Why Do Lagrange Multipliers Work?

In a first-year calculus optimization problem you have a function $f(x)$ and your goal is to find a point $x$ where $df/dx = 0$, because such a point could be a minimum or maximum. In a 2D optimization problem you have a function $f(x, y)$ and your goal is to find a point $(x, y)$ where all of the derivatives of $f$ equal zero, i.e. where $\nabla f = 0$. In a constrained 2D optimization problem you have a function $f(x, y)$ and a constraint $g(x, y) = k$. That constraint defines a curve, and now your goal is to find a point $(x, y)$ on that curve where the derivative of $f$ along that curve is zero. The objective function $f$ could reach a maximum or minimum at such a point as you move along the curve.

The directional derivative along the curve is zero when the gradient is perpendicular to the curve, so you are looking for a point where $\nabla f$ is perpendicular to the curve $g = k$. Recall, however, that the gradient points perpendicular to the level curves of a function, so $\nabla g$ is also perpendicular to that curve. That means you are looking for a point where $\nabla f$ is parallel to $\nabla g$. For two vectors to be parallel one must be a multiple of the other, so $\nabla f = \lambda \nabla g$ for some constant $\lambda$.

This argument can easily be extended to more dimensions. In 3D, for example, you want the derivative of $f$ to be zero in every direction along some surface $g(x, y, z) = k$, which means $\nabla f$ must point perpendicular to that level surface of $g$. Once again you are led to $\nabla f = \lambda \nabla g$. 