Seminar In Nonlinear Optimization

Constrained Optimization and Duality

Applications To Machine Learning

Technical University Of Berlin
Department Of Mathematics

Master in Scientific Computing

Students:

Moussa Atwi
ID: 464241
Surya Iyer
ID: 464252

Teacher:

Prof. Dr. Dietmar Hömberg

July 20, 2022

Contents

Introduction

In many Machine Learning applications such as Support Vector Machine (SVM) and nonnegative regression, the optimization problems are *constrained*. In other words, we seek to find solutions over a *feasible region* that is defined by the problem constraints. These settings present a challenge, since vanilla gradient descent and other related methods do not work without certain modifications. In general, there are two approaches to solving constrained optimization problems:

- **Primal approach**: We attempt to modify gradient descent so that we stay within the feasible region.
- **Dual approach**: We use *Lagrangian relaxation* and formulate the *dual problem* in which primal constraints are converted into dual variables. This can, in many cases, be easier to solve.

The complexity of the problem depends on the structure of its constraints. Fortunately, many Machine Learning applications involve two simple types of constraints:

- **Linear and convex constraints**: Linear constraints are of the form $f(x) \leq b$ or of the form $f(x) = c$ where $f(x)$ is a linear function; convex constraints are of the form $h(x) \leq d$, where $h(x)$ is convex.
- **Norm constraints**: In ML problems such as Principal Component Analysis (PCA), Spectral Value Decomposition (SVD) and Spectral Clustering, we seek to optimize the objective function $F(x)$ subject to constraints $\|x\|^2 = 1$.

Several methods for handing constrained optimization will be discussed, such as coordinate descent, Lagrangian relaxation and barrier method.

Our seminar report discusses both primal and dual methods for constrained optimization, as well as some Machine Learning applications.

1 The General Optimization Problem

Consider the minimization problem:

$$\begin{aligned} &\text{minimize} && F(\bar{\omega}) \\ &\text{subject to:} && \\ & && f_i(\bar{\omega}) \leq 0, \quad i = 1, \ldots, m \\ & && h_i(\bar{\omega}) = 0, \quad i = 1, \ldots p, \end{aligned}$$

where $\bar{\omega} = [\omega_1, \ldots, \omega_n]^\intercal \in \mathbb{R}^n$.

Note that, each equality constraint $h_i(\bar{\omega}) = 0$ can be expressed as the intersection of two **linear** inequality constraints $h_i(\bar{\omega}) \leq 0$ and $-h_i(\bar{\omega}) \leq 0$. For simplicity, we will drop equality constraints and the discussion will be centered around inequality constraints. Therefore, the general inequality constrained problem will be as follows:

$$\begin{aligned} &\text{minimize} && F(\bar{\omega}) \\ &\text{subject to:} && \\ & && f_i(\bar{\omega}) \leq 0, \quad i = 1, \ldots, m \end{aligned}$$

2 Lagrangian Relaxation and Duality

Lagrangian relaxation is an approach whereby the constraints of an optimization problem are relaxed, while penalizing their violation within the objective function. We will relax these constraints by associating a penalty parameter with each of these m-inequality constraints that will call $\bar{\alpha} = [\alpha_1 ... \alpha_m]^\intercal \in \mathbb{R}^m$. The relaxation of these constraints by moving them into the objective function will generate a function called Lagrangian denoted by $H(\bar{\omega}, \bar{\alpha})$. Here, $\bar{\omega}$ contains primal variables and $\bar{\alpha}$ contains Lagrangian multipliers or dual variables. [3]

$$H(\bar{\omega}, \bar{\alpha}) = F(\bar{\omega}) + \sum_{i=1}^{m} \alpha_i f_i(\bar{\omega})$$

2.1 Lagrange Relaxation Problem

The Lagrange dual function or the dual function is the function $L : \mathbb{R}^m \longrightarrow \mathbb{R}$ such that:

$$L(\bar{\alpha}) = \min_{\bar{\omega}} H(\bar{\omega}, \bar{\alpha}) = \min_{\bar{\omega}} F(\bar{\omega}) + \sum_{i=1}^{m} \alpha_i f_i(\bar{\omega})$$

subject to :

No constraints on $\bar{\omega}$

It is important to note that each α_i is **nonnegative** to ensure that violations of the constraints are penalized. We want to pay a penalty only when a constraint is violated, i.e. $f_i(\bar{\omega}) > 0$. In this case, we would like the "penalty" $\alpha_i f_i(\bar{\omega}) \geq 0$. In order to do so, the condition $\bar{\alpha} \geq 0$ must be satisfied.

