

Optimal Interpolation

Here we present the optimal interpolation (OI) equations for vectors of observations and background fields, to lead us into the 3DVAR discussions. Many concepts and notations are also used by 3DVAR and later 4DVAR.

The OI equations were originally derived by Eliassen (1954, reproduced in Bengtsson et al, 1981). However, Gandin (1963), derived the multivariate OI equations independently and applied them to objective analysis in the former Soviet Union.

Gandin's work had a profound influence upon the research and operational community, and OI became the operational analysis scheme of choice during the 1980's and early 1990's.

In our discussion, we follow the general notation proposed by Ide et al (1997) for data assimilation methods.

It can be shown that 3D-Var is equivalent to the OI problem, except that the method to solve the problem is quite different and advantageous for operational systems.

References:

Eliassen, A., 1954: Provisional report on calculation of spatial covariance and autocorrelation of the pressure field. *Dynamic Meteorology: Data Assimilation Methods*, L. Bengtsson, M. Ghil, and E. Kallen, Eds., Springer-Verlag, 319-330.

Bengtsson, L., M. Ghil, and E. Kallen, 1981: *Dynamic Meteorology: Data Assimilation Methods*. Springer-Verlag. 330.

Gandin, L. S., 1963: *Objective Analysis of Meteorological Fields*. (Translated by Israel Program for Scientific Translations).

Ide, K., P. Courtier, M. Ghil, and A. Lorenc, 1997: Unified notation for data assimilation: Operational, sequential and variational. *J. Meteor. Soc. Japan*, **75**, 181-189.

Variable Notations

Consider the complete NWP operational problem of finding

\mathbf{x}_a - an optimum analysis of a field of model variables,

given

\mathbf{x}_b - a background field available at grid points, and

\mathbf{y}_o - a set of p observations available at irregularly spaced points \mathbf{r}_i

\mathbf{x}_a , and \mathbf{x}_b are vectors of length n .

The unknown analysis and the known background can be 2D fields of a single variable like the surface temperature analysis $\mathbf{T}_a(x, y)$, or the 3D field of the initial conditions for all the model prognostic variables:

$\mathbf{x} = (\mathbf{p}, \mathbf{T}, \mathbf{q}_v, \mathbf{u}, \mathbf{v})$.

These model variables are ordered by grid point and by variable, forming a single vector of length n , the product of the number of points times the number of variables. The (unknown) "truth" \mathbf{x}_t , discretized at the model points, is also a vector of length n .

A different variable \mathbf{y}_o is for the observations than what we use of gridded variables.

The observed variables are, in general, different from the model variables by

- a) being located in different points, and
- b) by possibly being *indirect* measures of the model variables.

Examples of this are radar reflectivity and Doppler radar radial velocity, satellite radiances, and Global Positioning System (GPS) atmospheric refractivity.

For example, a simplified formula of the radar reflectivity given the mixing ratios of rain water, snow and hail species inside the cloud is

$$Z = 10\{3 + \log_{10}[17.3(\rho q_r)^{7/4} + 38(\rho(q_s + q_h))^{2.2}]\}$$

where Z is the reflectivity in dBZ, ρ is air density in kg/m^3 , and q_r , q_s and q_h are the mixing ratios in g/kg. It is a formula that brings model variables q_r , q_s and q_h (part of \mathbf{x}) to the observed values Z (part of \mathbf{y}), is therefore a forward operator, and it is a nonlinear function of q 's.

Tong and Xue (2005) use a more complicated (but still simplified) version of the reflectivity formula for WSR-88D type 10-cm wavelength S-band radars:

$$Z = 10\log_{10}\left(\frac{Z_{er} + Z_{es} + Z_{eh}}{1\text{mm}^6 \text{m}^{-3}}\right) \quad (\text{in dBZ})$$

For rain, $Z_{er} = \frac{10^{18} \times 720 (\rho q_r)^{1.75}}{\pi^{1.75} N_r^{0.75} \rho_r^{1.75}}$. $Z_{es} = \frac{10^{18} \times 720 K_i^2 \rho_s^{0.25} (\rho q_s)^{1.75}}{\pi^{1.75} K_r^2 N_s^{0.75} \rho_i^2}$ for dry snow and $Z_{es} = \frac{10^{18} \times 720 (\rho q_s)^{1.75}}{\pi^{1.75} N_s^{0.75} \rho_s^{1.75}}$ for wet snow.

For hail, $Z_{eh} = \left(\frac{10^{18} \times 720}{\pi^{1.75} N_h^{0.75} \rho_h^{1.75}}\right)^{0.95} (\rho q_h)^{1.6625}$. Detailed definition of the variables can be found in Tong and Xue (2005).

An example of more sophisticated observation operators for polarimetric radar data can be found in Jung et al. (2008).

In the case of satellite radiances, it is the radiative transfer equations that relate the atmospheric temperature and moisture profiles to the radiances that satellite sees at the top of the atmosphere.

