10. (5 points) Mark each of the following statements as true or false, and provide a correct – and brief – explanation for the validity of each of your answers:

    (a) The purpose of `step_size_fn` is to allow step sizes to increase with iteration index `k`, so that `sgd` and `gd` can converge faster.

    ◯ True √ **False**

    > **Solution:** The purpose of `step_size_fn` is to allow step sizes to get **smaller**, so that `sgd` converges.

    (b) If $\theta =$ `th0` were luckily at a local **minimum** of the objective function, then the output of gradient descent `gd` would always be θ, independent of `num_steps`.

    √ **True** ◯ False

    > **Solution:** At a local minimum, `dJ(th)` $= dJ/d\theta = 0$, so for `gd` the update rule `th = th - step_size_fn(k) * dJ(th)` would reduce to `th = th`

    (c) If $\theta =$ `th0` were luckily at a local **maximum** of the objective function, then the output of gradient descent `gd` would always be θ, independent of `num_steps`.

    √ **True** ◯ False

    > **Solution:** At a local maximum, `dJ(th)` $= dJ/d\theta = 0$, so for `gd` the update rule `th = th - step_size_fn(k) * dJ(th)` would reduce to `th = th`

    (d) If $\theta =$ `th0` were luckily at a local **minimum** of the objective function, then the output of stochastic gradient descent `sgd` would always be θ, independent of `num_steps`.

    ◯ True √ **False**

    > **Solution:** For SGD, the update rule `th = th - step_size_fn(k) * dJ(Xj, yj, th)` means that the gradient is dependent on which datapoint is chosen. In general, this gradient will be nonzero for some datapoints, even if the gradient is zero when averaged over all datapoints.

    (e) The gradient produced by `dJ` is a d-dimensional vector which points in the direction which maximizes the objective function.

    √ **True** ◯ False

    > **Solution:** $dJ/d\theta$ points θ in the direction of increasing J.