- $L(\bar{\alpha})$ may take on the value $-\infty$, for example when $F(\bar{\omega}) = \bar{\omega}$ and $\bar{\alpha}$ is zero.
- The dual function is always concave, pointwise min of affine functions. The Lagrangian objective function $H(\bar{\omega}, \bar{\alpha})$ is linear **in the dual variables**, therefore the concavity with respect to dual variables is always satisfied.

2.2 Dual Bound

Consider the original minimization problem:

$$\begin{aligned} P^* = \text{minimize} \quad & F(\bar{\omega}) \\ \text{subject to:} \quad & f_i(\bar{\omega}) \leq 0, \quad i = 1, \ldots, m \end{aligned}$$

This problem is referred to as the **primal problem** in optimization parlance. We introduce the notation P^* to denote **its optimal solution** i.e., $P^* = F(\bar{\omega}^*)$. **Let us examine why the Lagrangian relaxation problem provides a lower bound on** the solution to the original optimization problem P^*.

Theorem 2.1. *Let $\bar{\omega}^*$ be the optimal solution to the primal problem. [3] If $\bar{\alpha} \in \mathbb{R}^m, \bar{\alpha} \geq 0$ then,*

$$L(\bar{\alpha}) \leq P^*.$$

The Lagrangian relaxation always gives a lower bound on the optimal solution.

Proof. The optimal solution $\bar{\omega}^*$ is feasible, then the penalty $\alpha_i f_i(\bar{\omega}^*) \leq 0$.

$$\begin{aligned} L(\bar{\alpha}) = & \min_{\bar{\omega}} F(\bar{\omega}) + \sum_{i=1}^{m} \alpha_i f_i(\bar{\omega}) \\ \leq & F(\bar{\omega}^*) + \underbrace{\sum_{i=1}^{m} \alpha_i f_i(\bar{\omega}^*)}_{\leq 0} \quad [\omega^* \text{ might not be optimal for relaxation}] \\ \leq & F(\bar{\omega}^*) = P^* \end{aligned}$$

□

It follows that for any feasible point $\bar{\omega}$ the value of the dual is always the lower.

$$L(\bar{\alpha}) \leq P^* = F(\bar{\omega}^*) \leq F(\bar{\omega}).$$

2.3 Find The Best Lower Bound

One can tighten the above bound by maximizing $L(\bar{\alpha})$ over all nonnegative $\bar{\alpha}$ and formulating the **dual problem**:

$$\begin{aligned} D^* &= \max_{\bar{\alpha} \geq 0} L(\bar{\alpha}) \\ &= \max_{\bar{\alpha} \geq 0} \min_{\bar{\omega}} [F(\bar{\omega}) + \sum_{i=1}^{m} \alpha_i f_i(\bar{\omega})] \\ &= \max_{\bar{\alpha} \geq 0} \min_{\bar{\omega}} H(\bar{\omega}, \bar{\alpha}) \end{aligned}$$

Lagrange dual problem is a search for best lower bound on P^*. Note that, $\bar{\alpha}$ is **dual feasible** if $\bar{\alpha} \geq 0$ and $L(\bar{\alpha}) > -\infty$.

We summarize the relationship between the primal and the dual as follows:

$$D^* = L(\bar{\alpha}^*) \leq P^*$$

This result is referred to as that of **weak duality**. It is noteworthy that the Lagrangian optimization problem is a *minmax* problem containing disjoint **minimization and maximization variables**. Minimization and maximization is done in a specific order that **matters** in a minmax optimization problem.

Theorem 2.2. ***Weak Maxmin inequality*** *[2]*
For any $H : X \times Y \to \mathbb{R}$ *we have:*

$$\max_y \min_x H(x, y) \leq \min_x \max_y H(x, y).$$

Proof. Let $f(y) = \min_x H(x, y)$. Then, for any $x \in X$, $y \in Y$ we clearly have :

$$f(y) \leq H(x, y) \leq \max_y H(x, y)$$

that gives

$$\max_y f(y) \leq \max_y H(x, y)$$

hence, for any $x \in X$ we get

$$\max_y f(y) \leq \min_x \max_y H(x, y).$$

□

The above theorem states that "min-max" is an upper bound on "max-min" of a function containing both minimization and maximization variables. Furthermore, **strict equality** occurs when the function is **convex in its minimization variables and also concave in the maximization variables**.

