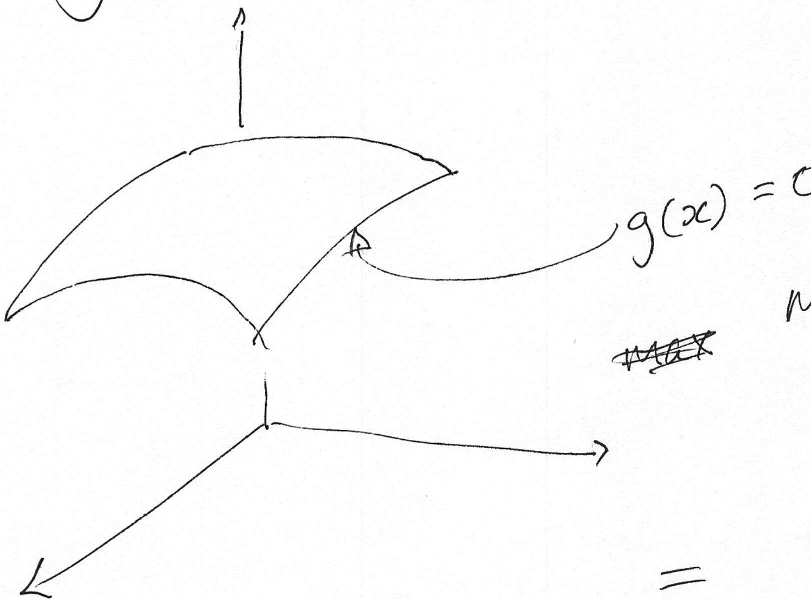
Constrained optimization :

- Two kinds of problem
  - equality constraints
  - ineq constraints.

- These are quite different, in important ways.

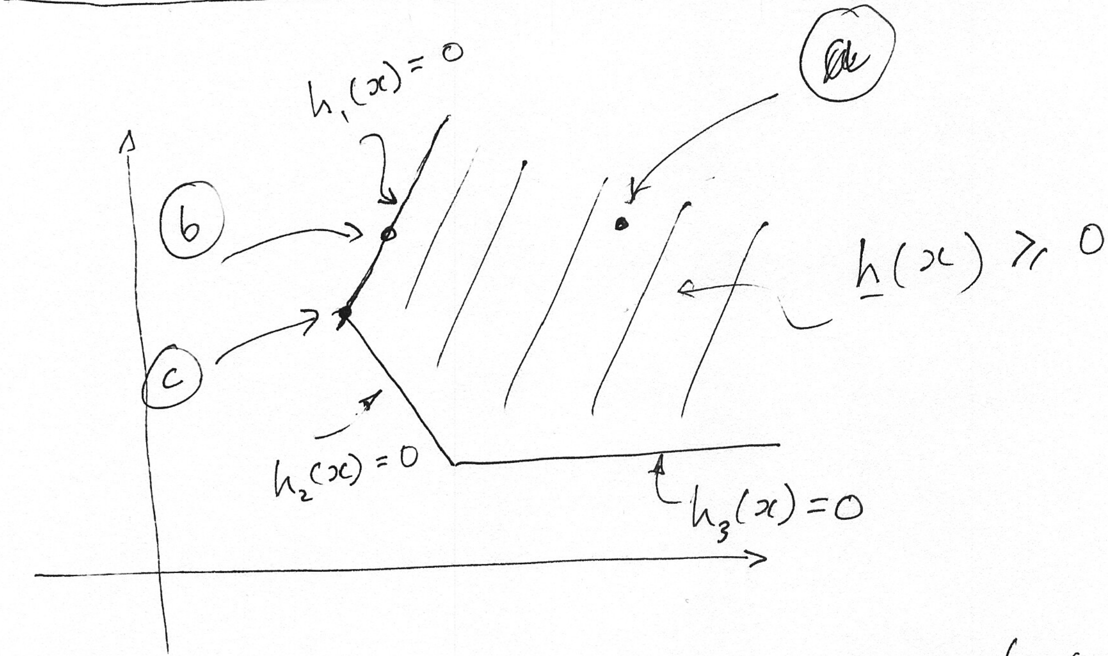


~~max~~ minize f subject to g = 0

≡ ∇f is ~~par~~ normal to the "surface" $g = 0$

(because otherwise, we could move along $g=0$ in a way that reduces f)

