1. The MLE, $\hat{x}_{\text{MLE}}$, is determined as

$$\hat{x}_{\text{MLE}} = \arg \min_x \{- \ln p(y|x)\} = \arg \min_x \sum_{i=1}^{2} \frac{1}{\sigma_i^2} (y_i - x)^2 = \arg \min_x \|y - Ax\|^2_W,$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad A = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad W = \begin{pmatrix} w_1 & 0 \\ 0 & w_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix}.$$  

We have that the adjoint operator is given by

$$A^* = A^T W = (w_1, w_2) = \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \end{pmatrix},$$

so that,

$$A^* A = w_1 + w_2 = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2},$$

which is invertible (as we already knew since $A$ has full column rank). Continuing, we obtain,

$$(A^* A)^{-1} = \frac{1}{w_1 + w_2} = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2},$$

and

$$A^+ = (A^* A)^{-1} A^* = \begin{pmatrix} \frac{w_1}{w_1 + w_2} & \frac{w_2}{w_1 + w_2} \\ \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} & \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \end{pmatrix}.$$

Thus,

$$\hat{x}_{\text{MLE}} = A^+ y = y_2 = \alpha_1 y_1 + \alpha_2 y_2,$$

where

$$0 \leq \alpha_i = \frac{w_i}{w_1 + w_2} = \frac{\sigma_i^2}{\sigma_1^2 + \sigma_2^2} \leq 1, \quad \text{for} \quad i = 1, 2,$$

and $\alpha_1 + \alpha_2 = 1$. This shows that the MLE is a so-called convex combination (i.e., a normalized weighted average) of the measurements $y_1$ and $y_2$. Recalling that $w_i = \frac{1}{\sigma_i^2}$ gives the precision of the measurement $y_i$, we see that the weight $\alpha_i$ is a measure of the normalized (or relative) precision.

(a) In the limit that $\sigma_i^2 \to 0$ ($w_i \to \infty$), we see from the general solution derived above that $\hat{x}_{\text{MLE}} \to y_1$. This makes sense as $\sigma_i^2 \to 0$ ($w_i \to \infty$) corresponds to $y_1$ becoming a perfect, non-noisy (infinitely precise) measurement of the unknown quantity $x$. In the limit that $\sigma_1^2 \to \infty$, ($w_1 \to 0$) the general solution shows that $\hat{x}_{\text{MLE}} \to y_2$. This makes sense as $\sigma_1^2 \to \infty$ ($w_1 \to 0$) corresponds to the first measurement becoming so noisy (so imprecise) that it is worthless relative to the second measurement.
In the case that \( \sigma_1^2 = \sigma_2^2 \equiv \sigma^2 \) \((w_1 = w_2 \equiv w)\) we have,

\[
\hat{x}_{\text{MLE}} = \arg \min_x \|y - Ax\|_W^2 = \arg \min_x \frac{1}{\sigma^2} \|y - Ax\|^2 = \arg \min_x \|y - Ax\|^2,
\]

the last expression being an unweighted least-squares problem. Using the condition \( \sigma_1^2 = \sigma_2^2 \) \((w_1 = w_2)\), the general solution derived earlier becomes,

\[
\hat{x}_{\text{MLE}} = \frac{1}{2} (y_1 + y_2).
\]

This solution makes sense because the condition \( \sigma_1^2 = \sigma_2^2 \) means that the two measurements are equally precise, \( w_1 = w_2 \), so that we have no rational reason to prefer one measurement over the other. This is because the errors in the two measurements are additive, independent, zero mean, and identically symmetrically-distributed measurement errors. This symmetry condition yields a linear estimator which is the symmetric sample-mean solution.  