2.4 The Min-max Primal Formulation

We have seen the importance of ordering of minimization and maximization in minimax problems. Consider the minmax objective function

$$H(\bar{\omega}, \bar{\alpha}) = F(\bar{\omega}) + \sum_{i=1}^{m} \alpha_i f_i(\bar{\omega}) \tag{1}$$

where, $\bar{\omega}$ contains the minimization variables and $\bar{\alpha}$ contains the maximization variables. While the dual computes $\underset{\bar{\alpha}\geq 0}{max}\ \underset{\bar{\omega}}{min}\ H(\bar{\omega}, \bar{\alpha})$ (which is a lower bound on the primal), reversing the order to $\underset{\bar{\omega}}{min}\ \underset{\bar{\alpha}\geq 0}{max}\ H(\bar{\omega}, \bar{\alpha})$ always yields the original (primal) optimization problem **irrespective of whether the original problem has a convex objective function or convex constraints**.

Lemma 2.3. ***Minmax Primal Formulation***
The unconstrained minimax problem $\underset{\bar{\omega}}{min}\ \underset{\bar{\alpha}\geq 0}{max}\ H(\bar{\omega}, \bar{\alpha})$ *is equivalent to the original, unrelaxed primal formulation irrespective of the convexity structure of the original problem.*

$$P^* = \underset{\bar{\omega}}{min}\ \underset{\bar{\alpha}\ \geq 0}{max}\ H(\bar{\omega}, \bar{\alpha}) \tag{2}$$

Proof. Consider the Lagrangian objective function $H(\bar{\omega}, \bar{\alpha})$. Suppose we have $\bar{\omega}$ that violates at least one of the original primal inequality constraints i.e., $f_i(\bar{\omega}) > 0,\ i = 1, 2, ...$ Then, $\underset{\bar{\alpha}\geq 0}{max}\ H(\bar{\omega}, \bar{\alpha}) = \infty$. . This is achieved by setting the corresponding α_i of the violated constraint to ∞.

$$\underset{\bar{\alpha}\geq 0}{max}\ H(\bar{\omega}, \bar{\alpha}) = \begin{cases} F(\bar{\omega}) & \text{when } f_i(\bar{\omega}) \leq 0 \text{ for all } i \\ \infty & \text{otherwise} \end{cases}$$

For $\bar{\omega}$ satisfying all the inequality constraints i.e., $f_i(\bar{\omega}) \leq 0$ then, $\bar{\alpha}_i f_i(\bar{\omega}) \leq 0$ for each i. Therefore, $H(\bar{\omega}, \bar{\alpha})$ will be maximized with respect to $\bar{\alpha}$ only when the value of α_i is set to zero for each i satisfying $f_i(\bar{\omega}) < 0$. □

Duality results of Lagrangian relaxation can be derived from the more **general minmax theorem** in mathematics that states: weak maxmin inequality implies weak duality

$$\begin{aligned} P^* &= \min_{\bar{\omega}} \max_{\bar{\alpha}\ \geq 0} H(\bar{\omega}, \bar{\alpha}) \\ &\geq \max_{\bar{\alpha}\ \geq 0} \min_{\bar{\omega}} H(\bar{\omega}, \bar{\alpha}) = D^* \end{aligned}$$

Furthermore, the minmax theorem also states that **strict equality occurs** when the optimization function is convex in the minimization (primal) variables and concave in the maximization (dual) variables.

2.5 Strong Duality

What types of optimization problems are such that their Lagrangian relaxations show strict equality between primal and dual solutions? Firstly, the function $H(\bar{\omega}, \bar{\alpha})$ is linear in the maximization variables, and therefore concavity with respect to maximization variables is always satisfied. Secondly, the function $H(\bar{\omega}, \bar{\alpha})$ is a sum of $F(\bar{\omega})$ and nonnegative multiples of the various $f_i(\bar{\omega})$ for $i \in \{1 \dots m\}$. Therefore, if $F(\bar{\omega})$ and each of $f_i(\bar{\omega})$ are convex in $\bar{\omega}$, then $H(\bar{\omega}, \bar{\alpha})$ will be convex in the minimization variables. This is the primary pre-condition for **strong duality:**