Inequality Picture:

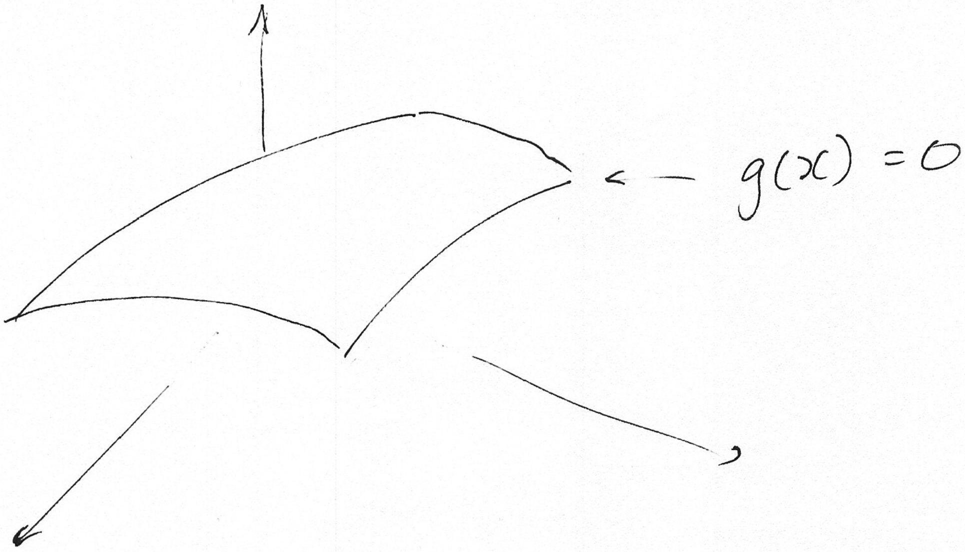


- at ⓐ, constraints are irrelevant; locally, problem involves min $f(x)$ w/o constraints
- at ⓑ, h_1 applies, but no others, so locally problem looks like min $f(x)$ s.t. $h_1(x)=0$
- at ⓒ, h_1, h_2 . . .

$$\min f(x) \quad \text{st} \quad h_1(x)=0, \quad h_2(x)=0$$

BUT: one step may mean picture changes!

Equality constraints :



- Simple picture: · 3D, one constraint $g(x) = 0$
- $\min f(x)$ st $g(x) = 0$
- answer occurs at points where

  ∇f is normal to $\{g(x) = 0\}$

- Normal of implicit surface $g(x) = 0$ is ∇g

$$\therefore \quad \nabla f = \underset{\uparrow \text{ some unknown constant.}}{\lambda} \nabla g$$

What if there are many constraints in N-D?

$g_1(x) = 0, \quad g_2(x) = 0,$ etc. . . .

∇f is normal

$|||$

$\nabla f \in \text{span}\{\nabla g_1, \nabla g_2, \nabla g_3, \dots\}$

$|||$

$\nabla f = \lambda_1 \nabla g_1 + \lambda_2 \nabla g_2 + \dots$

equivalently, write $\underline{g} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ \vdots \end{bmatrix}$

$$J_g = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots \\ \frac{\partial g_2}{\partial x_1} & & \\ \vdots & & \\ \vdots & & \end{bmatrix}$$

then

$$\nabla f = \lambda^T J_g .$$

This justifies writing the <u>Lagrangian</u>

$$\mathcal{L} = \underline{f} - \lambda^T \underline{g} .$$

at minimum :

$$\nabla_x \mathcal{L} = \nabla f - \lambda^T J_g = 0$$

$$\nabla_\lambda \mathcal{L} = - g = 0$$

These conditions must be true at a minimum.

I will deal with pages 6- 10 later - DAF

V. Important special cases for constrained optimization

① max $x^T \frac{A}{2} x$ s.t. $x^T x = 1$

Lagrangian

$$x^T \frac{A}{2} x - \lambda (x^T x - 1)$$

$$\therefore \quad \boxed{Ax = \lambda x}$$

eigenvalue problem

$$\max \quad x^T \underset{\sim}{A} x \qquad \text{st} \quad x^T B x = 1$$

Lagrangian

$$x^T \underset{\sim}{A} x - \lambda(x^T B x - 1)$$

$$\therefore \quad \boxed{Ax - \lambda B x = 0}$$

↑ generalized eigenvalue problem

Notice: NOT the same as

$$B^{-1} A x - \lambda x = 0$$ ← NAUGHTY!

because B^{-1} may not exist

- Any good Numerical linear Alg package can do these.