Reference: Tong, M. and M. Xue, 2005: Ensemble Kalman filter assimilation of Doppler radar data with a compressible nonhydrostatic model: OSS Experiments. *Mon. Wea. Rev.*, **133**, 1789-1807.

Jung, Y., G. Zhang, and M. Xue, 2008: Assimilation of simulated polarimetric radar data for a convective storm using ensemble Kalman filter. Part I: Observation operators for reflectivity and polarimetric variables. *Mon. Wea. Rev.*, 2228-2245.
(http://twister.ou.edu/papers/JungEtal_MWR2008a.pdf) .

Optimal Analysis

The optimal analysis is equal to the background plus the innovation weighted by optimal weights which are determined so as to minimize the analysis error variance.

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{W}(\mathbf{y}_o - H(\mathbf{x}_b)) = \mathbf{x}_b + \mathbf{W}\mathbf{d}.$$

The analysis and the background are vectors of length n (the total number of grid points times the number of model variables), the weights are given by a matrix of dimension $(n \times p)$.

H is the forward observational operator H that converts the background field into "observed first guesses."

H can be nonlinear (e.g., the reflectivity equation or the radiative transfer equations that go from temperature and moisture vertical profiles to the satellite observed radiances).

The observation field \mathbf{y}_o is a vector of length p , the number of observations.

The vector \mathbf{d} , also of length p , is called the "innovation" or "observational increments" vector:

$$\mathbf{d} = \mathbf{y}_o - H(\mathbf{x}_b),$$

which is defined as the difference between the observation and the background mapped to the observational point via forward operator H .

Remarks:

a) The weight matrix \mathbf{W} is also called the *gain matrix* \mathbf{K} , especially in the Kalman filter literature.

b) An error covariance matrix is obtained by multiplying a vector error $\boldsymbol{\varepsilon} = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ e_n \end{bmatrix}$ by its

transpose $\boldsymbol{\varepsilon}^T = [e_1 \ e_2 \ \cdot \ \cdot \ e_n]$, and averaging over many cases, to obtain the expected value:

$$\mathbf{P} = \overline{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T} = \begin{bmatrix} \overline{e_1 e_1} & \overline{e_1 e_2} & \cdot & \overline{e_1 e_n} \\ \overline{e_2 e_1} & \overline{e_2 e_2} & \cdot & \overline{e_2 e_n} \\ \cdot & \cdot & \cdot & \cdot \\ \overline{e_n e_1} & \overline{e_n e_2} & \cdot & \overline{e_n e_n} \end{bmatrix}$$

where the overbar represents the expected value (same as $E(\)$).

Therefore the background error covariance matrix is a huge $n \times n$ matrix where n can easily be 10^7 for an operational NWP model!

A covariance matrix is symmetric and positive definite (i.e., $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$ for all nonzero \mathbf{x}).

The diagonal elements are the variances of the vector error components $\overline{e_i e_i} = \sigma_i^2$.

If we normalize the covariance matrix, dividing each component by the product of the standard deviations, $\overline{e_i e_j} / \sigma_i \sigma_j = \text{corr}(e_i, e_j) = \rho_{ij}$, we obtain a correlation matrix

$$\mathbf{C} = \begin{bmatrix} 1 & \rho_{12} & \cdot & \rho_{1n} \\ \rho_{12} & 1 & \cdot & \rho_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \rho_{1n} & \rho_{12} & \cdot & 1 \end{bmatrix}$$

and if $\mathbf{D} = \begin{bmatrix} \sigma_1^2 & 0 & \cdot & 0 \\ 0 & \sigma_2^2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \sigma_n^2 \end{bmatrix}$ is the diagonal matrix of the variances, then we can write

$$\mathbf{P} = \mathbf{D}^{1/2} \mathbf{C} \mathbf{D}^{1/2}$$

$$\text{and } \mathbf{D}^{1/2} = \begin{bmatrix} \sigma_1 & 0 & \cdot & 0 \\ 0 & \sigma_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \sigma_n \end{bmatrix}.$$

$$\text{c) } [\mathbf{AB}]^T = \mathbf{B}^T \mathbf{A}^T; [\mathbf{AB}]^{-1} = \mathbf{B}^{-1} \mathbf{A}^{-1}$$

d) The general form of a quadratic function is $F(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{d}^T \mathbf{x} + c$ where \mathbf{A} is a symmetric matrix, \mathbf{d} is a vector and c a scalar.

