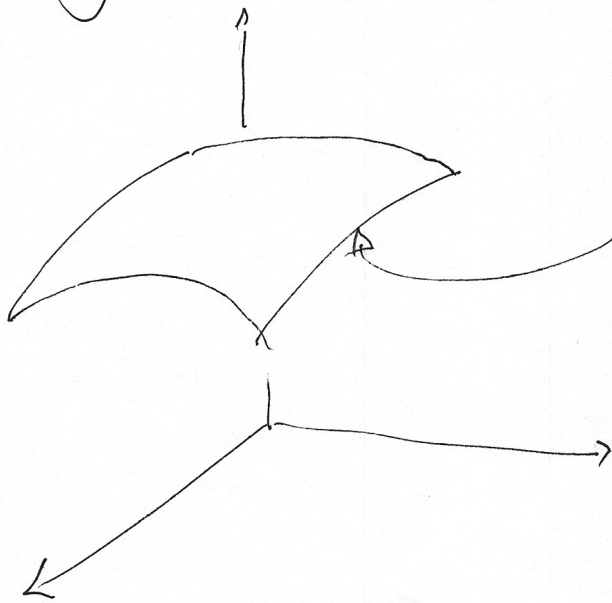


# Constrained Optimization:

①

Two kinds of problem - equality constraints  
- inequality constraints.

These are quite different, in important ways.

