Lecture 15

Nonlinear Problems

Newton's Method

Syllabus

Lecture 01	Describing Inverse Problems
Lecture 02	Probability and Measurement Error, Part 1
Lecture 03	Probability and Measurement Error, Part 2
Lecture 04	The L_2 Norm and Simple Least Squares
Lecture 05	A Priori Information and Weighted Least Squared
Lecture 06	Resolution and Generalized Inverses
Lecture 07	Backus-Gilbert Inverse and the Trade Off of Resolution and Variance
Lecture 08	The Principle of Maximum Likelihood
Lecture 09	Inexact Theories
Lecture 10	Nonuniqueness and Localized Averages
Lecture 11	Vector Spaces and Singular Value Decomposition
Lecture 12	Equality and Inequality Constraints
Lecture 13	L_1 , L_∞ Norm Problems and Linear Programming
Lecture 14	Nonlinear Problems: Grid and Monte Carlo Searches
Lecture 15	**Nonlinear Problems: Newton's Method**
Lecture 16	Nonlinear Problems: Simulated Annealing and Bootstrap Confidence Intervals
Lecture 17	Factor Analysis
Lecture 18	Varimax Factors, Empircal Orthogonal Functions
Lecture 19	Backus-Gilbert Theory for Continuous Problems; Radon's Problem
Lecture 20	Linear Operators and Their Adjoints
Lecture 21	Fréchet Derivatives
Lecture 22	Exemplary Inverse Problems, incl. Filter Design
Lecture 23	Exemplary Inverse Problems, incl. Earthquake Location
Lecture 24	Exemplary Inverse Problems, incl. Vibrational Problems

Purpose of the Lecture

Introduce Newton's Method

Generalize it to an Implicit Theory

Introduce the Gradient Method

Part 1

Newton's Method

grid search
Monte Carlo Method
are completely undirected

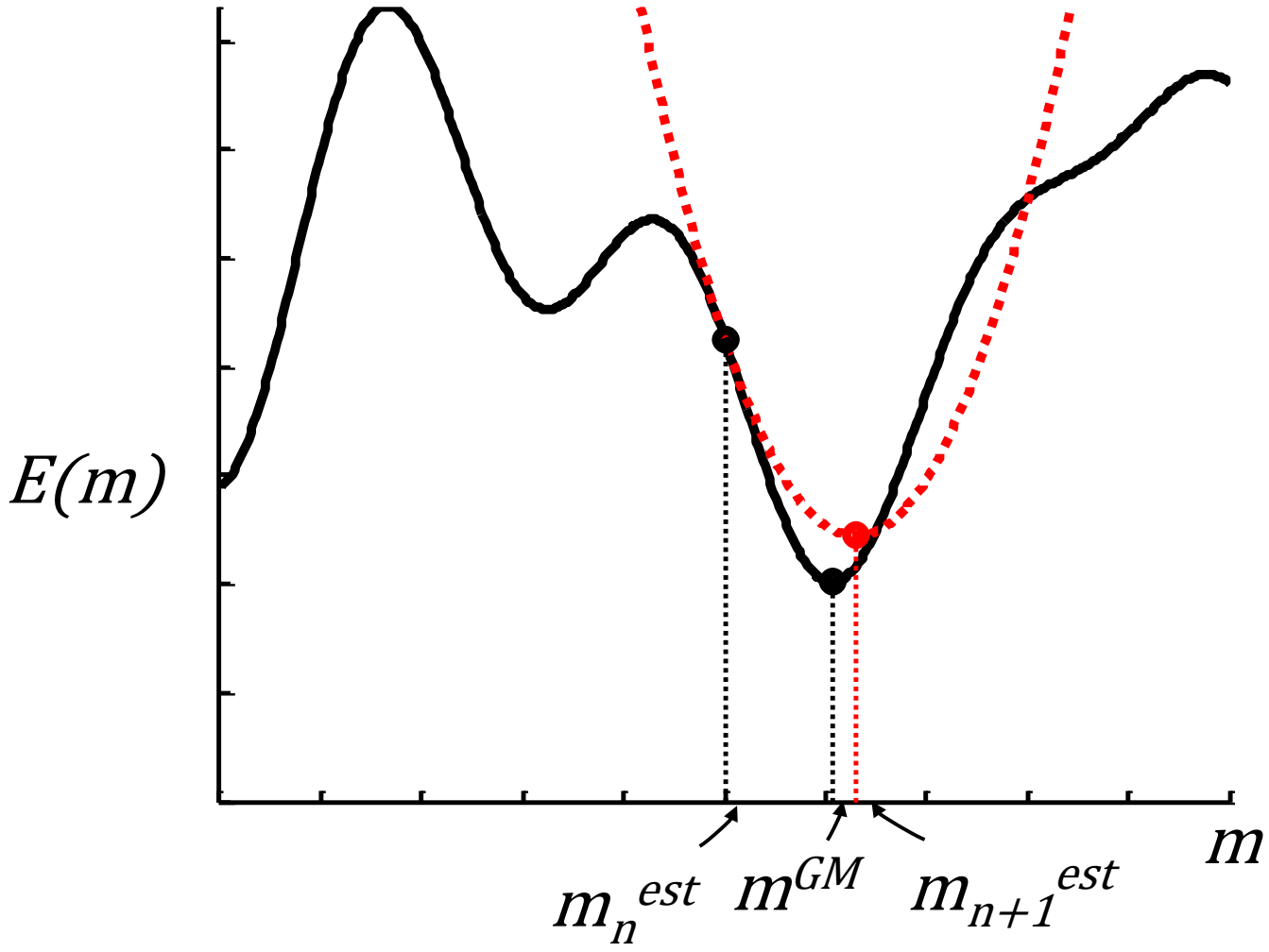
alternative
take directions from the
local properties
of the error function $E(\mathbf{m})$

Newton's Method

start with a guess $\mathbf{m}^{(p)}$

near $\mathbf{m}^{(p)}$, approximate $E(\mathbf{m})$ as a parabola and find its minimum

set new guess to this value and iterate


E(m)
m
m_n^{est}
m^{GM}
m_{n+1}^{est}

Taylor Series Approximation for $E(\mathbf{m})$

expand E around a point $\mathbf{m}^{(\mathrm{p})}$

$$E(\mathbf{m}) \approx E\left(\mathbf{m}^{(\mathrm{p})}\right) + \sum_{i=1}^{M} b_i \left(m_i - m_i^{(p)}\right) + \tfrac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} B_{ij} \left(m_i - m_i^{(p)}\right)\left(m_j - m_j^{(p)}\right)$$

$$\text{with} \quad b_i = \left.\frac{\partial E}{\partial m_i}\right|_{\mathbf{m}^{(\mathrm{p})}} \quad \text{and} \quad B_{ij} = \left.\frac{\partial^2 E}{\partial m_i \partial m_j}\right|_{\mathbf{m}^{(\mathrm{p})}}$$

differentiate and set result to zero to find minimum

$$E(\mathbf{m}) \approx E\left(\mathbf{m}^{(\mathrm{p})}\right) + \sum_{i=1}^{M} b_i \left(m_i - m_i^{(p)}\right) + \tfrac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} B_{ij} \left(m_i - m_i^{(p)}\right)\left(m_j - m_j^{(p)}\right)$$

$$\frac{\partial E(\mathbf{m})}{\partial m_q} = 0 = b_q + \sum_{j=1}^{M} B_{qj} \left(m_j - m_j^{(p)}\right)$$

$$\mathbf{m} - \mathbf{m}^{(\mathrm{p})} = -\mathbf{B}^{-1}\mathbf{b}$$

relate **b** and **B** to **g(m)**

$$E(\mathbf{m})=[\mathbf{d}-\mathbf{g}(\mathbf{m})]^T[\mathbf{d}-\mathbf{g}(\mathbf{m})]$$

