Finite Differencing

$$\frac{\partial f}{\partial x_i} \approx \frac{f(x+\epsilon e_i) - f(x-\epsilon e_i)}{2\varepsilon} \quad \text{— (1)}$$

* $\varepsilon > 0$ is "small"
* $e_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ 1 in i-th position only.

* (1) is called the <u>central-difference</u> formula.

Another formula is:

$$\frac{\partial f}{\partial x_i}(x) \approx \frac{f(x+\epsilon e_i) - f(x)}{\epsilon} \quad \text{— (2)}$$

(2) is called the <u>forward-difference</u> formula.

Note that (1) + (2) refer to one-dimensional functions of x_i, where all other coordinates are fixed.

Unit Roundoff (p. 60+)

Represent floating point numbers using $.d_1 d_2 \cdots d_t$ binary digits

The number is:

$$\sum_{i=1}^{t} d_i 2^{-i} \times 2^{e},$$

where e is an exponent.

Define $u = 2^{-t}$ to be the unit roundoff.

Now add 1 to a floating point number:

$$1 + \sum_{i=1}^{t} d_i 2^{-i} \times 2^{e}$$

and we don't have enough digits (in general) to represent the number.

Eg: $t=2$. Let $A = (0.11)_2$

$$\Rightarrow \quad A = 1 \times 2^{-1} + 1 \times 2^{-2} = \sum_{i=1}^{2} d_i 2^{-i} \times 2^{0}$$

$$1 + A = (1.11)_2 = (0.111)_2 \times 2^{+1}$$

If we truncate the least-significant bit:

$1+A \cong (0.11)_2 \times 2^{+1}$ with an error of $(2^{-3}) \times 2^{+1} = 2^{-t}$.

Now if we apply roundoff:

$1+A = (0.111)_2 \times 2^{+1} \cong 1 \times 2^{+1}$

with an error of $(2^{-3} \times 2^{+1}) = 2^{-t}$ again.

It this example: 2^{-t} is the worst error for roundoff arithmetic.

For truncation arithmetic:

$$2^{+1} \times (0.00\underbrace{111\cdots}_{\text{infinite number of 1s}})_2 = (0.01)_2 \times 2^{+1} = 2^{-t+1}$$

is the worst error.

For determining t, Dennis & Schnabel use the concept of a machine-epsilon (machep) which is computed as the smallest τ such that $1+\tau > 1$. Here is the code:

```
τ = 1;
while (1+τ*0.5>1) do
      τ = 0.5*τ;
```

which exits with τ.

Note that on some machines the first bit omitted, assumed to be 1.

Floating-Point Arithmetic (Conte & de Boor,

Ch7 $\frac{t}{24}$

For an n-digit number:

$$fl(x) = \pm \underbrace{(.d_1 d_2 \cdots d_n)_\beta}_{} \beta^e$$

mantissa, $\beta = 2$ (except calculators!)

and e is the exponent.

Assume: $m < e < M$

Overflow: $|x| > \beta^M$.

Underflow: $|x| \leq \beta^{m-n}$

Write: $fl(x) = x(1+\delta)$

where:

* $|\delta| < \frac{1}{2}\beta^{1-n}$ in rounding arithmetic,

* $-\beta^{1-n} < \delta \leq 0$ in truncating (chopping).

$$\max_x |\delta| = \text{unit roundoff} = u$$

Relationship between actual and floating-point approximations:

$$x \,\square^{*}\, y = (x \,\square\, y)(1+\delta)$$

<u>for</u> $\square$: $+, -, *, /$ operations
and some δ with $|\delta| \leq u$,
$\square^{*}$ = floating point operation,
$\square$ = exact operation.