2. Note that \( E\{y_i\} = \alpha t_i \).

(a) The least-squares problem to be solved is

\[
\min_{\alpha} \|y - A\alpha\|^2,
\]

where

\[
y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} t_1 \\ \vdots \\ t_m \end{pmatrix}.
\]

The least-squares solution is

\[
\hat{\alpha}(m) = A^+ y = (A^T A)^{-1} A^T y = \frac{\sum_{i=1}^{m} t_i y_i}{\sum_{i=1}^{m} t_i^2}.
\]

(b)

\[
E\{\hat{\alpha}(m)\} = \frac{\sum_{i=1}^{m} t_i E\{y_i\}}{\sum_{i=1}^{m} t_i^2} = \alpha \frac{\sum_{i=1}^{m} t_i^2}{\sum_{i=1}^{m} t_i^2} = \alpha.
\]

\footnote{In advanced courses we would say that “the solution to the optimal linear estimator is obvious from symmetry.”}
\[ \text{Var}\{\tilde{\alpha}(m)\} = \sigma^2 (A^T A)^{-1} = \frac{\sigma^2}{\sum_{i=1}^{m} t_i^2} \rightarrow 0 \quad \text{as} \quad m \rightarrow \infty. \]

3. (a) Note that for \( A \) one-to-one, \( A^+ A = (A^T A)^{-1} A^T A = I \). Let \( e \) be the vector of all ones, \( e = (1 \cdots 1)^T \). Note that

\[ E\{n\} = b \quad \text{where} \quad b = \beta e. \]

We have that

\[ \tilde{x} = \hat{x} - x = A^+ y - x = A^+(A x + n) - x = x + A^+ n - x = A^+ n. \]

Thus

\[ E\{\tilde{x}\} = A^+ E\{n\} = A^+ b \neq 0. \]

(b) We have

\[ y = A x + n = A x + b + (n - b) = A x + b + \nu \]

where \( \nu = n - b = n - \beta e \) has zero mean. Continuing,

\[ y = (A \ e)^T \begin{pmatrix} x \\ \beta \end{pmatrix} + \nu = A \ x + \nu, \]

where

\[ A = (A \ e) \quad \text{and} \quad x = \begin{pmatrix} x \\ \beta \end{pmatrix}. \]

Assuming that \( A \) is one-to-one, we have

\[ \hat{x} = A^+ y = (A^T A)^{-1} A^T y \]

and

\[ \tilde{x} = A^+ \nu. \]

Since \( \nu \) has zero mean, \( \hat{x} \) is unbiased.

4. **Vocabulary.** The relevant definitions can be found in your class lecture notes and/or in the lecture supplement handouts.

5. **Scalar Generalized Gradient Descent Algorithms.**

Note that in the scalar case the loss function is

\[ \ell(x) = \frac{1}{2} (y - h(x))^2. \]
Also the gradient of \( \ell(x) \) is just the derivative of \( \ell(x) \) with respect to \( x \),

\[
\ell'(x) = \nabla \ell(x) = \frac{d}{dx} \ell(x) = -h'(x) (y - h(x)) ,
\]
where \( h'(x) = \frac{d}{dx} h(x) \). Note also that the second derivative of the loss function is

\[
\ell''(x) = h'^2(x) - h''(x) (y - h(x)) .
\]

(a) **Gradient Descent Method.**

\[
\hat{x}_{k+1} = \hat{x}_k - \alpha_k \ell'(\hat{x}_k) = \hat{x}_k + \alpha_k h'(\hat{x}_k) (y - h(\hat{x}_k)) ,
\]
where the step size \( \alpha_k > 0 \) is used for convergence control. It is evident that \( Q_k = 1 \) for all \( k \).

(b) **Gauss’ Method (Gauss–Newton Method).**

To find a correction to the current estimate \( \hat{x} \), we linearize the nonlinear inverse problem \( y = h(x) \) about the point \( \hat{x} \),

\[
y \approx h(\hat{x}) + h'(\hat{x}) (x - \hat{x}) = h(\hat{x}) + h'(\hat{x}) \Delta x ,
\]
where \( \Delta x = x - \hat{x} \), and find a solution which is optimal in the least–squares sense for this linearized problem. Note that the loss function for this linearized problem is

\[
\ell_{gauss}(x) = \frac{1}{2} (y - [h(\hat{x}) + h'(\hat{x}) (x - \hat{x})])^2 = \frac{1}{2} (\Delta y - h'(\hat{x}) \Delta x)^2 = \ell_{gauss}(\Delta x) ,
\]
where \( \Delta y = y - h(\hat{x}) \). Also note that because

\[
x = \hat{x} + \Delta x ,
\]
where \( \hat{x} \) is given and fixed, it is evident that optimizing over \( x \) is equivalent to optimizing over \( \Delta \). An equivalent problem, then, is to find the correction \( \Delta x \) which is optimal in the least–squares sense by minimizing \( \ell_{gauss}(\Delta x) \) with respect to \( \Delta x \).