Lemma 2.4. ***Strong Duality*** *Consider the convex optimization problem:*

$$\begin{aligned} P^* = \; & \textit{minimize} && F(\bar{\omega}) \\ & \textit{subject to:} && f_i(\bar{\omega}) \leq 0, \quad i = 1, \dots, m \end{aligned}$$

where $F(\bar{\omega})$ and each $f_i(\bar{\omega})$ are convex functions. Then, strong duality $P^* = D^*$ ***usually*** *holds, but not always.*

For example, consider the convex problem without strong duality

$$\begin{aligned} \text{minimize} \quad & e^{-\omega_1} \\ \text{subject to:} \quad & \omega_1^2/\omega_2 \leq 0 \\ & \omega_2 > 0 \end{aligned}$$

The additional conditions needed under the strong duality holds are called **constraint qualifications**. [2]

- **Slater's constraint qualification:** Suppose that the primal problem is convex and there exists one strictly feasible point such that $f_i(\bar{\omega}) < 0$ for each i. Then, Slater's condition holds and strong duality holds.
- For any **affine inequality constraints**, $f_i(\bar{\omega}) \leq 0$ is sufficient. That means, strict inequality is **not** necessary if $f_i(\bar{\omega})$ is an affine constraint.

Many optimization problems in Machine Learning such as SVM and logistic regression satisfy strong duality. The constraint qualifications hold by default.

2.5.1 Complementary Slackness

Consider the primal problem (not necessarily convex). Suppose strong duality holds and let $\bar{\omega}^*$ be primal optimal and $\bar{\alpha}^*$ be dual optimal. [2] Then, we get an interesting relationship between the optimal Lagrange multiplier α_i^* and the ith constraint at optimum $f_i(\bar{\omega}^*)$

$$\alpha_i^* \, f_i(\bar{\omega}^*) = 0, \qquad i = 1, \dots, m.$$

Proof. Under strong duality we have

$$\begin{aligned} F(\bar{\omega}^*) = P^* = D^* = L(\bar{\alpha}^*) &= \min_{\bar{\omega}} H(\bar{\omega}, \bar{\alpha}^*) \\ &\leq H(\bar{\omega}^*, \bar{\alpha}^*) \\ &= F(\bar{\omega}^*) + \underbrace{\sum_{i=1}^{m} \alpha_i^* f_i(\bar{\omega}^*)}_{\leq 0} \\ &\leq F(\bar{\omega}^*). \end{aligned}$$

Each term of the sum $\sum_{i=1}^{m} \alpha_i^* f_i(\bar{\omega}^*)$ must be zero. □

Theorem 2.5. ***Kuhn-Tucker Optimality Conditions***
Consider a ***convex*** *optimization primal problem such that:*

- *Strong duality holds*
- *F and f_i are differentiable.*

Then, a solution $\bar{\omega}$ is optimal for the primal and a solution $\bar{\alpha}$ is optimal for the dual, if and only if:

- ***Feasibility:*** *$\bar{\omega}$ is feasible for the primal by satisfying each $f_i(\bar{\omega}) \leq 0$ and $\bar{\alpha}$ is feasible for the dual by being nonnegative.*
- ***Complementary slackness:*** *We have $\alpha_i f_i(\bar{\omega}) = 0$ for each $i \in \{1, \dots m\}$.*
- ***Stationarity:*** *The primal and the dual variables are related as follows:*

$$\nabla F(\bar{\omega}) + \sum_{i=1}^{m} \alpha_i \nabla f_i(\bar{\omega}) = \bar{0}$$

For a **convex** optimization problem, **KKT** conditions are **necessary and sufficient** for primal and dual optimal points with zero duality gap.

Remark 2.6. *The stationary conditions relate the primal and dual variables. We will refer to them as* ***primal-dual (PD)*** *constraints. They are often used to:*

- *Eliminating the primal variables $\bar{\omega}$ from the Lagrangian to obtain a pure maximization problem $L(\bar{\alpha})$ in terms of the dual variable $\bar{\alpha}$.*
- *Infer an optimal primal solution $\bar{\omega}$ from the optimal dual solution $\bar{\alpha}$.*

3 Coordinate Descent Method

Coordinate descent is an optimization algorithm that optimizes the multivariate objective function along one direction at a time, i.e., one variable at a time to find the minimum. For example if we consider the objective function $J(\bar{\omega})$ of $d-$dimensional vector variables. We can try to optimize a single component ω_i from the vector $\bar{\omega}$, while holding all the other components fixed to their values $\bar{\omega}^t$ in the tth iteration. This corresponds to the following univariate optimization problem:

$$\bar{\omega}^{t+1} = \underset{\omega_i \text{ varies only, ith component of } \bar{\omega}}{argmin\ J(\bar{\omega})} \qquad [\text{All parameters except } \omega_i \text{ are fixed to } \bar{\omega}^t]$$

One cycles through the variables one at a time, until convergence is achieved. For example, if no improvement occurs during a cycle of optimizing each variable, then it means that the solution is a global optimum.