<u>Eg.</u> $\underbrace{x \overset{*}{+} y}_{\text{floating point}} = \underbrace{(x+y)}_{\text{actual}}(1+\delta)$, $|\delta| \leq u$

Equivalently:

$$x \,\square^{*}\, y = (x \,\square\, y) / (1+\delta)$$
for some δ with $|\delta| \leq u$

Application:

Consider the code:

$$s_0 := c$$
$$s_i := s_{i-1} - a_i b_i, \quad i = 1, 2, \ldots, r$$
$$s := s_r / a_{r+1}$$

Apply the approximation model:

$$\hat{s}_0 = c$$
$$\hat{s}_i = \left[\hat{s}_{i-1} - a_i b_i (1+\delta) \right] (1+\delta)$$
$$i = 1, 2, \ldots, r$$

for multiplying

for subtracting

$$\hat{s} = \hat{s}_r / [a_{r+1}(1+\delta)]$$

Note that we have used the same δ w/out subscripts. This follows by first using δ_i, and then showing that this can be simplified to this form.

1.4 Loss of Significance and Error Propagation Condition and instability

ch 7 - 8
24

Write:

$$\text{Exact} = \text{approximation} + \text{Error}$$
$$x = x^* + e$$

Relative error is: $(x - x^*) / x$.

Loss of significance digits:

x^* approximates x to r significant β-digits provided that:

$$|x - x^*| \leq \tfrac{1}{2} \beta^{s-r+1}$$

where: * $\beta = 2$ for non-calculators

* s is given by: $\max_{s} \beta^{s} \leq |x|$, and s is integer.

* (r is the number of significant digits to be determined.)

E.g.

For $x^* = 3$, $x = \pi$, $\beta = 10$.

$$\Rightarrow \max_s 10^s \leq \pi \Rightarrow s = 0.$$

$$\Rightarrow |3 - \pi| \leq \tfrac{1}{2} 10^{1-r}$$

is $0.14\cdots \leq \tfrac{1}{2} 10^{1-r}$

for $r = 1$, but fails for $r = 2$.

($0.14\cdots \leq \tfrac{1}{2} \times \tfrac{1}{10} = 0.05$ fails)

$\Rightarrow$ 3 agrees to π to one-significant digit

For $x^* = 22/7$, we have agreement to three significant digits (verify)

Condition

From: $f(x) - f(x^*) \approx f'(x)(x - x^*)$,

consider the quotient of relative errors:

$$\frac{\left(\dfrac{f(x)-f(x^*)}{f(x)}\right)}{\left(\dfrac{x-x^*}{x}\right)} = \frac{f(x)-f(x^*)}{(x-x^*)} \times \frac{x}{f(x)}$$

$$\approx \frac{f'(x)}{f(x)} \times x \quad \text{for } |x - x^*| \text{ small.}$$

Formally, define:

$$\text{cond}(f)(x) = \max \left\{ \left| \frac{f(x)-f(x^*)}{f(x)} \right| \Big/ \left| \frac{x-x^*}{x} \right| \text{ such that: } |x - x^*| \text{ is small} \right\}$$

and write:

$$\text{cond}(f)(x) \approx \left| \frac{f'(x) \times x}{f(x)} \right|$$

Note that cond (f) measures how a small initial error gets amplified through a computation:

$$\text{max-relative-error-of}(f)(x) = \text{cond}(f) \cdot \left|\frac{x - x^*}{x}\right|$$

<u>Ex:</u> $f(x) = \dfrac{10}{1-x^2}$

$$\Rightarrow \left|\frac{f'(x)\, x}{f(x)}\right| = \frac{2x^2}{|1-x^2|}$$ which is <u>unstable</u> <u>for $x = \pm 1$</u>.

<u>For stable computation, require that each step in the computation has a small condition number for every value of x.</u> Else, we have instability.