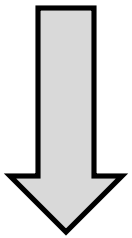
$$\mathbf{b} = -2\mathbf{G}^{(p)T}\left[\mathbf{d} - \mathbf{g}\left(\mathbf{m}^{(p)}\right)\right] \quad \text{and} \quad \mathbf{B} \approx 2\left[\mathbf{G}^{(p)T}\mathbf{G}^{(p)}\right]$$

$$\text{with} \quad G_{ij}^{(p)} = \left.\frac{\partial g_i}{\partial m_j}\right|_{\mathbf{m}^{(p)}}$$

linearized
data kernel

formula for approximate solution

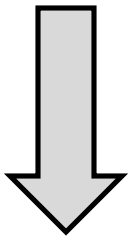
$$\mathbf{m} - \mathbf{m}^{(p)} = -\mathbf{B}^{-1}\mathbf{b}$$



$$\mathbf{m} - \mathbf{m}^{(p)} \approx \left[\mathbf{G}^{(p)T}\mathbf{G}^{(p)}\right]^{-1}\mathbf{G}^{(p)T}\left[\mathbf{d} - \mathbf{g}\left(\mathbf{m}^{(p)}\right)\right]$$

relate **b** and **B** to **g**(**m**)

$$\mathbf{m} - \mathbf{m}^{(p)} = -\mathbf{B}^{-1}\mathbf{b}$$



$$\mathbf{m} - \mathbf{m}^{(p)} \approx \left[\mathbf{G}^{(p)T}\mathbf{G}^{(p)}\right]^{-1}\mathbf{G}^{(p)T}\left[\mathbf{d} - \mathbf{g}\left(\mathbf{m}^{(p)}\right)\right]$$

very reminiscent
of least squares

what do you do if you can't analytically differentiate $\mathbf{g}(\mathbf{m})$?

use finite differences to numerically differentiate

$\mathbf{g}(\mathbf{m})$

or

$E(\mathbf{m})$

first derivative

$$\left.\frac{\partial E}{\partial m_i}\right|_{\mathbf{m}^{(p)}} \approx \frac{1}{\Delta m}\left\{E\left(\mathbf{m}+\Delta\mathbf{m}^{(\mathrm{i})}\right)-E(\mathbf{m})\right\}$$

first derivative

$$\left.\frac{\partial E}{\partial m_i}\right|_{\mathbf{m}^{(p)}} \approx \frac{1}{\Delta m}\left\{E\left(\mathbf{m} + \Delta\mathbf{m}^{(\mathrm{i})}\right) - E(\mathbf{m})\right\}$$

vector
$\Delta m\,[0, \ldots, 0, 1, 0, \ldots, 0]^{\mathrm{T}}$

need to evaluate $E(\mathbf{m})$ $M+1$ times

second derivative

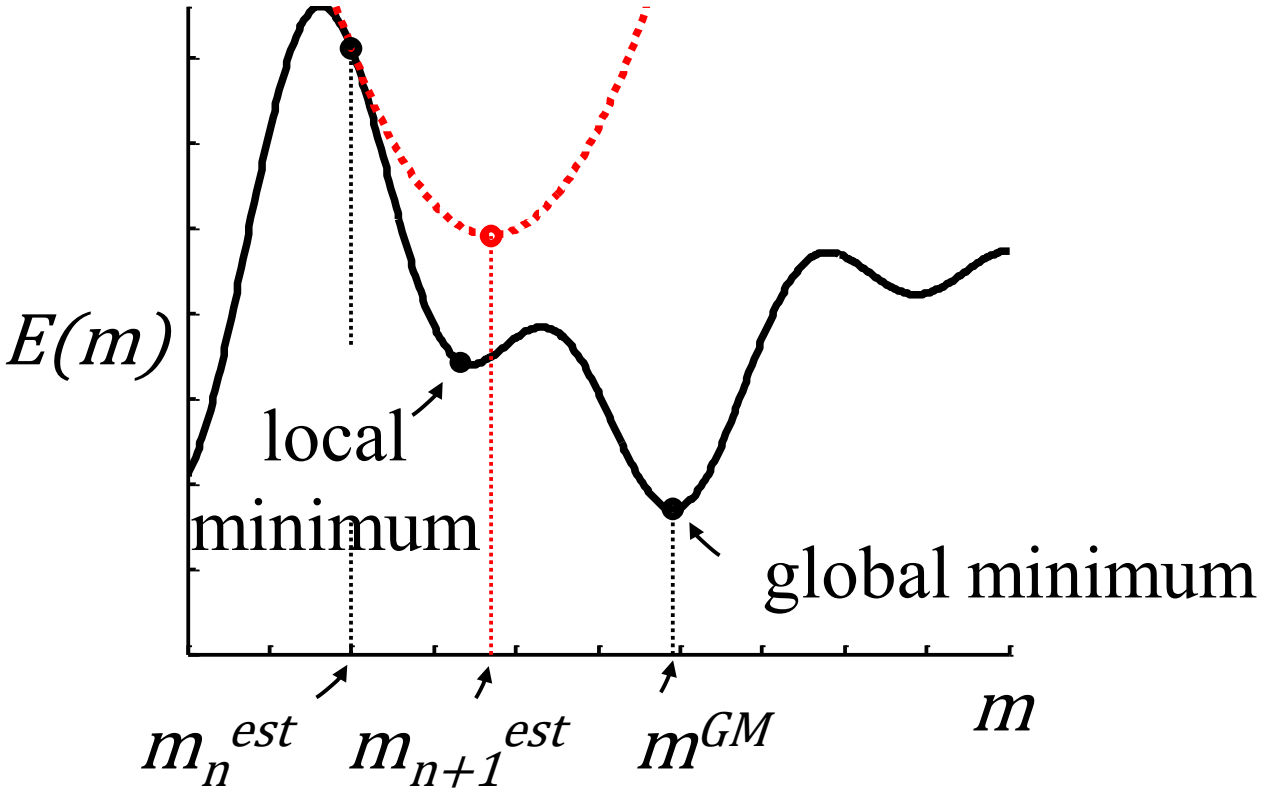
$$\left.\frac{\partial^2 E}{\partial m_i \partial m_j}\right|_{\mathbf{m}^{(p)}} \approx \frac{1}{(\Delta m)^2}\left\{E\left(\mathbf{m}+\Delta\mathbf{m}^{(\mathrm{i})}\right)-2E(\mathbf{m})+E\left(\mathbf{m}-\Delta\mathbf{m}^{(\mathrm{i})}\right)\right\} \quad \text{if} \quad i=j$$

