until stopping criterion is satisfied.

If $\| \;\|$ is the ℓ^2-norm, then we see immediately that $d_{\mathrm{sd},k} = -\nabla J_{u_k}$, so in this case the method *coincides* with the steepest descent method for the Euclidean norm as defined at the beginning of Section 49.6 in (3) and (4).

If P is a symmetric positive definite matrix, it is easy to see that $\|z\|_P = (z^\top P z)^{1/2} = \left\|P^{1/2} z\right\|_2$ is a norm. Then it can be shown that the normalized steepest descent direction is

$$d_{\mathrm{nsd,k}} = -(\nabla J_{u_k}^\top P^{-1} \nabla J_{u_k})^{-1/2} P^{-1} \nabla J_{u_k},$$

the dual norm is $\|z\|^D = \left\|P^{-1/2} z\right\|_2$, and the steepest descent direction with respect to $\| \;\|_P$ is given by

$$d_{\mathrm{sd},k} = -P^{-1} \nabla J_{u_k}.$$

A judicious choice for P can speed up the rate of convergence of the gradient descent method; see see Boyd and Vandenberghe [29] (Section 9.4.1 and Section 9.4.4).

If $\| \;\|$ is the ℓ^1-norm, then it can be shown that $d_{\mathrm{nsd,k}}$ is determined as follows: let i be any index for which $\|\nabla J_{u_k}\|_\infty = |(\nabla J_{u_k})_i|$. Then

$$d_{\mathrm{nsd,k}} = -\mathrm{sign}\left(\frac{\partial J}{\partial x_i}(u_k)\right) e_i,$$

where e_i is the ith canonical basis vector, and

$$d_{\mathrm{sd,k}} = -\frac{\partial J}{\partial x_i}(u_k) e_i.$$

For more details, see Boyd and Vandenberghe [29] (Section 9.4.2 and Section 9.4.4). It is also shown in Boyd and Vandenberghe [29] (Section 9.4.3) that the steepest descent method converges for any norm $\| \;\|$ and any strictly convex function J.

One of the main goals in designing a gradient descent method is to ensure that the convergence factor is as small as possible, which means that the method converges as quickly as possible. Machine learning has been a catalyst for finding such methods. A method discussed in Strang [171] (Chapter VI, Section 4) consists in adding a *momentum term* to the gradient. In this method, u_{k+1} and d_{k+1} are determined by the following system of equations:

$$\begin{aligned} u_{k+1} &= u_k - \rho d_k \\ d_{k+1} - \nabla J_{u_{k+1}} &= \beta d_k. \end{aligned}$$

Of course the trick is to choose ρ and β in such a way that the convergence factor is as small as possible. If J is given by a quadratic functional, say $(1/2)u^\top A u - b^\top u$, then $\nabla J_{u_{k+1}} = A u_{k+1} - b$ so we obtain a linear system. It turns out that the rate of convergence of