DRAFT

537 Final Exam

Fall 2025

Thursday, Dec 11

Name: SOLUTIONS

Person number:

0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9

EXAM INSTRUCTIONS

The exam is closed-book.

One page of handwritten notes in prescribed form is allowed.

You may continue answers on the backs of the pages if needed.

- Work steadily and carefully.
- For maximum credit, show all your work.
- Do not waste time writing information that is not asked for.
- About 5% will be awarded for "style":
  be sure your writing is easy to read, and
  make your reasoning, explanations, and calculations
  clear, explicit, easy to follow.
- Some questions are easier and/or shorter than others.
  In particular, Q6 is not just a computation and requires thought.

1

Machine numbers

(a) What is the IEEE754 64-bit floating-point code for the smallest machine number greater than 8 in hexadecimal format? About far above 8 is it, as a power of 10?

Hints: Start by finding the code for 8 itself. The exponent bias is $1023 = 2^{10} - 1$, and $\log_{10} 2 \approx .30$.

$$8 = 2^3 = 2^{1026 - \underset{\uparrow \text{ bias}}{1023}}$$

Exponent code

$$1026 = 1024 + 2 = 2^{10} + 2 = (1\underbrace{0000000010}_{\text{ten binary digits}})_2$$

Mantissa is all zeros $(8 = 1.\underline{0 \quad 0} \times 2^3)$.

Code for 8 is therefore

sign	biased exponent	mantissa
0	100 0000 0010	0 0

$\underbrace{0 \ldots\ldots 0}_{52 \text{ digits}}$

Hex

$$4 \quad 0 \quad 2 \quad \underbrace{0 \ldots\ldots 0}_{13 \text{ digits}}$$

The next machine number is attained by setting the last bit of the mantissa to 1.

Hex

$$4\,0\,2\,\underbrace{0000000000001}_{13 \text{ hex digits}}.$$

The difference between 8 and the next machine number is

$$8\varepsilon_{mach} = 2^3 \times 2^{-52} = 2^{-49} \approx 10^{-(4.9)(3)} = 10^{-14.7}$$

(The relative difference is ε_{mach}.)

2
Round-off error bound

Find the approximate maximum relative error in the floating-point evaluation of this expression

$$(x + c) - c$$

when x is small and not necessarily a machine number, and c is a machine number greater than 1. Answer in terms of $|x|$, c, and ϵ_{mach}, and just keep the term or terms that dominate as $|x|$ goes to zero.

Suggestion: to avoid mistakes, expand everything fully before cancelling anything.

$$|\text{relative error}| = \left| \frac{\left[\left(\overbrace{x(1+\delta_1)}^{fl(x)} + c\right)(1+\delta_2) - c\right](1+\delta_3) - x}{x} \right|$$

$$= \left| \frac{\left(x + x\delta_1 + \cancel{c} + x\delta_2 + x\delta_1\delta_2 + c\delta_2 - \cancel{c}\right)(1+\delta_3) - x}{x} \right|$$

$$= \left| \frac{\cancel{x} + x\delta_1 + x\delta_2 + x\delta_1\delta_2 + c\delta_2 + x\delta_3 + x\delta_1\delta_3 + x\delta_2\delta_3 + x\delta_1\delta_2\delta_3 + c\delta_2\delta_3 - \cancel{x}}{x} \right|$$

