14 Lagrange Multipliers

The Method of Lagrange Multipliers is a powerful technique for constrained optimization. While it has applications far beyond machine learning (it was originally developed to solve physics equations), it is used for several key derivations in machine learning.

The problem set-up is as follows: we wish to find extrema (i.e., maxima or minima) of a differentiable objective function

$$E(\mathbf{x}) = E(x_1, x_2, ...x_D) \tag{1}$$

If we have no constraints on the problem, then the extrema must necessarily satisfy the following system of equations:

$$\nabla E = 0 \tag{2}$$

which is equivalent to writing $\frac{dE}{dx_i} = 0$ for all i. This equation says that there is no way to infinitesimally perturb $\mathbf{x}$ to get a different value for E; the objective function is locally flat.

Now, however, our goal will be to find extrema subject to a single constraint:

$$g(\mathbf{x}) = 0 \tag{3}$$

In other words, we want to find the extrema among the set of points $\mathbf{x}$ that satisfy $g(\mathbf{x}) = 0$.

It is sometimes possible to reparameterize the problem in order to eliminate the constraints (i.e., so that the new parameterization includes all possible solutions to $g(\mathbf{x}) = 0$), however, this can be awkward in some cases, and impossible in others.

Given the constraint $g(\mathbf{x}) = 0$, we are no longer looking for a point where no perturbation in any direction changes E. Instead, we need to find a point at which perturbations that satisfy the constraints do not change E. This can be expressed by the following condition:

$$\nabla E + \lambda \nabla g \;\; = \;\; 0 \tag{4}$$

for some arbitrary scalar value λ. First note that, for points on the contour $g(\mathbf{x}) = 0$, the gradient ∇g is always perpendicular to the contour (this is a great exercise if you don't remember the proof). Hence the expression $\nabla E = -\lambda \nabla g$ says that the gradient of E must be parallel to the gradient of the contour at a possible solution point. In other words, any perturbation to $\mathbf{x}$ that changes E also makes the constraint become violated. Perturbations that do not change g, and hence still lie on the contour $g(\mathbf{x}) = 0$ do not change E either. Hence, our goal is to find a point $\mathbf{x}$ that satisfies this condition and also $g(\mathbf{x}) = 0$

In the Method of Lagrange Multipliers, we define a new objective function, called the **Lagrangian**:

$$L(\mathbf{x}, \lambda) = E(\mathbf{x}) + \lambda g(\mathbf{x}) \tag{5}$$

Now we will instead find the extrema of L with respect to both $\mathbf{x}$ and λ. The key fact is that **extrema of the unconstrained objective L are the extrema of the original constrained problem.** So we have eliminated the nasty constraints by changing the objective function and also introducing new unknowns.

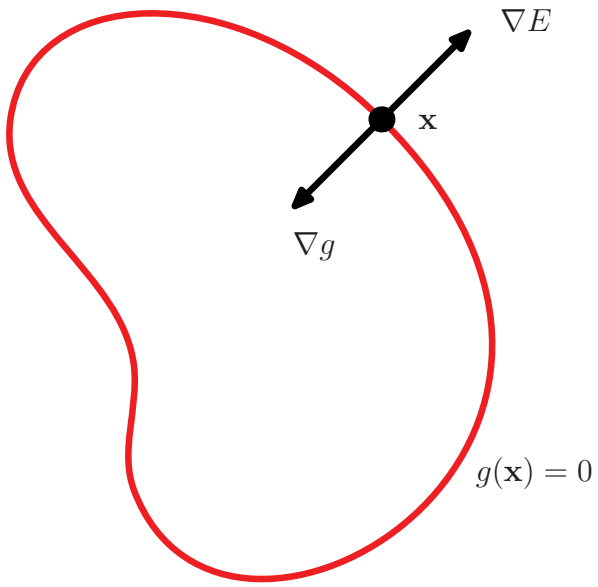

Figure 1: The set of solutions to $g(\mathbf{x}) = 0$ visualized as a curve. The gradient ∇g is always normal to the curve. At an extremal point, ∇E points is parallel to ∇g. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)

To see why, let's look at the extrema of L. The extrema to L occur when

$$\frac{dL}{d\lambda} = g(\mathbf{x}) = 0 \tag{6}$$

$$\frac{dL}{d\mathbf{x}} = \nabla E + \lambda \nabla g = 0 \tag{7}$$

which are exactly the conditions given above. In other words, first equation ensures that $g(\mathbf{x})$ is zero, as desired, and the second equation is our constraint that the gradients of E and g mucst be parallel. Using the Lagrangian is a convenient way of combining these two constraints into one unconstrained optimization.