3.1 Coordinate Descent for Convex Optimization Problem

When using coordinate descent over a convex set, a very useful observation is that any univariate convex set is a continuous interval, and the corresponding variable $\bar{\omega}$ can be expressed in the form of the box constraint $l_i \leq \bar{\omega} \leq u_i$. This follows from the fact that a convex set is defined as any set such that any line passing through it must have exactly one continuous region belonging to the set.

Example 3.1. *Consider the 3-dimensional optimization problem:*

$$\begin{aligned} &\textit{minimize} && F(\omega_1, \omega_2, \omega_3) \\ &\textit{subject to:} && \\ & && w_1^2 - w_1 \cdot w_2 + \frac{w_2^2}{4} + 3w_2 \cdot w_3 + 4w_3^2 \leq 4 \\ & && 2w_1 + w_2 - 3w_3 \leq 4 \end{aligned}$$

Now we seek to perform coordinate descent, and we are trying to compute the optimum value w_1 so that $F(w_1, w_2, w_3)$ is minimized while holding w_2 and w_3 fixed to 2 and 0, respectively. Plugging in these values we obtain the following:

$$\begin{cases} w_1^2 - 2w_1 - 3 = (w_1 - 3)(w_1 + 1) \leq 0 \\ w_1 \leq 1 \end{cases}$$

Note that the first and second constraints imply that $w_1 \in [-1, 3]$ and $(-\infty, 1]$ respectively, which means $w_1 \in [-1, +1]$. The subproblem reduces to optimizing a univariate convex function $G(w_1) = F(w_1, 2, 0)$ over an interval $[-1, +1]$.

$$\begin{aligned} &\textit{minimize} && G(\omega_1) \\ &\textit{subject to:} && \\ & && -1 \leq w_1 \leq 1 \end{aligned}$$

4 ML Application: Formulating the SVM Dual

In order to illustrate how duality is used in machine learning, we consider the support vector machine (SVM), expressed as a **convex** minimization problem with slack variables $\xi_1 \dots \xi_n$ for non-linearly separable datasets as follows:

$$\begin{aligned}
\text{Minimize} \quad & J = \frac{1}{2}\left\|\bar{W}\right\|^2 + C\sum_{i=1}^{n}\xi_i \\
& \text{subject to :} \\
& 1 - y_i\left[\bar{W}\cdot X_i^T\right] - \xi_i \leq 0 \quad \forall i \in \{1\dots n\} \quad \text{[Margin Constraints]} \\
& -\xi_i \leq 0 \quad \forall i \in \{1\dots n\} \quad \text{[Nonnegativity Constraints]}
\end{aligned}$$

The Lagrangian relaxation is then given by

$$L(\bar{\alpha},\bar{\gamma}) = \min_{\bar{W},\xi_i} J_r = \min_{\bar{W},\xi_i} \frac{1}{2}\|\bar{W}\|^2 + C\sum_{i=1}^{n}\xi_i + \underbrace{\sum_{i=1}^{n}\alpha_i\left(1 - y_i\left(\bar{W}\cdot\bar{X}_i^T\right) - \xi_i\right)}_{\text{Relax margin constraint}} - \underbrace{\sum_{i=1}^{n}\gamma_i\xi_i}_{\text{Relax}-\xi_i}$$

where J_r is the relaxed objective function, and the Lagrangian variables α_i and γ_i are nonnegative. The primal and dual problems are related by

$$\begin{aligned}
P^* &= \min_{\bar{W},\xi_i}\max_{\alpha_i,\gamma_i\geq 0} J_r \\
&\geq \max_{\alpha_i,\gamma_i\geq 0}\min_{\bar{W},\xi_i} J_r = D^*.
\end{aligned}$$

Strong duality holds as the objective function J is convex and we have affine constraints, which satisfy slater's constraint qualification for $\bar{W} = 0$ and $\xi_i = 2$.

