5 The Primal-Dual Method

Originally designed as a method for solving linear programs, where it reduces weighted optimization problems to simpler combinatorial ones, the primal-dual method (PDM) has received much attention over the last years, as it can be generalized to more complex optimization settings and can be used to derive approximation schemes for NP-hard problems. PDM is a vast topic, and we can only give a very basic idea here. There are many flavors, versions and extensions to it.[1] In its most basic form, the main principle is to improve a feasible dual solution until the primal satisfies complementary slackness conditions, indicating optimality in cases where strong duality holds, or an approximate solution where it doesn't.

5.1 The primal-dual method for linear programs

Assume we have a primal minimization and a dual maximization problem in standard form, i.e.

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & \mathbf{c}^\mathsf{T}\mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} \geq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{aligned}$$

and

$$\begin{aligned} \max_{\mathbf{y}\in\mathbb{R}^m} \quad & \mathbf{b}^\mathsf{T}\mathbf{y} \\ \text{s.t.} \quad & \mathbf{A}^\mathsf{T}\mathbf{y} \leq \mathbf{c} \\ & \mathbf{y} \geq \mathbf{0} \end{aligned}$$

Recall the property of *complimentary slackness* for optimal solutions $\mathbf{x}^*, \mathbf{y}^*$ in LP:

Primal complimentary slackness (PCS) : At least one of $x_j^* = 0$ or $\sum_i a_{ij} y_i^* = c_j$ must hold.

Dual complimentary slackness (DCS) : At least one of $y_i^* = 0$ or $\sum_j a_{ij} x_j^* = b_i$ must hold.

The central observation is that if strong duality holds for all constraints in the primal and the dual for some $\mathbf{y}$, that $\mathbf{y}$ is in fact an optimal solution. Hence, CS can be used as a certificate for optimality. This leads to the original version of the PDM for LPs, which can be summarized thusly:

[1] If you are eager to find out more, there is an extensive recent review (http://ieeexplore.ieee.org/document/7298566/), with a free preprint (i.e. not peer-reviewed) version (https://arxiv.org/abs/1406.5429).

1. Find some feasible dual solution $\mathbf{y}$.
2. Given $\mathbf{y}$, find some $\mathbf{x}$ that minimizes the violation of complementary slackness in the primal.
3. If CS holds, $\mathbf{y}$ is optimal, and PDM terminates.
4. Otherwise, change $\mathbf{y}$ so as to improve the dual objective $\mathbf{b}^\mathsf{T}\mathbf{y}$, and go to 2. Note that at this point, the solution $\mathbf{x}$ obtained in step 2. is *not* necessarily feasible and might violate primal constraints!

Obviously, we require some way to find $\mathbf{x}$ in step 2., and a way to measure whether complementary slackness holds, and if not, to what degree it is violated. Given some feasible dual solution $\mathbf{y}$, let

$$I := \left\{ i \,\middle|\, y_i = 0 \right\}$$

be the set of all indices for which the dual variables are zero, and

$$J := \left\{ j \,\middle|\, \sum_{i=1}^{m} a_{ij} y_i = c_j \right\}$$

the set of all indices for which the dual constraints are binding. Obviously, I serves as an index for those primal constraints for which DCS holds because their associated dual variable y_i is zero, and J denotes the dual constraints that are binding for a given $\mathbf{y}$. The complements of those sets are denoted by $I^\complement$ and $J^\complement$. $J^\complement$ is the set of indices j for which the dual constraints are not binding, and hence x_j would have to be zero for PCS to hold. Likewise, $I^\complement$ is the set of indices i for which $y_i > 0$, and this the primal constraints would have to be binding in order for DCS to hold. The idea is therefore to construct a new optimization problem called the *restricted primal*, in which we try to reduce the slackness in the primal constraints and the "non-zeroness" of the primal variables x_j, $j \in J^\complement$ as much as we can. If and only if they are both zero, complementary slackness holds and $\mathbf{y}$ was an optimal solution. Using slack variables s_i to capture primal constraint violations, the restricted primal (RP) is defined as

$$\begin{aligned} f_{RP} = \min \sum_{i\in I^\complement} s_i + \sum_{j\in J^\complement} x_j & \\ \text{s.t. } \forall i \in I: \quad \textstyle\sum_j a_{ij} x_j \geq b_i & \\ \forall i \in I^\complement: \textstyle\sum_j a_{ij} x_j - s_i = b_i & \\ \mathbf{s} \geq \mathbf{0} & \\ \mathbf{x} \geq \mathbf{0} & \end{aligned}$$