$$\left.\frac{\partial^2 E}{\partial m_i \partial m_j}\right|_{\mathbf{m}^{(p)}} \approx \frac{1}{4(\Delta m)^2}\left\{E\left(\mathbf{m}+\Delta\mathbf{m}^{(\mathrm{i})}+\Delta\mathbf{m}^{(\mathrm{j})}\right)-E\left(\mathbf{m}+\Delta\mathbf{m}^{(\mathrm{i})}-\Delta\mathbf{m}^{(\mathrm{j})}\right)\right.$$
$$\left.-\ E\left(\mathbf{m}-\Delta\mathbf{m}^{(\mathrm{i})}+\Delta\mathbf{m}^{(\mathrm{j})}\right)+E\left(\mathbf{m}-\Delta\mathbf{m}^{(\mathrm{i})}-\Delta\mathbf{m}^{(\mathrm{j})}\right)\right\} \quad \text{if} \quad i\neq j$$

need to evaluate $E(\mathbf{m})$ about $\frac{1}{2}M^2$ times

what can go wrong?

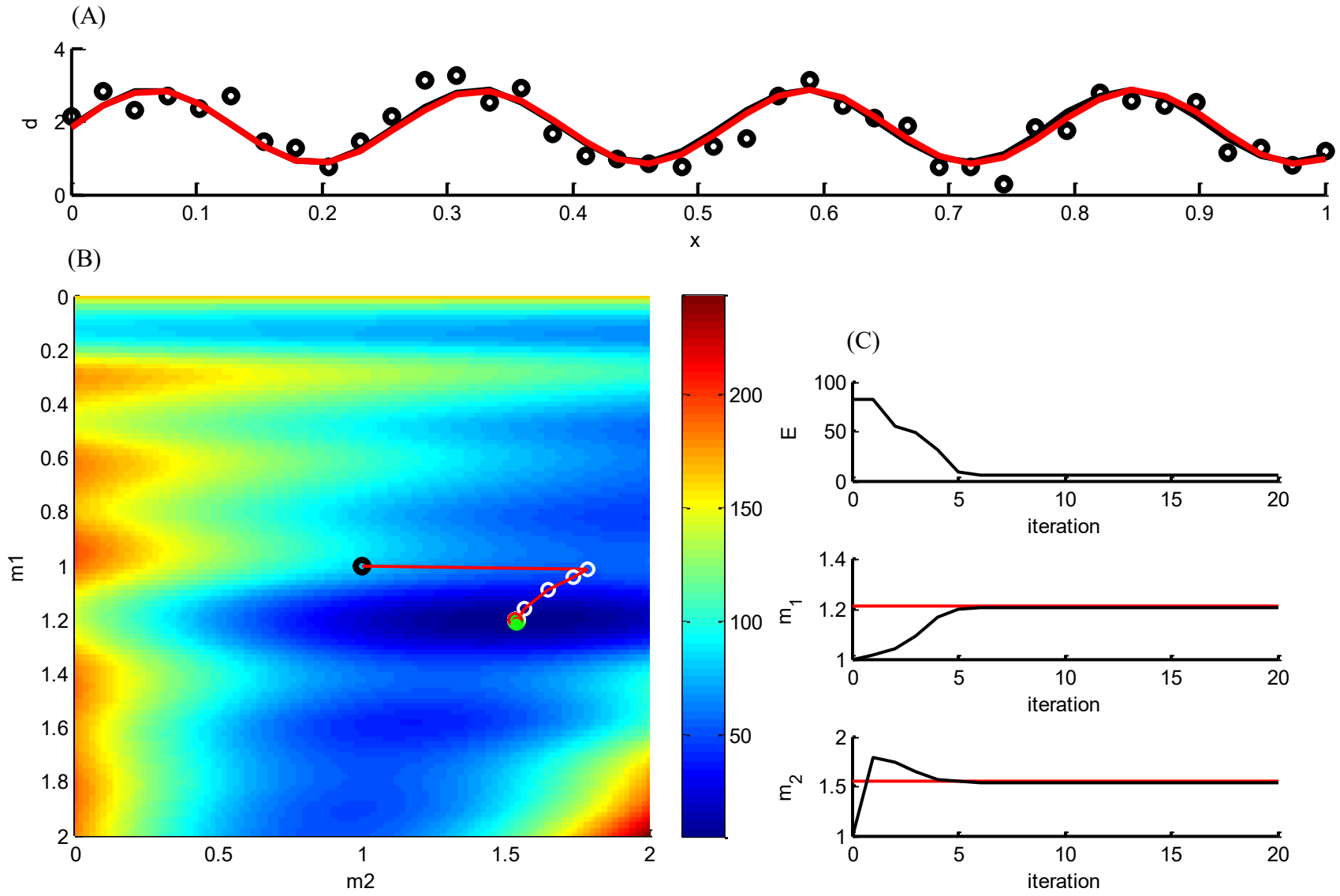
convergence to a *local minimum*


E(m)
local
minimum
global minimum
m
m_n^{est}
m_{n+1}^{est}
m^{GM}

analytically differentiate sample inverse problem

$$d_i(x_i) = \sin(\omega_0 m_1 x_i) + m_1 m_2$$

$$\mathbf{G}^{(\mathrm{p})} = \begin{bmatrix} \omega_0 x_1 \cos\left(\omega_0 x_1 m_1^{(p)}\right) + m_2^{(p)} & m_1^{(p)} \\ \omega_0 x_2 \cos\left(\omega_0 x_2 m_1^{(p)}\right) + m_2^{(p)} & m_1^{(p)} \\ \vdots & \vdots \\ \omega_0 x_N \cos\left(\omega_0 x_N m_1^{(p)}\right) + m_2^{(p)} & m_1^{(p)} \end{bmatrix}$$


(A)
d
x
(B)
m1
m2
(C)
E
iteration
m_1
iteration
m_2
iteration

often, the convergence is very rapid

often, the convergence is very rapid

but
sometimes the solution converges to a local minimum
and
sometimes it even diverges

```
mg = [1, 1]';
G = zeros(N,M);
for k = [1:Niter]

    dg = sin( w0*mg(1)*x) + mg(1)*mg(2);
    dd = dobs-dg;
    Eg=dd'*dd;

    G = zeros(N,2);
    G(:,1) = w0*x.*cos(w0*mg(1)*x) + mg(2);
    G(:,2) = mg(2)*ones(N,1);

    % least squares solution
    dm = (G'*G)\(G'*dd);

    % update
    mg = mg+dm;

end
```