$\textcircled{3}$

min $\frac{x^T x}{2}$ st $Ax = b$

(i.e. closest point on linear subspace to the origin)

<u>Lagrangian</u>:

$$\frac{x^T x}{2} - \lambda^T(Ax-b)$$

$$\therefore \quad x - A^T\lambda = 0 \quad \leftarrow \quad (\nabla_x L) \quad \textcircled{\beta}$$

<u>So</u>

$$AA^T\lambda = b \quad \leftarrow \quad \textcircled{\alpha}$$

<u>Alg</u> solve linear system α for λ, then subs for x in $\textcircled{\beta}$

(4)

$$\min \quad x^T \frac{A}{2} x + b^T x \qquad \text{s.t.} \quad Cx = d$$

<u>Lagrangian</u> :

$$x^T \frac{A}{2} x + b^T x - \lambda^T (Cx - d)$$

$$\nabla_x \mathcal{L} : \quad Ax + b - C^T \lambda = 0$$

$$\nabla_\lambda \mathcal{L} : \quad Cx - d = 0$$

$$\begin{bmatrix} A & -C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} -b \\ d \end{bmatrix}$$

↑ <u>Solve this</u> .

Notice how useful it has been to know the Lagrange multipliers.

Algorithms for other cases

Eliminating constraints

- sometimes, we can parametrize the constraint set and move on that.
- not usually a good idea

eg. $\min \; x^2 + (y-10)^2$

subject to

$y - \sin x = 0$

Notice the rich supply of local min

. we could rewrite as

$$\min \quad x^2 + (\sin x - 10)^2 \qquad \text{w/ no constraints}$$

then try to min this.

. Notice when we do this, we are confining our steps to the constraint space:

- <u>Problem</u> - don't see large scale structure of objective

. <u>Equivalent</u>:

. take step <u>Tangent</u> to constraint space

. project back.

(. i.e. make up local parametrization)

For example:

$$\min \quad x^T A x + b^T x$$

$$\text{s.t.} \quad \underline{\phi}(\underline{x}) = 0$$

↑ vector function.

Now consider a step Δx

$$\underline{\phi}(\underline{x} + \Delta \underline{x}) \approx \underline{\phi}(\underline{x}) + J_\phi \cdot \Delta x.$$

so we could try:

Step:

1) ~~min $x^T A x + b^T x$~~

st

$$\min_{\Delta x} \quad (x_k + \Delta x)^T A (x_k + \Delta x) + b^T(x_k + \Delta$$

$$\text{st} \quad J_\phi \cdot \Delta x = 0$$

2) correct by finding x_{k+1} s.t. $\phi(x_{k+1}) = 0$, start search at $x_k + \Delta x$.

Again, not usually a great plan, because we may have a hard time taking big steps

- **Quadratic penalty method**

$$\min \quad f(\underline{x}) \quad \text{st} \quad \underline{g}(\underline{x}) = 0$$

- approach by $\min \; f(x) + \frac{c}{2} g^T g$
- if c is big, this forces $g^T g$ to be small
- **advantage**: - we could take steps off the constraint space
  - now its unconstrained.

· <u>Disadvantages</u> (Big)

1) big c $\Rightarrow$ some big terms in Hessian

$$H = H_f + c\left[\cancel{\nabla} J_g^T J_g + \cdots \right]$$

(so we should see terms that look like $\underset{k}{\Sigma_i} \frac{\partial g_k}{\partial x_i}^2$ on Hess diag !

2) at soln, g <u>isn't</u> zero

at soln

$$\boxed{\nabla_x f + c\, g J_g = 0}$$

$\nabla_x f$ won't be zero, in general, so g can't be!

The method of multipliers (also, Augmented Lagrangian) method.

$$\min \quad f(x) \quad \text{st} \quad g(x) = 0$$

form: Augmented Lagrangian

$$A(x;\lambda) = \underbrace{f(x) - \lambda^T g(x)}_{\text{Lagrangian}} + \underset{\overset{\uparrow}{\text{augmentation}}}{\frac{c}{2}(g^T g)}$$

. Now, assume we have an estimate λ^K of the LM.s

Minimize $A(x;\lambda^K)$ to get $x^{(K)}$

at x^K we have

$$\nabla f(x^K_{\cancel{*}}) - \lambda^{K^T} g J_g + c g^T J_g = 0$$

Now, pattern match to conditions

$$\nabla_x \mathcal{L} = 0$$

$$\nabla_x \mathcal{L} = \nabla f - \lambda^T J_g$$

This suggests

$$\lambda^{K+1} = (\lambda^K - cg)$$

Notice: we could have a soln w/ $g=0$!