In case the RP has a non-zero optimal value, we need to find a new feasible dual variable that leads to a better dual objective, thus getting us closer to an optimal solution and thus closer to full CS. We derive the *restricted dual* (RD) by our usual scheme,

$$
\begin{array}{|c|ccc|ccc|cc|}
\hline
 & & & & \multicolumn{3}{c|}{\min \begin{bmatrix}\mathbf{0}\\ \mathbf{1}\\ \mathbf{1}\end{bmatrix}^\intercal \begin{bmatrix}\mathbf{x}_J\\ \mathbf{x}_{J^\complement}\\ \mathbf{s}\end{bmatrix}} & & \\
\hline
 & & & & \mathbf{0} & \mathbf{0} & \mathbf{0} & & \\
 & & & & \leq & \leq & \leq & & \\
 & & & & \mathbf{x}_J & \mathbf{x}_{J^\complement} & \mathbf{s} & & \\
\hline
 & \mathbf{0} & \leq & \mathbf{z}_I & \mathbf{A}_{IJ} & \mathbf{A}_{IJ^\complement} & \mathbf{O} & \geq & \mathbf{b}_I \\
\max \mathbf{b}^\intercal \mathbf{z} & & & \mathbf{z}_{I^\complement} & \mathbf{A}_{I^\complement J} & \mathbf{A}_{I^\complement J^\complement} & -\mathbf{E} & = & \mathbf{b}_{I^\complement} \\
\hline
 & & & & \leq & \leq & \leq & & \\
 & & & & \mathbf{0} & \mathbf{1} & \mathbf{1} & & \\
\hline
\end{array}
$$

This leads to

$$
\begin{aligned}
f_{RD} = \max \mathbf{b}^\intercal \mathbf{z} & \\
\text{s.t. } \forall j \in J &: \textstyle\sum_i a_{ij} z_i \leq \mathbf{0} \\
\forall j \in J^\complement &: \textstyle\sum_i a_{ij} z_i \leq \mathbf{1} \\
\forall i \in I &: z_i \geq \mathbf{0} \\
\forall i \in I^\complement &: z_i \geq -1
\end{aligned}
$$

Notice that the last constraint does not seem to appear in the scheme; while the primal equality constraint imposes no restrictions on y_i for $i \in I^\complement$, due to $-\mathbf{E}$ we also have $-z_i \leq 1$ and therefore $z_i \geq -1$.

If complementary slackness does not hold, we know that the restricted dual has a solution $\mathbf{b}^\intercal \mathbf{z} > 0$. Our goal is to find some value ϵ such that $\mathbf{b}^\intercal(\mathbf{y} + \epsilon \mathbf{z}) > \mathbf{b}^\intercal \mathbf{y}$, thus improving the dual objective. If

$$\epsilon \leq \min_{i \in I^\complement} \left\{ -\frac{y_i}{z_i} \,\middle|\, z_i < 0 \right\}$$

then $z \geq 0$, preserving the non-negativity constraint. Also, if

$$\epsilon \leq \min_{j \in J^\complement} \left\{ \frac{c_j - \sum_i a_{ij} y_i}{\sum_i a_{ij} z_i} \,\middle|\, \sum_i a_{ij} z_i > 0 \right\}$$

dual constraints are preserved. Therefore, choosing the smaller of those values takes us from a feasible dual solution $\mathbf{y}$ to a better feasible dual solution $\mathbf{y} + \epsilon \mathbf{z}$.

1. Find some feasible dual solution $\mathbf{y}$.
2. Given $\mathbf{y}$, formulate the restricted primal and find the minimum value of its objective f_{RP}.
3. If $f_{RP} = 0$, complementary slackness holds and $\mathbf{y}$ is optimal. Return.
4. Otherwise, formulate the restricted dual. Determine the best ϵ and change $\mathbf{y}$ to $\mathbf{y} + \epsilon \mathbf{z}$ so as to improve the dual objective $\mathbf{b}^\intercal \mathbf{y}$, and go to step 2.

One of the reasons to employ this sort of algorithm is that the cost $\mathbf{c}$ vanishes. This turns a weighted problem into an unweighted problem, and step 2 can potentially be solved using efficient combinatorial optimization algorithms that do *not* rely on linear programming.