Part 2

Newton's Method for an Implicit Theory

Implicit Theory

$\mathbf{f}(\mathbf{d},\mathbf{m})=0$

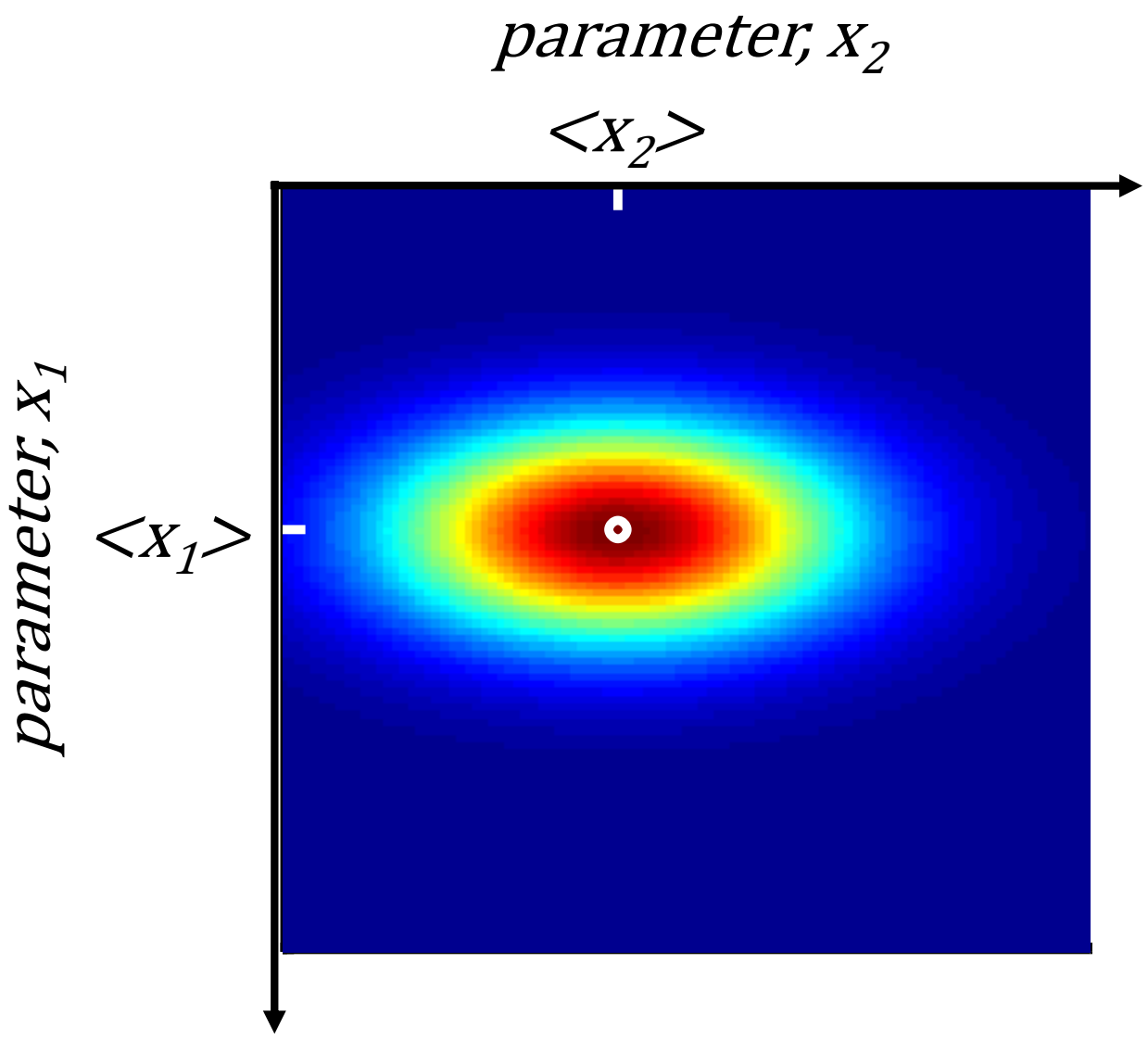
with Gaussian
prediction error
and
a priori information about $\mathbf{m}$

to simplify algebra
group **d**, **m** into a vector **x**

$$\mathbf{x} = [\mathbf{d}^T, \mathbf{m}^T]^T$$

$$\langle \mathbf{x} \rangle = [\mathbf{d}^{obsT}, \langle \mathbf{m} \rangle^T]^T$$

$$[\mathrm{cov}\, \mathbf{x}] = \begin{bmatrix} [\mathrm{cov}\, \mathbf{d}] & 0 \\ 0 & [\mathrm{cov}\, \mathbf{m}]_A \end{bmatrix}$$