Example:

$$f(x) = \sqrt{x+1} - \sqrt{x} \sim \circledast$$

is computed using:

$$\begin{aligned}
x_0 &:= x & &\Rightarrow f_0(x) = x \\
x_1 &:= x_0 + 1 & &\Rightarrow f_1(x) = x + 1 \\
x_2 &:= \sqrt{x_1} & &\Rightarrow f_2(x) = \sqrt{x} \\
x_3 &:= \sqrt{x_0} & &\Rightarrow f_3(x) = \sqrt{x} \\
x_4 &:= x_2 - x_3 & &\Rightarrow f_4(x,y) = x - y
\end{aligned}$$

Fix x_2, to get $g(t) = x_2 - t$

This gives:

$$\left| \frac{g'(t) \times t}{g(t)} \right| = \left| \frac{t}{x_2 - t} \right|$$

For $x = 12345$, $\left| \frac{t}{x_2 - t} \right| \sim \underline{\underline{22{,}222}}$

while $\text{cond}(f)(12345) = \left| \frac{f'(x)}{f(x)} \times x \right| \simeq \underline{\underline{1/2}}$

$\Rightarrow$ Original f is ok, but the algorithm is 44,000 worse in relative error.

This kind of error can usually be fixed.

Observe:

$$f(x) = \left(\sqrt{x+1} - \sqrt{x}\right) \times \frac{\left(\sqrt{x+1} + \sqrt{x}\right)}{\left(\sqrt{x+1} + \sqrt{x}\right)} = \frac{\cancel{x}+1-\cancel{x}}{\sqrt{x+1} + \sqrt{x}}$$

which gives the following algorithm:

$$x_0 := x$$
$$x_1 := x_0 + 1$$
$$x_2 := \sqrt{x_1}$$
$$x_3 := \sqrt{x_0}$$
$$\begin{cases} x_4 := x_2 + x_3 \\ x_5 := 1/x_4 \end{cases}$$

Now the last two steps compute:

$$g(t) = \frac{1}{(x_2+t)} \Rightarrow \left|\frac{g'(t)\,t}{g(t)}\right| = \left|\frac{t}{x_2+t}\right|$$

where $\left|\frac{t}{x_2+t}\right| \approx 1/2$ for $t \approx x_2$,

which is more stable, as stable as $f(x)$! $(\approx 1/2)$.

For matrices, we use:

$$\boxed{\text{cond}(A) = \|A\| \cdot \|A^{-1}\|}$$

where:

$$\|A\| = \max \frac{\|Ax\|}{\|x\|}$$

if they exist!

For the Euclidean norm (p.173):

$$\|A\|_2 = \max_x \frac{\|Ax\|_2}{\|x\|_2} \quad \left(= \frac{\lambda_1 \leftarrow \text{largest}}{\lambda_n \leftarrow \text{smallest eigenvalue}} \right),$$

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}, \text{ while:}$$

the maximum norm:

$$\|A\|_\infty = \max_x \frac{\|Ax\|_\infty}{\|x\|_\infty}$$

where $\|x\|_\infty = \max\limits_{1 \le i \le n} |x_i|$

Eg: $\left\| \begin{bmatrix} 1 \\ -2 \end{bmatrix} \right\|_\infty = 2$

$$\boxed{\|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|}$$

p.173

Eg: For $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$.

$$\|A\|_\infty = \max(1+2, 3+4) = 7$$

Then, similar to before:

for $Ax = b$, $e = x - \hat{x}$ is the error for $\hat{x}$ approximating x and for the residual error: $\boxed{r = Ax - A\hat{x} = Ae}$

We can show:

$$\frac{1}{\text{cond}(A)} \cdot \frac{\|r\|}{\|b\|} \leq \underbrace{\frac{\|e\|}{\|x\|}} \leq \text{cond}(A) \cdot \frac{\|r\|}{\|b\|}$$

relative error in solving $Ax = b$.

Determining the step-size for finite-differencing

To provide optimal values for ϵ for finite differencing, we will follow:

Conte and de Boor, "Elementary Numerical Analysis", 3rd ed.

Finite-differencing approximates f using polynomials, and then bounds the error.

Eg:

Forward-differencing:

$$f'(x) \approx \frac{1}{\varepsilon}\left[f(x+\varepsilon) - f(x)\right]$$

with an error of: $E(f) = -\frac{1}{2}\varepsilon f''(\xi)$,

for some $\xi \in (x, x+\varepsilon)$.

This is an $O(\varepsilon)$ approximation.