5.2 PDM for approximation schemes

The primal dual method can be used to derive approximation schemes for NP-hard problems. As a motivating example, we consider the following pair of discrete optimization problem: Let $G = (V, E)$ be an undirected graph with vertex set V and edge set E. The *vertex cover problem* asks to find the smallest subset $V_{\text{opt}} \subseteq V$ such that each edge in G is incident to at least one node in V_{opt}. Its ILP formulation is straightforward:

$$
\begin{aligned}
\min \quad & \textstyle\sum_{v \in V} x_v \\
\text{s.t.} \quad & \forall (v, w) \in E : x_v + x_w \geq 1 \\
& x_v \in \{0, 1\}
\end{aligned}
$$

Its dual is the *maximum matching problem*, which asks to find a maximum set of edges such that no two edges share a node. It can be written as

$$
\begin{aligned}
\max \quad & \textstyle\sum_{e \in E} y_e \\
\text{s.t.} \quad & \forall v \in V : \sum_{w:(v,w) \in E} y_{vw} \leq 1 \\
& y_e \in \{0, 1\}
\end{aligned}
$$

If we were to write this as a scheme, the matrix $\mathbf{A}$ would be the transpose of the node-edge incidence matrix. A node or edge is selected if its variable is 1, and deselected if it is 0.

In bipartite graphs, $\mathbf{A}$ is TUM, and thus we can solve both vertex cover as well as maximum matching in polynomial time. Specifically, maximum matching can be seen as a special case of s-t-flow, by adding source and sink nodes as well as putting directionality on the edges such that they all point towards the second vertex partition. Strong duality therefor provides a simple proof for the following important result:

Theorem 1 (Kőnig's theorem)**.** *In a bipartite graph, the number of edges in a maximum matching equals the number of nodes in a minimum vertex cover.*

As a non-mandatory exercise, try and see how the PDM for LPs applies in this case. Unfortunately, for general graphs, vertex cover is an NP-hard problem, whereas maximum matching is solvable in polynomial time, e.g. using Edmond's algorithm. We can not expect strong duality to hold. As we want to cover as many edges with as few nodes as possible, there is an obvious greedy heuristic we could try in order to find a good approximation. Let's call it the *naïve heuristic*:

1. Pick the node v with the highest degree $\deg v$, i.e. the highest number of incident edges.
2. Add v to the solution, then delete v and all its incident edges.
3. Repeat until no edges are left.

The question arises whether the naïve heuristic comes with any performance guarantees. To that end, let us assume w.l.o.g. that the nodes and edges are numbered in the order in which they are selected, so v_1 is selected before v_2, and e_1 before e_2. v_k and its incident edges are selected in the k-th iteration. We are trying to minimize the the total number of selected nodes, hence our objective function is a cost function, with optimal value f^*. Selecting a node adds 1 to the total cost. Equivalently, we can distribute the cost of selecting a node equally among all remaining edges incident to that node, hence in iteration k, each edge that gets deleted incurs a cost of

$$\frac{1}{\deg(v_k)}.$$

Assume there are n nodes and m edges. Assume edge e_j is incident to v_k and removed during the k-th iteration. We want to obtain a bound on how much cost is incurred by removing e_j. Since larger node degrees means lower edge costs, we want to find a lower bound on the largest degree still in the graph: Before entering the k-th iteration, there are at least $m-j+1$ uncovered edges left in the graph. Obviously, if we could solve vertex cover optimally, we would require no more than f^* nodes to cover these remaining edges. We therefore have to distribute $m-j+1$ edges among (at most) f^* nodes. This leads to a very basic but important result from Ramsey theory: if we were to have n pigeonholes and m pigeons, what is the minimum number of pigeons in the most crowded hole?

Theorem 2 (Pigeonhole principle)**.** *In any partition of a set of m elements into n blocks, there exists a block with at least $\lceil \frac{m}{n} \rceil$ elements.*

This means that somewhere in the graph, there exists a node v with

$$\deg(v) \geq \left\lceil \frac{m-j+1}{f^*} \right\rceil \geq \frac{m-j+1}{f^*},$$

so the cost of removing e_j is at most

$$\frac{f^*}{m-j+1}.$$

Over all m iterations, the cost incurred by the naïve heuristic, f_{NH}, is therefore bounded as

$$f_{NH} \leq \sum_{j=1}^{m} \frac{f^*}{m-j+1} = \sum_{j=1}^{m} \frac{f^*}{j} = H_m f^*$$

where H_m is the m-th harmonic number. Since[2] $H_m \in O\{\log m\} \subseteq O\{\log(n^2)\} = O\{2\log n\} = O\{\log n\}$, the *naïve heuristic* is an $O(\log n)$-approximation of vertex cover. That's not very exciting news: even for moderately sized problems with 1000 nodes, the approximation ratio can be almost seven! Even worse, the ratio grows with the problem size.