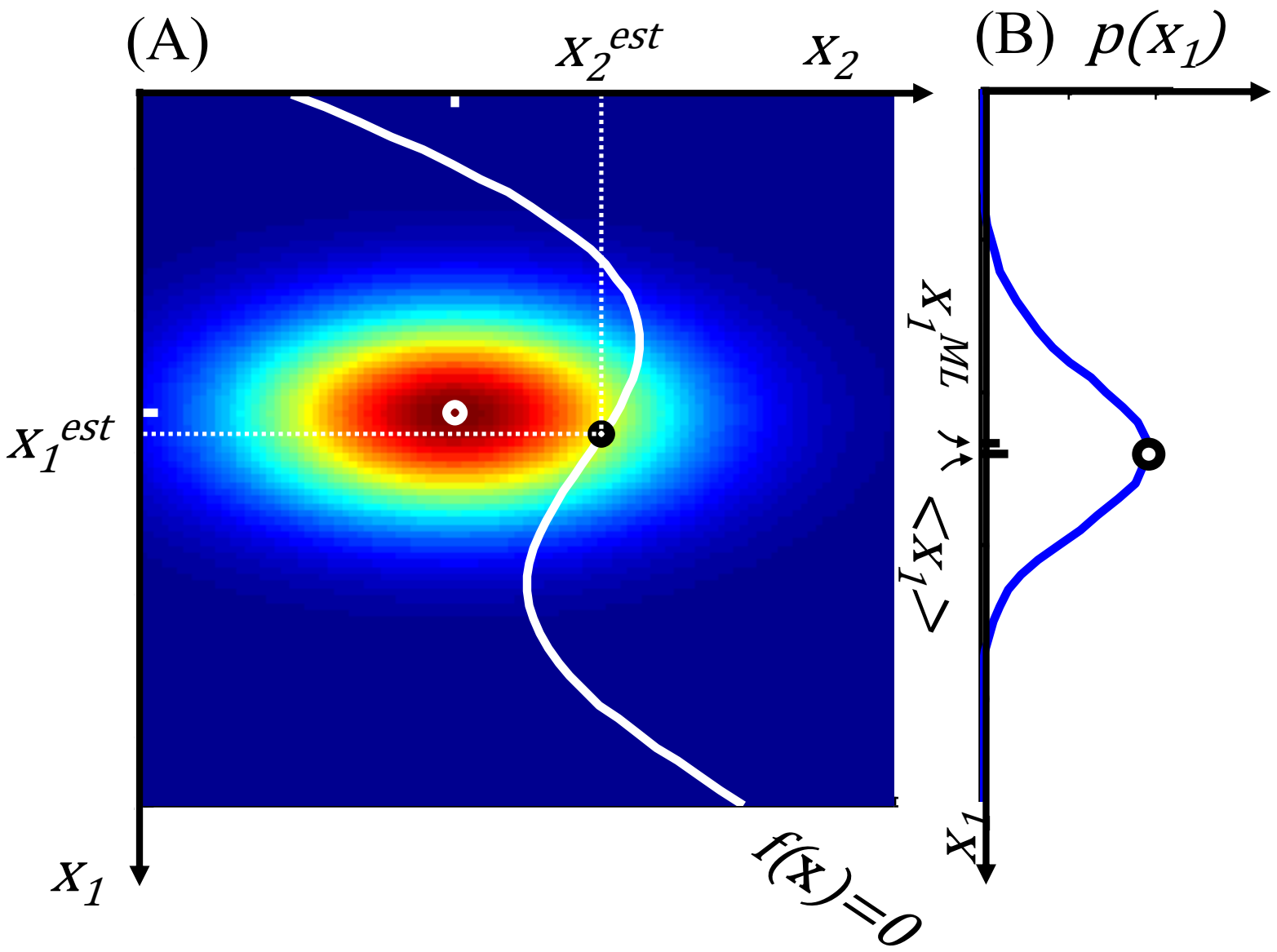

parameter, x_2
$\langle x_2 \rangle$
parameter, x_1
$\langle x_1 \rangle$

represent data and a priori model parameters as a Gaussian $p(\mathbf{x})$

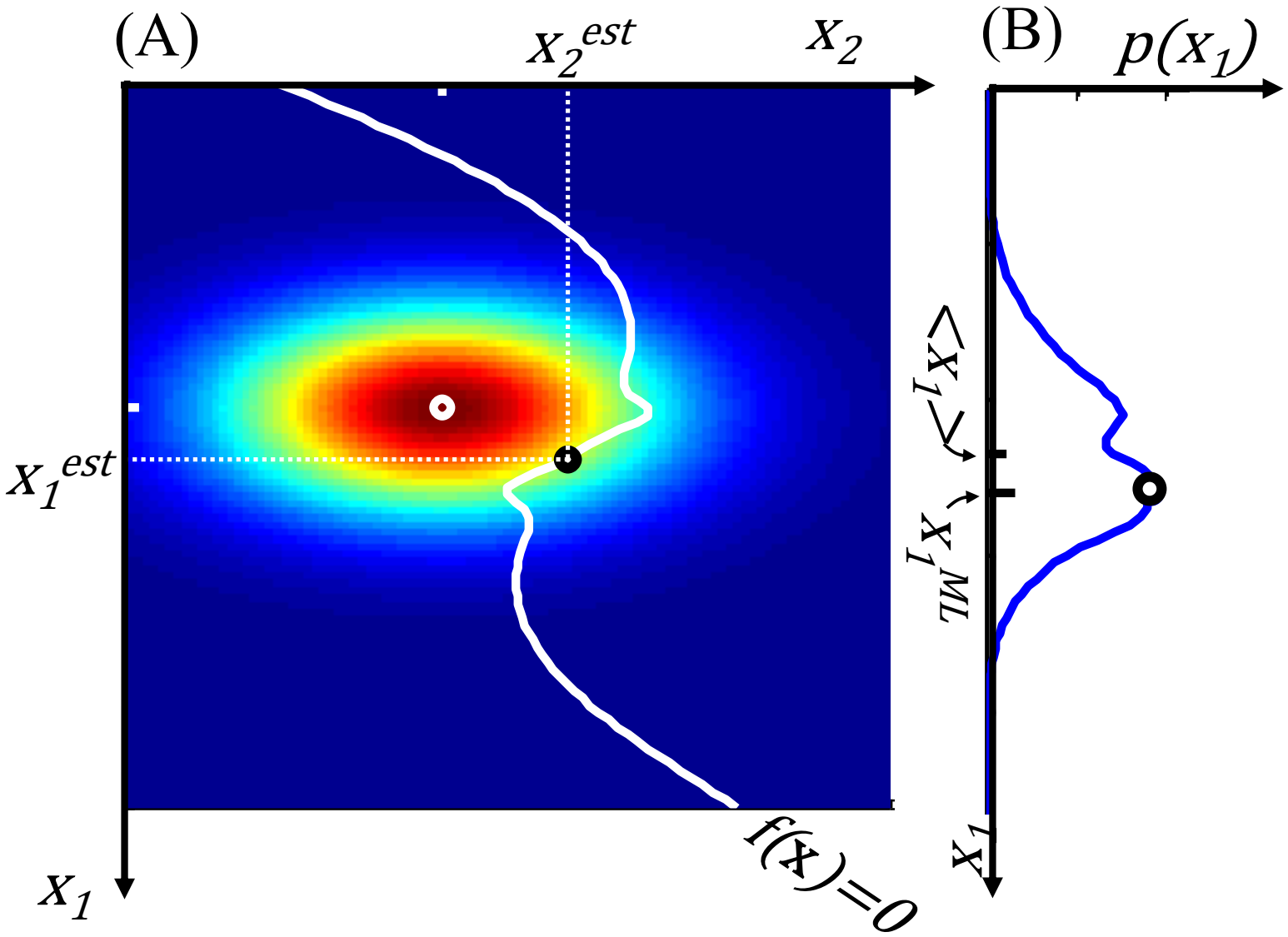
$\mathbf{f}(\mathbf{x})=0$ defines a surface in the space of $\mathbf{x}$

maximize $p(\mathbf{x})$ on this surface

maximum likelihood point is $\mathbf{x}^{\text{est}}$


(A)
x_2^{est}
x_2
x_1^{est}
x_1
$f(x)=0$
(B) $p(x_1)$
x_1^{ML}
$\langle x_1 \rangle$
x_1

can get local maxima if $\mathbf{f}(\mathbf{x})$ is very non-linear


(A)
x_2^{est}
x_2
x_1^{est}
x_1
f(x)=0
(B)
$p(x_1)$
$\langle x_1 \rangle$
x_1^{ML}
x_1

mathematical statement of the problem

$$\text{minimize} \quad \Phi = [\mathbf{x} - \langle \mathbf{x} \rangle]^T [\text{cov}\, \mathbf{x}]^{-1} [\mathbf{x} - \langle \mathbf{x} \rangle] \quad \text{subject to} \quad \mathbf{f}(\mathbf{x}) = 0$$

