15 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

$$E(\mathbf{x}) \;=\; E(x_1, x_2, ...x_D)\;. \tag{15.1}$$

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

$$\nabla E \;=\; 0\;. \tag{15.2}$$

This is equivalent to requiring that $\frac{\partial E}{\partial x_i} = 0$ for all i. This equation says that there is no way to infinitesimally perturb $\mathbf{x}$ to get a different value for E. That is, the objective function is locally flat.

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

$$g(\mathbf{x}) \;=\; 0\;. \tag{15.3}$$

In other words, we want to find the extrema among the set of points $\mathbf{x}$, all of which satisfy $g(\mathbf{x}) = 0$. It is sometimes possible to reparameterize the problem to eliminate the constraints (i.e., so that the new parameterization includes all possible solutions to $g(\mathbf{x}) = 0$). But 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:

$$\nabla E + \lambda \nabla g \;\;=\;\; 0\;, \tag{15.4}$$

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 how to prove that this is true). 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 gradient condition and also $g(\mathbf{x}) = 0$

In the method of Lagrange multipliers, we change the constrained optimization above into an unconstrained optimization with a new objective function, called the **Lagrangian**:

$$L(\mathbf{x}, \lambda) \;=\; E(\mathbf{x}) + \lambda g(\mathbf{x})\;. \tag{15.5}$$

Now, our goal is to find 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

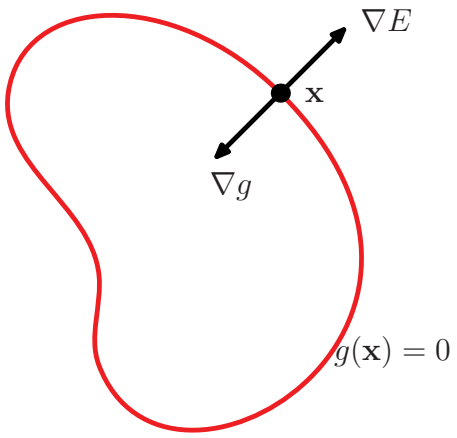

Figure 15.1: The set of solutions to $g(\mathbf{x}) = 0$ visualized as a curve. The gradient ∇g is always normal to the curve. At an extremal point, ∇E points is parallel to ∇g. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)

have eliminated the nasty constraints by changing the objective function and also introducing new unknowns.

To see why, let's look at the extrema of L. Because L depends on two parameters its extrema must necessarily satisfy two gradient constraints, i.e.,

$$\frac{\partial L}{\partial \lambda} = g(\mathbf{x}) = 0 \tag{15.6}$$

$$\frac{\partial L}{\partial \mathbf{x}} = \nabla E + \lambda \nabla g = 0 \,. \tag{15.7}$$

One can immediately see that these gradient constraints are exactly the conditions given above. The first equation ensures that $g(\mathbf{x})$ is zero, and the second is our constraint that the gradients of E and g mucst be parallel. Using the Lagrangian is a convenient way to combine these two constraints into one unconstrained optimization.

15.1 Examples

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

$$\arg\min_{x,y} \; x + y \tag{15.8}$$

$$\text{subject to } x^2 + y^2 = 1 \tag{15.9}$$

In words, we want to find the point on a unit circle that minimizes $x + y$. The problem is depicted in Fig. 15.2. Here, $E(x, y) = x + y$ and $g(x, y) = x^2 + y^2 - 1$. The Lagrangian for this problem is given by

$$L(x, y, \lambda) = x + y + \lambda(x^2 + y^2 - 1) \,. \tag{15.10}$$

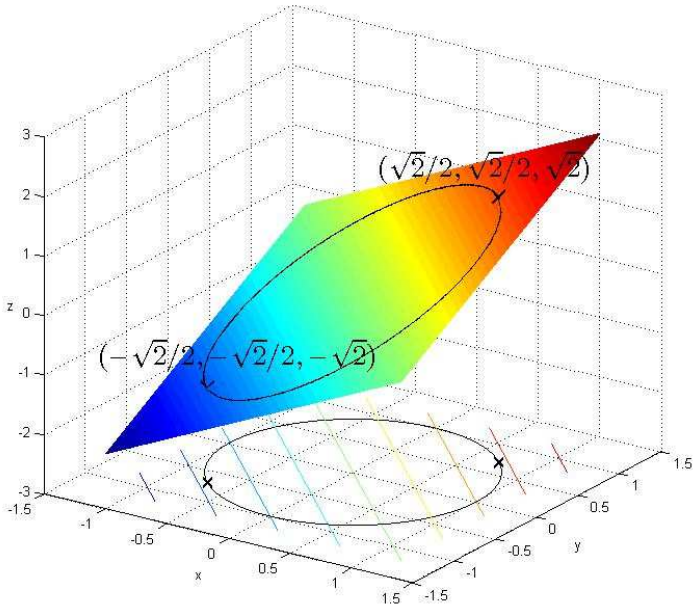

Figure 15.2: Illustration of the maximization on a circle problem. (Image from Wikipedia.)

