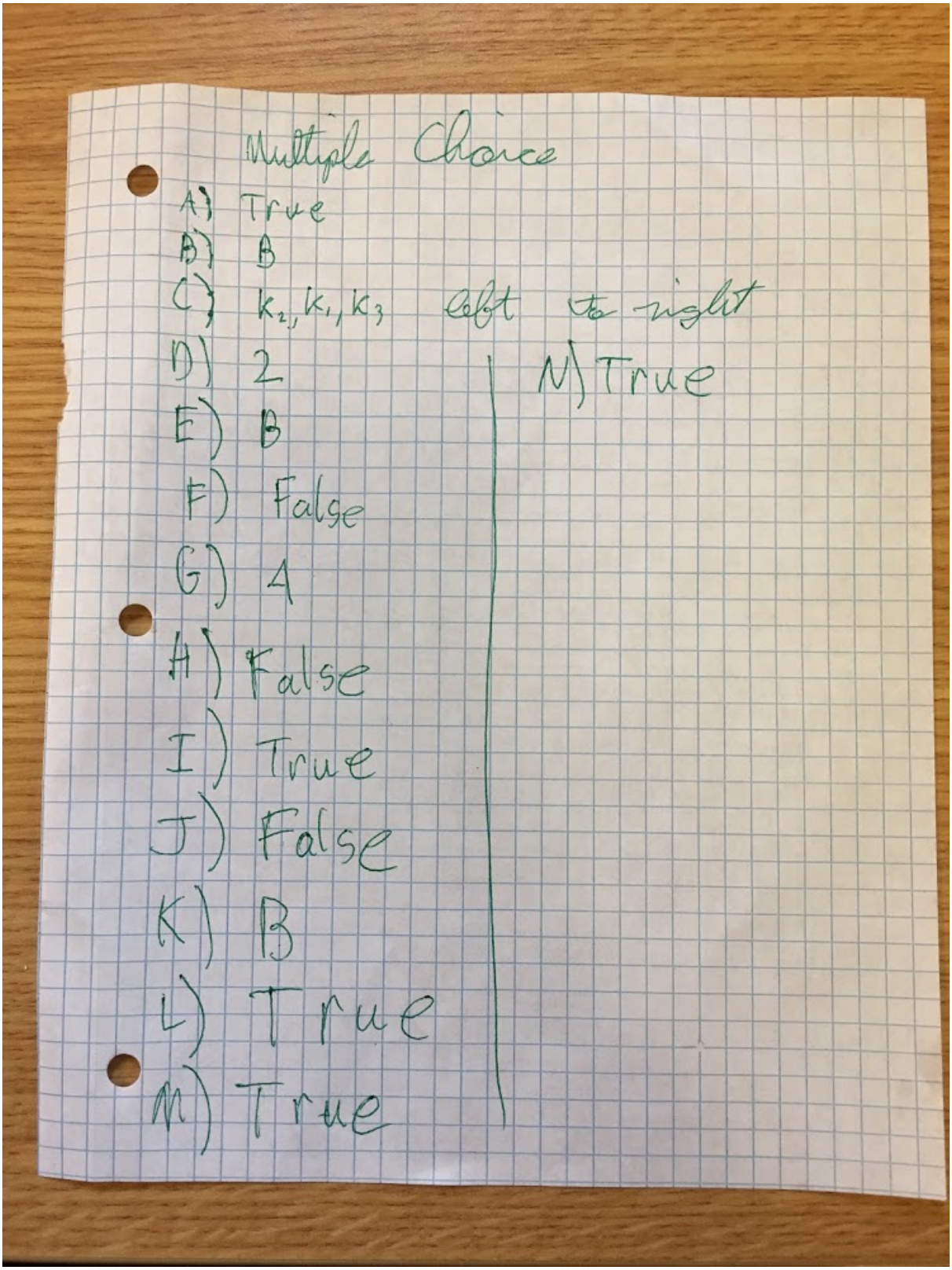