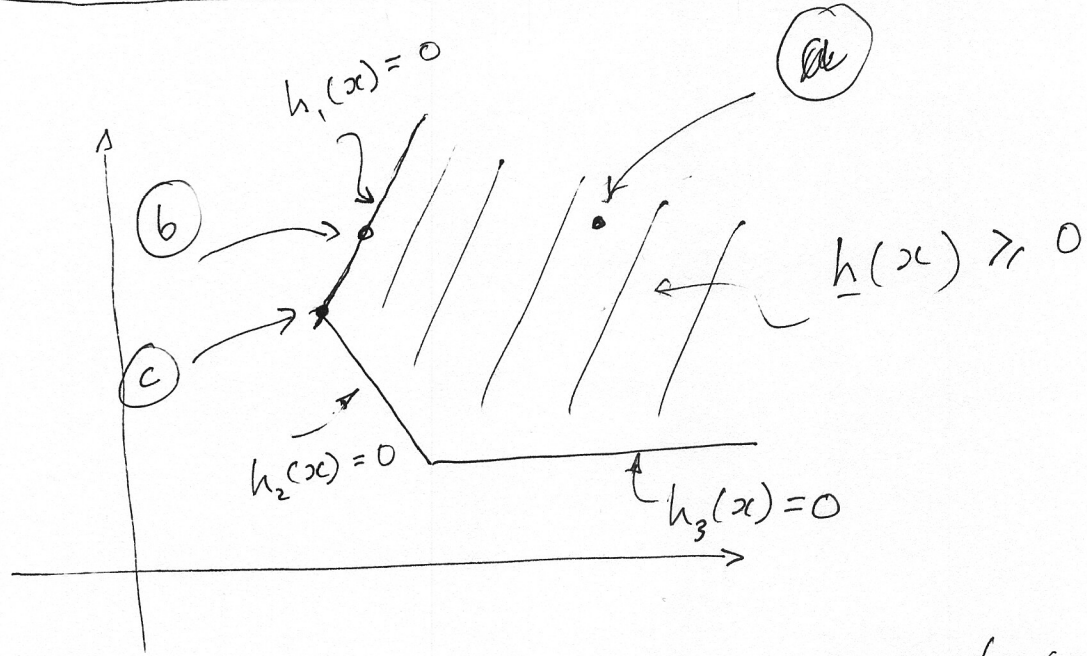


~~minimize~~ minimize  $f$  subject  
to  $g = 0$

$\equiv \nabla f$  is ~~not~~  
normal to the

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

# Inequality Picture:

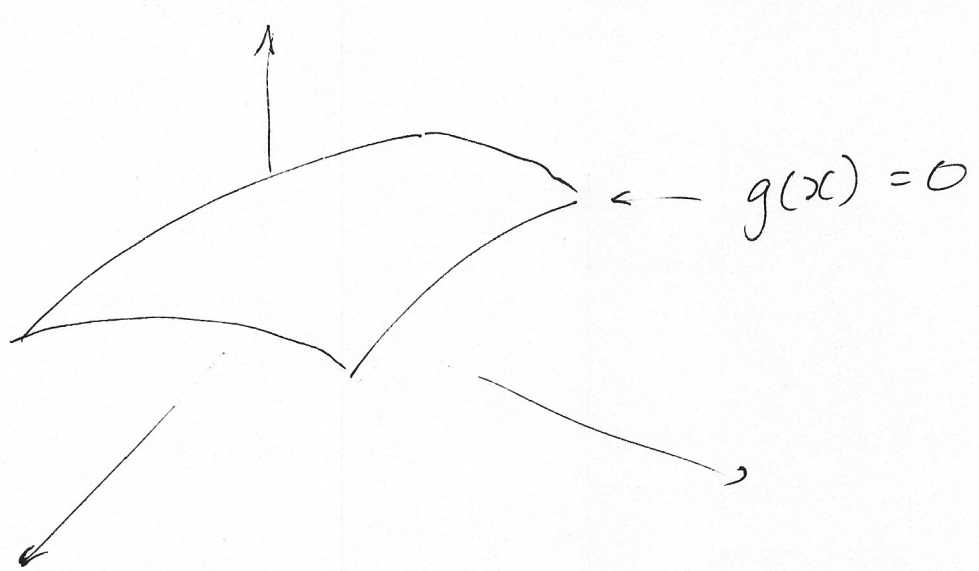


- at (a), constraints are irrelevant; locally, problem involves min  $f(x)$  w/o constraints
- at (b),  $h_1$  applies, but no others, so locally problem looks like min  $f(x)$  s.t.  $h_1(x) = 0$
- at (c),  $h_1, h_2$  ...

$$\min f(x) \quad \text{s.t. } \begin{cases} h_1(x) = 0 \\ h_2(x) = 0 \end{cases}$$

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  $\nabla f$  is normal to  $\{g(x) = 0\}$
- Normal of implicit surface  $g(x) = 0$  is  $\nabla g$
- $\therefore \nabla f = \lambda \nabla g$   
    ↑ some unknown constant.

(A)

What if there are many constraints in  $N-D$ ?

$$g_1(x) = 0, g_2(x) = 0, \text{ etc.}$$

$\nabla f$  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  $g = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \end{bmatrix}$

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

(5)

then

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

This justifies writing the Lagrangian

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

at minimum:

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

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

These conditions must be true at  
a minimum.

good estimates of some aspect

Interesting  
example

Variational calc.

(6)

Problem: find a PDF that has  
a fixed set of Expectations  
(i.e.  $\{ E_p(f_i) = m_i \}$  ← 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

$$\max - \int p \log p \, dx$$

st.

$$\int p \, dx = 1$$
$$\int p \cdot f_i \, dx = m_i$$

Variational  
problem  
with constraint.

# Lagrangian

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

~~recall~~ we want to form 2 gradients

$\nabla_x \mathcal{L}$  is easy

$\nabla_p \mathcal{L}$  follows the case we saw earlier.  
(i.e. at  $p^*$ ,  $\left. \frac{d}{d\varepsilon} \mathcal{L}(p^* + \varepsilon \varphi) \right|_{\varepsilon=0} = 0$  for any  $\varphi$ .)

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

this must be zero for any  $\varphi$ ,

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

⑧

Fitting:

. (old way)

. adjust  $\lambda_i$  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  $\lambda_0$  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

$$\max_j \sum \log p^*(x_j)$$

$$\text{st. } \int p^*(x) dx = 1$$

(problem in  $\lambda_0, \lambda_i$ )

$$\begin{aligned} \text{Now } \int p^*(x) dx = 1 &= \int e^{-\lambda_0} e^{-\sum_i \lambda_i f_i(x)} dx \\ &= e^{-\lambda_0} \int e^{-\sum_i \lambda_i f_i(x)} dx \end{aligned}$$

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

So we must solve:

(10)

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

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

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

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

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

So we must solve

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

# V. Important special cases for constrained optimization

---

①①

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

Lagrangian

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

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

↖ eigenvalue problem

---

(2)

(12)

$$\max \quad \frac{x^T A x}{2}$$

$$\text{st } x^T B x = 1$$

Lagrangian

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

$\therefore$

$$Ax - \lambda Bx = 0$$

↑ generalized eigenvalue problem

Notice:

NOT the same as

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

because  $B^{-1}$  may not exist

• Any good Numerical linear Alg package can do these.

③

⑬

min

$$\frac{x^T x}{2}$$

st

$$Ax = b$$

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

Lagrangian :

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

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

$(\nabla_x I)$   $\beta$

So

$$AA^T \lambda = b \quad \leftarrow \quad \alpha$$

Alg solve linear system  $\alpha$  for  $\lambda$ , then subs for  $x$ . in  $\beta$

4

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

Lagrangian:

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

$\nabla_x$

$$: Ax + b - C^T \lambda = 0$$

$\nabla_\lambda$

$$: Cx - d = 0$$

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

Solve this.

Notice how useful it has been to know the Lagrange multipliers. (15)

## 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

(16)

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

then try to min this.

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

- Problem - don't see large scale  
structure of objective

• Equivalent:

- take step Tangent to  
constraint space
- project back.