its solution (using Lagrange Multipliers)

$$[\mathbf{x} - \langle \mathbf{x} \rangle] = [\text{cov}\, \mathbf{x}]\mathbf{F}^{\mathrm{T}} \{\mathbf{F}[\text{cov}\, \mathbf{x}]\mathbf{F}^{\mathrm{T}}\}^{-1} \{\mathbf{F}[\mathbf{x} - \langle \mathbf{x} \rangle] - \mathbf{f}(\mathbf{x})\}$$

with $F_{ij} = \partial f_i / \partial x_j$

mathematical statement of the problem

$$\text{minimize} \quad \Phi = [\mathbf{x} - \langle \mathbf{x} \rangle]^T [\text{cov}\, \mathbf{x}]^{-1} [\mathbf{x} - \langle \mathbf{x} \rangle] \quad \text{subject to} \quad \mathbf{f}(\mathbf{x}) = 0$$

its solution (using Lagrange Multipliers)

$$[\mathbf{x} - \langle \mathbf{x} \rangle] = [\text{cov}\, \mathbf{x}]\mathbf{F}^T \{\mathbf{F}[\text{cov}\, \mathbf{x}]\mathbf{F}^T\}^{-1} \{\mathbf{F}[\mathbf{x} - \langle \mathbf{x} \rangle] - \mathbf{f}(\mathbf{x})\}$$

reminiscent of
minimum length solution

mathematical statement of the problem

$$\text{minimize} \quad \Phi = [\mathbf{x} - \langle \mathbf{x} \rangle]^T [\text{cov}\, \mathbf{x}]^{-1} [\mathbf{x} - \langle \mathbf{x} \rangle] \quad \text{subject to} \quad \mathbf{f}(\mathbf{x}) = 0$$

its solution (using Lagrange Multipliers)

$$[\mathbf{x} - \langle \mathbf{x} \rangle] = [\text{cov}\, \mathbf{x}]\mathbf{F}^T \{\mathbf{F}[\text{cov}\, \mathbf{x}]\mathbf{F}^T\}^{-1} \{\mathbf{F}[\mathbf{x} - \langle \mathbf{x} \rangle] - \mathbf{f}(\mathbf{x})\}$$

oops!

x appears in 3 places

solution
iterate !

$$\mathbf{x}^{(p+1)} = \langle \mathbf{x} \rangle + [\mathrm{cov}\,\mathbf{x}]\mathbf{F}^{(p)T}\left\{\mathbf{F}^{(p)}[\mathrm{cov}\,\mathbf{x}]\mathbf{F}^{(p)T}\right\}^{-1}\left\{\mathbf{F}^{(p)}\left[\mathbf{x}^{(p)} - \langle \mathbf{x} \rangle\right] - \mathbf{f}(\mathbf{x}^{(p)})\right\}$$

new value for $\mathbf{x}$ is $\mathbf{x}^{(p+1)}$

old value for $\mathbf{x}$ is $\mathbf{x}^{(p)}$

special case of an explicit theory

$\mathbf{f(x) = d\text{-}g(m)}$

$$\mathbf{m}^{(p+1)} = \langle \mathbf{m} \rangle + \mathbf{G}_{(p)}^{-g}\left\{\mathbf{d} - \mathbf{g}\left(\mathbf{m}^{(p)}\right) + \mathbf{G}^{(p)}\left[\mathbf{m}^{(p)} - \langle \mathbf{m} \rangle\right]\right\}$$

$$\mathbf{G}_{(p)}^{-g} = \left\{[\mathrm{cov}\,\mathbf{m}]_A^{-1} + \mathbf{G}^{(p)T}[\mathrm{cov}\,\mathbf{d}]^{-1}\mathbf{G}^{(p)}\right\}^{-1}\mathbf{G}^{(p)T}[\mathrm{cov}\,\mathbf{d}]^{-1}$$

equivalent to solving

$$\begin{bmatrix} [\mathrm{cov}\,\mathbf{d}]^{-½}\mathbf{G}^{(p)} \\ [\mathrm{cov}\,\mathbf{m}]_A^{-½}\mathbf{I} \end{bmatrix}\mathbf{m}^{(p+1)} = \begin{bmatrix} [\mathrm{cov}\,\mathbf{d}]^{-½}\{\mathbf{d} - \mathbf{g}\left(\mathbf{m}^{(p)}\right) + \mathbf{G}^{(p)}\mathbf{m}^{(p)}\} \\ [\mathrm{cov}\,\mathbf{m}]_A^{-½}\langle \mathbf{m} \rangle \end{bmatrix}$$

using simple least squares

special case of an explicit theory

$\mathbf{f(x) = d\text{-}g(m)}$

$$\mathbf{m}^{(p+1)} = \langle \mathbf{m} \rangle + \mathbf{G}_{(p)}^{-g}\left\{\mathbf{d} - \mathbf{g}\left(\mathbf{m}^{(p)}\right) + \mathbf{G}^{(p)}\left[\mathbf{m}^{(p)} - \langle \mathbf{m} \rangle\right]\right\}$$

$$\mathbf{G}_{(p)}^{-g} = \left\{[\mathrm{cov}\,\mathbf{m}]_A^{-1} + \mathbf{G}^{(p)T}[\mathrm{cov}\,\mathbf{d}]^{-1}\mathbf{G}^{(p)}\right\}^{-1}\mathbf{G}^{(p)T}[\mathrm{cov}\,\mathbf{d}]^{-1}$$

weighted least squares generalized inverse
with a linearized data kernel

special case of an explicit theory

$\mathbf{f(x) = d\text{-}g(m)}$

$$\mathbf{m}^{(p+1)} = \langle \mathbf{m} \rangle + \mathbf{G}_{(p)}^{-g}\left\{\mathbf{d} - \mathbf{g}\left(\mathbf{m}^{(p)}\right) + \mathbf{G}^{(p)}\left[\mathbf{m}^{(p)} - \langle \mathbf{m} \rangle\right]\right\}$$