To find the gradient of this scalar function

$$\nabla_{\mathbf{x}} F = \frac{\partial F}{\partial \mathbf{x}} = \begin{bmatrix} \partial F / \partial x_1 \\ \partial F / \partial x_2 \\ \cdot \\ \partial F / \partial x_n \end{bmatrix},$$

a column vector, we use the following properties of the gradient with respect to \mathbf{x} :

$$\nabla(\mathbf{d}^T \mathbf{x}) = \nabla(\mathbf{x}^T \mathbf{d}) = \mathbf{d} \quad (\text{since } \nabla_{\mathbf{x}} \mathbf{x}^T = \mathbf{I}, \text{ the identity matrix), or}$$

$$\nabla(\mathbf{d}^T \mathbf{x}) = \nabla(d_1 x_1 + d_2 x_2 + \dots d_n x_n) = \begin{bmatrix} \partial(d_1 x_1 + d_2 x_2 + \dots d_n x_n) / \partial x_1 \\ \partial(d_1 x_1 + d_2 x_2 + \dots d_n x_n) / \partial x_2 \\ \cdot \\ \partial(d_1 x_1 + d_2 x_2 + \dots d_n x_n) / \partial x_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \cdot \\ d_n \end{bmatrix} = \mathbf{d}.$$

And $\nabla(\mathbf{x}^T \mathbf{A} \mathbf{x}) = 2 \mathbf{A} \mathbf{x}.$

We can see this by taking, say, time derivative of \mathbf{x} :

$$\frac{d(\mathbf{x}^T \mathbf{A} \mathbf{x})}{dt} = \dot{\mathbf{x}}^T \mathbf{A} \mathbf{x} + \mathbf{x}^T \mathbf{A} \dot{\mathbf{x}} = (\dot{\mathbf{x}}^T \mathbf{A} \mathbf{x})^T + \mathbf{x}^T \mathbf{A} \dot{\mathbf{x}} = \mathbf{x}^T \mathbf{A} \dot{\mathbf{x}} + \mathbf{x}^T \mathbf{A} \dot{\mathbf{x}} = 2\mathbf{x}^T \mathbf{A} \dot{\mathbf{x}}$$

$$d(\mathbf{x}^T \mathbf{A} \mathbf{x}) = 2\mathbf{x}^T \mathbf{A} d\mathbf{x} = 2(\mathbf{A} \mathbf{x})^T d\mathbf{x}$$

Because the differential of function $F(\mathbf{x}) = F(x_1, x_2, \dots, x_n)$ is

$$dF = \frac{\partial F}{\partial x_1} dx_1 + \frac{\partial F}{\partial x_2} dx_2 + \dots + \frac{\partial F}{\partial x_n} dx_n = \begin{bmatrix} \frac{\partial F}{\partial x_1} & \frac{\partial F}{\partial x_2} & \dots & \frac{\partial F}{\partial x_n} \end{bmatrix} \begin{bmatrix} dx_1 \\ dx_2 \\ \vdots \\ dx_n \end{bmatrix} = [\nabla_x F]^T d\mathbf{x}$$

In our current case, $F = \mathbf{x}^T \mathbf{A} \mathbf{x}$, therefore

$$\nabla(\mathbf{x}^T \mathbf{A} \mathbf{x}) = 2\mathbf{A} \mathbf{x}$$

Therefore, $F(x) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{d}^T \mathbf{x} + c$,

$$\nabla F(\mathbf{x}) = \mathbf{A} \mathbf{x} + \mathbf{d}, \quad \nabla^2 F(\mathbf{x}) = \mathbf{A} \quad \text{and} \quad \delta F = (\nabla F)^T \delta \mathbf{x}$$

The above will be used when deriving the gradient of quadratic-form cost function with respect to the analysis variable \mathbf{x} , which is needed when we try to find 3DVAR or 4DVAR solution.

Refer to http://twister.ou.edu/OBAN2008/Intro_FEM_files/IFEM.AppD.pdf for further discussions on matrix calculus.

Statistical assumptions

We define the background error and the analysis error as vectors of length n :

$$\varepsilon_b(x, y) = \mathbf{x}_b(x, y) - \mathbf{x}_t(x, y)$$

$$\varepsilon_a(x, y) = \mathbf{x}_a(x, y) - \mathbf{x}_t(x, y)$$

The p observations available at irregularly spaced points $\mathbf{y}_o(\mathbf{r}_i)$ have observational errors

$$\varepsilon_{oi} = \mathbf{y}_o(\mathbf{r}_i) - \mathbf{y}_t(\mathbf{r}_i) = \mathbf{y}_o(\mathbf{r}_i) - H(\mathbf{x}_t).$$

We do not know the truth \mathbf{x}_t , thus we do not know the errors of the available background and observations, but we can make a number of assumptions about their statistical properties.

The background and observations are assumed to be unbiased:

$$E\{\varepsilon_b(x, y)\} = E\{\mathbf{x}_b(x, y)\} - E\{\mathbf{x}_t(x, y)\} = 0$$

$$E\{\varepsilon_o(\mathbf{r}_i)\} = E\{\mathbf{y}_o(\mathbf{r}_i)\} - E\{\mathbf{y}_t(\mathbf{r}_i)\} = 0$$

If the forecasts (background) and the observations are biased, in principle we can and should correct the bias before proceeding.