Setting the gradient to zero (with respect to x, y and λ) gives us the following system of equations:

$$\frac{\partial L}{\partial x} = 1 + 2\lambda x = 0 \tag{15.11}$$

$$\frac{\partial L}{\partial y} = 1 + 2\lambda y = 0 \tag{15.12}$$

$$\frac{\partial L}{\partial \lambda} = x^2 + y^2 - 1 = 0 \tag{15.13}$$

The first two equations ensure 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 find that the minimum occurs at $x = y = -\frac{1}{\sqrt{2}}$.

Estimating a Categorial distribution. A Categorial distribution over a random variable c with K possible discrete, disjoint states (or outcomes). Accordingly it is specified by K probabilities, denoted here by p_k:

$$P(c = k) \equiv p_k \,, \tag{15.14}$$

for $k = 1...K$, and let $\mathbf{p} = (p_1, ..., p_K)$. For example, in coin-flipping the outcome of a coin flip follows a Bernoulli distribution, which is the special case of a Categorical distribution when $K = 2$, and $c = 1$ might indicate that the coin lands heads side up.

Suppose we observe N independent draws from such a random process, i.e., we observe the sequence $c_{1:N}$. The likelihood of the observed data is therefore the product of the N independent likelihoods:

$$P(c_{1:N} \,|\, \mathbf{p}) = \prod_{i=1}^{K} P(c_i \,|\, \mathbf{p}) = \prod_k p_k^{N_k} \,, \tag{15.15}$$

where N_k is the number of times that $c = k$, i.e., the number of occurrences of the k-th state.

To estimate this Categorical distribution, we minimize the negative log-likelihood of the observed data,

$$\min - \sum_k N_k \ln p_k \tag{15.16}$$

$$\text{subject to } \sum_k p_k = 1 \text{ and } p_k \geq 0, \text{ for all } k\,. \tag{15.17}$$

The constraints here are required to ensure that the p's form a valid probability distribution. (One way to optimize this problem is to reparameterize the probabilities, i.e., replace p_K in the likelihood by $1 - \sum_{k=1}^{K-1} p_k$, 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

$$L(p, \lambda) \;=\; -\sum_k N_k \ln p_k \;+\; \lambda \left(\sum_k p_k - 1 \right) \,. \tag{15.18}$$

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

$$\frac{\partial L}{\partial p_k} \;=\; -\frac{N_k}{p_k} + \lambda = 0 \quad \text{for all } k \tag{15.19}$$

$$\frac{\partial L}{\partial \lambda} \;=\; \sum_k p_k - 1 = 0 \tag{15.20}$$

Multiplying $\partial L / \partial p_k = 0$ by p_k, and summing over k yields

$$0 \;=\; -\sum_{k=1}^{K} N_k + \lambda \sum_k p_k \;=\; -N + \lambda \,, \tag{15.21}$$

since $\sum_k N_k = N$ and $\sum_k p_k = 1$. Hence, the optimal $\lambda = N$. Substituting this into $\partial L / \partial p_k$ and solving yields the estimated probabilities

$$p_k \;=\; \frac{N_k}{N} \,, \tag{15.22}$$

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

Maximum variance PCA. In the original formulation of PCA, the goal is to find a low-dimensional projection of N data points $\mathbf{y}$. Here, suppose we just want to find a one-dimensional subspace spanned by the vector $\mathbf{w}$. In that case the subspace projection is given by

$$x \;=\; \mathbf{w}^T (\mathbf{y} - \mathbf{b}) \,. \tag{15.23}$$

One way to formulate PCA is as a optimization to find the direction $\mathbf{w}$ which maximizes the variance of the projection, subject to the constraint that $\mathbf{w}^T\mathbf{w} = 1$. The Lagrangian can be expressed as

$$\begin{aligned}
L(\mathbf{w}, \mathbf{b}, \lambda) &= \frac{1}{N}\sum_i \left(x_i - \frac{1}{N}\sum_i x_i \right)^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\
&= \frac{1}{N}\sum_i \left(\mathbf{w}^T(\mathbf{y}_i - \mathbf{b}) - \frac{1}{N}\sum_i \mathbf{w}^T(\mathbf{y}_i - \mathbf{b}) \right)^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\
&= \frac{1}{N}\sum_i \left(\mathbf{w}^T \left((\mathbf{y}_i - \mathbf{b}) - \frac{1}{N}\sum_i (\mathbf{y}_i - \mathbf{b}) \right) \right)^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\
&= \frac{1}{N}\sum_i \left(\mathbf{w}^T(\mathbf{y}_i - \bar{\mathbf{y}}) \right)^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\
&= \frac{1}{N}\sum_i \mathbf{w}^T(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T\mathbf{w} + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\
&= \mathbf{w}^T \left(\frac{1}{N}\sum_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T \right) \mathbf{w} + \lambda(\mathbf{w}^T\mathbf{w} - 1) \,, \qquad (15.24)
\end{aligned}$$

where $\bar{\mathbf{y}} = \sum_i \mathbf{y}_i / N$.