(i.e. make up local parametrization)



For example:

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

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

↑ vector function.

Now consider a step  $\Delta x$

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

so we could try:

Step:

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

$$\min_{\Delta x} \text{s.t. } (x_k + \Delta x)^T A (x_k + \Delta x) + b^T (x_k + \Delta x)$$

$$\text{s.t. } J_{\phi} \cdot \Delta x = 0$$

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. (18)

## Quadratic penalty method

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

• approach by  $\min f(\underline{x}) + \frac{c}{2} g^T g$

• if  $c$  is big, this forces  $g^T g$  to be small

• advantage: - we could take steps off the constraint space  
- now it's unconstrained.

# Disadvantages (Big)

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

$$H = H_f + c \left[ J_g^T J_g + \dots \right]$$

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

2) at soln,  $g$  isn't zero

at soln

$$\nabla_x f + c g J_g = 0$$

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

The method of multipliers (also, Augmented Lagrangian) method. (20)

$$\min f(x) \quad \text{s.t.} \quad g(x) = 0$$

form: Augmented Lagrangian

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

↑  
augmentation

• Now, assume we have an estimate  $\lambda^k$  of the LM's

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

at  $x^k$  we have

(21)

$$\nabla f(x_{**}^k) - \lambda^{kT} g J_g + c g^T J_g = 0$$

Now, pattern match to conditions

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

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

This suggests

$$\lambda^{k+1} = (\lambda^k - c g)$$

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

# ALM:

- start w  $x^0, d^0, c^0$
- min  $A(x, \lambda^k) = f + \lambda^{kT} g + \frac{c}{2} g^T g$
- to get  $x^k$
- $\lambda^{k+1} = \lambda^k - \frac{c}{2} g^T(x^k)$
- $c^{k+1} = \tau c^k$

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  $\lambda$  ests help.  
(formally, there is some bound on the  $c$  required to get exact soln.)

# First glimpse of Duality:

we have  $\min f(x)$  st  $g(x) = 0$

$$J = f(x) - \lambda^T g(x) = L(x, \lambda)$$

we have solution when

$$\left. \begin{aligned} \nabla_x L &= 0 \\ \nabla_\lambda L &= 0 \end{aligned} \right\} \text{so solution is at a critical point of } L.$$

→ what kind of c.p.?

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

(b) but for fixed  $x_* = \hat{x}$ ,  $L(\hat{x}, \lambda)$  is

(c) linear at  $\frac{x^c, \lambda^c}{\uparrow \text{solu}}$ ,  $\frac{\partial^2 L(x, \lambda)}{\partial \lambda_i \partial \lambda_j}$  is zero.

i.e think about

$$H_L = \text{Hessian of } L \text{ in } x \text{ and } \lambda$$

at  $x^c, \lambda^c$ , there are some dirs  
(the  $x$  dirs)  $\delta_x$  s.t.

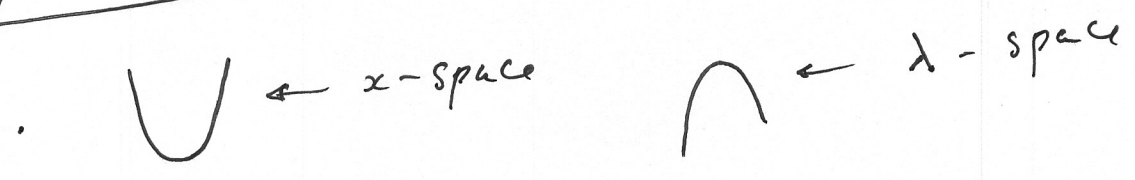
$$\delta_x^T H_{\delta_x} \geq 0$$

AND some dirs (in  $\lambda$  dirs)

~~st~~  $\delta_\lambda$  st  $\delta_\lambda^T H_{\delta_\lambda} = 0$

So  $x^c, \lambda^c$  must be a

Saddle point





This means we could think about

(25)

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

Notice:

$$q(\lambda) \leq f(x^c)$$

(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 it's less than zero; but then  $\inf_x f(x) + \lambda g(x) < \inf_{s.t. g(x)=0} f(x)$

multiple dimensions follow

This is powerful because we could consider

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

if we have  $q, \lambda^k, x^k$ ,  
and  $q(\lambda^k) - f(x^k)$  is small,  
then  $f(x^k) - f(x^c)$  is also small

This could help us track progress.

Simple Duals:

①  ~~$\frac{x^T A x}{2}$~~

① min  $\frac{x^T x}{2}$

st.  $Ax + b = 0$

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

Now  $\inf_x \mathcal{L}(x, \lambda)$  occurs when

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

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

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

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

(It's not always this easy)

Notice  $\max_{\lambda} q(\lambda)$

occurs when

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

(i.e. at soln).