Algorithm 6.6 Sparse Greedy Algorithm

Require: Set of functions X, Precision ϵ, Criterion $Q[\cdot]$
  Set $\tilde{X} = \emptyset$
  repeat
    Choose random subset X' of size m' from $X \backslash \tilde{X}$.
    Pick $\hat{x} = \text{argmax}_{x \in X'} Q[X' \cup \{x\}]$
    $X' = X' \cup \{\hat{x}\}$
    If needed, (re)compute bound on $Q[X]$.
  until $Q[\tilde{X}] + \epsilon \geq$ Bound on $Q[X]$
Output: $\tilde{X}, Q[\tilde{X}]$

(iv) The set of functions X is typically very large (i.e. more than 10^5 elements), yet the individual improvements by f_i via $Q[\tilde{X} \cup \{x_i\}]$ do not differ too much, meaning that specific x_i for which $Q[\tilde{X} \cup \{x_i\}]$ deviate by a large amount from the rest of $Q[\tilde{X} \cup \{x_i\}]$ do not exist.

Iterative Enlargement of $\tilde{X}$

In this case we may use a sparse greedy algorithm to find near optimal solutions among the remaining $X \backslash \tilde{X}$ elements. This combines the idea of an iterative enlargement of $\tilde{X}$ by one more element at a time (which is feasible since we can compute $Q[\tilde{X} \cup \{f_i\}]$ cheaply) with the idea that we need not consider all f_i as possible candidates for the enlargement. This uses the reasoning in Section 6.5.1 combined with the fact that the distribution of the improvements is not too long tailed (cf. (iv)). The overall strategy is described in Algorithm 6.6.

Problems 6.9 and 6.10 contain more examples of sparse greedy algorithms.

6.6 Summary

This chapter gave an overview of different optimization methods, which form the basic toolbox for solving the problems arising in learning with kernels. The main focus was on convex and differentiable problems, hence the overview of properties of convex sets and functions defined on them.

The key insights in Section 6.1 are that *convex sets* can be defined by *level sets of convex functions* and that convex optimization problems have *one global minimum*. Furthermore, the fact that the solutions of convex maximization over polyhedral sets can be found on the vertices will prove useful in some unsupervised learning applications (Section 14.4).

Basic tools for unconstrained problems (Section 6.2) include interval cutting methods, the Newton method, Conjugate Gradient descent, and Predictor-Corrector methods. These techniques are often used as building blocks to solve more advanced constrained optimization problems.

Since constrained minimization is a fairly complex topic, we only presented a selection of fundamental results, such as necessary and sufficient conditions in the general case of nonlinear programming. The KKT conditions for differentiable

convex functions then followed immediately from the previous reasoning. The main results are dualization, meaning the transformation of optimization problems via the Lagrangian mechanism into possibly simpler problems, and that optimality properties can be estimated via the KKT gap (Theorem 6.27).

Interior point algorithms are practical applications of the duality reasoning; these seek to find a solution to optimization problems by satisfying the KKT optimality conditions. Here we were able to employ some of the concepts introduced at an earlier stage, such as predictor corrector methods and numerical ways of finding roots of equations. These algorithms are robust tools to find solutions on moderately sized problems ($10^3 - 10^4$ examples). Larger problems require decomposition methods, to be discussed in Section 10.4, or randomized methods. The chapter concluded with an overview of randomized methods for maximizing functions or finding the best subset of elements. These techniques are useful once datasets are so large that we cannot reasonably hope to find exact solutions to optimization problems.

6.7 Problems

6.1 (Level Sets •) *Given the function $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(x) := |x_1|^p + |x_2|^p$, for which p do we obtain a convex function?*

Now consider the sets $\{x|f(x) \leq c\}$ for some $c > 0$. Can you give an explicit parametrization of the boundary of the set? Is it easier to deal with this parametrization? Can you find other examples (see also [489] and Chapter 8 for details)?

6.2 (Convex Hulls •) *Show that for any set X, its convex hull* $\operatorname{co} X$ *is convex. Furthermore, show that* $\operatorname{co} X = X$ *if X is convex.*

6.3 (Method of False Position [334] •••) *Given a unimodal (possessing one minimum) differentiable function $f : \mathbb{R} \to \mathbb{R}$, develop a quadratic method for minimizing f.*

Hint: Recall the Newton method. There we used $f''(x)$ to make a quadratic approximation of f. Two values of $f'(x)$ are also sufficient to obtain this information, however.

What happens if we may only use f? What does the iteration scheme look like? See Figure 6.8 for a hint.

6.4 (Convex Minimization in one Variable ••) *Denote by f a convex function on $[a, b]$. Show that the algorithm below finds the minimum of f. What is the* rate *of convergence in x to* $\operatorname{argmin}_x f(x)$*? Can you obtain a bound in $f(x)$ wrt.* $\min_x f(x)$*?*

input *a, b, f and threshold ε*
  $x_1 = a, x_2 = \frac{a+b}{2}, x_3 = b$ and compute $f(x_1), f(x_2), f(x_3)$
  repeat
    if *$x_3 - x_2 > x_2 - x_1$* ***then***