3] for a characterization of such problems). In practice, problems that exhibit strong duality give us an alternate route for designing efficient solvers for the problem by considering the dual problem instead of the primal one. Below, we give examples of dual problems for some standard constrained optimization problems.

Example 2. *Consider the following linear optimization problem over $\mathbb{R}^d$ with n linear equality constraints:*

$$\min_{\mathbf{x}\in\mathbb{R}^d} \mathbf{c}^\top \mathbf{x} \quad \textit{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \tag{P1}$$

where $\mathbf{c} \in \mathbb{R}^d$, $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^n$. Clearly this is a convex optimization problem and admits strong duality. Introducing Lagrange multipliers $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n] \in \mathbb{R}^n$, the Lagrangian associated with the above problem is given by $L(\mathbf{x}, \boldsymbol{\mu}) = \mathbf{c}^\top \mathbf{x} + \boldsymbol{\mu}^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$. The corresponding dual function takes the form $D(\boldsymbol{\mu}) = \min_{\mathbf{x}\in\mathbb{R}^d} L(\mathbf{x}, \boldsymbol{\mu}) = \min_{\mathbf{x}\in\mathbb{R}^d} (\mathbf{c}+\mathbf{A}^\top\boldsymbol{\mu})^\top \mathbf{x} - \boldsymbol{\mu}^\top\mathbf{b}$. Notice that if $\mathbf{c}+\mathbf{A}^\top\boldsymbol{\mu} \neq \mathbf{0}$, then $D(\boldsymbol{\mu}) = -\infty$. Otherwise, $D(\boldsymbol{\mu}) = -\boldsymbol{\mu}^\top\mathbf{b}$. Hence, the only values of $\boldsymbol{\mu}$ that are (dual) feasible are those for which $\mathbf{c} + \mathbf{A}^\top\boldsymbol{\mu} = \mathbf{0}$. The dual problem involving maximization of D over all such feasible $\boldsymbol{\mu}$ then becomes:

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^n} -\boldsymbol{\mu}^\top\mathbf{b} \quad \textit{s.t.} \quad \mathbf{A}^\top\boldsymbol{\mu} = -\mathbf{c}. \tag{D1}$$

Clearly by strong duality, the optimal value of the above dual problem (D1) is same that of the primal problem (P1). Moreover, note that there are as many variables in (D1) as there are constraints in (P1), and as many constraints in (D1) as the there are variables in (P1).

Example 3. *Let us consider a slight variant of the previous problem, now with additional non-negativity constraints on $\mathbf{x}$; this a linear program in standard form that occurs often in practice.*

$$\min_{\mathbf{x}\in\mathbb{R}^d} \mathbf{c}^\top \mathbf{x} \quad \textit{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \;\; \mathbf{x} \geq \mathbf{0}. \tag{P2}$$

where $\mathbf{c} \in \mathbb{R}^d$, $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^n$. Introducing Lagrange multipliers $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n] \in \mathbb{R}^n$ for the equality constraints and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_d] \in \mathbb{R}^d$ for the inequality constraints, the Lagrangian for the above problem takes the form $L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \mathbf{c}^\top\mathbf{x} + \boldsymbol{\mu}^\top(\mathbf{A}\mathbf{x} - \mathbf{b}) - \boldsymbol{\lambda}^\top\mathbf{x}$. The corresponding dual function is $D(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \min_{\mathbf{x}\in\mathbb{R}^d} L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \min_{\mathbf{x}\in\mathbb{R}^d} (\mathbf{c} + \mathbf{A}^\top\boldsymbol{\mu} - \boldsymbol{\lambda})^\top\mathbf{x} - \boldsymbol{\mu}^\top\mathbf{b}$. As before, $D(\boldsymbol{\mu}) = -\infty$ if $\mathbf{c} + \mathbf{A}^\top\boldsymbol{\mu} - \boldsymbol{\lambda} \neq \mathbf{0}$ and $D(\boldsymbol{\mu}) = -\boldsymbol{\mu}^\top\mathbf{b}$ otherwise; this gives us the following dual problem:

$$\max_{\boldsymbol{\mu},\boldsymbol{\lambda}\,\in\,\mathbb{R}^n} -\boldsymbol{\mu}^\top\mathbf{b} \quad \textit{s.t.} \quad \mathbf{A}^\top\boldsymbol{\mu} - \boldsymbol{\lambda} = -\mathbf{c}, \;\; \boldsymbol{\lambda} \geq \mathbf{0}.$$

