

The Lagrange Multiplier Method

An Introduction to Constrained Optimization

Feng Li
Shandong University
fli@sdu.edu.cn

September 28, 2026

Abstract

The Lagrange multiplier method finds candidates for extrema of a function subject to equality constraints. This article develops the method from feasible directions, states the regularity assumptions, and explains how to check the resulting candidates. Worked examples illustrate constrained minima, maxima, and sensitivity to a constraint. Applications include probability estimation and principal component analysis, followed by a brief introduction to inequality constraints. The assumed background is multivariable differentiation and basic linear algebra.

Contents

1	Constrained Optimization	2
2	Geometry and the Lagrangian	3
3	Worked Examples	6
4	Applications in Machine Learning	9

1 Constrained Optimization

1.1 Why Do We Need Constraints?

Many learning problems restrict the values that parameters can take. For example, probabilities must sum to one, $\sum_{k=1}^K p_k = 1$; a direction vector may have unit length, $w^\top w = 1$; and a model may satisfy a linear condition, $Ax = b$.

We seek the best objective value *among feasible points*:

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad h_j(x) = 0, \quad j = 1, \dots, m.$$

The same method finds candidates for constrained maxima. Throughout this article, we use the convention $\mathcal{L} = f + \sum_j \lambda_j h_j$ for equality constraints.

1.2 A Simple Constrained Minimum

$$\min_{x,y} x^2 + y^2 \quad \text{subject to} \quad x + y = 1.$$

The unconstrained minimum $(0, 0)$ is infeasible.

Substitution gives $y = 1 - x$ and

$$x^2 + (1 - x)^2 = 2 \left(x - \frac{1}{2} \right)^2 + \frac{1}{2}.$$

Thus $(x^*, y^*) = (\frac{1}{2}, \frac{1}{2})$ and $f^* = \frac{1}{2}$.

At this constrained optimum,

$$\nabla f(x^*, y^*) = (1, 1)^\top \neq 0.$$

Setting the objective gradient to zero would miss the solution. Figure 1 shows the smallest objective contour that reaches the feasible line. It touches the line at the solution.

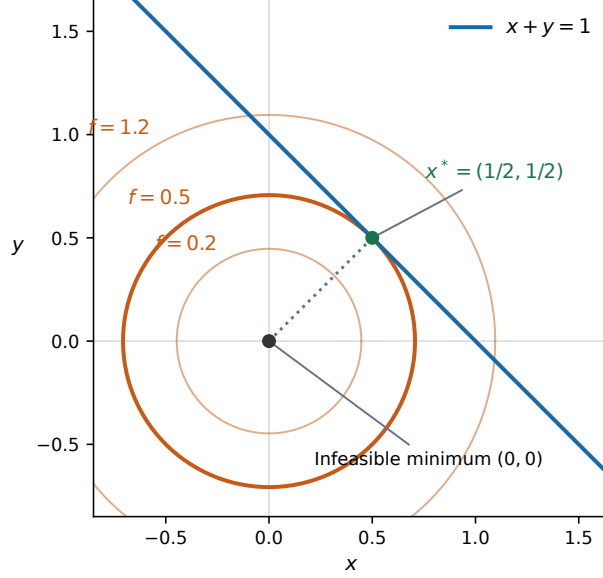


Figure 1: Minimizing $f = x^2 + y^2$ on $x + y = 1$. Orange circles are objective level sets; the blue line is the feasible set. The contour $f = 1/2$ is tangent to the line at $(1/2, 1/2)$. Smaller contours contain no feasible points.

2 Geometry and the Lagrangian

2.1 Feasible Directions at an Optimum

Suppose $f, h : \mathbb{R}^n \rightarrow \mathbb{R}$ are continuously differentiable near a feasible point x^* , and the only restriction is $h(x) = 0$. An unconstrained optimum requires zero first-order change in every direction. With a constraint, we only need to consider motions that stay in the feasible set.