14.1 Examples

Minimizing on a circle. We begin with a simple geometric example. We have the following constrained optimization problem:

$$\arg\min_{x,y} x + y \tag{8}$$

$$\text{subject to } x^2 + y^2 = 1 \tag{9}$$

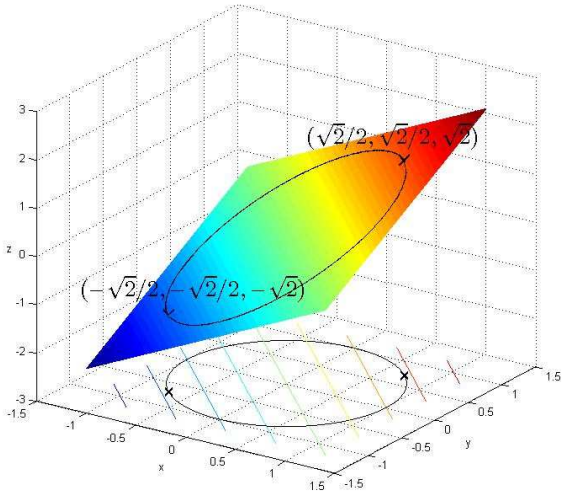

Figure 2: Illustration of the maximization on a circle problem. (Image from Wikipedia.)

In other words, we want to find the point on a circle that minimizes $x+y$; the problem is visualized in Figure 2. Here, $E(x, y) = x + y$ and $g(x, y) = x^2 + y^2 - 1$. The Lagrangian for this problem is:

$$L(x, y, \lambda) = x + y + \lambda(x^2 + y^2 - 1) \tag{10}$$

Setting the gradient to zero gives this system of equations:

$$\frac{dL}{dx} = 1 + 2\lambda x = 0 \tag{11}$$

$$\frac{dL}{dy} = 1 + 2\lambda y = 0 \tag{12}$$

$$\frac{dL}{d\lambda} = x^2 + y^2 - 1 = 0 \tag{13}$$

From the first two lines, we can see that $x = y$. Substituting this into the constraint and solving gives two solutions $x = y = \pm\frac{1}{\sqrt{2}}$. Substituting these two solutions into the objective, we see that the minimum is at $x = y = -\frac{1}{\sqrt{2}}$.

Estimating a multinomial distribution. In a multinomial distribution, we have an event e with K possible discrete, disjoint outcomes, where

$$P(e = k) = p_k \tag{14}$$

For example, coin-flipping is a binomial distribution where $N = 2$ and $e = 1$ might indicate that the coin lands heads.

Suppose we observe N events; the likelihood of the data is:

$$\prod_{i=1}^{K} P(e_i|p) = \prod_k p_k^{N_k} \tag{15}$$

where N_k is the number of times that $e = k$, i.e., the number of occurrences of the k-th event. To estimate this distribution, we can minimize the negative log-likelihood:

$$\begin{aligned} \text{arg min} \quad & -\textstyle\sum_k N_k \ln p_k & (16)\\ \text{subject to} \quad & \textstyle\sum_k p_k = 1, p_k \geq 0, \text{ for all } k & (17) \end{aligned}$$

The constraints are required in order to ensure that the p's form a valid probability distribution. One way to optimize this problem is to reparameterize: set $p_K = 1 - \sum_{k=1}^{K-1} p_k$, substitute in, and then optimize the unconstrained problem in closed-form. While this method does work in this case, it breaks the natural symmetry of the problem, resulting in some messy calculations. Moreover, this method often cannot be generalized to other problems.

The Lagrangian for this problem is:

$$L(p, \lambda) = -\sum_k N_k \ln p_k + \lambda \left(\sum_k p_k - 1 \right) \tag{18}$$

Here we omit the constraint that $p_k \geq 0$ and hope that this constraint will be satisfied by the solution (it will). Setting the gradient to zero gives:

$$\begin{aligned} \frac{dL}{dp_k} &= -\frac{N_k}{p_k} + \lambda = 0 \text{ for all } k & (19)\\ \frac{dL}{d\lambda} &= \sum_k p_k - 1 = 0 & (20) \end{aligned}$$

Multiplying $dL/dp_k = 0$ by p_k and summing over k gives:

$$0 = -\sum_{k=1}^{K} N_k + \lambda \sum_k p_k = -N + \lambda \tag{21}$$

since $\sum_k N_k = N$ and $\sum_k p_k = 1$. Hence, the optimal $\lambda = N$. Substituting this into dL/dp_k and solving gives:

$$p_k = \frac{N_k}{N} \tag{22}$$

which is the familiar maximum-likelihood estimator for a multinomial distribution.