We define the **error covariance matrices** for the analysis, background and observations respectively:

$$\mathbf{P}_a = \mathbf{A} = E\{\boldsymbol{\varepsilon}_a \boldsymbol{\varepsilon}_a^T\}$$

$$\mathbf{P}_b = \mathbf{B} = E\{\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_b^T\}$$

$$\mathbf{P}_o = \mathbf{R} = E\{\boldsymbol{\varepsilon}_o \boldsymbol{\varepsilon}_o^T\}$$

The nonlinear observation operator H that transforms model variables into observed variables can be linearized as

$$H(\mathbf{x} + \delta\mathbf{x}) = H(\mathbf{x}) + \mathbf{H}\delta\mathbf{x}$$

where \mathbf{H} is a $p \times n$ matrix with elements $h_{i,j} = \partial H_i / \partial x_j$ called the linear observation operator.

Let's see why \mathbf{H} is defined as above. First, for a scalar function of multiple independent variables, such as

$y = y(x_1, x_2, \dots, x_n)$ or written in a vector form $y = y(\mathbf{x})$, using Taylor series expansion, we have

$$\begin{aligned} y(\mathbf{x} + \delta\mathbf{x}) &= y(x_1 + \delta x_1, x_2 + \delta x_2, \dots, x_n + \delta x_n) \\ &= y(x_1, x_2, \dots, x_n) + \frac{\partial y}{\partial x_1} \delta x_1 + \frac{\partial y}{\partial x_2} \delta x_2 + \dots + \frac{\partial y}{\partial x_n} \delta x_n + O(\delta x_1^2, \delta x_2^2, \dots, \delta x_n^2) \\ &\approx y(\mathbf{x}) + \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \dots & \frac{\partial y}{\partial x_n} \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \\ \vdots \\ \delta x_n \end{bmatrix} = y(\mathbf{x}) + [\nabla_{\mathbf{x}} y]^T \delta\mathbf{x} \end{aligned}$$

Now the function we have is a vector function $\mathbf{y} = H(\mathbf{x})$ where \mathbf{y} is a vector of length p , the number of observations and H is a vector representing p number of functions. Writing out the vector function explicitly, we have

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ y_p \end{bmatrix} = \begin{bmatrix} H_1(\mathbf{x}) \\ H_2(\mathbf{x}) \\ \cdot \\ H_p(\mathbf{x}) \end{bmatrix}$$

therefore

$$\mathbf{y}(\mathbf{x} + \delta\mathbf{x}) = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ y_p \end{bmatrix} = \begin{bmatrix} H_1(\mathbf{x} + \delta\mathbf{x}) \\ H_2(\mathbf{x} + \delta\mathbf{x}) \\ \cdot \\ H_p(\mathbf{x} + \delta\mathbf{x}) \end{bmatrix} = \begin{bmatrix} H_1(\mathbf{x}) \\ H_2(\mathbf{x}) \\ \cdot \\ H_p(\mathbf{x}) \end{bmatrix} + \begin{bmatrix} \frac{\partial H_1}{\partial x_1} & \frac{\partial H_1}{\partial x_2} & \cdots & \frac{\partial H_1}{\partial x_n} \\ \frac{\partial H_2}{\partial x_1} & \frac{\partial H_2}{\partial x_2} & \cdots & \frac{\partial H_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial H_p}{\partial x_1} & \frac{\partial H_p}{\partial x_2} & \cdots & \frac{\partial H_p}{\partial x_n} \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \\ \cdot \\ \delta x_n \end{bmatrix} = H(\mathbf{x}) + \mathbf{H}\delta\mathbf{x},$$

and the elements of matrix \mathbf{H} are $h_{i,j} = \partial H_i / \partial x_j$, where $i=1, 2, \dots, p$, and $j=1, 2, \dots, n$.

In practice, \mathbf{H} is often a sparse matrix because each element of \mathbf{y} is often not dependent on all elements of \mathbf{x} . For example, if H operator is a bilinear interpolation operator, only 4 instead of n grid point values are involved, i.e., this particular \mathbf{y} element here is only a function of 4 elements of \mathbf{x} . In the example of radar reflectivity, only three

variables (q_r , q_s and q_h , the rainwater, snow and hail mixing ratios) are involved instead of 10 or so model variables (not u , v , w , T or p) (although T -dependency can also appear in the formula).

We also assume that the background (usually a model forecast) is a good approximation of the truth, so that the analysis and the observations are equal to the background values plus small increments. Therefore, the innovation vector \mathbf{d} defined earlier can be written as

$$\begin{aligned}\mathbf{d} &= \mathbf{y}_o - H(\mathbf{x}_b) = \mathbf{y}_o - H[\mathbf{x}_t + (\mathbf{x}_b - \mathbf{x}_t)] \\ &\approx \mathbf{y}_o - H(\mathbf{x}_t) - \mathbf{H}(\mathbf{x}_b - \mathbf{x}_t) = \boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b\end{aligned}$$

The \mathbf{H} matrix transforms variables in model space into their corresponding values in observation space. Its transpose or adjoint \mathbf{H}^T transforms variables in observation space to variables in model space (more on this later).