1. Describe a feasible motion. Let $x(t)$ be a continuously differentiable curve defined for $-\varepsilon < t < \varepsilon$ such that

$$x(0) = x^*, \quad h(x(t)) = 0 \quad \text{for every } t.$$

Its velocity $d = x'(0)$ is a *feasible tangent direction* at x^* . The curve satisfies

$$x(t) = x^* + td + o(t),$$

where $o(t)$, read “little-o of t ,” denotes a remainder that becomes negligible compared with $|t|$ as $t \rightarrow 0$. More precisely, writing $r(t) = x(t) - x^* - td$, we have

$$\frac{\|r(t)\|}{|t|} \rightarrow 0 \quad (t \rightarrow 0).$$

For a scalar remainder, this means, for example, that t^2 is $o(t)$ because $|t^2|/|t| = |t| \rightarrow 0$, whereas $2t$ is not because its ratio is always 2. The notation describes how small the remainder is; it does not specify a particular function. Thus td captures the first-order motion, while $o(t)$ includes smaller corrections that allow the curve to bend and remain feasible. In particular, the straight line $x^* + td$ need not satisfy the constraint.

For example, on the unit circle, the feasible curve $x(t) = (\cos t, \sin t)^\top$ passes through $(1, 0)^\top$ with $d = (0, 1)^\top$. Here the remainder is

$$r(t) = (\cos t - 1, \sin t - t)^\top.$$

Its components have leading terms $-t^2/2$ and $-t^3/6$, respectively, so $\|r(t)\|/|t| \rightarrow 0$. These smaller corrections keep the curve on the circle, whereas the straight-line step $(1, t)^\top$ leaves it because $1 + t^2 > 1$ whenever $t \neq 0$.

2. Differentiate the constraint. Since $h(x(t))$ is identically zero, the chain rule gives

$$0 = \left. \frac{d}{dt} h(x(t)) \right|_{t=0} = \nabla h(x^*)^\top x'(0) = \nabla h(x^*)^\top d.$$

Every feasible tangent direction is therefore perpendicular to the constraint gradient. If $\nabla h(x^*) \neq 0$, the constraint is locally a smooth surface, and the converse also holds: every vector orthogonal to $\nabla h(x^*)$ is the velocity of a feasible curve through x^* . This consequence of the implicit function theorem identifies the *tangent space* as

$$T_{x^*} = \{d \in \mathbb{R}^n : \nabla h(x^*)^\top d = 0\}.$$

It is a line for a regular constraint in $\mathbb{R}^2$ and a plane for a regular constraint in $\mathbb{R}^3$.

3. Differentiate the objective along the curve. If x^* is a constrained local minimum, the one-variable function $\phi(t) = f(x(t))$ has a local minimum at $t = 0$. Because the curve is defined for both positive and negative t , the usual one-variable necessary condition gives $\phi'(0) = 0$. The same argument applies to a local maximum. Consequently,

$$0 = \phi'(0) = \left. \frac{d}{dt} f(x(t)) \right|_{t=0} = \nabla f(x^*)^\top d \quad \text{for every } d \in T_{x^*}.$$

To see why zero is necessary, write

$$f(x(t)) = f(x^*) + t \nabla f(x^*)^\top d + o(t).$$

If the coefficient of t were nonzero, one sign of a sufficiently small t would decrease the objective and the opposite sign would increase it. Such a point could be neither a local minimum nor a local maximum on the constraint.

4. Convert orthogonality into a multiplier equation. At an optimum, $\nabla f(x^*)$ is orthogonal to the entire tangent space, not just to one chosen tangent vector. For a single regular equality constraint, the normal space is the one-dimensional span of $\nabla h(x^*)$. Hence $\nabla f(x^*) = \alpha \nabla h(x^*)$ for some scalar α . Writing $\lambda = -\alpha$ gives

$$\boxed{\nabla f(x^*) + \lambda \nabla h(x^*) = 0, \quad h(x^*) = 0.}$$

The objective gradient need not vanish: its component along the constraint must vanish. The case $\nabla f(x^*) = 0$ is included by taking $\lambda = 0$. The assumption $\nabla h(x^*) \neq 0$ is essential for this argument; without it, the linearized constraint may fail to describe the actual feasible tangent directions.

Example: moving around a circle. For $f(x, y) = x + y$ and $h(x, y) = x^2 + y^2 - 1$, parameterize the circle by $x(\theta) = (\cos \theta, \sin \theta)^\top$. Then

$$d(\theta) = (-\sin \theta, \cos \theta)^\top, \quad \frac{d}{d\theta} f(x(\theta)) = \nabla f(x(\theta))^\top d(\theta) = -\sin \theta + \cos \theta.$$

At $\theta = 0$, this derivative is 1: moving slightly in the positive direction increases f , while moving in the negative direction decreases it. Thus $(1, 0)$ is not an optimum. At $\theta = \pi/4$, the derivative is zero and

$$x^* = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)^\top, \quad \nabla f(x^*) = (1, 1)^\top, \quad \nabla h(x^*) = (\sqrt{2}, \sqrt{2})^\top.$$

The multiplier equation holds with $\lambda = -1/\sqrt{2}$. Here $f(x(\theta)) = \sqrt{2} \cos(\theta - \pi/4)$ confirms that this point is a global maximum. In general, vanishing tangent derivatives give only a necessary first-order condition; further checks are needed to distinguish minima, maxima, and other stationary points.