So, if we have a bound for $f''(\cdot)$, L

then: $\left| f'(x) - \frac{1}{\varepsilon}\left[f(x+\varepsilon) - f(x)\right]\right| < \frac{1}{2}\varepsilon L$

For <u>central differencing</u>:

$$f'(x) \approx \frac{1}{2\varepsilon}\left[f(x+\varepsilon) - f(x-\epsilon)\right]$$

with an error of $E(f) = -\varepsilon^2/6 \times f'''(\xi)$ where $\xi \in (x-\varepsilon, x+\varepsilon)$. The method is of $O(\varepsilon^2)$.

<u>For the second-derivative:</u>

$$f''(x) \approx \frac{1}{\varepsilon^2}\left[f(x-\varepsilon) - 2f(x) + f(x+\varepsilon)\right]$$

with an error of: $E(f) = -\frac{\varepsilon^2}{12} \cdot f^{(iv)}(\xi)$ where $\xi \in (x-\varepsilon, x+\varepsilon)$. It is $O(\varepsilon^2)$.

<u>What should ε be?</u>

* We need to take quantization artifacts into account.

Replace $f(x)$ by $f(x) + E$.

for the error in evaluating f.

Ex: Compute optimal ε for central-differencing:

Set:

$$f_c(x-\varepsilon) = f(x-\varepsilon) + E_-$$

computer value — true value — error in computation.

and: $f_c(x+\varepsilon) = f(x+\varepsilon) + E_+$

Then:

$$f'(x) \simeq \frac{1}{2\varepsilon}\left[\left(f(x+\varepsilon) + E_+\right) - \left(f(x-\varepsilon) + E_-\right)\right]$$

$$\simeq \frac{1}{2\varepsilon}\left[\left(f(x+\varepsilon) - f(x-\varepsilon)\right) + (E_+ - E_-)\right]$$

For the error, consider the original error and the extra error term $(E_+ - E_-)$.

The maximum error is thus bounded by

$$\left| -\frac{\varepsilon^2}{6} \cdot f'''(\xi) \right| + \frac{1}{2\varepsilon}\left| E_+ - E_- \right|$$

Define:

$$A = \max |E_+ - E_-|,$$

$$B = \max_x |f'''(x)|$$

For the optimal value ε^*, we have:

$$\frac{d}{d\varepsilon}\left[\frac{A}{2\varepsilon} + \frac{B\varepsilon^2}{6}\right] = 0$$

$$\Rightarrow \quad \frac{-A}{2\varepsilon^2} + \frac{2B\varepsilon}{6} = 0$$

$\Rightarrow$ For $\varepsilon \neq 0$:

$$-\frac{A}{2} + \frac{B\varepsilon^3}{3} = 0 \Rightarrow \boxed{\varepsilon = \left(\frac{3A}{2B}\right)^{1/3}} \sim (*)$$

* For infinite precision:

$$A \to 0 \quad \text{and} \quad \varepsilon \to 0$$

* For low precision, then we need to evaluate (*).

* Same for small $f'''(\cdot)$, occuring for noisy data.

Ch 7

(I) Approximating the Jacobian

For the residuals:

$$\frac{\partial r}{\partial x_i}(x) \simeq \frac{r(x+\epsilon e_i)-r(x)}{\epsilon}$$

(II) Approximating the Hessian

$$\nabla^2 f(x)\, e_i = \frac{\partial(\nabla f)}{\partial x_i}(x) \simeq \frac{\nabla f(x+\epsilon e_i)-\nabla f(x)}{\varepsilon}$$

with an error that is $O(\varepsilon)$

Similarly:

