Since $u_{k+1} = u_k - \rho_k \nabla J_{u_k}$ and $\nabla J_{u_k} = Au_k - b$, we have

$$u_{k+1} - u = (u_k - u) - \rho_k(Au_k - Au) = (I - \rho_k A)(u_k - u),$$

so we get

$$\|u_{k+1} - u\| \leq \|I - \rho_k A\|_2 \, \|u_k - u\| \, .$$

However, since $I - \rho_k A$ is a symmetric matrix, $\|I - \rho_k A\|_2$ is the largest absolute value of its eigenvalues, so

$$\|I - \rho_k A\|_2 \leq \max\{|1 - \rho_k \lambda_1|, |1 - \rho_k \lambda_n|\}.$$

The function

$$\mu(\rho) = \max\{|1 - \rho\lambda_1|, |1 - \rho\lambda_n|\}$$

is a piecewise affine function, and it is easy to see that if we pick a, b such that

$$0 < a \leq \rho_k \leq b < \frac{2}{\lambda_n},$$

then

$$\max_{\rho \in [a,b]} \mu(\rho) \leq \max\{\mu(a), \mu(b)\} < 1.$$

Therefore, the upper bound $2\lambda_1/\lambda_n^2$ can be replaced by $2/\lambda_n$, which is typically much larger. A "good" pick for ρ_k is $2/(\lambda_1 + \lambda_n)$ (as opposed to λ_1/λ_n^2 for the first version). In this case

$$|1 - \rho_k \lambda_1| = |1 - \rho_k \lambda_n| = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1},$$

so we get

$$\beta = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} = \frac{\frac{\lambda_n}{\lambda_1} - 1}{\frac{\lambda_n}{\lambda_1} + 1} = \frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1},$$

where $\mathrm{cond}_2(A) = \lambda_n/\lambda_1$ is the condition number of the matrix A with respect to the spectral norm. Thus we see that the larger the condition number of A is, the slower the convergence of the method will be. This is not surprising since we already know that linear systems involving ill-conditioned matrices (matrices with a large condition number) are problematic and prone to numerical instability. One way to deal with this problem is to use a method known as preconditioning.

We only described the most basic gradient descent methods. There are numerous variants, and we only mention a few of these methods.

The method of *scaling* consists in using $-\rho_k D_k \nabla J_{u_k}$ as descent direction, where D_k is some suitably chosen symmetric positive definite matrix.

In the *gradient method with extrapolation*, u_{k+1} is determined by

$$u_{k+1} = u_k - \rho_k \nabla J_{u_k} + \beta_k(u_k - u_{k-1}).$$