Figure 2 compares an ordinary feasible point with a constrained maximum. Away from the maximum, the objective has a nonzero derivative in a tangent direction. At the maximum, its gradient is normal to the constraint, and the first-order change along every feasible tangent direction is zero.

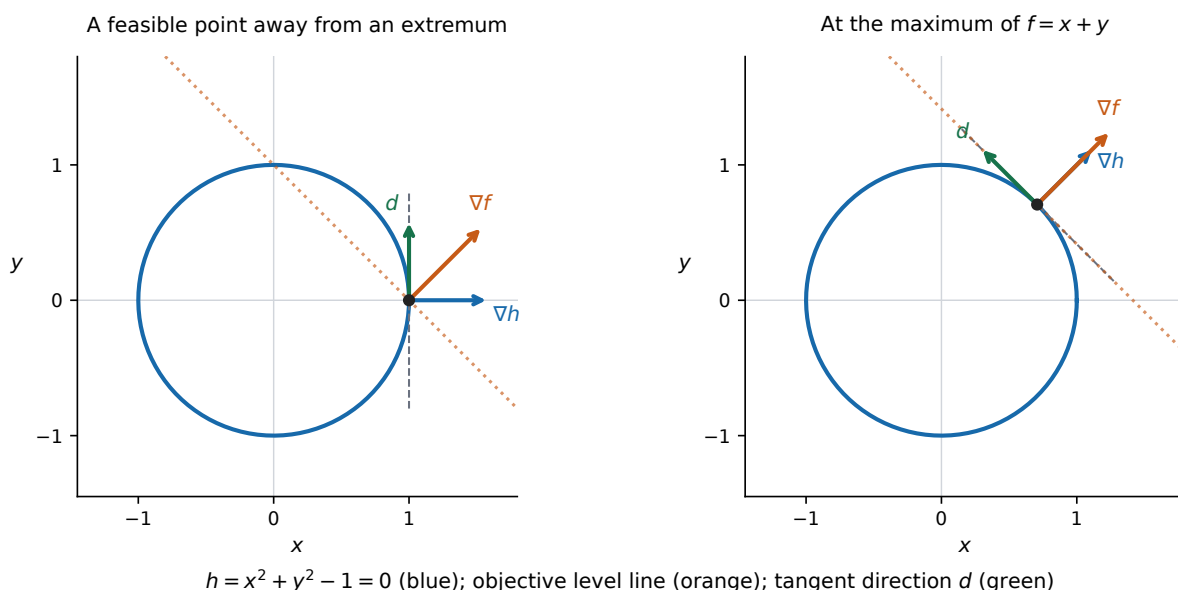


Figure 2: Tangent and normal directions for maximizing $f = x + y$ on the unit circle $h = x^2 + y^2 - 1 = 0$. At $(1, 0)$ (left), $\nabla f^\top d \neq 0$. At $(1/\sqrt{2}, 1/\sqrt{2})$ (right), ∇f and ∇h are parallel and perpendicular to d . Arrow lengths are scaled for readability; only their directions are being compared.

2.2 The Lagrangian

Combine the objective and the equality constraint:

$$\mathcal{L}(x, \lambda) = f(x) + \lambda h(x).$$

x contains the original decision variables.

$\lambda \in \mathbb{R}$ is a **Lagrange multiplier**; it may be positive, negative, or zero.

Differentiating with respect to both variables gives

$$\nabla_x \mathcal{L} = \nabla f + \lambda \nabla h = 0, \quad \frac{\partial \mathcal{L}}{\partial \lambda} = h(x) = 0.$$

These are stationarity and feasibility equations. We do not minimize $\mathcal{L}$ jointly over x and λ . For a fixed infeasible x with $h(x) \neq 0$, the expression is affine and unbounded below as a function of

the unrestricted multiplier. The multiplier is an auxiliary unknown that enforces the first-order balance at a feasible candidate.