Maximum variance PCA. In the original formulation of PCA, the goal is to find a low-dimensional projection of N data points $\mathbf{y}$

$$x = \mathbf{w}^T(\mathbf{y} - \mathbf{b}) \tag{23}$$

such that the variance of the $x_i's$ is maximized, subject to the constraint that $\mathbf{w}^T\mathbf{w} = 1$. The Lagrangian is:

$$\begin{aligned}
L(\mathbf{w}, \mathbf{b}, \lambda) &= \frac{1}{N}\sum_i \left(x_i - \frac{1}{N}\sum_i x_i\right)^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (24)\\
&= \frac{1}{N}\sum_i \left(\mathbf{w}^T(\mathbf{y}_i - \mathbf{b}) - \frac{1}{N}\sum_i \mathbf{w}^T(\mathbf{y}_i - \mathbf{b})\right)^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (25)\\
&= \frac{1}{N}\sum_i \left(\mathbf{w}^T\left((\mathbf{y}_i - \mathbf{b}) - \frac{1}{N}\sum_i (\mathbf{y}_i - \mathbf{b})\right)\right)^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (26)\\
&= \frac{1}{N}\sum_i \left(\mathbf{w}^T(\mathbf{y}_i - \bar{\mathbf{y}})\right)^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (27)\\
&= \frac{1}{N}\sum_i \mathbf{w}^T(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T\mathbf{w} + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (28)\\
&= \mathbf{w}^T\left(\frac{1}{N}\sum_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T\right)\mathbf{w} + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (29)
\end{aligned}$$

where $\bar{\mathbf{y}} = \sum_i \mathbf{y}_i/N$. Solving $dL/d\mathbf{w} = 0$ gives:

$$\left(\frac{1}{N}\sum_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T\right)\mathbf{w} = \lambda\mathbf{w} \tag{30}$$

This is just the eigenvector equation: in other words, $\mathbf{w}$ must be an eigenvector of the sample covariance of the $\mathbf{y}'s$, and λ must be the corresponding eigenvalue. In order to determine which one, we can substitute this equality into the Lagrangian to get:

$$\begin{aligned}
L &= \mathbf{w}^T\lambda\mathbf{w} + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (31)\\
&= \lambda && (32)
\end{aligned}$$

since $\mathbf{w}^T\mathbf{w} = 1$. Since our goal is to maximize the variance, we choose the eigenvector $\mathbf{w}$ which has the largest eigenvalue λ.

We have not yet selected $\mathbf{b}$, but it is clear that the value of the objective function does not depend on $\mathbf{b}$, so we might as well set it to be the mean of the data $\mathbf{b} = \sum_i \mathbf{y}_i/N$, which results in the $x's$ having zero mean: $\sum_i x_i/N = 0$.

14.2 Least-Squares PCA in one-dimension

We now derive PCA for the case of a one-dimensional projection, in terms of minimizing squared error. Specifically, we are given a collection of data vectors $\mathbf{y}_{1:N}$, and wish to find a bias $\mathbf{b}$, a single unit vector $\mathbf{w}$, and one-dimensional coordinates $x_{1:N}$, to minimize:

$$\arg \min_{\mathbf{w}, x_{1:N}, \mathbf{b}} \sum_i ||\mathbf{y}_i - (\mathbf{w}x_i + \mathbf{b})||^2 \tag{33}$$

$$\text{subject to } \mathbf{w}^T\mathbf{w} = 1 \tag{34}$$

The vector $\mathbf{w}$ is called the first principal component. The Lagrangian is:

$$L(\mathbf{w}, x_{1:N}, \mathbf{b}, \lambda) = \sum_i ||\mathbf{y}_i - (\mathbf{w}x_i + \mathbf{b})||^2 + \lambda(||\mathbf{w}||^2 - 1) \tag{35}$$

There are several sets of unknowns, and we derive their optimal values each in turn.