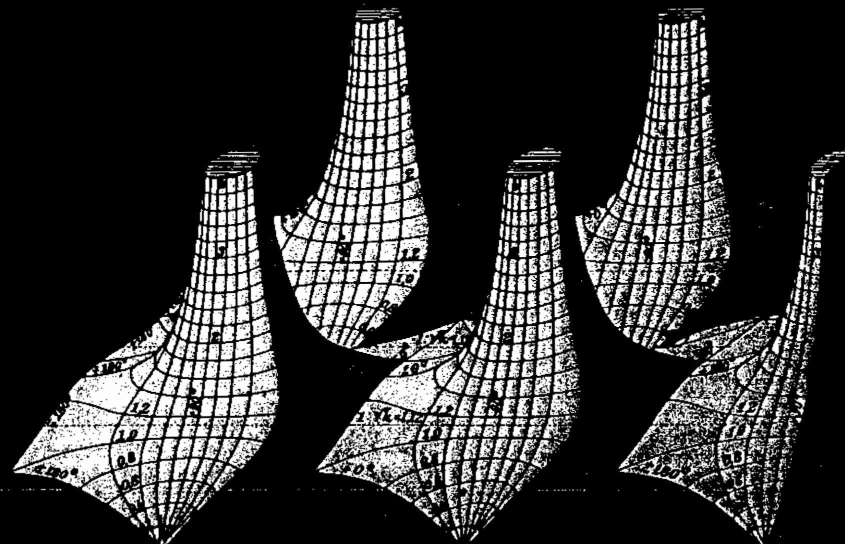
$$\nabla^2 f(x)\underline{p} \simeq \frac{\nabla f(x+\epsilon \underline{p})-\nabla f(x)}{\epsilon}$$

that is $O(\epsilon)$.

Also:

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{f(x+\epsilon e_i+\epsilon e_j)-f(x+\epsilon e_i)-f(x+\epsilon e_j)-f(x)}{\epsilon^2}$$

with an error of $O(\varepsilon)$



HANDBOOK OF MATHEMATICAL FUNCTIONS

with Formulas, Graphs, and Mathematical Tables

Edited by Milton Abramowitz and Irene A. Segun

Powers and roots n^k • Common logarithms • Circular sines and cosines for radian arguments • Exponential Integrals $E_n(x)$ • Tetragamma and pentagamma functions • Gamma function for complex arguments • Derivatives of the Legendre Function • Bessel functions–orders 0, 1 and 2, orders 10, 11, 20 and 21, etc. • Spherical Bessel functions • Struve functions • Confluent hypergeometric functions $M(a, b, x)$ • Coulomb wave functions of order zero • Jacobian zeta function $Z(\phi\backslash\alpha)$ • Heuman's lambda function • Table for obtaining periods for invariants g_2 and g_3 • Invariants and values at half-periods • Parabolic cylinder functions • Mathieu functions: characteristic values, joining factors, some critical values • Oblate radial functions – first and second kinds • Sums of reciprocal powers • Bernoulli and Euler numbers • Stirling numbers of the first and second kinds

Ch7$^+$-21
24

Ch7-2
24

25.3.3

$$R_n'(x)=\frac{f^{(n+1)}}{(n+1)!}(\xi)\pi_n'(x)+\frac{\pi_n(x)}{(n+1)!}\frac{d}{dx}f^{(n+1)}(\xi)$$

$$\xi=\xi(x)\quad(x_0<\xi<x_n)$$

Equally Spaced Abscissas

Three Points

25.3.4

$$f_p'=f'(x_0+ph)$$
$$=\frac{1}{h}\{(p-\tfrac{1}{2})f_{-1}-2pf_0+(p+\tfrac{1}{2})f_1\}+R_2'$$

Four Points

25.3.5

$$f_p'=f'(x_0+ph)=\frac{1}{h}\Bigg\{-\frac{3p^2-6p+2}{6}f_{-1}+\frac{3p^2-4p-1}{2}f_0-\frac{3p^2-2p-2}{2}f_1+\frac{3p^2-1}{6}f_2\Bigg\}+R_3'$$

Five Points

25.3.6

$$f_p'=f'(x_0+ph)=\frac{1}{h}\Bigg\{\frac{2p^3-3p^2-p+1}{12}f_{-2}-\frac{4p^3-3p^2-8p+4}{6}f_{-1}+\frac{2p^3-5p}{2}f_0-\frac{4p^3+3p^2-8p-4}{6}f_1+\frac{2p^3+3p^2-p-1}{12}f_2\Bigg\}+R_4'$$

