① • 更大的batch会计算更精确的梯度估计，但是回报却是小于线性的。

② • 极小batch通常难以充分利用多核架构。这促使我们使用一些绝对最小batch，低于这个值的小batch处理不会减少计算时间。

③ • 如果batch处理中的所有样本可以并行地处理（通常确是如此），那么内存消耗和batch大小会成比增长。对于很多硬件设施，这是batch大小的限制因素。

④ • 有些硬件在特定大小的数列上效果更好。尤其是在使用GPU时，通常使用 2 的幂数作为batch大小来获得更少的运行时间。一般，2 的幂数的取值范围是 32 到 256，16 有时用于尝试大的模型。

⑤ • 可能是由于小batch处理在学习过程中加入了噪扰，它们会有一些正则化效果 (Wilson and Martinez, 2003)。泛化误差通常在batch大小为 1 时最好。因为梯度估计的高方差，小batch训练需要较小的学习速率以维持稳定性。由于降低的学习速率和消耗更多步骤来观察整个训练集而产生更多的步骤，会导致总的运行时间非常大。