We seek to eliminate the primal variables in order to formulate the dual problem purely in terms of dual variables using the (PD) constraints by setting the gradient of J_r with respect to the primal variables to zero, as seen here:

$$\frac{\partial J_r}{\partial \bar{W}} = \bar{W} - \sum_{i=1}^{n}\alpha_i y_i \bar{X}_i^T = \bar{0}, \tag{3a}$$

$$\frac{\partial J_r}{\partial \xi_i} = C - \alpha_i - \gamma_i = 0, \quad \forall i \in \{1\dots n\} \tag{3b}$$

Substituting the (PD) constraints, we get:

$$J_r = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \bar{X}_i\cdot\bar{X}_j$$

And finally, Equation (3b) gives:

$$\gamma_i = C - \alpha_i \geq 0$$

Note that the variables α_i satisfy the box constraints $0 \leq \alpha_i \leq C$. This gives us the **SVM dual problem**:

$$D^* = \max_{0 \leq \alpha_i \leq C} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \bar{X}_i \cdot \bar{X}_j$$
$$= \min_{0 \leq \alpha_i \leq C} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \bar{X}_i \cdot \bar{X}_j - \sum_{i=1}^{n} \alpha_i$$

Beside the fact that the dual problem (in minimization form) is always convex, the quadratic term can be expressed as $\bar{\alpha}^T B B^T \bar{\alpha}$, where B is $n \times d$ matrix. Matrices of the form BB^T are always positive semi-definite, and hence this is a **convex optimization problem**.

4.1 Optimization Algorithms for the SVM Dual

With the box constraint structure of the variables α_i in mind, we now consider two algorithms for solving the SVM dual: Gradient descent and Coordinate descent.

4.1.1 Gradient Descent

We state the dual problem in minimization form with box constraints:

$$\text{Minimize} \quad L_D = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \bar{X}_i \cdot \bar{X}_j - \sum_{i=1}^{n} \alpha_i$$
$$\text{subject to:}$$
$$0 \leq \alpha_i \leq C$$

The partial derivative of L_D with respect to α_k is as follows:

$$\frac{\partial L_D}{\partial \alpha_k} = y_k \sum_{s=1}^{n} y_s \alpha_s \bar{X}_k \cdot \bar{X}_s - 1 \quad \forall k \in \{1 \ldots n\} \tag{4}$$

One can use the standard gradient-descent procedure at iteration t:

$$\bar{\alpha}_t \Leftarrow \bar{\alpha}_t - \eta \left[\frac{\partial L_D}{\partial \bar{\alpha}_t} \right]$$

This update might result in some intermediate values of α_k being infeasible. To solve this, we project the infeasible components of α_k to the feasible box. In other words, if the components are negative, they are reset to 0. Similarly, if they exceed C, then they are reset to C. Therefore, we initialize the vector of Lagrangian parameters $\bar{\alpha} = [\alpha_1 \ldots \alpha_n]$ to an n-dimensional vector of 0s and perform the following update procedure with learning rate η:

repeat

Update $\alpha_k \Leftarrow \alpha_k + \eta \left[1 - y_k \sum_{s=1}^{n} y_s \alpha_s \bar{X}_k \cdot \bar{X}_s\right]$ for each $k \in \{1 \dots n\}$;

foreach $k \in \{1 \dots n\}$:

$\alpha_k \Leftarrow \min\{\alpha_k, C\}$;

$\alpha_k \Leftarrow \max\{\alpha_k, 0\}$;

endfor;

until convergence

The parameter vectors are updated all at once, which is not the case in coordinate descent.

4.1.2 Coordinate Descent

In coordinate descent, each intermediate value of α_k must be an optimum, i.e, the partial derivative of L_D with respect to each α_k should be 0. Setting Equation (4) to zero provides us with the following condition:

$$y_k \sum_{s=1}^{n} y_s \alpha_s \bar{X}_k \cdot \bar{X}_s - 1 = 0$$

Separating the component α_k from the others on the LHS, we get:

$$\alpha_k \left\|\bar{X}_k\right\|^2 y_k^2 = 1 - y_k \sum_{s \neq k} y_s \alpha_s \bar{X}_k \cdot \bar{X}_s$$