ALM:

- start w x^0, λ^0, c^0
- min $A(x, \lambda^k) = f - \lambda^{k^T} g + \frac{c^0}{2} g^T g$

  to get x^k
- $\lambda^{k+1} = \lambda^k - \frac{c^k}{2} g^T(x^k)$.

  $c^{k+1} = r c^k$

  $\uparrow$ $\boxed{\text{often } 2}$

Q: How do we know its converged?

A: In ALM, usually nothing to do — we don't reject steps — but issue for future.

Q: do we have Hessian probs?

A: No, because λ ests help.

(formally, there is some bound on the c required to get exact soln.)

First glimpse of Duality:

we have

$$\min \quad f(x) \qquad \text{st} \quad \underline{g}(x) = 0$$

$$\mathcal{L} = f(x) - \lambda^T g(x) = \mathcal{L}(x, \lambda)$$

we have solution when

$$\left.\begin{array}{l} \nabla_x \mathcal{L} = 0 \\ \nabla_\lambda \mathcal{L} = 0 \end{array}\right] - \text{so solution is at a critical point of } \mathcal{L}.$$

→ what kind of c.p.?

(a) fix $\lambda = \hat{\lambda}$; then $\mathcal{L}(x, \hat{\lambda})$ is (locally) at a min

(b) <u>but</u> for fixed $x_{\cancel{a}} = \hat{x}$, $\mathcal{L}(\hat{x}, \lambda)$ is <u>linear</u>

(c) ~~So~~ at $\underset{\uparrow \text{soln}}{\underline{x^c, \lambda^c}}$, $\dfrac{\partial^2 \mathcal{L}(x, \lambda)}{\partial \lambda_i \partial \lambda_j}$ is zero.

i.e think about

$H_{\mathcal{L}}$ = Hessian of $\mathcal{L}$ in x and λ

at x^c, λ^c, there are some dirns (the x dirns) δ_x s.t.

$$\delta_x^T H \delta_x \geq 0$$

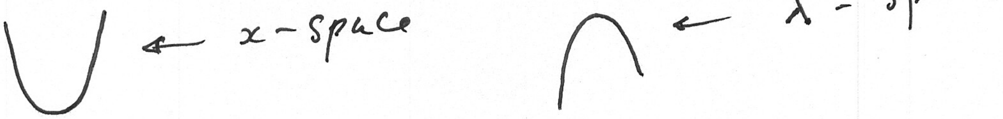
AND some dirns (λ dirns)

~~δλ~~ δ_λ st

$$\delta_\lambda^T H \delta_\lambda = 0$$

So x^c, λ^c must be a

saddle point



This means we could think about

$$q(\lambda) = \inf_{x} \mathcal{L}(x, \lambda)$$

Notice:

$$q(\lambda) \leqslant f(x^c).$$

(fairly easy: consider 1 constraint, then

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

now for $q(\lambda)$ to be finite, we must have $\lambda g(x)$ bounded below; if ~~$\lambda g(x)$~~ bound is greater than zero, no feasible point, so its less than zero; but then $\inf_{x} f(x) + \lambda g(x) \leqslant \inf_{\text{st } g(x)=0} f(x)$

and we're done

multiple dimensions follow)

This is powerful because we could consider

$$\underbrace{\max_{\lambda} q(\lambda)} \leq f(x^c).$$

if we have q, λ^k, x^k,

<u>and</u> $q(\lambda^k) - f(x^k)$ is small,

then $f(x^k) - f(x^c)$ is also small

this could help us track progress.