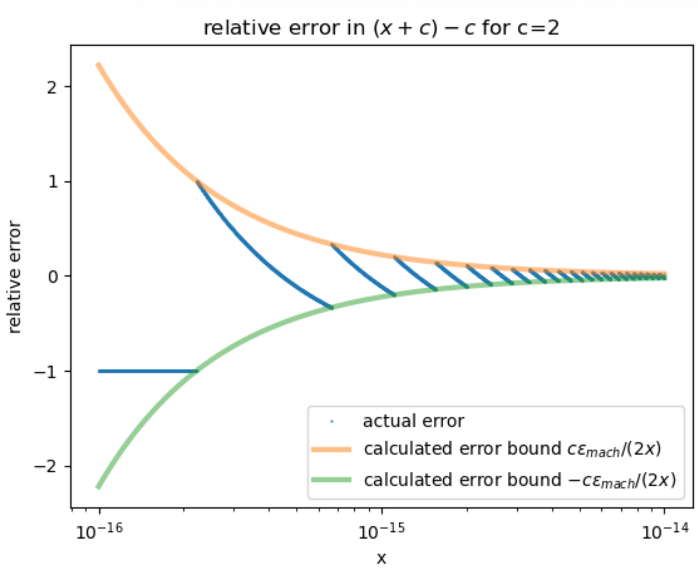
$$= \left| \frac{\cancel{x}}{\cancel{x}}\left(\delta_1 + \delta_2 + \delta_1\delta_2 + \delta_3 + \delta_1\delta_3 + \delta_2\delta_3 + \delta_1\delta_2\delta_3\right) + \frac{c\delta_2 + c\delta_2\delta_3}{x} \right|$$

$$\approx \left| (\delta_1 + \delta_2 + \delta_3) + \frac{c\delta_2}{x} \right|$$

$$\leq \frac{3\varepsilon_{mach}}{2} + \frac{c\,\varepsilon_{mach}}{2|x|} \xrightarrow{x \to 0} \frac{c\,\varepsilon_{mach}}{2|x|}.$$

(catastrophic error as $x \to 0$).

Empirical check:



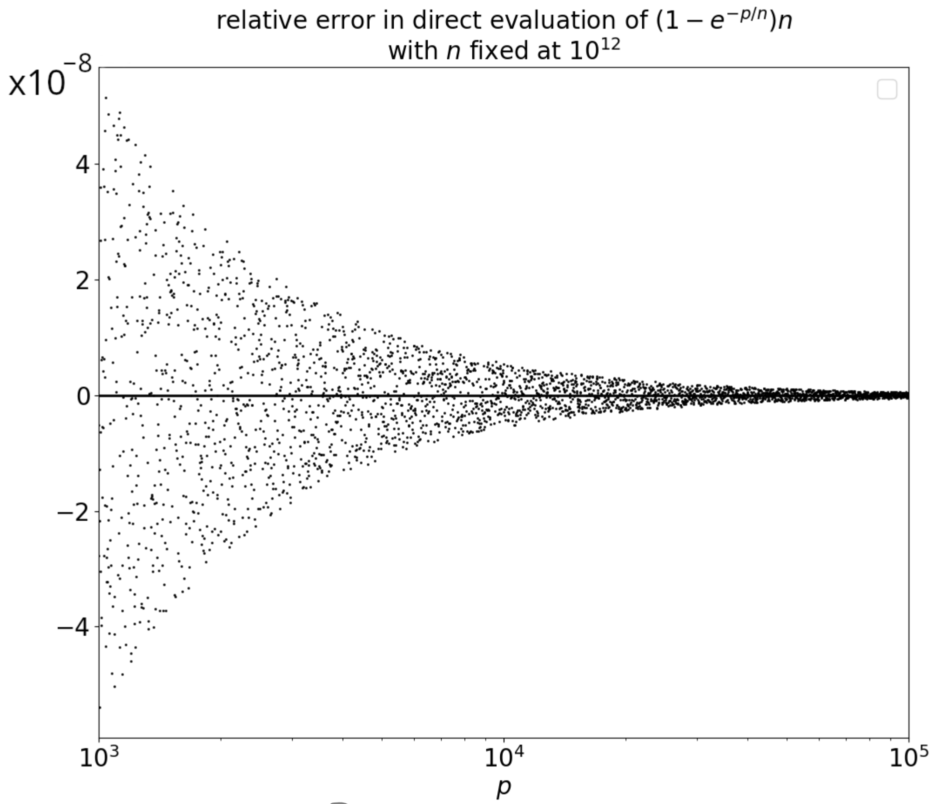
3

Round-off error avoidance

(a) Explain briefly and qualitatively why we obtain the results pictured below when evaluating the expression

$$n\left(1 - e^{-\frac{p}{n}}\right)$$

in floating-point arithmetic when $0 < p \ll n$.



If $0 < p \ll n$, $e^{-p/n} \approx 1$.