Newton's Method, but making $E+L$ small
not just E small

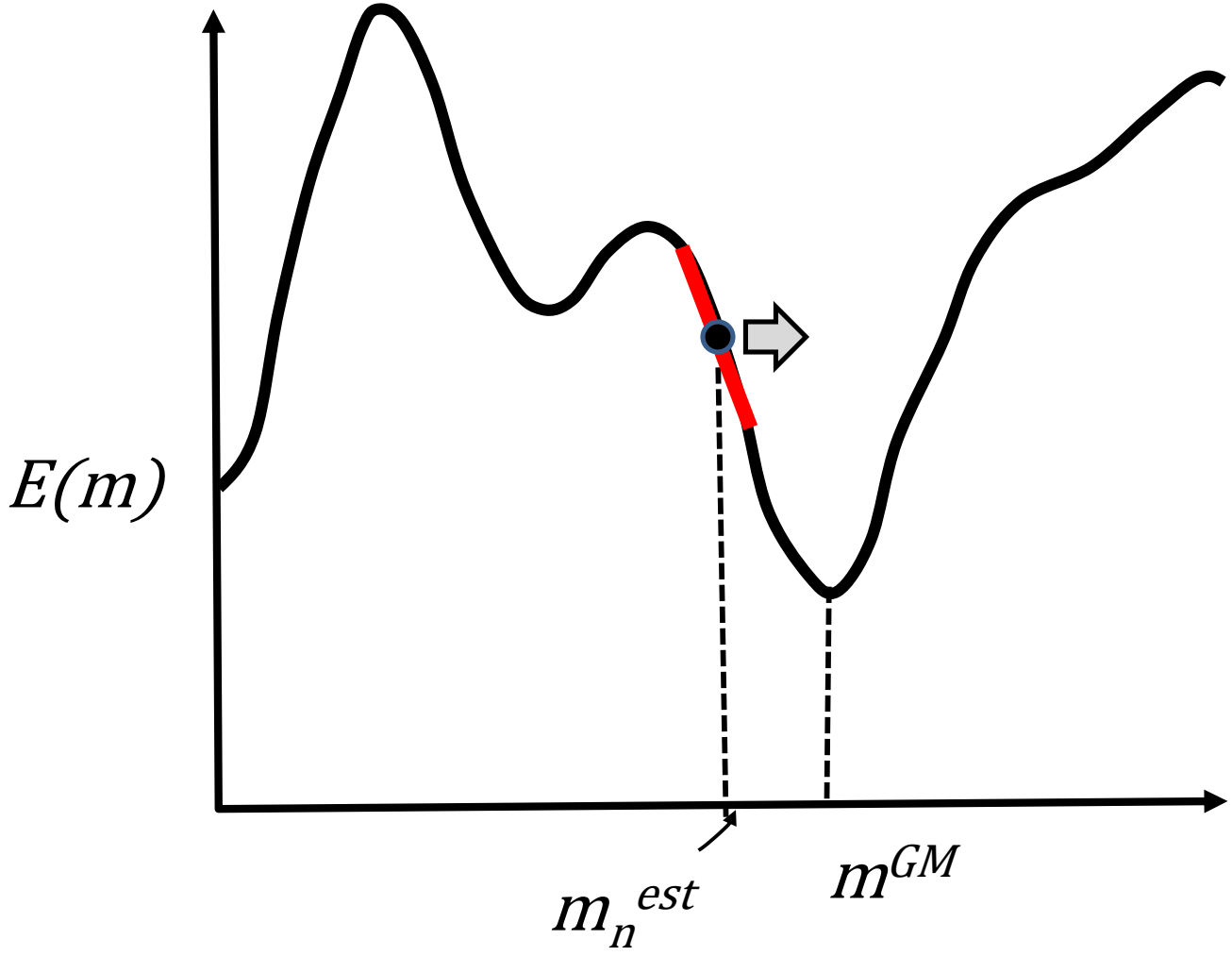
$$\mathbf{G}_{(p)}^{-g} = \left\{[\mathrm{cov}\,\mathbf{m}]_A^{-1} + \mathbf{G}^{(p)T}[\mathrm{cov}\,\mathbf{d}]^{-1}\mathbf{G}^{(p)}\right\}^{-1}\mathbf{G}^{(p)T}[\mathrm{cov}\,\mathbf{d}]^{-1}$$

Part 3

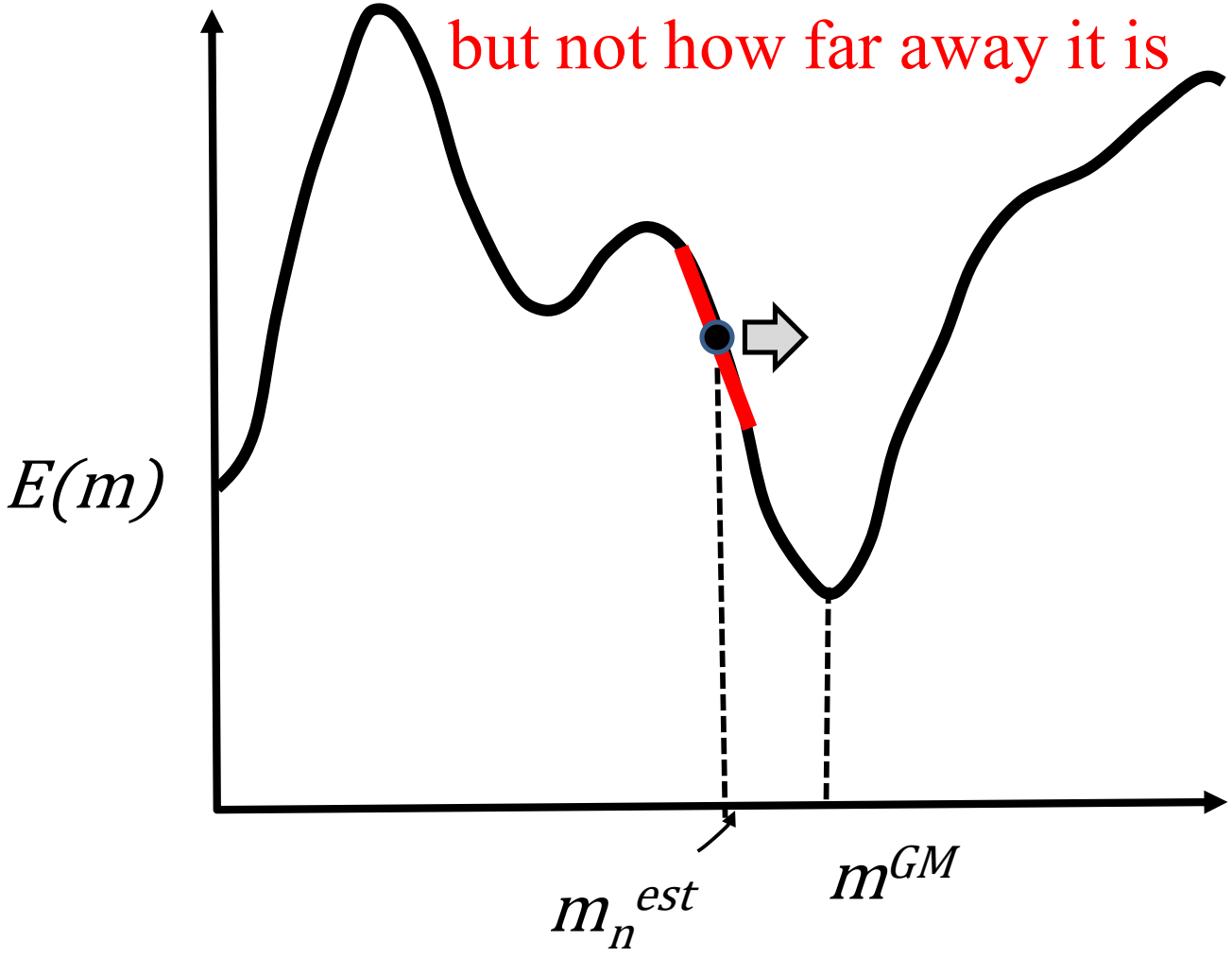
The Gradient Method

What if you can compute
$E(\mathbf{m})$ and $\partial E/\partial m_p$

but you can't compute
$\partial g/\partial m_p$ or $\partial^2 E/\partial m_p \partial\, m_q$