For numerical values of differentiation coefficients see **Table 25.2.**

Markoff's Formulas

(Newton's Forward Difference Formula Differentiated)

25.3.7

$$f'(a_0+ph)=\frac{1}{h}\left[\Delta_0+\frac{2p-1}{2}\Delta_0^2+\frac{3p^2-6p+2}{6}\Delta_0^3+\dots+\frac{d}{dp}\binom{p}{n}\Delta_0^n\right]+R_n'$$

25.3.8

$$R_n'=h^nf^{(n+1)}(\xi)\frac{d}{dp}\binom{p}{n+1}+h^{n+1}\binom{p}{n+1}\frac{d}{dx}f^{(n+1)}(\xi)$$
$$(a_0<\xi<a_n)$$

25.3.9 $\quad hf_0'=\Delta_0-\frac{1}{2}\Delta_0^2+\frac{1}{3}\Delta_0^3-\frac{1}{4}\Delta_0^4+\dots$

25.3.10 $\quad h^2f_0^{(2)}=\Delta_0^2-\Delta_0^3+\frac{11}{12}\Delta_0^4-\frac{5}{6}\Delta_0^5+\dots$

25.3.11

$$h^3f_0^{(3)}=\Delta_0^3-\frac{3}{2}\Delta_0^4+\frac{7}{4}\Delta_0^5-\frac{15}{8}\Delta_0^6+\dots$$

25.3.12

$$h^4f_0^{(4)}=\Delta_0^4-2\Delta_0^5+\frac{17}{6}\Delta_0^6-\frac{7}{2}\Delta_0^7+\dots$$

25.3.13

$$h^5f_0^{(5)}=\Delta_0^5-\frac{5}{2}\Delta_0^6+\frac{25}{6}\Delta_0^7-\frac{35}{6}\Delta_0^8+\dots$$

Everett's Formula

25.3.14

$$hf'(x_0+ph)\approx-f_0+f_1-\frac{3p^2-6p+2}{6}\delta_0^2+\frac{3p^2-1}{6}\delta_1^2-\frac{5p^4-20p^3+15p^2+10p-6}{120}\delta_0^4+\frac{5p^4-15p^2+4}{120}\delta_1^4+\dots-\left[\binom{p+n-1}{2n+1}\right]'\delta_0^{2n}+\left[\binom{p+n}{2n+1}\right]'\delta_1^{2n}$$

25.3.15

$$hf_0'\approx-f_0+f_1-\frac{1}{3}\delta_0^2-\frac{1}{6}\delta_1^2+\frac{1}{20}\delta_0^4+\frac{1}{30}\delta_1^4$$

Differences in Terms of Derivatives

25.3.16

$$\Delta_0\approx hf_0'+\frac{h^2}{2!}f_0^{(2)}+\frac{h^3}{3!}f_0^{(3)}+\frac{h^4}{4!}f_0^{(4)}+\frac{h^5}{5!}f_0^{(5)}$$

25.3.17

$$\Delta_0^2\approx h^2f_0^{(2)}+h^3f_0^{(3)}+\frac{7}{12}h^4f_0^{(4)}+\frac{1}{4}h^5f_0^{(5)}$$

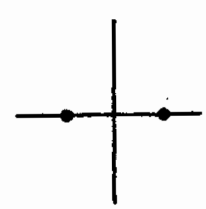
25.3.18 $\quad \Delta_0^3\approx h^3f_0^{(3)}+\frac{3}{2}h^4f_0^{(4)}+\frac{5}{4}f_0^{(5)}$