Solving $\partial L / \partial \mathbf{w} = 0$ yields

$$\left(\frac{1}{N}\sum_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T \right) \mathbf{w} \;=\; \lambda \mathbf{w} \qquad (15.25)$$

This is the eigenvector equation. That is, $\mathbf{w}$ must be an eigenvector of the sample covariance matrix 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 obtain

$$\begin{aligned}
L &= \mathbf{w}^T \lambda \mathbf{w} + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\
&= \lambda \,, \qquad (15.26)
\end{aligned}$$

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, i.e., $\sum_i x_i / N = 0$.

15.2 Least-Squares PCA in 1D

Let's now consider a different way to formulate PCA. Instead of finding the direction of maximum variance, let's find the one-dimensional projection which minimizes the squared error in the subspace approximation. 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:

$$\arg \min_{\mathbf{w}, x_{1:N}, \mathbf{b}} \sum_i ||\mathbf{y}_i - (\mathbf{w}x_i + \mathbf{b})||^2 \tag{15.27}$$

$$\text{subject to } \mathbf{w}^T\mathbf{w} = 1 \tag{15.28}$$

Here, x_i specifies position along a line with direction $\mathbf{w}$ and distance from the origin $||\mathbf{b}||$. The total error is the sum of squared Euclidean distances between data points $\mathbf{y}_i$ and their corresponding points on the model line.[1] The vector $\mathbf{w}$ is called the first principal component. The Lagrangian is:

$$L(\mathbf{w}, x_{1:N}, \mathbf{b}, \lambda) = \sum_i ||\mathbf{y}_i - (\mathbf{w}x_i + \mathbf{b})||^2 + \lambda(||\mathbf{w}||^2 - 1) \tag{15.29}$$

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

Projections (x_i). We first derive the projections:

$$\frac{\partial L}{\partial x_i} = -2\mathbf{w}^T(\mathbf{y}_i - (\mathbf{w}x_i + \mathbf{b})) = 0 \tag{15.30}$$

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

$$x_i = \mathbf{w}^T(\mathbf{y}_i - \mathbf{b}) \tag{15.31}$$

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

$$\frac{\partial L}{\partial \mathbf{b}} = -2\sum_i (\mathbf{y}_i - (\mathbf{w}x_i + \mathbf{b})) \tag{15.32}$$

Substituting in Equation 15.31 gives

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{b}} &= -2\sum_i (\mathbf{y}_i - (\mathbf{w}\mathbf{w}^T(\mathbf{y}_i - \mathbf{b}) + \mathbf{b})) \\
&= -2\sum_i \mathbf{y}_i + 2\mathbf{w}\mathbf{w}^T\sum_i \mathbf{y}_i - 2N\mathbf{w}\mathbf{w}^T\mathbf{b} + 2N\mathbf{b} \\
&= -2(\mathbf{I} - \mathbf{w}\mathbf{w}^T)\sum_i \mathbf{y}_i + 2(\mathbf{I} - \mathbf{w}\mathbf{w}^T)N\mathbf{b} = 0
\end{aligned} \tag{15.33}$$

Factoring out $2(\mathbf{I} - \mathbf{w}\mathbf{w}^T)$ from both terms, one can see that we obtain

$$\mathbf{b} = \frac{1}{N}\sum_i \mathbf{y}_i \tag{15.34}$$

[1] It is important to note that this optimization problem differs in subtle ways from the linear regression earlier in the notes. With linear regression we had multi-dimensional inputs and a scalar output. Here we have vector-valued data $\mathbf{y}$ and we are trying to find a scalar input x. In linear regression we minimized the error in the predicted y (i.e., the vertical distance of each point to the curve), while here the error is the Euclidean distant from each 2D data point $\mathbf{y}$ to a location on the model line.