The background error covariance \mathbf{B} (a matrix of size $n \times n$) and the observation error covariance \mathbf{R} (a matrix of size $p \times p$) are assumed to be known in IO analysis.

In addition, we assume that **the observation and background errors are uncorrelated.**

The analysis error covariance \mathbf{P}_a is not known, but **we will minimize it through the optimal choice of weights.**

Derivation for analysis error covariance

In order to be able to minimize the analysis error, we first derive the formula for analysis error covariance.

From the above, $\mathbf{d} = \mathbf{y}_o - H(\mathbf{x}_b) = \boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b$ and $\mathbf{x}_a = \mathbf{x}_b + \mathbf{W}(\mathbf{y}_o - H(\mathbf{x}_b)) = \mathbf{x}_b + \mathbf{W}\mathbf{d}$, and therefore the analysis error is

$$\boldsymbol{\varepsilon}_a = \mathbf{x}_a - \mathbf{x}_t = (\mathbf{x}_b + \mathbf{W}\mathbf{d}) - \mathbf{x}_t = (\mathbf{x}_b - \mathbf{x}_t) + \mathbf{W}\mathbf{d} = \boldsymbol{\varepsilon}_b + \mathbf{W}(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)$$

(so we wrote the analysis errors in terms of the background error and observation error)

We want to choose the weight matrix \mathbf{W} (a.k.a. gain matrix \mathbf{K}) so as to minimize the total analysis error variance (**note that minimizing the total analysis error variance is the key or essence of optimal statistical analysis**):

$$\mathbf{P}_a = E\{(\mathbf{x}_a - \mathbf{x}_t)(\mathbf{x}_a - \mathbf{x}_t)^T\} = E\{[\boldsymbol{\varepsilon}_b + \mathbf{W}(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)][\boldsymbol{\varepsilon}_b + \mathbf{W}(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)]^T\}$$

Expanding the above equation and using the rules for matrix multiplication from the earlier review, we get

$$\begin{aligned}\mathbf{P}_a &= E\{\boldsymbol{\varepsilon}_a \boldsymbol{\varepsilon}_a^T\} = E\{[\boldsymbol{\varepsilon}_b + \mathbf{W}(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)][\boldsymbol{\varepsilon}_b^T + (\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)^T \mathbf{W}^T]\} \\ &= E\{\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_b^T + \boldsymbol{\varepsilon}_b (\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)^T \mathbf{W}^T + \mathbf{W}(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b) \boldsymbol{\varepsilon}_b^T + \mathbf{W}(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b) (\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)^T \mathbf{W}^T\} \\ &= E\{\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_b^T\} + E\{\boldsymbol{\varepsilon}_b (\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)^T \mathbf{W}^T\} + E\{\mathbf{W}(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b) \boldsymbol{\varepsilon}_b^T\} + E\{\mathbf{W}(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b) (\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)^T \mathbf{W}^T\} \\ &= E\{\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_b^T\} + E\{\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_o^T\} \mathbf{W}^T - E\{\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_b^T\} \mathbf{H}^T \mathbf{W}^T + \mathbf{W} E\{\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_o^T\} - \mathbf{W} \mathbf{H} E\{\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_b^T\} \\ &\quad + \mathbf{W}[E\{\boldsymbol{\varepsilon}_o \boldsymbol{\varepsilon}_o^T\} - \mathbf{H} E\{\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_o^T\} - E\{\boldsymbol{\varepsilon}_o \boldsymbol{\varepsilon}_b^T\} \mathbf{H}^T + \mathbf{H} E\{\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_b^T\} \mathbf{H}^T] \mathbf{W}^T \\ &= \mathbf{B} - \mathbf{B} \mathbf{H}^T \mathbf{W}^T - \mathbf{W} \mathbf{H} \mathbf{B} + \mathbf{W}[\mathbf{R} + \mathbf{H} \mathbf{B} \mathbf{H}^T] \mathbf{W}^T\end{aligned}$$

In the above, we used the assumption that the *background errors are not correlated with the observational errors*, i.e., that their covariance is equal to zero. We used \mathbf{B} and \mathbf{R} for the background and observational error covariance matrices, respectively.

The optimal solution we want to obtain is a solution that minimizes the total variance of the analysis, which is a measure of the analysis error and is a scalar quantity.