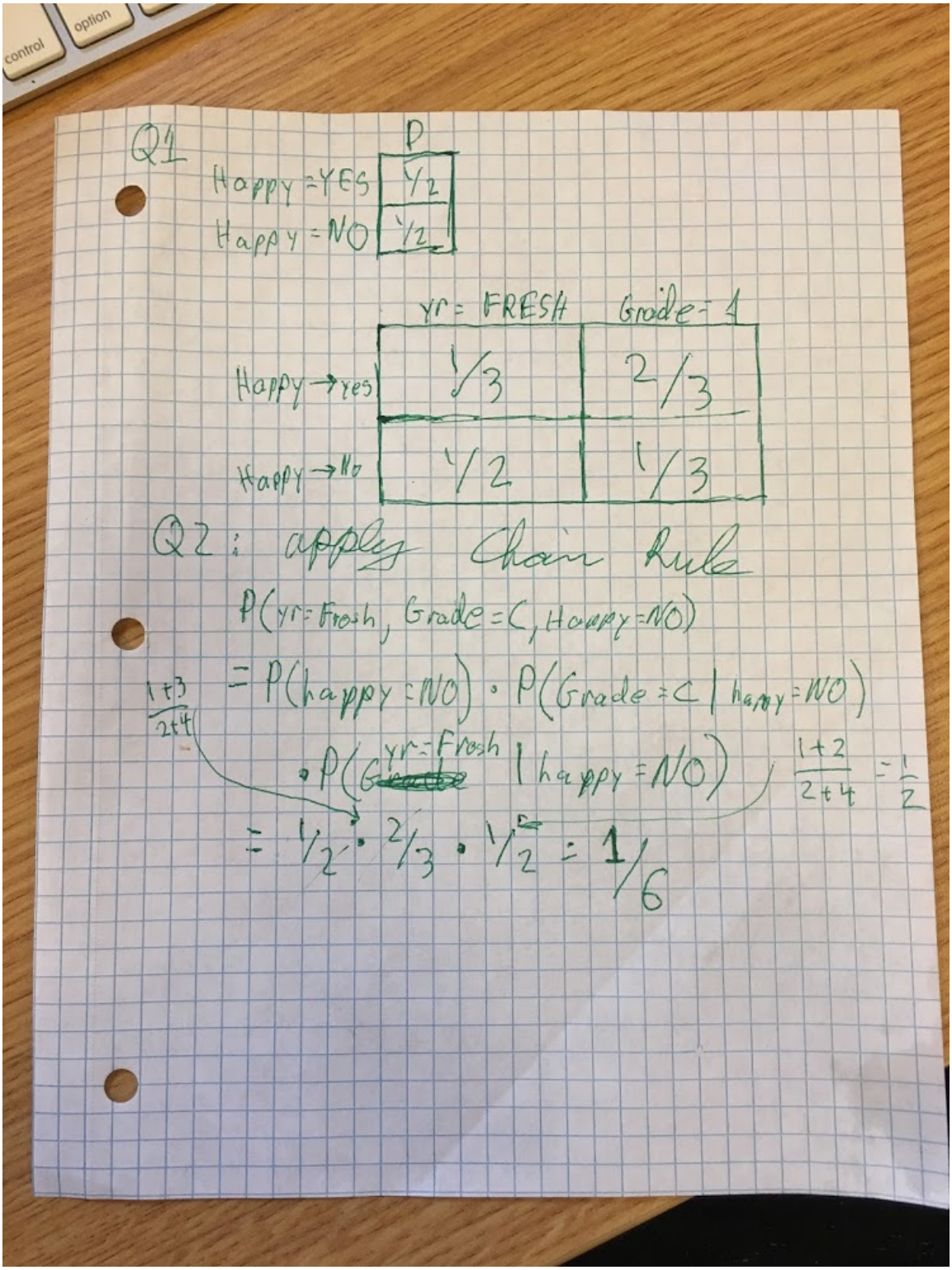
Eli Pinkus
155 Final