Note that $y_k^2 = 1$ because each $y_i \in \{-1, +1\}$:

$$\alpha_k = \frac{1 - y_k \sum_{s \neq k} y_s \alpha_s \bar{X}_k \cdot \bar{X}_s}{\left\|\bar{X}_k\right\|^2} = \alpha_k + \frac{1 - y_k \sum_{s=1}^{n} y_s \alpha_s \bar{X}_k \cdot \bar{X}_s}{\left\|\bar{X}_k\right\|^2}$$

And now we can reinterpret the above equation as an iterative update, in which α_k is updated with learning rate $\eta_k = 1/\left\|X_k\right\|^2$.

$$\alpha_k \Leftarrow \alpha_k + \eta_k \left[1 - y_k \sum_{s=1}^{n} y_s \alpha_s \bar{X}_k \cdot \bar{X}_s\right]$$

The update for coordinate and gradient descent is similar, except that it is done in a component-wise fashion with a component-specific learning rate:

repeat
foreach $k \in \{1 \dots n\}$:

$$\textbf{Update } \alpha_k \Leftarrow \alpha_k + \eta_k \left[1 - y_k \sum_{s=1}^{n} y_s \alpha_s \bar{X}_k \cdot \bar{X}_s\right]$$

$\alpha_k \Leftarrow \min\{\alpha_k, C\}$;
$\alpha_k \Leftarrow \max\{\alpha_k, 0\}$;
endfor;
until convergence

The coordinate descent procedure always yields faster convergence than gradient descent.

5 ML Application: Norm Constrained Optimization

Norm-constrained Optimization appears repeatedly in different types of ML problems, such as PCA, and SVD. One such formulation of these problems is of the form

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{k} \bar{x_i}^T A \bar{x_i} \\ & \text{subject to:} \\ & \|\bar{x_i}\|^2 = 1, \qquad \forall i \in \{1 \dots k\} \\ & \bar{x_1} \dots \bar{x_k} \quad \text{are mutually orthogonal} \end{aligned}$$

where A is a symmetric $d \times d$ matrix, and $\bar{x_1} \dots \bar{x_k}$ correspond to the d-dimensional vectors containing the optimization variables. We also assume $k \leq d$, as the problem will not have feasible solutions otherwise.

We introduce the Lagrangian multiplier $-\alpha_i$ for each equality constraint, in order to reinterpret the optimality condition as an eigenvector relation. Subsequently, one can write the Lagrangian relaxation as follows:

$$L(\bar{\alpha}) = \text{ Minimize }_{[\bar{x_1} \dots \bar{x_k} \text{ are orthogonal}]} \sum_{i=1}^{k} \bar{x_i}^T A \bar{x_i} - \sum_{i=1}^{k} \alpha_i \left(\|\bar{x_i}\|^2 - 1\right)$$

where the Lagrange multipliers are not necessarily nonnegative (because we have equality constraints). Setting the gradient of the Lagrangian with respect to each $\bar{x_i}$ to zero, we get the (PD) constraints:

$$A\bar{x_i} = \alpha_i \bar{x_i}, \qquad \forall i \in \{1 \dots k\}$$

Note that the constrains $A\bar{x_i} = \alpha_i \bar{x_i}$ implies that the feasible space for α_i is restricted to the d eigenvalues of A. This is the reason that we chose **not to relax** the orthogonality constraints on $\bar{x_1} \dots \bar{x_k}$, because the eigenvectors of a symmetric matrix are orthogonal. Substituting the (PD) constraints into the Lagrangian gives:

$$\begin{aligned} L(\bar{\alpha}) = & \text{ Minimize }_{[\bar{x}_1 \dots \bar{x}_k \text{ are orthogonal}]} \sum_{i=1}^{k} \alpha_i \bar{x}_i^T \bar{x}_i - \sum_{i=1}^{k} \alpha_i \left(\|\bar{x}_i\|^2 - 1 \right) \\ = & \text{ Minimize }_{[\text{Eigenvalues of } A]} \sum_{i=1}^{k} \alpha_i \end{aligned}$$

Clearly, $L(\bar{\alpha})$ is minimized over the smallest eigenvalues of A.
Therefore, one obtains the following trivial dual problem:

$$\begin{aligned} \text{Maximize} \quad & L(\bar{\alpha}) = \sum_{i=1}^{k} \alpha_i \\ & \text{subject to:} \\ & \alpha_1 \dots \alpha_k \quad \text{are smallest eigenvalues of } A \end{aligned}$$

