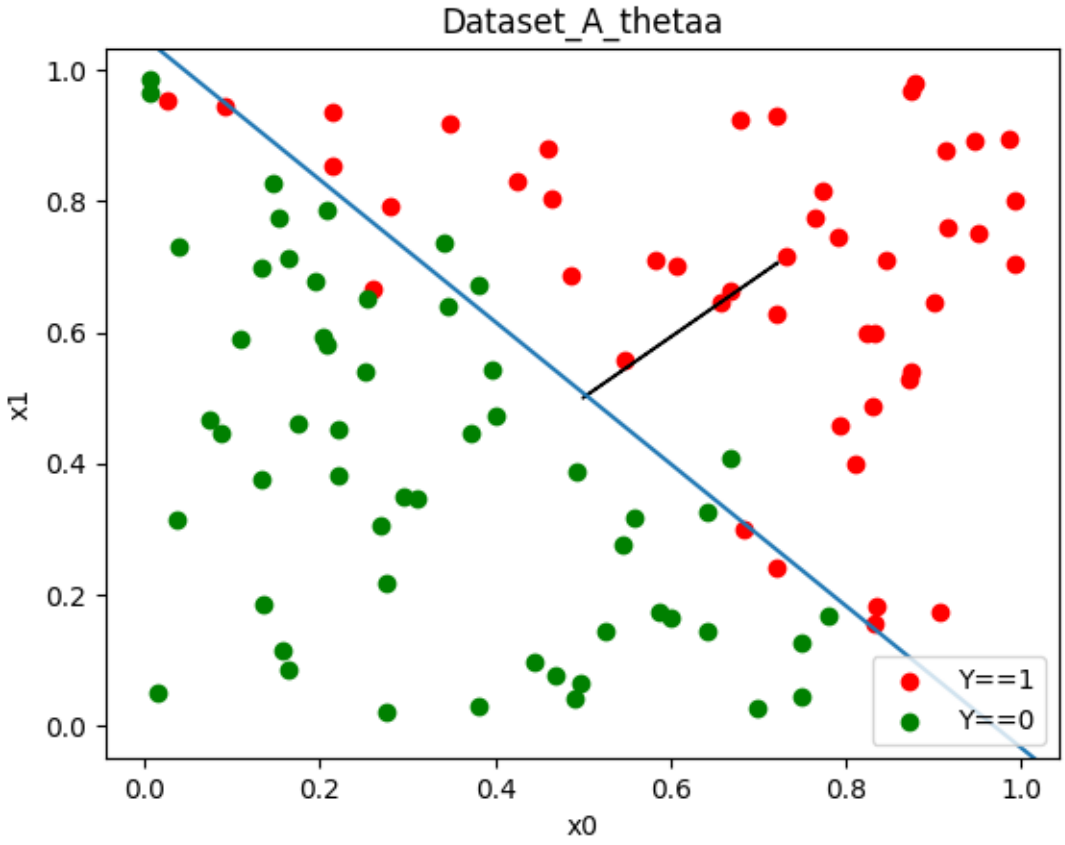
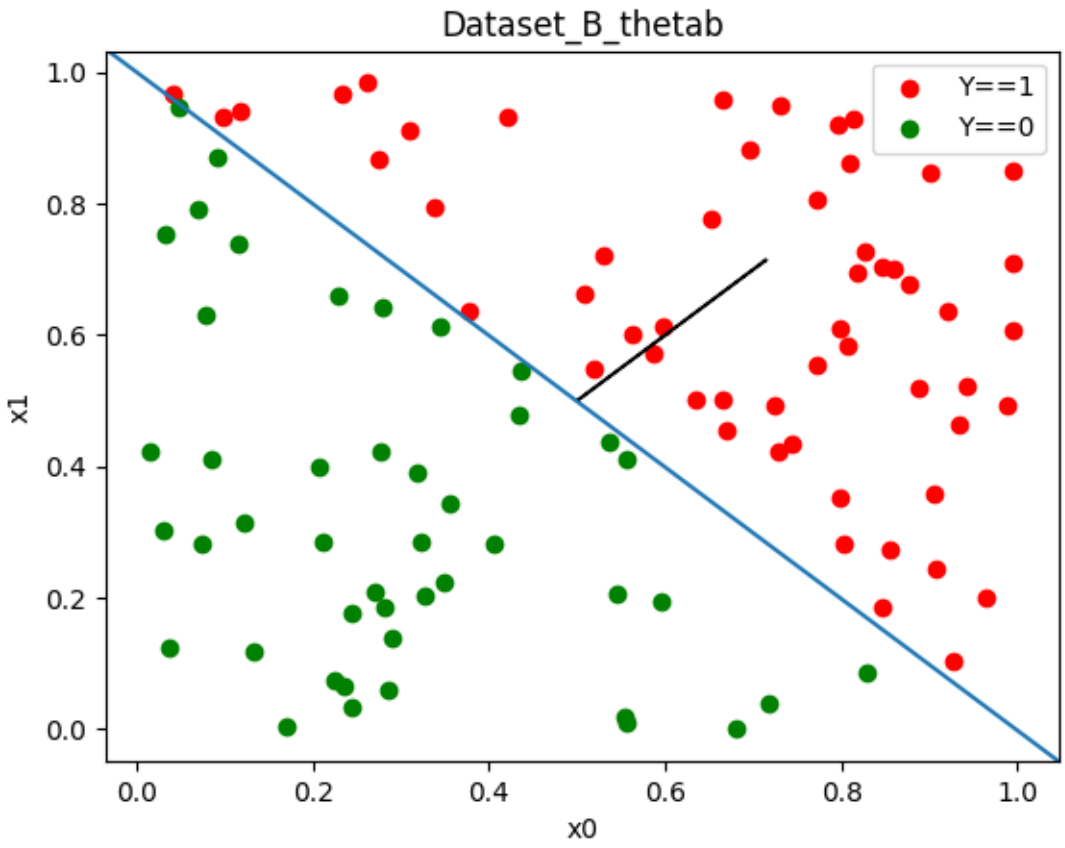
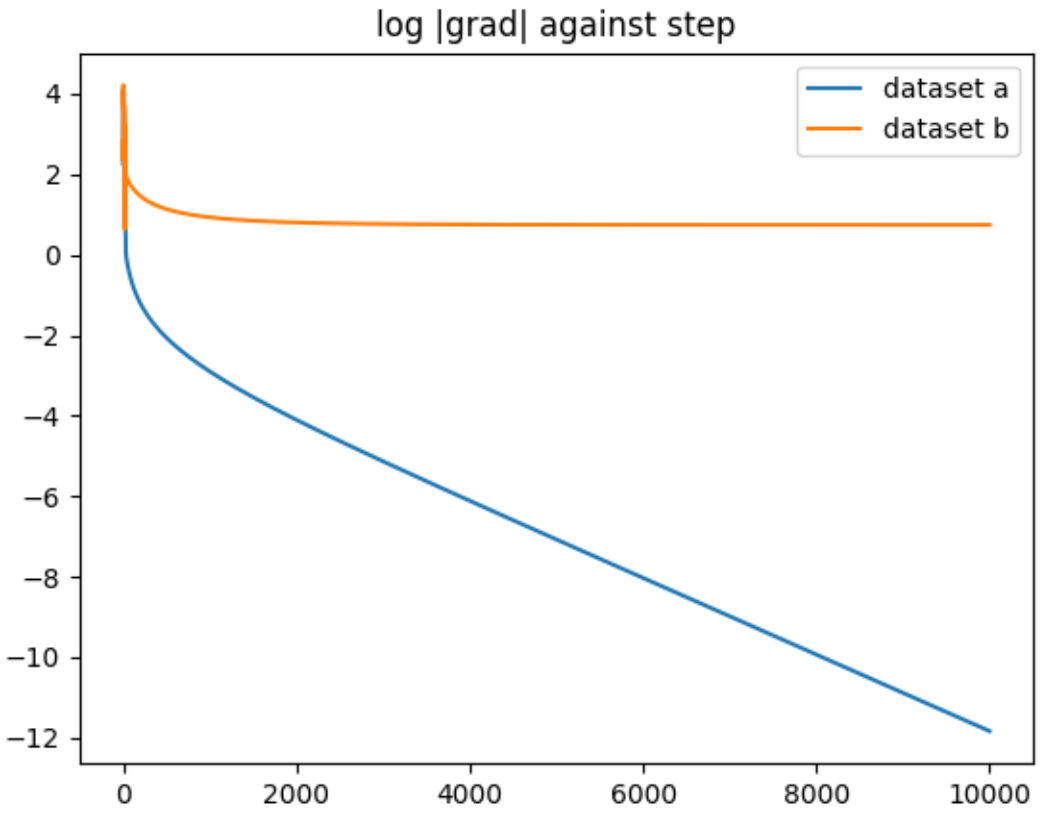
Answer:





We observe that dataset A seems less well differentiated than dataset B, the former has a few points mixing around the boundaries, whereas the latter can almost perfectly be separated by a dividing line. Otherwise the two are very comparable.

The classification line shown are derived many iterations on through gradient descent. For Dataset B we achieve perfect classification, whereas for Dataset A there are a few points that are misclassified, but this is unavoidable.



Our hypothesis is confirmed by this figure which shows the log size of the log likelihood gradient vector, for both they seem to eventually decrease exponentially quickly (linearly in log), but for B it's much shallower a slope than A. This would seem to be because perfect classification is possible with B, such that further improvements can only be made by increasing the magnitude of θ without changing direction. Indeed if $h_\theta(x) = g(z)$ then $h_{2\theta}(x) = g(2z)$. If classification is good, then $z >> 0$ or $z << 0$ s.t. writing $e^{-z} = y$, have $g(z) \approx 1 - y,\ g(2z) \approx 1 - y^2$ or $g(z) \approx y^{-1},\ g(2z) \approx y^{-2}$, i.e. the size $|y - g(z)| = \epsilon \mapsto \epsilon^2$.

Hence the grad step $\frac{\partial l}{\partial \theta} = \sum (y_i - h_\theta(x_i))x_i$ roughly gets smaller as we step. I.e. if $\theta \mapsto 2\theta$ then $\epsilon \mapsto \epsilon^2$. I.e. the size of $\frac{\partial l}{\partial \theta}$ is approximately proportional to ϵ^{2^k} where the parameter θ itself has size $2^k\theta$. The time of doubling, $n_k \propto \theta 2^k \epsilon^{-2^k}$, and so