The optimal weight matrix is obtained by differentiating the total analysis variance, or the trace of matrix \mathbf{P}_a , with respect to weight matrix \mathbf{W} and setting the derivative to zero. The trace of \mathbf{P}_a is

$$\begin{aligned}\text{Tr}(\mathbf{P}_a) &= \text{Tr}(\mathbf{B}) - \text{Tr}(\mathbf{B}\mathbf{H}^T\mathbf{W}^T) - \text{Tr}(\mathbf{W}\mathbf{H}\mathbf{B}) + \text{Tr}(\mathbf{W}[\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T]\mathbf{W}^T) \\ &= \text{Tr}(\mathbf{B}) - \text{Tr}(\mathbf{B}\mathbf{H}^T\mathbf{W}^T) - \text{Tr}(\mathbf{B}\mathbf{H}^T\mathbf{W}^T)^T + \text{Tr}(\mathbf{W}[\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T]\mathbf{W}^T) \quad . \\ &= \text{Tr}(\mathbf{B}) - 2 \text{Tr}(\mathbf{B}\mathbf{H}^T\mathbf{W}^T) + \text{Tr}(\mathbf{W}[\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T]\mathbf{W}^T)\end{aligned}$$

One way to find the derivative or gradient is to define a differential, following that for a scalar function $f = f(x)$:

$$\delta f = \frac{\delta f}{\delta x} \delta x = f(x + \delta x) - f(x)$$

In the limit that δx goes to zero,

$$\nabla_x f \delta x = \lim_{\delta x \rightarrow 0} [f(x + \delta x) - f(x)] \quad (\text{where } \nabla_x f \equiv \lim_{\delta x \rightarrow 0} \frac{\delta f}{\delta x}).$$

We apply the above formula to a small variation in \mathbf{W} ,

$$\begin{aligned}
\nabla_{\mathbf{W}} \text{Tr}(\mathbf{P}_a) \delta \mathbf{W} &= -2 \text{Tr}(\mathbf{B}\mathbf{H}^T \delta \mathbf{W}^T) + \text{Tr}\{(\mathbf{W} + \delta \mathbf{W})[\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T](\mathbf{W}^T + \delta \mathbf{W}^T)\} - \text{Tr}\{\mathbf{W}[\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T]\mathbf{W}^T\} \\
&= -2 \text{Tr}(\mathbf{B}\mathbf{H}^T \delta \mathbf{W}^T) + \text{Tr}\{\delta \mathbf{W}[\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T]\mathbf{W}^T\} + \text{Tr}\{\mathbf{W}[\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T]\delta \mathbf{W}^T\} \\
&= -2 \text{Tr}(\mathbf{B}\mathbf{H}^T \delta \mathbf{W}^T) + 2\text{Tr}(\mathbf{W}[\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T]\delta \mathbf{W}^T) \\
&= 2\text{Tr}\{(\mathbf{W}[\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T] - \mathbf{B}\mathbf{H}^T)\delta \mathbf{W}^T\}
\end{aligned}$$

where the second order term in $\delta \mathbf{W}$ has been neglected.

If we let $\mathbf{W}[\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T] - \mathbf{B}\mathbf{H}^T \equiv \mathbf{E}$ and

$$\mathbf{E} = \begin{bmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \\ \vdots \\ \mathbf{e}_n^T \end{bmatrix}, \text{ where } \mathbf{e}_i^T \text{ are row vectors (of } 1 \times p) \text{ making up matrix } \mathbf{E}.$$

Further write

$$\delta \mathbf{W} = \begin{bmatrix} \delta \mathbf{w}_1^T \\ \delta \mathbf{w}_2^T \\ \vdots \\ \delta \mathbf{w}_n^T \end{bmatrix} \text{ and } \delta \mathbf{W}^T = [\delta \mathbf{w}_1 \quad \delta \mathbf{w}_2 \quad \dots \quad \delta \mathbf{w}_n],$$

where \mathbf{w}_i are vectors of $p \times 1$, that make up matrix \mathbf{W} , and \mathbf{W} is an $n \times p$ matrix, then

$$\nabla_{\mathbf{w}} \text{Tr}(\mathbf{P}_a) \delta \mathbf{W} = 2 \text{Tr}\{\mathbf{E} \delta \mathbf{W}^T\} = 2 \text{Tr} \left(\begin{bmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \\ \vdots \\ \mathbf{e}_n^T \end{bmatrix} \begin{bmatrix} \delta \mathbf{w}_1 & \delta \mathbf{w}_2 & \dots & \delta \mathbf{w}_n \end{bmatrix} \right) = 2 \sum_{i=1}^n \mathbf{e}_i^T \delta \mathbf{w}_i.$$

Here the definition of trace has been used.

Since $\delta \mathbf{W}$ can take any value, for the above gradient to be zero, \mathbf{e}_i^T has to be zero for all i . This gives $\mathbf{E} = 0$, or

$$\mathbf{W}[\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T] - \mathbf{B}\mathbf{H}^T = 0.$$

We thus obtain the **optimal weight matrix as given by**

$$\mathbf{W} = \mathbf{B}\mathbf{H}^T(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1}.$$