Hence we are subtracting near-equals, which leads to loss of significance and large error relative to the result.

(b) Calculate or write down the Taylor approximation for e^{-x} with remainder at $O(x^3)$, and use it to obtain a two-term approximation and error bound for $n\left(1-e^{-\frac{p}{n}}\right)$. Give a *relative* error bound for the values at the right end of the plot (where it is worst). Very briefly compare with the round-off error seen in (a).

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \frac{f'''(\xi)}{6}x^3 \quad , \quad \xi \in [0,x].$$

$f(x) = e^{-x}$

$$e^{-x} = e^{-0} + (-e^{-0})x + \frac{(+e^{-0})}{2}x^2 + \frac{(-e^{-\xi})}{6}x^3$$

$$= 1 - x + \frac{1}{2}x^2 - \frac{e^{-\xi}}{6}x^3$$

$$n(1-e^{-\frac{p}{n}}) = n\left(\cancel{1} - \left[\cancel{1} - \frac{p}{n} + \frac{1}{2}\left(\frac{p}{n}\right)^2 - \frac{e^{-\xi}}{6}\left(\frac{p}{n}\right)^3\right]\right)$$

$$= n\left(\frac{p}{n} - \frac{1}{2}\left(\frac{p}{n}\right)^2 + \frac{e^{-\xi}}{6}\left(\frac{p}{n}\right)^3\right)$$

Use approximation

$$n(1-e^{-\frac{p}{n}}) \approx p - \frac{n}{2}\left(\frac{p}{n}\right)^2$$

with $|\text{truncation error}| \leq \frac{n \cdot 1}{6}\left(\frac{p}{n}\right)^3$.

At $n = 10^{12}$, $p = 10^5$, $\frac{p}{n} = 10^{-7}$

$$\frac{|\text{truncation error}|}{\text{value}} \lesssim \frac{\frac{10^{12}(10^{-7})^3}{6}}{10^5} = \frac{10^{-14}}{6}$$

$$\approx 10^{-15} \sim \varepsilon_{mach}.$$

4

Sensitivity of solution of a linear system

Suppose

$$Ax = b,$$

where A is an invertible matrix and b is a non-zero vector, and define δx by

$$A(x + \delta x) = b + \delta b.$$

That is, δx is the change in the solution when the right hand side is changed by δb.

Using just algebra and the properties of a vector norm and the induced matrix norm, obtain a bound on $\frac{||\delta x||}{||x||}$, explicitly citing which property of a norm you are using at each step. Do NOT cite or use Theorem 3.10. Express your answer in terms of the condition number $\kappa(A)$.

$$Ax = b \qquad (1)$$

$$A(x + \delta x) = b + \delta b \qquad (2)$$

$(2) - (1)$: $\quad A\delta x = \delta b$

$$A^{-1}A\delta x = A^{-1}\delta b$$

$$\delta x = A^{-1}\delta b$$

$$\|\delta x\| \le \|A^{-1}\delta b\| \overset{\text{compat}}{\le} \|A^{-1}\|\,\|\delta b\| \qquad (3)$$

$(1) \rightarrow \quad \|b\| = \|Ax\| \le \|A\|\,\|x\| \qquad$ prop (v)

So $\quad \|x\| \ge \|b\| / \|A\| \qquad (4)$

(3) & $(4) \rightarrow$

$$\frac{\|\delta x\|}{\|x\|} \le \frac{\|A^{-1}\|\,\|\delta b\|}{\frac{\|b\|}{\|A\|}}$$

$$= \underbrace{\|A\|\,\|A^{-1}\|}_{\equiv K(A)} \frac{\|\delta\|}{\|b\|}$$

$$\frac{\|\delta x\|}{\|x\|} \le K(A)\frac{\|\delta b\|}{\|b\|} \quad \text{where } K(A) = \|A\|\,\|A^{-1}\|.$$

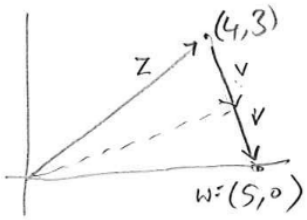
5 Householder QR decomposition

Consider the over-determined linear system

$$Ax = b,$$
$$\begin{bmatrix} 4 \\ 3 \end{bmatrix} [x_1] = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Observe that b is not in the column space of A.