Note that we did not assume the matrix A to be positive semidefinite, and this raises a question: Is there a duality gap? The answer is no: When we substitute the derived primal solution into the primal objective function, we find that the primal objective function is also the sum over the smallest eigenvalues.

$$P^* = \min \sum_{i=1}^{k} \bar{x_i}^T A \bar{x_i} = \min \sum_{i=1}^{k} \bar{x_i}^T \alpha_i \bar{x_i} = \min \sum_{i=1}^{k} \alpha_i \|\bar{x_i}\|^2 = \sum_{i=1}^{k} \alpha_i = D^*$$

6 Barrier methods

In contrast to the **Lagrangian relaxation** method, Barrier methods always maintain strict feasibility of intermediate solutions. Therefore, barrier methods are designed only for **inequality constraints** of the form $f_i(\bar{w}) \geq 0$. Consider the following optimization problem:

$$\begin{aligned} \text{Minimize} \quad & F(\bar{w}) \\ & \text{subject to:} \\ & f_i(\bar{w}) \geq 0, \qquad \forall i \in \{1 \dots m\} \end{aligned}$$

Then, the *logarithmic* barrier function $B(\bar{w}, \alpha)$ is well-defined only for feasible values of the parameter vector $\bar{w}$, and it is defined as follows:

$$B(\bar{w}, \alpha) = F(\bar{w}) - \alpha \sum_{i=1}^{m} \log \left(f_i(\bar{w}) \right)$$

The barrier function is convex if $F(\bar{w})$ is convex, and each $f_i(\bar{w})$ is **concave**. The logarithm is an increasing and concave function. Thus, its composition with a concave function is concave, and the negative logarithm is therefore convex.

Since the barrier function cannot be evaluated at $f_i(\bar{w}) = 0$, the method starts with points in the interior of the feasible region, and hence the Barrier method is also called the *Interior Point method*. Furthermore, one starts with large values of α in early iterations, and this value is reduced over time. One might be tempted to start with a small value of α to begin with, as this allows us to approach closer to the boundary. However, this can cause problems in gradient descent, because the barrier function becomes ill-conditioned near the boundary, and using small values of α early is bad for convergence.

If the optimal solution does not lie near the boundary, one often approaches the optimal solution fairly quickly, with smooth convergence. However, in these cases, the logarithmic term does not contribute to the barrier function, and so it reduces to just the primal function. The harder case to handle would be when the optimum lies near the boundary, since $\log(f_i(\bar{w}))$ increases rapidly to ∞, acting as a "barrier" and preventing approach to the optimum. Since we start with high values of α, convergence is rather slow in this case.

At any fixed value of α, gradient descent is performed on $\bar{w}$ to optimize the weight vector. For gradient descent, the gradient of the objective function is

$$\nabla_{\bar{w}} B(\bar{w}, \alpha) = \nabla F(\bar{w}) - \alpha \sum_{i=1}^{m} \frac{\nabla(f_i(\bar{w}))}{f_i(\bar{w})}$$

and setting this equation to zero yields the optimality condition, which we can compare to the (PD) constraint of the Lagrangian

$$\nabla_{\bar{w}} L(\bar{w}, \bar{\alpha}) = \nabla F(\bar{w}) - \sum_{i=1}^{m} \alpha_i \nabla(f_i(\bar{w})) = \bar{0},$$

where the Lagrangian relaxation is given by

$$L(\bar{w}, \bar{\alpha}) = F(\bar{w}) - \sum_{i=1}^{m} \alpha_i f_i(\bar{w}).$$

Upon comparing the equations, we observe that $\alpha / f_i(\bar{w})$ provides an estimate for the dual variables α_i. This means that we have $\alpha_i f_i(\bar{w}) = \alpha$, which is almost equivalent to the **complementary-slackness** condition of Lagrangian relaxation, except that instead of zero we have α. Therefore, at small values of α, the optimality conditions of the dual relaxation are **nearly satisfied** when one views the barrier function as a Lagrangian relaxation.

References

[1] Charu C. Aggarwal, Linear Algebra and Optimization for Machine Learning, Book, Springer, Cham.

[2] Boyd, S. and Vandenberghe, L. (2004) Convex Optimization, Book, Cambridge University Press, Cambridge.

[3] Bierlaire, M. (2015) Optimization: Principles and Algorithms, Book, EPFL Press, Lausanne.