E(m)
m_n^{est}
m^{GM}

you know the direction
towards the minimum
but not how far away it is



unit vector pointing towards the minimum

$$\mathbf{v} = -\nabla E / |\nabla E|$$

so improved solution would be

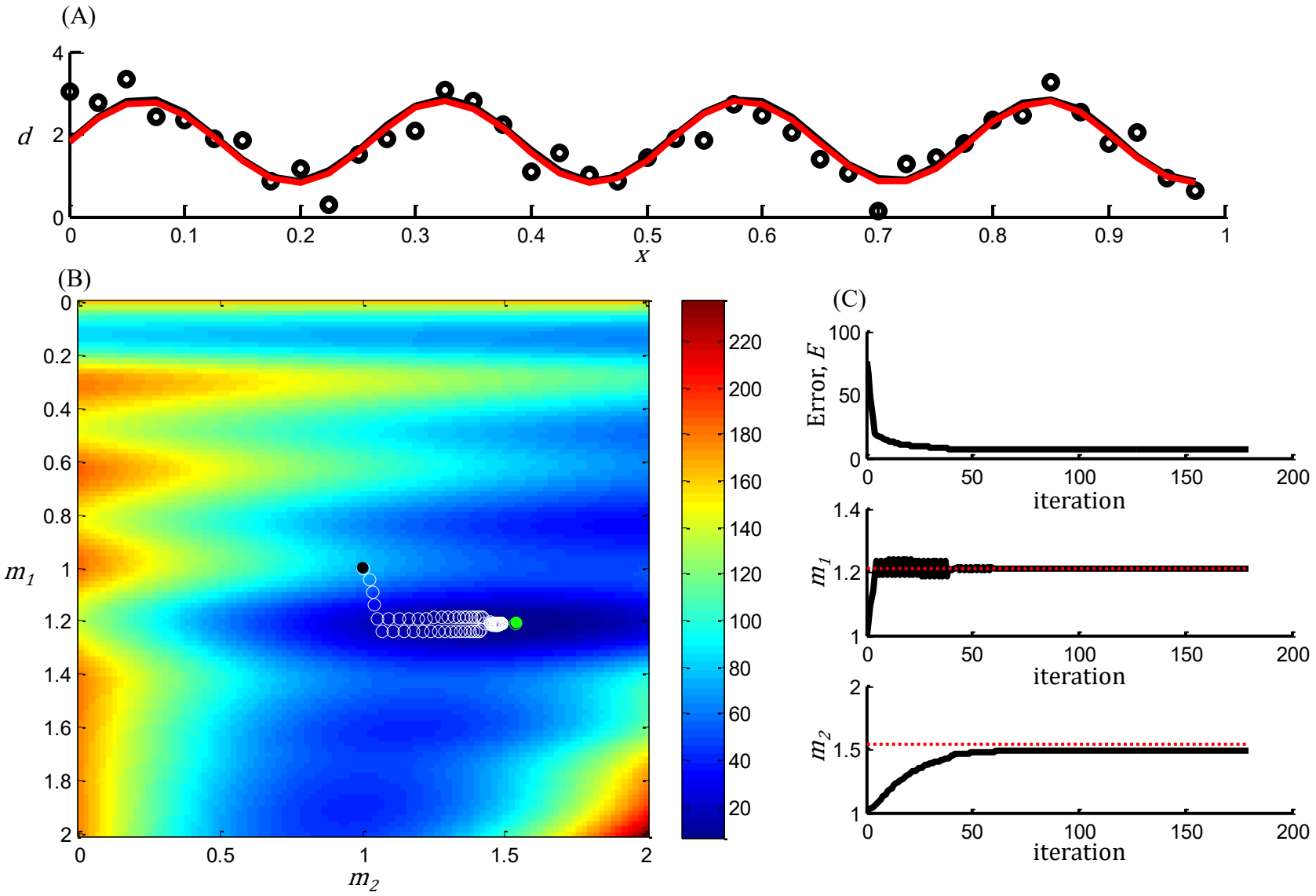
$$\mathbf{m}^{(j+1)} = \mathbf{m}^{(j)} + \alpha \mathbf{v}$$

if we knew how big to make α

Armijo's rule
provides an acceptance criterion for α

$$E\left(\mathbf{m}^{(k+1)}\right) \leq E\left(\mathbf{m}^{(k)}\right) + c\alpha\mathbf{v}^{\mathrm{T}}\nabla E|_{\mathbf{m}^{(k)}}$$ with $c \approx 10^{-4}$

simple strategy
start with a largish α
divide it by 2 whenever it fails Armijo's Rule


(A)
d
x
(B)
m_1
m_2
(C)
Error, E
iteration
m_1
iteration
m_2
iteration

```
% error and its gradient at the trial solution
mgo=[1,1]';
ygo = sin( w0*mgo(1)*x) + mgo(1)*mgo(2);
Ego = (ygo-y)'*(ygo-y);
dydmo = zeros(N,2);
dydmo(:,1) = w0*x.*cos(w0*mgo(1)*x) + mgo(2);
dydmo(:,2) = mgo(2)*ones(N,1);
dEdmo = 2*dydmo'*(ygo-y);

alpha = 0.05;
c1 = 0.0001;
tau = 0.5;

Niter=500;
for k = [1:Niter]
    v = -dEdmo / sqrt(dEdmo'*dEdmo);
```

```
% backstep
    for kk=[1:10]
        mg = mgo+alpha*v;
        yg = sin(w0*mg(1)*x)+mg(1)*mg(2);
        Eg = (yg-y)'*(yg-y);
        dydm = zeros(N,2);
        dydm(:,1) = w0*x.*cos(w0*mg(1)*x)+ mg(2);
        dydm(:,2) = mg(2)*ones(N,1);
        dEdm = 2*dydm'*(yg-y);
        if( (Eg<=(Ego + c1*alpha*v'*dEdmo)) )
            break;
        end
        alpha = tau*alpha;
    end
```

```
    % change in solution
    Dmg = sqrt( (mg-mgo)'*(mg-mgo) );

    % update
    mgo=mg;
    ygo = yg;
    Ego = Eg;
    dydmo = dydm;
    dEdmo = dEdm;

    if( Dmg < 1.0e-6 )
        break;
    end
end
```

often, the convergence is reasonably rapid

often, the convergence is reasonably rapid

exception

when the minimum is in along a long shallow valley