Finally, we obtain the analysis error covariance:

$$\mathbf{P}_a = \mathbf{B} - \mathbf{B}\mathbf{H}^T\mathbf{W}^T - \mathbf{W}\mathbf{H}\mathbf{B} + \mathbf{W}\mathbf{R}\mathbf{W}^T + \mathbf{W}\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{W}^T$$

and substituting in \mathbf{W} obtained above, we have

$$\begin{aligned}
\mathbf{P}_a &= \mathbf{B} - \mathbf{B}\mathbf{H}^T\mathbf{W}^T - \mathbf{W}\mathbf{H}\mathbf{B} + \mathbf{W}(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)\mathbf{W}^T \\
&= \mathbf{B} - \mathbf{B}\mathbf{H}^T\mathbf{W}^T - \mathbf{W}\mathbf{H}\mathbf{B} + \mathbf{B}\mathbf{H}^T\mathbf{W}^T \\
&= \mathbf{B}(\mathbf{I} - \mathbf{W}\mathbf{H})
\end{aligned}$$

For convenience, we repeat the basic equations of OI:

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{W}[\mathbf{y}_o - H(\mathbf{x}_b)] = \mathbf{x}_b + \mathbf{W}\mathbf{d} \quad (1)$$

$$\mathbf{W} = \mathbf{B}\mathbf{H}^T(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1} \quad (2a)$$

We will see later when we derive the variational approach or 3D-Var that the weight matrix (2a) can be written in an alternative equivalent form as

$$\mathbf{W} = (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1} \quad (2b)$$

$$\mathbf{P}_a = (\mathbf{I}_n - \mathbf{W}\mathbf{H})\mathbf{B} \quad (3)$$

where the subindex n is a reminder that the identity matrix is in the analysis or model space.

The interpretation of these equations is:

Equation (1) says: $(\mathbf{x}_a = \mathbf{x}_b + \mathbf{W}[\mathbf{y}_o - H(\mathbf{x}_b)] = \mathbf{x}_b + \mathbf{W}\mathbf{d})$

“The analysis is obtained by adding to the first guess (background) the product of the optimal weight (or gain) matrix and the innovation (difference between the observation and first guess).

The first guess of the observations is obtained by applying the observation operator H to the background vector.”

Also, note that from earlier, $H(\mathbf{x}_b) = H(\mathbf{x}_t) + \mathbf{H}(\mathbf{x}_b - \mathbf{x}_t) = H(\mathbf{x}_t) + \mathbf{H}\varepsilon_b$, where the matrix \mathbf{H} is the linear tangent perturbation of H .

Equation (2a) says: $(\mathbf{W} = \mathbf{B}\mathbf{H}^T(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1})$

"The optimal weight (or gain) matrix is given by the background error covariance in the observation space ($\mathbf{B}\mathbf{H}^T$) multiplied by the inverse of the total error covariance (sum of the background and the observation error covariances)."

Note that the larger the background error covariance compared to the observation error covariance, the larger the correction to the first guess.

Equation (3) says: $(\mathbf{P}_a = (\mathbf{I}_n - \mathbf{W}\mathbf{H})\mathbf{B})$

"The error covariance of the analysis is given by the error covariance of the background, reduced by a matrix equal to the identity matrix ($n \times n$) minus the optimal weight matrix".

Finally we derive **an alternative formulation for the analysis error covariances** showing the additive properties of the "precision " (if all the statistical assumptions hold true):

From equations $\varepsilon_a = \mathbf{x}_a - \mathbf{x}_t = \mathbf{x}_b + \mathbf{W}\mathbf{d} - \mathbf{x}_t = \varepsilon_b + \mathbf{W}(\varepsilon_o - \mathbf{H}\varepsilon_b)$ and $\mathbf{W} = (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}$ (obtained using the 3DVAR approach later and can be shown to be equivalent to the \mathbf{W} we obtained above), we can show that

$$\begin{aligned}
\varepsilon_a &= \varepsilon_b + [\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}]^{-1} \mathbf{H}^T \mathbf{R}^{-1} (\varepsilon_0 - \mathbf{H} \varepsilon_b) \\
&= [\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}]^{-1} \{ [\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}] \varepsilon_b + \mathbf{H}^T \mathbf{R}^{-1} (\varepsilon_0 - \mathbf{H} \varepsilon_b) \} \\
&= [\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}]^{-1} [\mathbf{B}^{-1} \varepsilon_b + \mathbf{H}^T \mathbf{R}^{-1} \varepsilon_0]
\end{aligned}$$

If we compute again $\mathbf{P}_a = E\{\varepsilon_a \varepsilon_a^T\}$ from the above, and make use of $E\{\varepsilon_b \varepsilon_o^T\} = 0$, $\mathbf{P}_b = \mathbf{B} = E\{\varepsilon_b \varepsilon_b^T\}$, $\mathbf{P}_o = \mathbf{R} = E\{\varepsilon_o \varepsilon_o^T\}$, we obtain

$$\mathbf{P}_a^{-1} = \mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}. \quad (4)$$