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

$$\begin{aligned} L &= \sum_i ||\tilde{\mathbf{y}}_i - \mathbf{w}x_i||^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\ &= \sum_i ||\tilde{\mathbf{y}}_i - \mathbf{w}\mathbf{w}^T\tilde{\mathbf{y}}_i||^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\ &= \sum_i (\tilde{\mathbf{y}}_i - \mathbf{w}\mathbf{w}^T\tilde{\mathbf{y}}_i)^T(\tilde{\mathbf{y}}_i - \mathbf{w}\mathbf{w}^T\tilde{\mathbf{y}}_i) + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\ &= \sum_i (\tilde{\mathbf{y}}_i^T\tilde{\mathbf{y}}_i - 2\tilde{\mathbf{y}}_i^T\mathbf{w}\mathbf{w}^T\tilde{\mathbf{y}}_i + \tilde{\mathbf{y}}_i^T\mathbf{w}\mathbf{w}^T\mathbf{w}\mathbf{w}^T\tilde{\mathbf{y}}_i) + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\ &= \sum_i \tilde{\mathbf{y}}_i^T\tilde{\mathbf{y}}_i - \sum_i (\tilde{\mathbf{y}}_i^T\mathbf{w})^2 + \lambda(\mathbf{w}^T\mathbf{w} - 1) \end{aligned} \tag{15.35}$$

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

$$\frac{\partial L}{\partial \mathbf{w}} = -2\sum_i \tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T\mathbf{w} + 2\lambda\mathbf{w} = 0 \tag{15.36}$$

We can rearrange this to get:

$$\left(\sum_i \tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T\right)\mathbf{w} = \lambda\mathbf{w} \tag{15.37}$$

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:

$$\left(\frac{1}{N}\sum_i (\mathbf{y}_i - \mathbf{b})(\mathbf{y}_i - \mathbf{b})^T\right)\mathbf{w} = \frac{\lambda}{N}\mathbf{w} \tag{15.38}$$

Now we must determine which eigenvector to use. To this end, we rewrite Eqn. (15.35) as

$$\begin{aligned} L &= \sum_i \tilde{\mathbf{y}}_i^T\tilde{\mathbf{y}}_i - \sum_i \mathbf{w}^T\tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T\mathbf{w} + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\ &= \sum_i \tilde{\mathbf{y}}_i^T\tilde{\mathbf{y}}_i - \mathbf{w}^T\left(\sum_i \tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T\right)\mathbf{w} + \lambda(\mathbf{w}^T\mathbf{w} - 1)\,, \end{aligned} \tag{15.39}$$

and substitute in Eqn. (15.37):

$$\begin{aligned} L &= \sum_i \tilde{\mathbf{y}}_i^T\tilde{\mathbf{y}}_i - \lambda\mathbf{w}^T\mathbf{w} + \lambda(\mathbf{w}^T\mathbf{w} - 1) \\ &= \sum_i \tilde{\mathbf{y}}_i^T\tilde{\mathbf{y}}_i - \lambda\,, \end{aligned} \tag{15.40}$$

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.

15.3 Multiple Constraints

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

$$\arg\min_{\mathbf{x}} E(\mathbf{x}) \tag{15.41}$$

$$\text{subject to } g_k(\mathbf{x}) = 0 \text{ for } k = 1...K \tag{15.42}$$

Extrema occur when:

$$\nabla E + \sum_k \lambda_k \nabla g_k \;\; = \;\; 0 \tag{15.43}$$

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

$$L(\mathbf{x}, \lambda_{1:K}) = E(\mathbf{x}) + \sum_k \lambda_k g_k(\mathbf{x}) \tag{15.44}$$

15.4 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:

$$\max_{\mathbf{w},\lambda} \quad E(\mathbf{x}) + \lambda g(\mathbf{x}) \tag{15.45}$$

$$\text{such that} \quad g(\mathbf{x}) \geq 0 \tag{15.46}$$

$$\lambda \geq 0 \tag{15.47}$$

$$\lambda g(\mathbf{x}) = 0 \tag{15.48}$$

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 $E + \lambda g$.

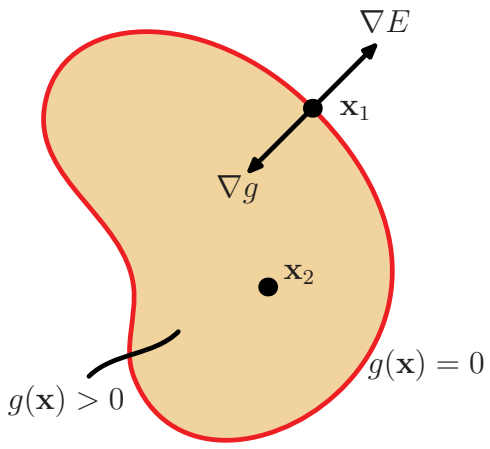

Figure 15.3: Illustration of the condition for inequality constraints: the solution may lie on the boundary of the constraint region, or in the interior. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)