<u>Simple duals</u>:

~~① $\frac{x^T A x}{2}$~~

① $\underline{\min}$ $\frac{x^T x}{2}$ s.t. $Ax + b = 0$

$\mathcal{L}$: $\frac{x^T x}{2} - \lambda^T (Ax + b)$

now $\inf_{x} \mathcal{L}(x, \lambda)$ occurs when

$$x - A^T\lambda = 0$$

~~so $q(\lambda) = \lambda^T A A^T \lambda$~~

$$q(\lambda) = -\lambda^T \frac{AA^T}{2} \lambda - b^T \lambda$$

(subs. into $\mathcal{L}$)

(It's not always this easy)

<u>Notice</u> $\max_{\lambda} q(\lambda)$

occurs when

$$AA^T\lambda - b = 0$$

(i.e. at soln).

Interesting example: Variational calc.

Problem: find a P.D.F. that has a fixed set of Expectations (i.e. $\forall_i \; E_p(f_i) = m_i$ ← known number) while maximizing entropy.

→ useful modelling idea. We observe good estimates of some expectations in Data, and want model to respect these. But we know nothing else, so max entropy.

So
$$
\left.
\begin{aligned}
&\max -\int p \log p \, dx \\
&\text{s.t.} \quad \int p \, dx = 1 \\
&\int p \cdot f_i \, dx = m_i
\end{aligned}
\right]
$$

Variational problem, with constraint.

Lagrangian

$$\mathcal{L}(p) = -\int p \log p \, dx - \lambda_0\left[\int p \, dx - 1\right] - \sum_i \lambda_i \left[\int p f_i \, dx - m_i\right]$$

~~Recall~~ we want to form 2 gradients

$\nabla_\lambda \mathcal{L}$ is easy

$\nabla_p \mathcal{L}$ follows the case we saw earlier.

(i.e at p^*, $\left[\frac{d}{d\varepsilon}\mathcal{L}(p^* + \varepsilon\varphi)\right]\Big|_{\varepsilon=0} = 0$ for any φ.)

$$\frac{d}{d\varepsilon}\mathcal{L}(p^* + \varepsilon\varphi)\Big|_{\varepsilon=0} = \int \phi\left[-\log p^* - 1 - \lambda_0 - \sum_i \lambda_i f_i\right] dx.$$

this must be zero for any φ,

so

$$p^* \propto e^{-\lambda_0} \cdot e^{-\sum_i \lambda_i f_i(x)}$$

This class of model used to be called a maximum entropy model

Fitting:

- (Old way)
  - adjust λ_i so that

$$\int p^* f_i \, dx = m_i$$

$$\uparrow$$

$$\frac{1}{N}\sum_j f_i(x_j)$$ - an estimate from data of this expectation

- and λ_0 so that

$$\int p^* \, dx = 1.$$

But

Imagine we have a model of the form

$$p^*(x) = e^{-\lambda_0} e^{-\sum_i \lambda_i f_i(x)}$$

and we fit with Max likelihood

We must solve

$$\max \sum_j \log p^*(x_j) \qquad \text{s.t. } \int p^*(x)\,dx = 1$$

(problem in λ_0, λ_i)

now $\int p^*(x)\,dx = 1 = \int e^{-\lambda_0} e^{-\sum_i \lambda_i f_i(x)}\,dx$

$$= e^{-\lambda_0} \int e^{-\sum_i \lambda_i f_i(x)}\,dx$$

So $\lambda_0 = \log\left[\int e^{-\lambda_i f_i(x)}\,dx\right] = \log Z(\lambda_i)$

So we must solve :

$$\max_{\lambda_i} \sum_j \left[-\log Z - \sum_i \lambda_i f_i(x_j) \right] = Q(\lambda)$$

$$\frac{\partial Q}{\partial \lambda_k} = \sum_j \left[-\frac{1}{Z} \frac{\partial Z}{\partial \lambda_k} - f_k(x_j) \right]$$

$$Z = e^{\lambda_0} = \left[\int e^{-\sum_i \lambda_i f_i(x)} \, dx \right]$$

$$\frac{\partial Z}{\partial \lambda_k} = -\int e^{-\sum_i \lambda_i f_i(x)} \cdot f_k(x) \, dx$$

So $$-\frac{1}{Z} \frac{\partial Z}{\partial \lambda_k} = \int e^{-\lambda_0} \cdot e^{-\Sigma_i \lambda_i f_i(x)} \cdot f_k(x) \, dx$$

So we must solve

$$\boxed{\underset{\substack{\uparrow \\ \text{Exact expectations}}}{\int p^* f_k(x) dx} = \underset{\substack{\uparrow \\ \text{empirical expectations}}}{\frac{1}{N} \sum_j f_k(x_j)}}$$