25.3.19 $\quad \Delta_0^4\approx h^4f_0^{(4)}+2h^5f_0^{(5)}$

25.3.20 $\quad \Delta_0^5\approx h^5f_0^{(5)}$

Partial Derivatives

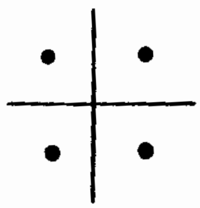
25.3.21



$$\frac{\partial f_{0,0}}{\partial x}=\frac{1}{2h}(f_{1,0}-f_{-1,0})+O(h^2)$$

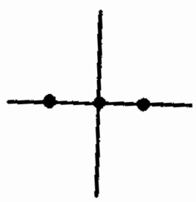
Ch7+ -23/24

25.3.22



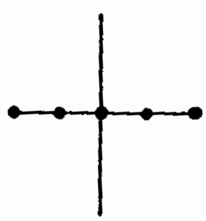
$$\frac{\partial f_{0,0}}{\partial x}=\frac{1}{4h}(f_{1,1}-f_{-1,1}+f_{1,-1}-f_{-1,-1})+O(h^2)$$

25.3.23



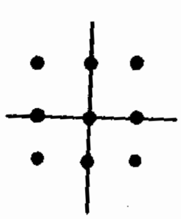
$$\frac{\partial^2 f_{0,0}}{\partial x^2}=\frac{1}{h^2}(f_{1,0}-2f_{0,0}+f_{-1,0})+O(h^2)$$

25.3.24



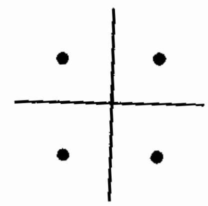
$$\frac{\partial^2 f_{0,0}}{\partial x^2}=\frac{1}{12h^2}(-f_{2,0}+16f_{1,0}-30f_{0,0}+16f_{-1,0}-f_{-2,0})+O(h^4)$$

25.3.25



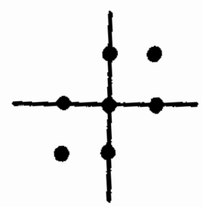
$$\frac{\partial^2 f_{0,0}}{\partial x^2}=\frac{1}{3h^2}(f_{1,1}-2f_{0,1}+f_{-1,1}+f_{1,0}-2f_{0,0}+f_{-1,0}+f_{1,-1}-2f_{0,-1}+f_{-1,-1})+O(h^2)$$

25.3.26



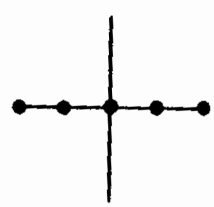
$$\frac{\partial^2 f_{0,0}}{\partial x\partial y}=\frac{1}{4h^2}(f_{1,1}-f_{1,-1}-f_{-1,1}+f_{-1,-1})+O(h^2)$$

25.3.27



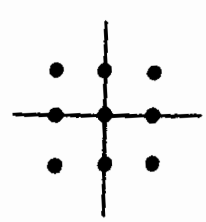
$$\frac{\partial^2 f_{0,0}}{\partial x\partial y}=\frac{-1}{2h^2}(f_{1,0}+f_{-1,0}+f_{0,1}+f_{0,-1}-2f_{0,0}-f_{1,1}-f_{-1,-1})+O(h^2)$$

25.3.28



$$\frac{\partial^4 f_{0,0}}{\partial x^4}=\frac{1}{h^4}(f_{2,0}-4f_{1,0}+6f_{0,0}-4f_{-1,0}+f_{-2,0})+O(h^2)$$

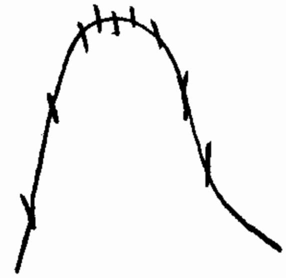
25.3.29



$$\frac{\partial^4 f_{0,0}}{\partial x^2\partial y^2}=\frac{1}{h^4}(f_{1,1}+f_{-1,1}+f_{1,-1}+f_{-1,-1}-2f_{1,0}-2f_{-1,0}-2f_{0,1}-2f_{0,-1}+4f_{0,0})+O(h^2)$$

Suggested Approaches

* Experiment with different precisions: float, double, etc.
* Experiment with different values for ε, guided by *
* Consider adaptive step methods, using smaller ε for high curvature points:



(see * ...)

* Use Maple / Mathematica for exact derivative evaluation.