The minimum norm solution in this case becomes the solution of the following optimization problem:

$$\text{Minimize } J = \mathbf{m}^T \mathbf{m}$$

$$\text{Subject to } \; ||\mathbf{Gm} - \mathbf{d}||_2^2 = \epsilon$$

It is clear that now rather than having one constraint per observation we have a single global constraint for the complete set of observations.

Please, notice that in the previous equation we have used the $l_2$ norm as a measure of distance for the errors in the data; we will see that this also implies that the errors are considered to be distributed according to the normal law (Gaussian errors). Before continuing with the analysis, we recall that

$$||\mathbf{Gm} - \mathbf{d}||_2^2 = ||\mathbf{e}||_2^2$$

which in matrix/vector notation can be also expressed as

$$||\mathbf{e}||_2^2 = \mathbf{e}^T \mathbf{e} \,.$$

Coming back to our optimization problem, we now minimize the cost function $J'$ given by

$$\begin{aligned} J' &= \mu \text{Model Norm} + \text{Misfit} \\ &= \mu \mathbf{m}^T \mathbf{m} + \mathbf{e}^T \mathbf{e} \\ &= \mu \mathbf{m}^T \mathbf{m} + (\mathbf{Gm} - \mathbf{d})^T (\mathbf{Gm} - \mathbf{d}) \end{aligned}$$

The solution is now obtained by minimizing $J'$ with respect to the unknown $\mathbf{m}$. This requires some algebra and I will give you the final solution:

$$\begin{aligned}\frac{d\,J'}{d\mathbf{m}} &= 0\\ &= (\mathbf{G}^T\mathbf{G}+\mu\mathbf{I})\mathbf{m}-\mathbf{G}^T\mathbf{d}=\mathbf{0}\,.\end{aligned}$$

The minimizer is then given by

$$\mathbf{m}=(\mathbf{G}^T\mathbf{G}+\mu\mathbf{I})^{-1}\mathbf{G}^T\mathbf{d}\,. \tag{13}$$

This solution is often called the *damped least squares solution*. Notice that the structure of the solution looks like the solution we obtain when we solve a least squares problem. A simple identity permits one to make equation (13) look like a minimum norm solution:

Identity $\quad (\mathbf{G}^T\mathbf{G}+\mathbf{I})^{-1}\mathbf{G}^T = \mathbf{G}^T(\mathbf{G}\mathbf{G}^T+\mathbf{I})^{-1}\,.$

Therefore, equation (13) can be re-expressed as

$$\mathbf{m} = \mathbf{G}^T(\mathbf{G}\mathbf{G}^T+\mu\mathbf{I})^{-1}\mathbf{d}\,. \tag{14}$$

It is important to note that the previous expression reduces to the minimum norm solution for exact data when $\mu = 0$.

## About $\mu$

The importance of $\mu$ can be seen from the cost function $J'$

$$J' = \mu \text{Model Norm} + \text{Misfit}$$

- Large $\mu$ means more weight (importance) is given to minimizing the misfit over the model norm.
- Small $\mu$ means that the model norm is the main term entering in the minimization; the misfit becomes less important.
- You can think that we are trying to simultaneously achieve two *goals*:

  <u>Norm Reduction</u> (Stability - we don't want higly oscillatory solutions)

  <u>Misfit Reduction</u> (We want to honor our observations)

We will explore the fact that these two goals cannot be simultaneously achieved, and, this is why we often call $\mu$ a trade-off parameter.

The parameter $\mu$ receives different names according to the scientific background of the user:

1. Statisticians: Hyper-parameter
2. Mathematicians: Regularization parameter, Penalty parameter
3. Engineers: Damping term, Damping factor, Stabilization parameter
4. Signal Processing: Ridge regression parameter, Trade-off parameter