Obtain a QR factorization of A (Q orthogonal, R upper triangular) using a Householder reflector, H. Show all the steps of the computation, including specifying the projector P from which H is built.



$$z = \begin{bmatrix} 4 \\ 3 \end{bmatrix} \quad w = \begin{bmatrix} \|z\|_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$$

$$v = w - z = \begin{bmatrix} 5 \\ 0 \end{bmatrix} - \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

$$v^T v = [1, -3] \begin{bmatrix} 1 \\ -3 \end{bmatrix} = 1 + 9 = 10$$

$$v v^T = \begin{bmatrix} 1 \\ -3 \end{bmatrix} [1 \ -3] = \begin{bmatrix} 1 & -3 \\ -3 & 9 \end{bmatrix}$$

$$P = \frac{v v^T}{v^T v} = \frac{1}{10} \begin{bmatrix} 1 & -3 \\ -3 & 9 \end{bmatrix}$$

$$H = I - 2P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{5} & \frac{-3}{5} \\ \frac{-3}{5} & \frac{9}{5} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{4}{5} & +\frac{3}{5} \\ \frac{3}{5} & -\frac{4}{5} \end{bmatrix}$$

$$\overbrace{\begin{bmatrix} \frac{4}{5} & \frac{3}{5} \\ \frac{3}{5} & -\frac{4}{5} \end{bmatrix}}^{H} A = \begin{bmatrix} 5 \\ 0 \end{bmatrix}, \quad H^{-1} = H^T = H$$

$$A = \begin{bmatrix} \frac{4}{5} & \frac{3}{5} \\ \frac{3}{5} & -\frac{4}{5} \end{bmatrix} \begin{bmatrix} 5 \\ 0 \end{bmatrix} = QR.$$

Solution

$$Rx = [Q^T b]_{:1}$$

$$[5][x_1] = \left(\begin{bmatrix} \frac{4}{5} & \frac{3}{5} \\ \frac{3}{5} & -\frac{4}{5} \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} \right)_{:1} = [11/5] \longrightarrow x_1 = \frac{11}{25}$$

6
Root finding

Recall that in Chapter 2 we defined a *contraction* on an interval $G = [a, b] \subseteq \mathbb{R}$ as a function $g : G \to \mathbb{R}$ that satisfies a Lipschitz condition on G with Lipschitz constant L that is strictly less than 1.

Suppose g is a contraction, and furthermore that g maps G into G.

(a) Prove that g has a fixed point in G.

If $g(a) = a$ or $g(b) = b$, we are done.

Otherwise $g(a) > a$ and $g(b) < b$ since $g: [a,b] \to [a,b]$,

so that with $h(x) \equiv g(x) - x$,

$h(a) > 0$ and $h(b) < 0$.

g Lipschitz $\Rightarrow$ g continuous, so h continuous.

Thus by intermediate value theorem,

$\exists\ x^* \in (a,b)$ such that $h(x^*) = 0$,

that is $g(x^*) - x^* = 0$, $g(x^*) = x^*$. That is

x^* is a fixed point of g.

(b) Prove that g has no more than one fixed point in G.

Suppose x_1, x_2 are fixed points of g in G:

$g(x_1) = x_1$

$g(x_2) = x_2$

So $g(x_1) - g(x_2) = x_1 - x_2$

$|g(x_1) - g(x_2)| = |x_1 - x_2|$

This contradicts $|g(x_1) - g(x_2)| \le L|x_1 - x_2|$

with $L < 1$ unless $x_1 = x_2$.