The solution is

\[
\Delta x = \frac{1}{h'(\hat{x})} \Delta y = \frac{1}{h'(\hat{x})} (y - h(\hat{x})) ,
\]
assuming that \( h'(\hat{x}) \neq 0 \). Incorporating a step size \( \alpha > 0 \) for convergence control, yields

\[
\Delta x = \alpha(\hat{x}) \frac{1}{h'(\hat{x})} (y - h(\hat{x})) = \alpha(\hat{x}) Q(\hat{x}) h'(\hat{x}) (y - h(\hat{x})) ,
\]
where

$$Q(x) = \frac{1}{h''(x)}.$$  

This yields the iterative algorithm,

$$\hat{x}_{k+1} = \hat{x}_k + \alpha_k \frac{1}{h'(\hat{x}_k)} (y - h(\hat{x}_k)) = \hat{x}_k + \alpha_k Q_k h'(\hat{x}_k) (y - h(\hat{x}_k))$$  

where

$$Q_k = \frac{1}{h''(\hat{x}_k)}.$$  

(c) **Newton’s Method.**

To find a correction to the current estimate $\hat{x}$ we expand the loss function $\ell(x)$ about $\hat{x}$ to second order and then minimize this quadratic approximation to $\ell(x)$. With $\Delta x = x - \hat{x}$, the quadratic approximation is

$$\ell(x) \approx \ell_{\text{quad}}(x) = \ell(\hat{x}) + \ell'(\hat{x}) \Delta x + \frac{1}{2} \ell''(\hat{x})(\Delta x)^2 = \ell_{\text{quad}}(\Delta x).$$

Note that $\ell_{\text{quad}}(x) = \ell_{\text{quad}}(\Delta x)$ can be equivalent minimized either with respect to $x$ or with respect to $\Delta x$. If we take the derivative of $\ell_{\text{quad}}(\Delta x)$ with respect $\Delta x$ and set it equal to zero we get

$$\Delta x = -\frac{\ell'(\hat{x})}{\ell''(\hat{x})} = \frac{h'(\hat{x}) (y - h(\hat{x}))}{h''(\hat{x}) - h''(\hat{x}) (y - h(\hat{x}))}.$$  

If we multiply the correction $\Delta x$ by a step size $\alpha > 0$ in order to control convergence we obtain the iterative algorithm

$$\hat{x}_{k+1} = \hat{x}_k + \alpha_k Q_k h'(\hat{x}_k) (y - h(\hat{x}_k)),$$

where

$$Q_k = \frac{1}{h''(\hat{x}_k) - h''(\hat{x}) (y - h(\hat{x}))}.$$  

Comparing the values of $Q_k$ for the Gauss method and the Newton method, we see that when $y - h(\hat{x}_k) \approx 0$ the two methods become essentially equivalent.\(^2\) Also note that when $h'' \equiv 0$ the two methods become exactly equivalent.

\(^2\)This fact also holds for the more general vector case. For example, because this latter situation holds for the GPS computer assignment, the Gauss-Newton method used in that assignment is essentially equivalent to the Newton method and therefore has the very fast convergence rate associated with the Newton method.

Let $y$ denote the known number for which you wish to determine the cube root. Let $x$ denote the unknown cube root of $y$. We need a relationship $y = h(x)$, which is obviously given by $h(x) = x^3$. This relationship defines an inverse problem that we wish to solve. (I.e., given $y = h(x) = x^3$ determine $x$.) The relevant derivatives needed to implement the descent algorithms are

$$h'(x) = 3x^2 \quad \text{and} \quad h''(x) = 6x.$$