Projections (x_i). We first derive the projections:

$$\frac{dL}{dx_i} = -2\mathbf{w}^T(\mathbf{y}_i - (\mathbf{w}x_i + \mathbf{b})) = 0 \tag{36}$$

Using $\mathbf{w}^T\mathbf{w} = 1$ and solving for x_i gives:

$$x_i = \mathbf{w}^T(\mathbf{y}_i - \mathbf{b}) \tag{37}$$

Bias ($\mathbf{b}$). We begin by differentiating:

$$\frac{dL}{d\mathbf{b}} = -2\sum_i(\mathbf{y}_i - (\mathbf{w}x_i + \mathbf{b})) \tag{38}$$

Substituting in Equation 37 gives

$$\frac{dL}{d\mathbf{b}} = -2\sum_i(\mathbf{y}_i - (\mathbf{w}\mathbf{w}^T(\mathbf{y}_i - \mathbf{b}) + \mathbf{b})) \tag{39}$$

$$= -2\sum_i \mathbf{y}_i + 2\mathbf{w}\mathbf{w}^T\sum_i \mathbf{y}_i - 2N\mathbf{w}\mathbf{w}^T\mathbf{b} + 2N\mathbf{b} \tag{40}$$

$$= -2(\mathbf{I} - \mathbf{w}\mathbf{w}^T)\sum_i \mathbf{y}_i + 2(\mathbf{I} - \mathbf{w}\mathbf{w}^T)N\mathbf{b} = 0 \tag{41}$$

Dividing both sides by $2(\mathbf{I} - \mathbf{w}\mathbf{w}^T)N$ and rearranging terms gives:

$$\mathbf{b} = \frac{1}{N}\sum_i y_i \tag{42}$$

Basis vector ($\mathbf{w}$). To make things simpler, we will define $\tilde{\mathbf{y}}_i = (\mathbf{y}_i - \mathbf{b})$ as the mean-subtracted data points, and the reconstructions are then $x_i = \mathbf{w}^T\tilde{\mathbf{y}}_i$, and the objective function is:

$$\begin{aligned} L &= \sum_i ||\tilde{\mathbf{y}}_i - \mathbf{w}x_i||^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (43)\\ &= \sum_i ||\tilde{\mathbf{y}}_i - \mathbf{w}\mathbf{w}^T\tilde{\mathbf{y}}_i||^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (44)\\ &= \sum_i (\tilde{\mathbf{y}}_i - \mathbf{w}\mathbf{w}^T\tilde{\mathbf{y}}_i)^T(\tilde{\mathbf{y}}_i - \mathbf{w}\mathbf{w}^T\tilde{\mathbf{y}}_i) + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (45)\\ &= \sum_i (\tilde{\mathbf{y}}_i^T\tilde{\mathbf{y}}_i - 2\tilde{\mathbf{y}}_i^T\mathbf{w}\mathbf{w}^T\tilde{\mathbf{y}}_i + \tilde{\mathbf{y}}_i^T\mathbf{w}\mathbf{w}^T\mathbf{w}\mathbf{w}^T\tilde{\mathbf{y}}_i) + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (46)\\ &= \sum_i \tilde{\mathbf{y}}_i^T\tilde{\mathbf{y}}_i - \sum_i (\tilde{\mathbf{y}}_i^T\mathbf{w})^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (47) \end{aligned}$$

where we have used $\mathbf{w}^T\mathbf{w} = 1$. We then differentiate and simplify:

$$\frac{dL}{d\mathbf{w}} = -2\sum_i \tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T\mathbf{w} + 2\lambda\mathbf{w} = 0 \quad (48)$$

We can rearrange this to get:

$$\left(\sum_i \tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T\right)\mathbf{w} = \lambda\mathbf{w} \quad (49)$$

This is exactly the eigenvector equation, meaning that extrema for L occur when $\mathbf{w}$ is an eigenvector of the matrix $\sum_i \tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T$, and λ is the corresponding eigenvalue. Multiplying both sides by $1/N$, we see this matrix has the same eigenvectors as the data covariance:

$$\left(\frac{1}{N}\sum_i (\mathbf{y}_i - \mathbf{b})(\mathbf{y}_i - \mathbf{b})^T\right)\mathbf{w} = \frac{\lambda}{N}\mathbf{w} \quad (50)$$