This case may serve as an example that sometimes the obvious approach is not as good a choice as you might think. Instead, we will turn to using the PDM to derive a much better heuristic. For approximation schemes for problems like vertex cover, we don't have strong duality, and complementary slackness does not hold for a primal and its dual simultaneously. Instead, the PDM is modified so that only primal complimentary slackness is enforced, whereas dual complimentary slackness is *relaxed*: instead of requiring constraints for non-zero variables to hold with equality, one can require the slack to be within a certain bounded interval. This leads to the following *relaxed complimentary slackness conditions*:

Relaxed PCS : At least one of $x_j = 0$ or $\sum_i a_{ij} y_i \geq \frac{c_j}{\alpha}$ must hold.

Relaxed DCS : At least one of y_i or $\sum_j a_{ij} x_j \geq \beta\, b_i$ must hold.

So we have

$$\frac{\mathbf{c}}{\alpha} \leq \mathbf{A}^\mathsf{T}\mathbf{y} \leq \mathbf{c}$$
$$\mathbf{c}^\mathsf{T}\mathbf{x} \leq \alpha \mathbf{y}^\mathsf{T}\mathbf{A}\mathbf{x}$$

[2]We use $O\{f\}$ for the set of functions growing not faster than f, and $O(f)$ for an element from this set. In practice, the two are often not distinguished.

as well as

$$\mathbf{b} \leq \mathbf{Ax} \leq \beta\mathbf{b}$$

$$\mathbf{y}^\mathsf{T}\mathbf{Ax} \leq \beta\mathbf{b}^\mathsf{T}\mathbf{y}$$

$$\alpha\mathbf{y}^\mathsf{T}\mathbf{Ax} \leq \alpha\beta\mathbf{b}^\mathsf{T}\mathbf{y}$$

As before, for a primal minimization problem, weak duality implies that $\mathbf{b}^\mathsf{T}\mathbf{y} \leq \mathbf{c}^\mathsf{T}\mathbf{x}$. Combining the above yields

$$\mathbf{b}^\mathsf{T}\mathbf{y} \leq \mathbf{c}^\mathsf{T}\mathbf{x} \leq \alpha\beta\mathbf{b}^\mathsf{T}\mathbf{y}$$

In other words, if relaxed complimentary slackness holds, the optimal primal solution is between the optimal dual solution and a constant factor of that solution, hence $\alpha\beta$ serves as an approximation ratio. We now have a principled way to derive an approximation heuristic for vertex cover: The primary complementary slackness condition involves primary node variables and dual constraints, so if $x_v > 0$, then PCS is enforced by requiring

$$\sum_{w:(v,w)\in E} y_{vw} = 1$$

On the other hand, dual complimentary slackness, involving the dual edge variables and primal constraints, says that if $y_e > 0$ for some $e = (v, w)$, we must have

$$x_v + x_w = 1$$

However, as we said before, in the primal dual method, only the PCS is enforced, but the DCS is relaxed if necessary. The PDM proceeds like this:

1. Set $\mathbf{y} = 0$, except for one entry $y_e = 1$. This is a feasible matching, as it obeys all dual constraints. Also, set $\mathbf{x} = 0$; note this is not a feasible primal solution, as it is not a vertex cover and violates primal constraints.

2. We now have $x_e = 1 > 0$ for some $e := (v, w)$. This means that the dual constraints v, w are binding ($\alpha = 1$) and their PCS is satisfied. This allows us to set $x_v = x_w = 1$, as they don't have to be zero anymore for PCS to hold. We thereby decrease the number of violated primal constraints, as we added covering nodes and $x_v + x_w \geq 1$ now holds. However, DCS holds only in its relaxed form, as $x_v + x_w = 2 \geq 1$, so $\beta = 2$. This means the algorithm is a 2-approximation of vertex cover!

3. If no more edge can be added, return.

4. Otherwise, pick another edge that does not share a node with one we selected before. This improves the dual objective, since the matching gets larger. The primal solution might still not be a vertex cover and thus be infeasible. Go to step 2.