2.3 Several Equality Constraints

For m constraints, introduce one multiplier per constraint:

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{j=1}^m \lambda_j h_j(x).$$

If $f, h_1, \dots, h_m$ are continuously differentiable and the constraint gradients are linearly independent at a local optimum x^* , there is a multiplier vector λ^* satisfying

$$\nabla f(x^*) + \sum_{j=1}^m \lambda_j^* \nabla h_j(x^*) = 0,$$

$$h_j(x^*) = 0, \quad j = 1, \dots, m.$$

The objective gradient lies in the span of the constraint normals. To see why, let $J_h(x^*)$ be the matrix whose rows are $\nabla h_j(x^*)^\top$. Feasible tangent directions satisfy $J_h(x^*)d = 0$. At an optimum, $\nabla f(x^*)^\top d = 0$ for every such direction. Therefore $\nabla f(x^*)$ is orthogonal to the null space of $J_h(x^*)$, so it belongs to the range of $J_h(x^*)^\top$. This yields the multiplier equation above.

2.4 How to Apply the Method

1. Write every equality in the form $h_j(x) = 0$.
2. Form $\mathcal{L} = f + \sum_j \lambda_j h_j$.
3. Solve $\nabla_x \mathcal{L} = 0$ together with all constraints.
4. Check feasibility and examine every candidate.
5. Classify the candidates using objective values, convexity, or second-order information.

Also check domain boundaries and points where the constraint gradients are dependent: the regularity assumption may fail there.

Using $f - \sum_j \lambda_j h_j$ instead is equivalent for equalities; it reverses the signs of the multipliers.

3 Worked Examples

3.1 Example 1: Minimum Distance to a Line

$$\min_{x,y} x^2 + y^2 \quad \text{subject to} \quad x + y - 1 = 0.$$

Form the Lagrangian:

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

Solve the three equations:

$$2x + \lambda = 0, \quad 2y + \lambda = 0, \quad x + y - 1 = 0.$$

Therefore

$$\boxed{x^* = y^* = \frac{1}{2}, \quad \lambda^* = -1.}$$

The objective is strictly convex and the constraint is affine, so this is the unique global minimum, with $f^* = \frac{1}{2}$.

3.2 Example 2: Extrema on a Circle

Find the maximum and minimum of $f(x, y) = xy$ subject to $x^2 + y^2 = 1$.

$$\mathcal{L} = xy + \lambda(x^2 + y^2 - 1).$$

Stationarity and feasibility give

$$y + 2\lambda x = 0, \quad x + 2\lambda y = 0, \quad x^2 + y^2 = 1.$$

Neither x nor y can be zero at a stationary point. Thus

$$4\lambda^2 = 1.$$

$\lambda = -\frac{1}{2}$: $y = x$, giving $(x, y) = \pm(1, 1)/\sqrt{2}$ and $f = \frac{1}{2}$.

$\lambda = \frac{1}{2}$: $y = -x$, giving $(x, y) = \pm(1, -1)/\sqrt{2}$ and $f = -\frac{1}{2}$.

Checking global optimality. On the unit circle, the inequalities

$$(x - y)^2 \geq 0 \implies xy \leq \frac{1}{2},$$

$$(x + y)^2 \geq 0 \implies xy \geq -\frac{1}{2}$$

establish global bounds.

The candidates attain both bounds, so

$$\boxed{f_{\max} = \frac{1}{2}, \quad f_{\min} = -\frac{1}{2}.}$$

Lagrange equations identify candidates; this additional argument certifies their type and global optimality.

Figure 3 shows all four tangencies. A contour with $|f| > 1/2$ cannot meet the unit circle.

