Lagrange Multipliers

We tackle the problem of maximizing a function $H(P)$ subject to constraints $F_j(P)=0$ for $j=1,\ldots,J$. and present three important examples.

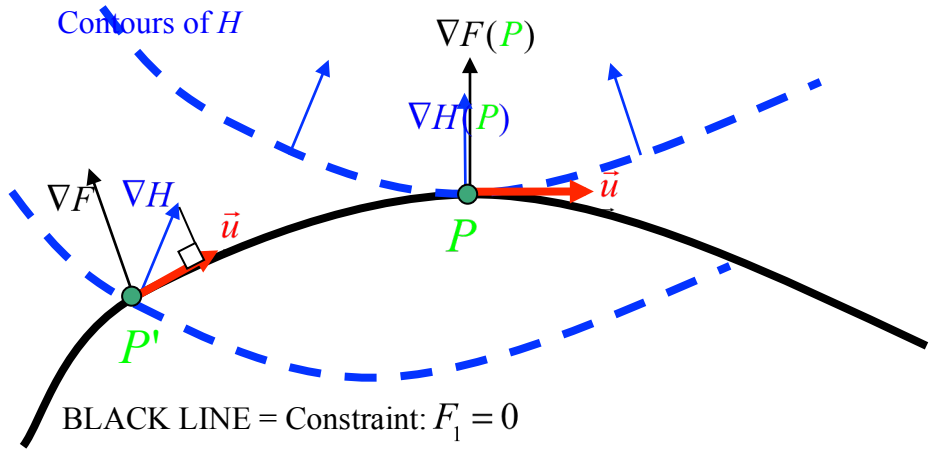
To find local maxima, we want to find a P which is a ***stationary point*** in the constraint surface: if you moved slightly away from P while staying in the surface, H would decrease no matter what the direction is. Staying in the surface means that all the constraints are still true (to second order in ε) as you move in direction $\vec{u}$. Therefore if you replace replace P by $P+\varepsilon\vec{u}$, taking a very small step in the direction of the unit vector $\vec{u}$, then $\vec{u}$ is tangent to each of the constraint surfaces, which means that: $\vec{u}$ is perpendicular to ∇F_j for each j:

$$\vec{u}\cdot\nabla F_j(P)=0, \quad \text{for } j=1,\ldots J. \qquad (1)$$

On the other hand, if H is locally stationary, then the <u>*directional*</u> derivative in this direction, $\vec{u}\cdot\nabla H(P)$ (the dot product of the direction vector with the gradient vector), is zero. That means $\vec{u}$ and ∇H are perpendicular.

$$\vec{u}\cdot\nabla H(P)=0. \qquad (2)$$

So we want a P such that $\nabla H(P)$ **lies in the subspace Ω spanned by the** $\nabla F_j(P)$.
If we find it, then any $\vec{u}$ satisfying the constraint (1) will also satisfy (2), so no direction in the constrained set will allow you increase H. The converse is a little harder to see, but if we are at a P' where $\nabla H(P')$ does <u>*not*</u> lie in Ω, then $\vec{u}=$ [*projection* of $\nabla H(P')$ onto Ω] is a non-zero vector which will satisfy (1) but violate (2), so that you can increase H by moving along $\vec{u}$ either in the positive or negative direction.



Therefore, to find such a stationary P, solve

$$\nabla H(P)=\sum_{j=1}^{J}\lambda_j\nabla F_j(P) \qquad (3)$$

The number of unknowns is the dimension of P (call it M), plus the number of $\lambda's$ (which is J). The number of equations is also $M + J$, because the gradient equation (3) is really M equations, and there are J constraints. So usually there is a unique solution.

The "Lagrangian" is $\mathfrak{L}(P,\lambda)=H(P)-\sum_{j=1}^{J}\lambda_j F_j(P)$.

Eq. 3 corresponds to $\partial\mathfrak{L}/\partial P=0$ and the constraints correspond to $\partial\mathfrak{L}/\partial\lambda=0$.
It seems odd that adding constraints makes Ω bigger . That seems to give more freedom: (3) is easier to fulfill. How can adding a constraint give more freedom? But that makes sense. Suppose we add a new constraint, going from J=1 to 2. The previous solution P still satisfies (3) but no longer satisfies the constraint, so to find the new P we have to relax (3). The new P didn't satisfy the previous (3) with J=1. But with J=2 it does.
See also http://www.slimy.com/~steuard/teaching/tutorials/Lagrange.html .

Examples:

A. Optimal weighting for weighted averages
B. Penalized likelihoods: ridge, lasso, elastic net
C. Constrained maximum entropy, applied to priors and statistical physics

A. Application to optimal weighting

Suppose we have estimates X_i of a parameter, each with a different variance v_i, and we want to combine them into a single estimate $\hat{\mu} = \sum w_i X_i$ with smallest variance possible. So our objective function is

$$H(w) = \text{var}(\sum w_i X_i) = \sum w_i^2 v_i$$

subject to the constraint $\sum w_i = 1$. The constraint $F_1(w) = \sum w_i = K_1 = 1$ guarantees that if the X_i are unbiased, so is $\hat{\mu}$. (Without the constraint, setting all the weights to zero would make the variance zero.)

We want to find *w* solving $\nabla H\left(w\right) = \sum_{j=1}^{J} \lambda_j \nabla F_j\left(w\right)$,

Taking the derivatives,

$$0 = \partial / \partial w_i (\text{var}(\sum w_i X_i) - \lambda(\sum w_i - 1))$$
$$0 = 2 w_i v_i - \lambda$$
$$w_i = \lambda / (2 v_i)$$

Thus all $w_i \propto 1 / v_i$. Finally, applying the constraint, we get our optimal weights:

$$w_i = \frac{1 / v_i}{\sum_{i'} 1 / v_{i'}} .$$

B. Penalized likelihoods. ***See Hastie Tibshirani Friedman's book.***

The Lagrange method converts the constraint problem into a Lagrangian form.

	As a constraint on parameter space	As a penalized likelihood (log likelihood + log prior)
$\hat{\beta}^{\text{ridge}}$ Ridge Regression (L2 norm)	$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2,$ subject to $\sum_{j=1}^{p} \beta_j^2 \le t,$	$\underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\}$
$\hat{\beta}^{\text{lasso}}$ Lasso (L1 norm)	$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$ subject to $\sum_{j=1}^{p} \lvert\beta_j\rvert \le t.$	$\underset{\beta}{\operatorname{argmin}}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\lvert\beta_j\rvert\right\}.$

(In addition, the elastic net combines the L1 and L2 penalties. Instead of two multipliers, there is a single multiplier with a user-chosen ratio mixing the two penalties.)

C. Constrained maximum entropy

One way to obtain a prior distribution $\pi(\theta) : \theta \in \Theta$ for a parameter θ is to specify a small number of constraints, then pick the distribution which maximizes the entropy $S = -\sum_{\theta} \pi(\theta)\log\pi(\theta)$ subject to those constraints. This approach was favored by E.T.Jaynes, a major proponent of the "objective Bayes" school, which favors using non-informative or minimally informative priors. In this case, we interpret *P* as the vector of probabilities, π. The constraints are typically moments:

$$F_j(\pi) = E(g_j(\theta)) - K_j = \sum_{\theta} \pi(\theta) g_j(\theta) - K_j$$

See the document "maximum entropy and lagrange multipliers.doc" for details.