It is easy to verify that this is equivalent to:

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^n} -\boldsymbol{\mu}^\top\mathbf{b} \quad \textit{s.t.} \quad \mathbf{A}^\top\boldsymbol{\mu} \geq -\mathbf{c}. \tag{D2}$$

Notice that unlike the primal problem, the dual problem here has no equality constraints.

Example 4. *The final example that we consider is a quadratic program:*

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\mathbf{x}^\top\mathbf{x} \quad \textit{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \tag{P3}$$

where $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^n$. Here again strong duality holds. Introducing Lagrange multipliers $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n] \in \mathbb{R}^n$, we have $L(\mathbf{x}, \boldsymbol{\mu}) = \frac{1}{2}\mathbf{x}^\top\mathbf{x} + \boldsymbol{\mu}^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$. In this case all values of $\boldsymbol{\mu}$ are (dual) feasible. Hence the value of $\mathbf{x}$ that minimizes the Lagrangian can be obtained by setting $\nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\mu}) = \mathbf{0}$ (note that the Hessian of L w.r.t. $\mathbf{x}$ is p.d. satisfying the sufficient condition for optimality). This give us $D(\boldsymbol{\mu}) = -\frac{1}{2}\boldsymbol{\mu}^\top\mathbf{A}\mathbf{A}^\top\boldsymbol{\mu} - \mathbf{b}^\top\boldsymbol{\mu}$. The dual optimization problem is then simply:

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^n} -\frac{1}{2}\boldsymbol{\mu}^\top\mathbf{A}\mathbf{A}^\top\boldsymbol{\mu} - \mathbf{b}^\top\boldsymbol{\mu}. \tag{D3}$$

Interestingly, while the primal problem is constrained, the associated dual problem has no constraints and moreover, does not depend on the number of variables in the primal problem. Clearly, in settings where the primal problem has a large large number of variables but fewer number of constraints, it would prove more efficient to solve the equivalent dual problem instead of the primal problem.

3 Conclusion

We have thus come to the end of this three part of tutorial on optimization. While the material covered in the tutorial aimed at giving a high-level working knowledge of basic tools in this field, we believe that equipped with the preview provided here, the reader would now be able to move to other advanced topics essential for solving optimization problems that arise in today's engineering and computer science applications.

References

[1] D. Luenberger and Y. Ye. *Linear and Nonlinear Programming.* Springer, 3nd edition, 2008.

[2] S. Boyd. E364b, Stanford University, Lecture Slides: Subgradient Methods for Constrained Problems. `http://stanford.edu/class/ee364b/lectures/constr_subgrad_slides.pdf`.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

Practice Exercise Questions

1. **Lagrangian dual and projected gradient descent.** Consider the following constrained optimization problem over $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^n$:

$$\min_{\mathbf{x}\in\mathbb{R}^d, \mathbf{y}\in\mathbb{R}^n} \frac{1}{2}\mathbf{x}^\top \mathbf{x} + \frac{1}{n}\sum_{i=1}^{n} y_i$$

s.t.

$$\mathbf{a}_i^\top \mathbf{x} + y_i \geq 1, \;\; i = 1, \dots, n,$$

$$y_i \geq 0, \;\; i = 1, \dots, n,$$

where each $\mathbf{a}_i \in \mathbb{R}^d$.

(a) Determine the Lagrangian dual of the above problem. What is the difference between the optimal value of the above primal problem and that of the dual problem (i.e., the *duality gap*)?

(b) Design a projected gradient descent solver for the dual optimization problem with a closed-form expression for the projection step. Are there values of d and n for which solving the dual problem using your proposed projected gradient method will be computationally more efficient than solving the primal problem using a suitable projected gradient solver?