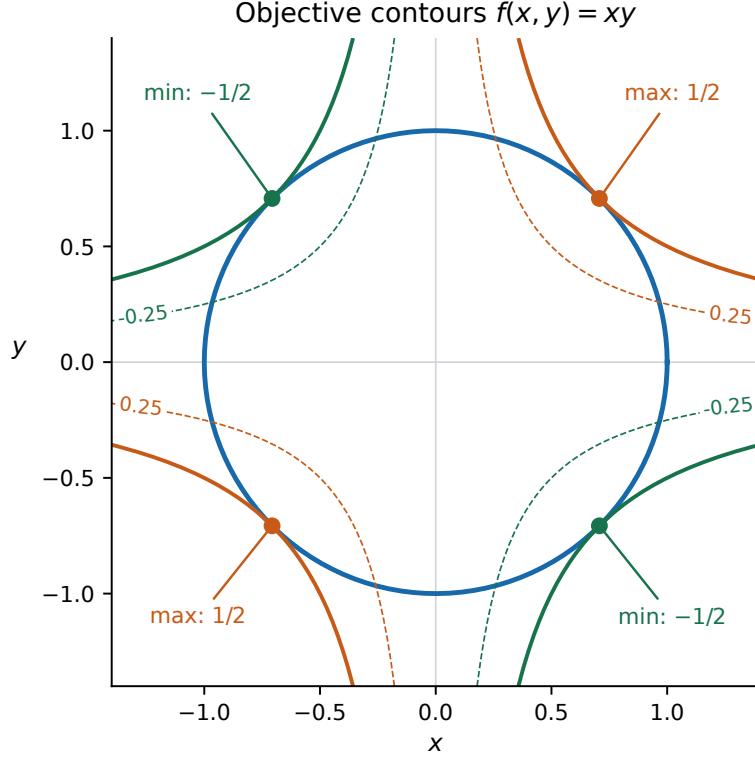


Figure 3: The objective $f = xy$ has hyperbolic level sets. The contours $f = 1/2$ (orange) and $f = -1/2$ (green) are tangent to the unit circle at the two maxima and two minima, respectively. Dashed contours at $f = \pm 1/4$ cross the circle and do not represent extrema.

3.3 When the Regularity Assumption Fails

Consider

$$\min_x f(x) = x \quad \text{subject to} \quad h(x) = x^2 = 0.$$

The only feasible point is $x^* = 0$, so it is a global minimum. However,

$$\mathcal{L}(x, \lambda) = x + \lambda x^2, \quad \frac{\partial \mathcal{L}}{\partial x} = 1 + 2\lambda x.$$

At $x = 0$, stationarity would require $1 = 0$.

No multiplier exists because $h'(0) = 0$. The gradient independence assumption is essential to the theorem.

3.4 Multipliers and Sensitivity

Let the right-hand side of a constraint vary:

$$v(b) = \min_{x,y} \{x^2 + y^2 : x + y = b\}.$$

The solution is $x^* = y^* = b/2$, so

$$v(b) = \frac{b^2}{2}, \quad \mathcal{L} = x^2 + y^2 + \lambda(x + y - b), \quad \lambda^* = -b.$$

Hence $v'(b) = b = -\lambda^*$.

More generally, when the optimal value and solution vary smoothly, the convention $\mathcal{L} = f + \lambda(h-b)$ gives

$$\boxed{\frac{dv}{db} = -\lambda^*}.$$

A multiplier measures the local change in the optimal value as the constraint changes.

4 Applications in Machine Learning

4.1 Estimating a Categorical Distribution

Observe K categories with counts $n_1, \dots, n_K$ and $N = \sum_k n_k > 0$.

$$\max_p \ell(p) = \sum_{k=1}^K n_k \log p_k, \quad \sum_{k=1}^K p_k = 1, \quad p_k \geq 0.$$

First assume every $n_k > 0$, so an optimum must have every $p_k > 0$.

$$\mathcal{L}(p, \lambda) = \sum_{k=1}^K n_k \log p_k + \lambda \left(\sum_{k=1}^K p_k - 1 \right).$$

Stationarity gives

$$\frac{n_k}{p_k} + \lambda = 0 \quad \implies \quad n_k = -\lambda p_k.$$

Solving for the probabilities. Sum $n_k = -\lambda p_k$ over all categories:

$$N = -\lambda \sum_{k=1}^K p_k = -\lambda.$$

Therefore

$$\boxed{\hat{p}_k = \frac{n_k}{N}}.$$

The log-likelihood is strictly concave when every $n_k > 0$; the stationary point is the unique global maximum.

For counts $(2, 3, 5)$, the estimate is $(0.2, 0.3, 0.5)$.

If some counts are zero, the same formula gives boundary estimates; treat those categories separately, with $0 \log 0 = 0$.

This normalization appears in categorical Naive Bayes parameter estimation and mixture-weight updates in EM.