(c) Prove that if $x_0 \in G$, and $x_{k+1} = g(x_k)$, $k = 0, 1, 2, \ldots$, then x_k converges to the fixed point x^* as $k \to \infty$.

$$x_{k+1} = g(x_k)$$

$$\begin{aligned} x_{k+1} - x^* &= g(x_k) - x^* \\ &= g(x_k) - g(x^*) \qquad (x^* \text{ fixed}) \\ &\leq L|x_k - x^*| \qquad \text{with } L < 1. \end{aligned}$$

Thus $|x_k - x^*| \leq L^k |x_0 - x^*| \xrightarrow{k \to \infty} 0$

(d) If $f : G \to \mathbb{R} \in C^2(G)$, f has a root $z \in G$, and $f'(z) \neq 0$, show that Newton's method applied to f is a contraction on some neighborhood N of z and maps N into itself. Hint: g' is small near z.

Newton's $\quad g(x) = x - \dfrac{f(x)}{f'(x)}$

We've seen that as long as $f'(z) \neq 0$, $g'(z) = 0$:

$$g'(z) = \cancel{1} - \frac{\cancel{f'(z)^2} - f(z)f''(z)}{f'(z)^2} = \frac{f(z)f''(z)}{f'(z)^2} = 0 .$$

Thus $g'(x)$ is small for x near z.

In particular, with g' continuous and $g'(z) = 0$,

$\forall \; 0 < L < 1$, $\exists \; \delta_L > 0$ s.t. $|g'(x)| \leq L \;\; \forall \; x \in [z - \delta_L, z + \delta_L]$.

Thus, on $[z - \delta_L, z + \delta_L]$, g is L-Lipschitz, for any $L < 1$ we like.

And since $g(z) = z$, g maps $[z - \delta_L, z + \delta_L]$ into itself:

because g shrinks distance from z by a factor of L or more

(e) Prove that for $x_0 \in N$, Newton's method converges faster than any linearly convergent process.

From parts (a) & (c), iteration of g from any x_0 in $[z - \delta_L, z + \delta_L]$, $L < 1$, converges to z.

Then from (d), $\forall \; 0 < M < 1$ no matter how small, there therefore exists n_M such that for $k \geq n_M$, $|x_k - z| < \delta_M$

So $|x_{k+1} - z| \leq M |x_k - z|$ for $k = n_M, n_M + 1, \ldots$.

That is convergence is faster than linear at arbitrary rate M.

7 Quadrature

Derive the 3-point Gauss-Legendre quadrature rule for the interval $[-1, 1]$ from the requirement that it has polynomial degree 5. Minimize your labor by guessing the symmetry.

Guess $\{x_0, x_1, x_2\} = [-\alpha, 0, \alpha]$, $\alpha \in (0,1)$.

and guess $w_0 = w_2$, so that

$$Q(f) = w_0 f(-\alpha) + w_1 f(0) + w_0 f(\alpha).$$

For poly deg. 5 need

$$Q(1) = 2w_0 + w_1 = \int_{-1}^{1} 1\,dx = 2 \; : \quad 2w_0 + w_1 = 2 \; ①$$

$$Q(x) = -\alpha w_0 + \alpha w_0 = 0 = \int_{-1}^{1} x\,dx = 0 \; \checkmark \text{ by symmetry}$$

$$Q(x^2) = \alpha^2 w_0 + \alpha^2 w_0 = \int_{-1}^{1} x^2\,dx = \frac{2}{3} \; : \quad \alpha^2 w_0 = \frac{1}{3} \; ②$$

$$Q(x^3) = \int_{-1}^{1} x^3\,dx = 0 \; \checkmark \text{ by symmetry}$$

$$Q(x^4) = \alpha^4 w_0 + \alpha^4 w_0 = \int_{-1}^{1} x^4\,dx = \frac{2}{5} \; : \quad \alpha^4 w_0 = \frac{1}{5} \; ③$$

$$\frac{③}{②} \rightarrow \alpha^2 = \frac{1/5}{1/3} = \frac{3}{5} \; : \quad \boxed{\alpha = \sqrt{\frac{3}{5}}}$$

$$② \rightarrow \frac{3}{5} w_0 = \frac{1}{3} \; : \quad \boxed{w_0 = \frac{5}{9}}$$

$$① \rightarrow \frac{10}{9} + w_1 = \frac{18}{9} \; : \quad \boxed{w_1 = \frac{8}{9}}$$

And $Q(x^5) = \int_{-1}^{1} x^5\,dx = 0$ by symmetry.