Here is proof (noting that $\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}$ is symmetric):

$$\begin{aligned}
\mathbf{P}_a &= \frac{E\{[\mathbf{B}^{-1} \varepsilon_b + \mathbf{H}^T \mathbf{R}^{-1} \varepsilon_o][\varepsilon_b^T \mathbf{B}^{-1T} + \varepsilon_o^T \mathbf{R}^{-1T} \mathbf{H}]\}}{[\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}][\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}]^T} \\
&= \frac{\mathbf{B}^{-1} E\{\varepsilon_b \varepsilon_b^T\} \mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} E\{\varepsilon_o \varepsilon_o^T\} \mathbf{R}^{-1} \mathbf{H}}{[\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}][\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}]^T} \\
&= \frac{\mathbf{B}^{-1} \mathbf{B} \mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{R} \mathbf{R}^{-1} \mathbf{H}}{[\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}][\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}]^T} \\
&= \frac{\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}}{[\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}][\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}]^T} \\
&= \frac{1}{[\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}]^T}
\end{aligned}$$

Equation (4) says:

"The analysis precision, defined as the inverse of the analysis error covariance, is the sum of the background precision and the observation precision projected into the model space."

Note that **all these statements are dependent on the assumption that the statistical estimates of the errors are accurate.**

If

- 1) the observations and/or background error covariances are poorly known,
- 2) there are biases, or
- 3) the observations and background errors are correlated,

the analysis accuracy can be considerably worse than implied by (3) or (4).

Appendix: Derivative of vector with respect to vector and the chain rule of vector functions

Earlier, we defined the derivative of $\mathbf{y} = H(\mathbf{x})$ with respect to vector \mathbf{x} . What if \mathbf{x} is a function of another vector \mathbf{z} (or length m), *i.e.*, $\mathbf{x} = \mathbf{x}(\mathbf{z})$, and we want to find the derivative of \mathbf{y} with respect to \mathbf{z} ? In this case, $\mathbf{y} = \mathbf{y}(\mathbf{z})$.

Because

$$\frac{\partial \mathbf{y}}{\partial \mathbf{z}} = \begin{bmatrix} \frac{\partial y_1}{\partial z_1} & \frac{\partial y_1}{\partial z_2} & \cdots & \frac{\partial y_1}{\partial z_m} \\ \frac{\partial y_2}{\partial z_1} & \frac{\partial y_2}{\partial z_2} & \cdots & \frac{\partial y_2}{\partial z_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_p}{\partial z_1} & \frac{\partial y_p}{\partial z_2} & \cdots & \frac{\partial y_p}{\partial z_m} \end{bmatrix}$$

where, using scalar chain rule,

$$\frac{\partial y_i}{\partial z_j} = \sum_{k=1}^n \frac{\partial y_i}{\partial x_k} \frac{\partial x_k}{\partial z_j} \quad \text{for } i = 1, \dots, n; \quad j = 1, \dots, m.$$

Therefore,

$$\begin{aligned}
\frac{\partial \mathbf{y}}{\partial \mathbf{z}} &= \begin{bmatrix} \sum_{k=1}^n \frac{\partial y_1}{\partial x_k} \frac{\partial x_k}{\partial z_1} & \sum_{k=1}^n \frac{\partial y_1}{\partial x_k} \frac{\partial x_k}{\partial z_2} & \cdots & \sum_{k=1}^n \frac{\partial y_1}{\partial x_k} \frac{\partial x_k}{\partial z_m} \\ \sum_{k=1}^n \frac{\partial y_2}{\partial x_k} \frac{\partial x_k}{\partial z_1} & \sum_{k=1}^n \frac{\partial y_2}{\partial x_k} \frac{\partial x_k}{\partial z_2} & \cdots & \sum_{k=1}^n \frac{\partial y_2}{\partial x_k} \frac{\partial x_k}{\partial z_m} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^n \frac{\partial y_p}{\partial x_k} \frac{\partial x_k}{\partial z_1} & \sum_{k=1}^n \frac{\partial y_p}{\partial x_k} \frac{\partial x_k}{\partial z_2} & \cdots & \sum_{k=1}^n \frac{\partial y_p}{\partial x_k} \frac{\partial x_k}{\partial z_m} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_p}{\partial x_1} & \frac{\partial y_p}{\partial x_2} & \cdots & \frac{\partial y_p}{\partial x_n} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial z_1} & \frac{\partial x_1}{\partial z_2} & \cdots & \frac{\partial x_1}{\partial z_m} \\ \frac{\partial x_2}{\partial z_1} & \frac{\partial x_2}{\partial z_2} & \cdots & \frac{\partial x_2}{\partial z_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial z_1} & \frac{\partial x_n}{\partial z_2} & \cdots & \frac{\partial x_n}{\partial z_m} \end{bmatrix} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{z}}
\end{aligned}$$

The chain rule for the vector to vector derivative is similar to that of scalar function.

However, some authors definite the derivatives as transposes of our definition given here, in that case, the chain rule will look like $\frac{\partial \mathbf{y}}{\partial \mathbf{z}} = \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \frac{\partial \mathbf{y}}{\partial \mathbf{x}}$, i.e., the chain of matrices builds “towards the left” instead to right.