Now we must determine which eigenvector to use. We rewrite Equation 47 as:

$$\begin{aligned} L &= \sum_i \tilde{\mathbf{y}}_i^T\tilde{\mathbf{y}}_i - \sum_i \mathbf{w}^T\tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T\mathbf{w} + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (51)\\ &= \sum_i \tilde{\mathbf{y}}_i^T\tilde{\mathbf{y}}_i - \mathbf{w}^T\left(\sum_i \tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T\right)\mathbf{w} + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (52)\\ & && (53) \end{aligned}$$

and substitute in Equation 49:

$$\begin{aligned} L &= \sum_i \tilde{\mathbf{y}}_i^T\tilde{\mathbf{y}}_i - \lambda\mathbf{w}^T\mathbf{w} + \lambda(\mathbf{w}^T\mathbf{w} - 1) && (54)\\ &= \sum_i \tilde{\mathbf{y}}_i^T\tilde{\mathbf{y}}_i - \lambda && (55) \end{aligned}$$

again using $\mathbf{w}^T\mathbf{w} = 1$. We must pick the eigenvalue λ that gives the smallest value of L. Hence, we pick the largest eigenvalue, and set $\mathbf{w}$ to be the corresponding eigenvector.

14.3 Multiple constraints

When we wish to optimize with respect to multiple constraints $\{g_k(\mathbf{x})\}$, i.e.,

$$\arg\min_{\mathbf{x}} E(\mathbf{x}) \tag{56}$$

$$\text{subject to } g_k(\mathbf{x}) = 0 \text{ for } k = 1...K \tag{57}$$

Extrema occur when:

$$\nabla E + \sum_k \lambda_k \nabla g_k \quad = \quad 0 \tag{58}$$

where we have introduced K Lagrange multipliers λ_k. The constraints can be combined into a single Lagrangian:

$$L(\mathbf{x}, \lambda_{1:K}) = E(\mathbf{x}) + \sum_k \lambda_k g_k(\mathbf{x}) \tag{59}$$

14.4 Inequality constraints

The method can be extended to inequality constraints of the form $g(\mathbf{x}) \geq 0$. For a solution to be valid and maximal, there two possible cases:

- The optimal solution is inside the constraint region, and, hence $\nabla E = 0$ and $g(\mathbf{x}) > 0$. In this region, the constraint is "inactive," meaning that λ can be set to zero.
- The optimal solution lies on the boundary $g(\mathbf{x}) = 0$. In this case, the gradient ∇E must point in the *opposite* direction of the gradient of g; otherwise, following the gradient of E would cause g to become positive while also modifying E. Hence, we must have $\nabla E = -\lambda \nabla g$ for $\lambda \geq 0$.

Note that, in both cases, we have $\lambda g(\mathbf{x}) = 0$. Hence, we can enforce that one of these cases is found with the following optimization problem:

$$\max_{\mathbf{w},\lambda} \quad E(\mathbf{x}) + \lambda g(\mathbf{x}) \tag{60}$$

$$\text{such that} \quad g(\mathbf{x}) \geq 0 \tag{61}$$

$$\lambda \geq 0 \tag{62}$$

$$\lambda g(\mathbf{x}) = 0 \tag{63}$$

These are called the Karush-Kuhn-Tucker (KKT) conditions, which generalize the Method of Lagrange Multipliers.

When minimizing, we want ∇E to point in the same direction as ∇g when on the boundary, and so we minimize $E - \lambda g$ instead of $E + \lambda g$.

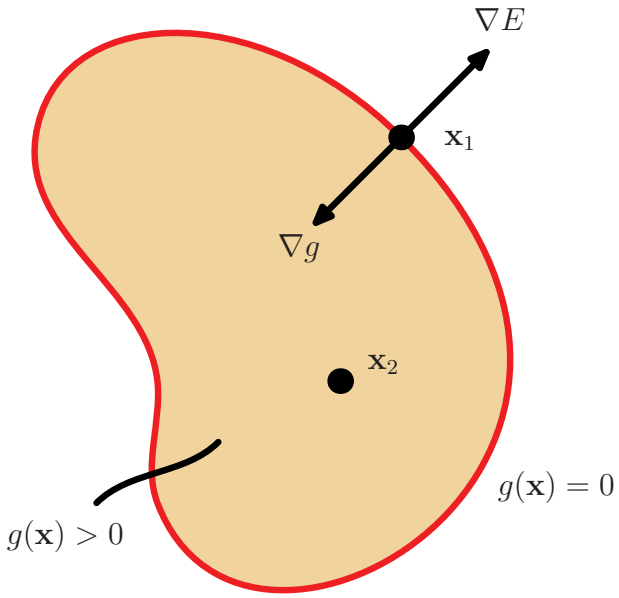

Figure 3: Illustration of the condition for inequality constraints: the solution may lie on the boundary of the constraint region, or in the interior. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)

