O. PENALTY FUNCTION METHOD

Very often we are interested in the minimization of a ("target") function,[1] i.e. in finding such values of variables, which ensure a minimum of the function when some constraints are satisfied. Just imagine hiking in the Smoky Mountains: we want to find the point of the lowest ground elevation provided that we hike along a straight line from, say, Gatlinburg to Cherokee.

Suppose the target function for minimization (which corresponds to the elevation of the ground in the Smoky Mountains region) is the function $f(x_1, x_2, \ldots, x_{n+m})$, but the variables x_i have to fulfil m equations ("constraints"):

$$\phi_i(x_1, x_2, \ldots, x_{n+m}) = 0 \quad \text{for } i = 1, 2, \ldots, m.$$

For such tasks we have at least three possibilities. The first is to eliminate m variables (by using the conditions) and express them by others. In this way the target function f takes into account all the constraints and depends only on n independent variables. Then the target function is to be minimized. The second possibility is to use the Lagrange multipliers method (see Appendix N). In both cases there is, however, the complication that the conditions to be satisfied might be quite complex and therefore solution of the corresponding equations may be difficult to achieve. An easier solution may be to choose a penalty method. The idea behind the penalty method is quite simple. Why go to the trouble of trying to satisfy the conditions $\phi_i = 0$, when we could propose the following: instead of function f let us minimize its modification

$$F = f + \sum_{i=1}^{m} K_i \phi_i^2,$$

where the penalty coefficients $K_i > 0$ are chosen to be large.[2] When minimizing F we *admit that the conditions* $\phi_i = 0$ *could be non-satisfied*, but any attempt to violate them introduces to F a positive contribution $\sum_{i=1}^{m} K_i \phi_i^2$. This means that, for minimization of F, it would always be better to explore such points in space (Fig. O.1) for which $\sum_{i=1}^{m} K_i \phi_i^2 = 0$. If the K's are large enough, the procedure will force the choice $\phi_i^2 = 0$, or $\phi_i = 0$ for $i = 1, 2, \ldots, m$, and this is what has to be satisfied during minimization.

Note that the task would be much more difficult if ϕ_i^2 had more than one minimum that corresponds to $\phi_i = 0$. This penalty method is worth keeping in our tool

[1] If we change the sign of the target function, the task is equivalent to maximization.

[2] This means a high penalty.

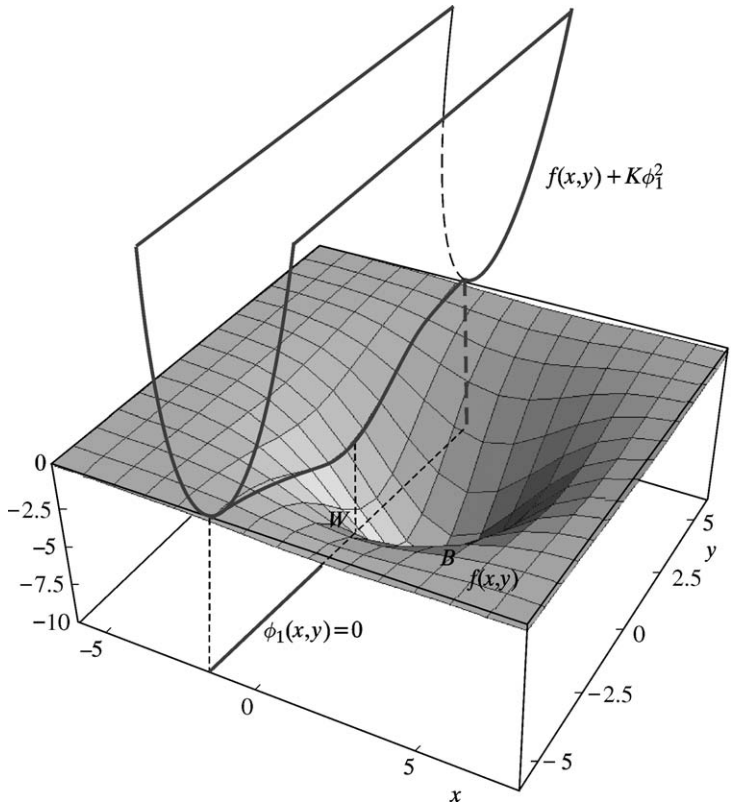

Fig. O.1. How does the penalty method work? We have to minimize $f(x, y)$, but under the condition that x and y satisfy the equation $\phi_1(x, y) = 0$ (black line). Function $f(x, y)$ exhibits a single minimum at point B, but this minimum is of no interest to us, because we are looking for a conditional minimum. To find it we minimize the sum $f(x, y) + K\phi_1^2$ with the penalty function $K\phi_1^2 \geqslant 0$ *allowing* any deviation from the black line $\phi_1(x, y) = 0$. However, going off this line does not pay, because this is precisely what switches the penalty on. As a result, at sufficiently large K we obtain the conditional minimum W. This is what the game is all about.

box, because it is general and easily applicable. For the method to work, it has to have a sufficiently large K. However, if K is too large, the numerical results might be of poor quality, since the procedure would first of all take care of the penalty, paying little attention to f. It is recommended that we take a few values of K and check whether the results depend on this.

As an example of the penalty function method, let us take the docking of two molecules. Our goal is to give such values of the atomic coordinates of both molecules as to ensure the contacts of some particular atoms of both molecules within some precise distance limits for the contacting atoms. The task sounds trivial, until we try to accomplish it in practice (especially for large molecules). The goal can be rather easily achieved when the penalty function method is used. We do the following. To the existing force field (i.e. an approximate electronic energy, Chapter 7)

we simply add a penalty for not satisfying the desired contacts. For a single pair of the atoms (a contact) the penalty could be set as

$$K(r - r_0)^2,$$

where r stands for the distance of the atoms, and r_0 is the optimum (desired) contact distance. At a chosen starting geometry the atoms are far from achieving the optimum distance, and therefore the force field energy is supplemented by a large distance-dependent penalty. The energy is so high that the minimization procedure tries to remove the penalty and relax the system. Often this can be done in only one way: by docking the molecules in such a way as to achieve the proper contact distance.