Name: _____

All Greek to me!

2. (8 points) Let's consider solving a ridge regression problem using stochastic gradient descent. For simplicity, we will ignore the offset. Our hypothesis has the form

$$h(x;\theta) = \theta^T x \ ;$$

our objective function has the form

$$J(\theta) = \left(\frac{1}{n} \sum_{i=1}^{n} \left(h(x^{(i)};\theta) - y^{(i)} \right)^2 \right) + \lambda \|\theta\|^2 \ ;$$

and we will do T steps of gradient descent using a rule of the form

$$\theta = \theta - \eta \nabla_\theta J(\theta) \ ,$$

where η has a fixed value throughout the execution.

What is with all these Greek letters!? Each of θ, λ, and η has a role in what happens.

In the following questions, mark all answers that apply.

(a) Which parameter(s) would be included when using the hypothesis to make predictions?
√ θ ◯ λ ◯ η ◯ none

(b) Which parameter(s) are primarily intended to improve generalization?
◯ θ √ λ ◯ η ◯ none

(c) Can T play a similar role to λ ?
√ **yes** ◯ no
Explain your answer.

> **Solution:** By stopping the optimization early, we force the algorithm to use less of the training data, hence we can prevent it from being overly specialized (overfit) to the training data.

(d) Can η play a similar role to λ ?
√ **yes** ◯ no
Explain your answer.

> **Solution:** Using a very small step size has a behavior somewhat like stopping early; using a large step size (or not decaying it appropriately) will also not allow the optimization to go into a "narrow valley," which might also help to prevent overfitting.