4.2 Unit Norm Constraints and Eigenvectors

For a symmetric matrix S , consider

$$\max_w w^\top S w \quad \text{subject to} \quad w^\top w = 1.$$

Using the same plus-sign convention,

$$\mathcal{L}(w, \lambda) = w^\top S w + \lambda(w^\top w - 1).$$

Stationarity gives

$$2Sw + 2\lambda w = 0 \implies Sw = (-\lambda)w.$$

Every stationary direction is a unit eigenvector of S . Its objective value is the corresponding eigenvalue.

The maximum uses a largest-eigenvalue eigenvector. With S a covariance matrix, this gives the first principal component.

For a concrete illustration, take

$$S = \begin{pmatrix} 5/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & 3/2 \end{pmatrix}.$$

Its eigenvalues are 3 and 1, with corresponding unit eigenvectors $(\sqrt{3}/2, 1/2)^\top$ and $(-1/2, \sqrt{3}/2)^\top$. Figure 4 shows the quadratic objective contours touching the unit circle at these directions and their negatives.

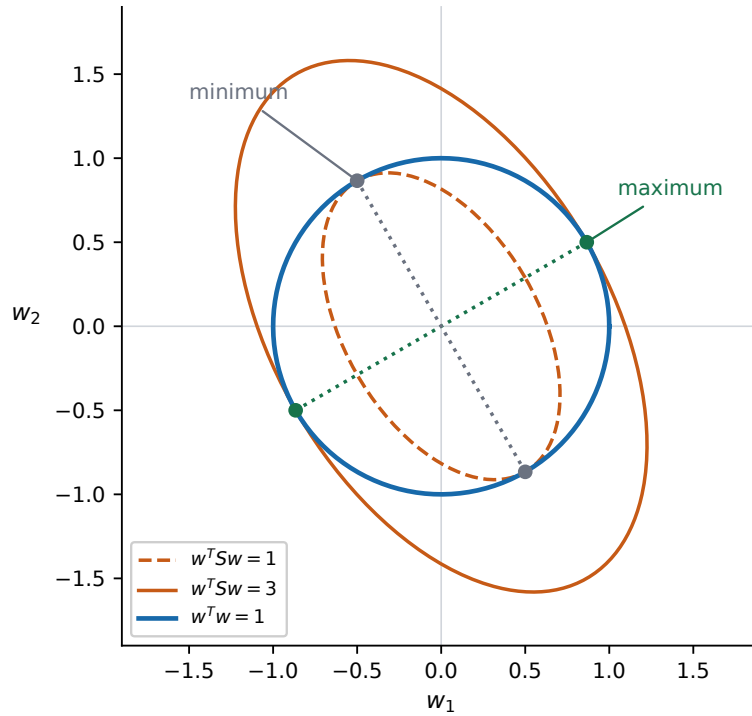


Figure 4: Geometry of the Rayleigh quotient on the unit circle. Orange ellipses are level sets of $w^\top S w$ for the matrix given above. The maximum value is 3 in the green eigenvector direction; the minimum is 1 in the gray direction. The constraint and objective contours are tangent at each extremum.

4.3 Extension to Inequalities: KKT Conditions

For a minimization problem with $g_i(x) \leq 0$ and $h_j(x) = 0$, use

$$\mathcal{L} = f + \sum_i \alpha_i g_i + \sum_j \lambda_j h_j.$$

Under suitable constraint qualifications, a local optimum satisfies

$$\begin{array}{ll}
 \nabla_x \mathcal{L} = 0 & \text{(stationarity),} \\
 g_i(x) \leq 0, \quad h_j(x) = 0 & \text{(primal feasibility),} \\
 \alpha_i \geq 0 & \text{(dual feasibility),} \\
 \alpha_i g_i(x) = 0 & \text{(complementary slackness).}
 \end{array}$$

An inactive inequality has zero multiplier. Equality multipliers remain unrestricted. These conditions underlie SVM optimization.

4.4 Practice Problem

Practice: Minimize $x^2 + 2y^2$ subject to $x + y = 3$.

Form $\mathcal{L} = x^2 + 2y^2 + \lambda(x + y - 3)$ and solve

$$2x + \lambda = 0, \quad 4y + \lambda = 0, \quad x + y = 3.$$

Answer: $(x^*, y^*) = (2, 1)$, $\lambda^* = -4$, and $f^* = 6$. Strict convexity establishes a unique global minimum.

The central idea is that constraints restrict the directions in which a point can move. Multipliers express the resulting balance between the objective gradient and the constraint normals. Solving this balance produces candidates, whose optimality must then be checked.