Thus

$$Q(f) = \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right)$$

has polynomial degree 5.

8

Optimization

Consider finding the minimum of $f(x, y) = x^2 - 4x + xy - y + 6$ on the square $D = \{(x, y) \in \mathbb{R}^2 : 0 \le x, y \le 3\}$. Apply one iteration of the method of steepest descent with $(x^{(0)}, y^{(0)}) = \overset{(2,1)}{\cancel{(1,2)}}$, to find the point $(x^{(1)}, y^{(1)}) \in D$ that minimizes f in the descent direction. Also find the value of f before and after the line minimization.

$$\nabla f(x,y) = (2x-4+y,\ x-1)$$

$$\nabla f(x^{(0)}, y^{(0)}) = \nabla f(2,1) = (2\cdot 2-4+1,\ 2-1) = (1,1)$$

Steepest descent line at $(x^{(0)}, y^{(0)})$ is therefore

$$(x(t), y(t)) = (2,1) + t(1,1) = (2+t,\ 1+t).$$

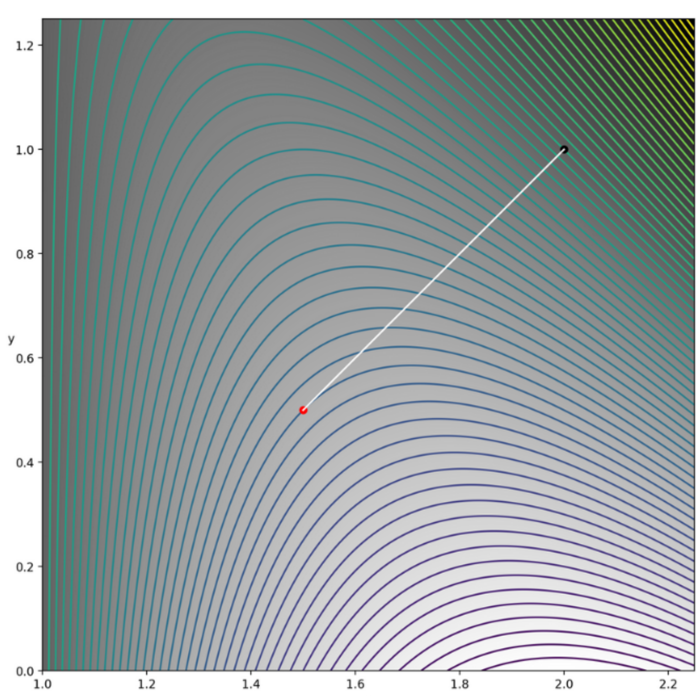
$$\begin{aligned} f(x(t),y(t)) &= (2+t)^2 - 4(2+t) + (2+t)(1+t) - (1+t) + 6 \\ &= 4+4t+t^2 - 8 - 4t + 2 + 3t + t^2 - 1 - t + 6 \\ &= (4-8+2-1+6) + (4-4+3-1)t + (1+1)t^2 \\ &= 3 + 2t + 2t^2 \end{aligned}$$

$$f'(x(t), y(t)) = 2 + 4t \overset{\text{set}}{=} 0 \longrightarrow 4t = -2, \quad \boxed{t^* = -\tfrac{1}{2}}$$

$$(x(t^*), y(t^*)) = (2 + (-\tfrac{1}{2}),\ 1 + (-\tfrac{1}{2})) = \boxed{(1\tfrac{1}{2}, \tfrac{1}{2})}$$ line minimizer

$$f(x(0), y(0)) = f(2,1) = 3 + 2\cdot 0 + 2 \cdot 0^2 = \boxed{3}$$ initial f-value

$$f(x(t^*), y^*(0)) = f(\tfrac{3}{2}, \tfrac{1}{2}) = (\tfrac{3}{2})^2 - 4(\tfrac{3}{2}) + (\tfrac{3}{2})(\tfrac{1}{2}) - (\tfrac{1}{2}) + 6$$



$$= \frac{9 - \cancel{24} + 3 - 2 + \cancel{24}}{4}$$

$$= \frac{10}{4} = \boxed{\frac{5}{2}}.$$ f after 1 line minimization.