1 Multiple Choice:



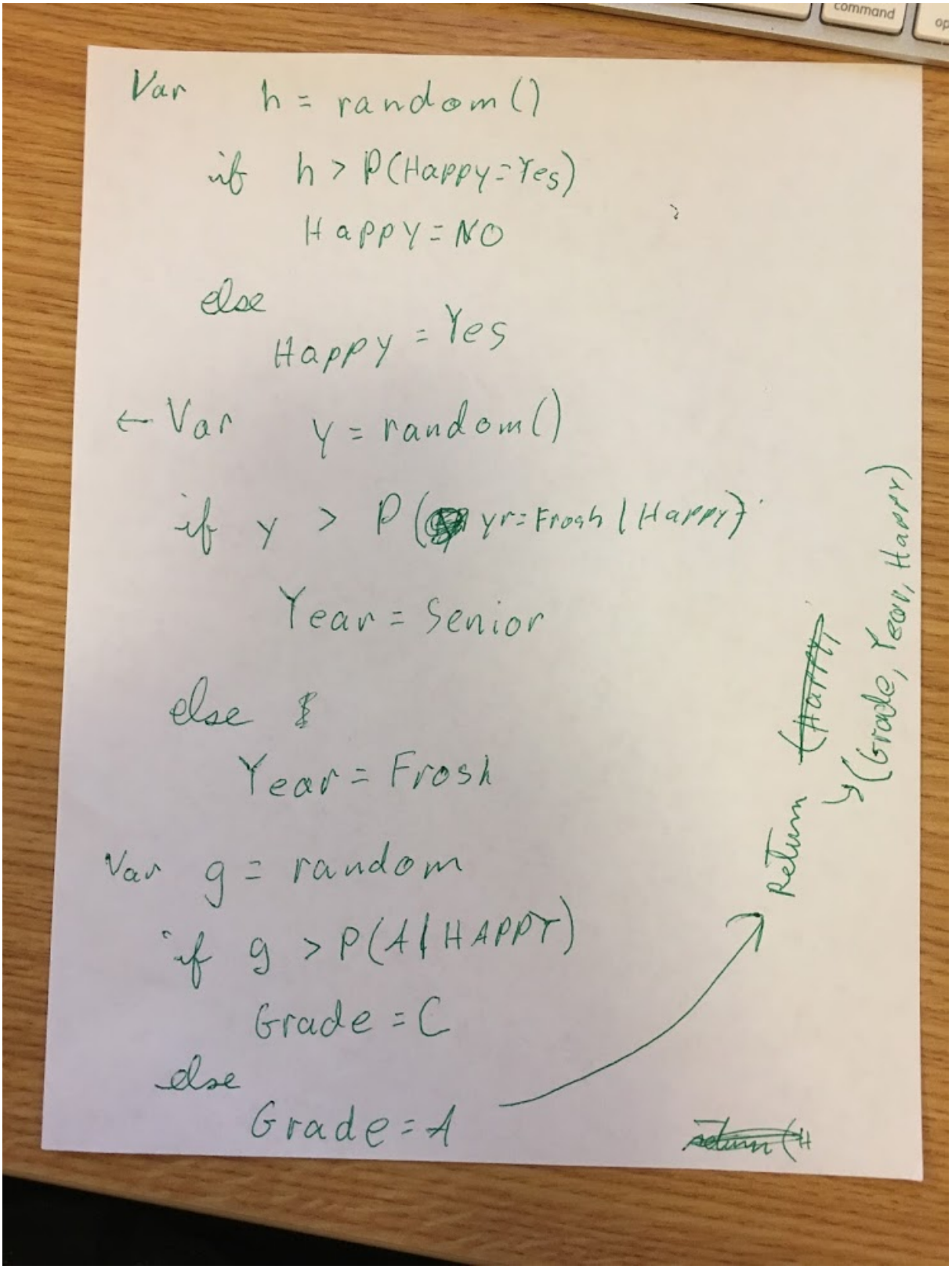
Eli Pinkus
155 Final

2 Naïve Bayes



Eli Pinkus
155 Final

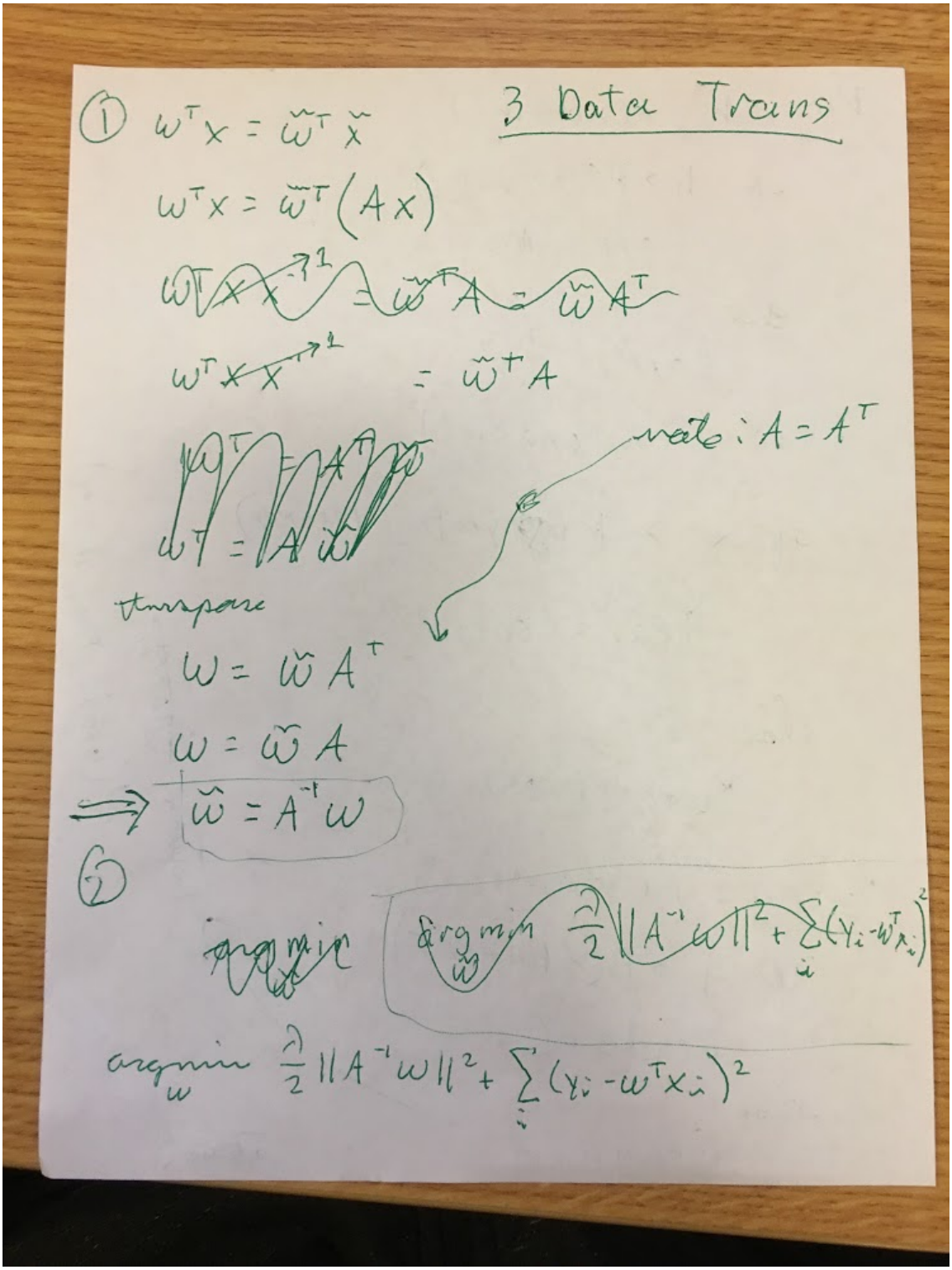
Part 3



Eli Pinkus
155 Final

3 Data Transformation

Part 1/2



Eli Pinkus
155 Final

Part 3

We know that our answer from part b

$$\arg\min_{w} \frac{\lambda}{2}\left|\left|A^{-1}w\right|\right|^2 + \sum_i (y_i - w^T x_i)^2$$

$$= A\left[\arg\min_{\tilde{w}} \frac{\lambda}{2}\left|\left|\tilde{w}\right|\right|^2 + \sum_i (y_i - \tilde{w}^T \tilde{x}_i)^2\right]$$

What is effectively going on is that the data is transformed into a new rescaled space and then ridge regression is performed and then the data is transformed back. Since the transformation in question is simply a rescaling, all that transforming really means is multiplying by a diagonal matrix A. Because of this, the model parameter, w can either be optimized directly by training the first objective or $\tilde{w}$ can be optimized and then transformed after the fact. The point is basically that transforming in this case is a well behaved 1 to 1 linear scaling transform within the same space that doesn't really change the result of the optimizing problem. It is possible that such a scaling could help avoid numerical issues in practice, however, apart from that, this transformed model class is exactly equally expressive as regular ridge regression and the model class isn't changed in any functional way.

Eli Pinkus
155 Final

4 Latent Markov Embedding

Question 1

We begin by noting that the dual point model is no less expressive than the single point model.

We can see this simply by noting that dual point is a generalization of single point or single point is a special case of dual point (the thinking is equivalent).

Consider some matrix X as in a single point model optimum. We can see that a dual point model with $U = V = X$ is the same model as the single point model and therefore the dual point model as a whole can't have a $P(S)$ less than that of the single point model.

Worst case the two are just the same, otherwise dual point will train to a greater maximum. Thus, we have shown $P(S)$ for the dual-point model is *never less* than $P(S)$ for the single point model.

Question 2

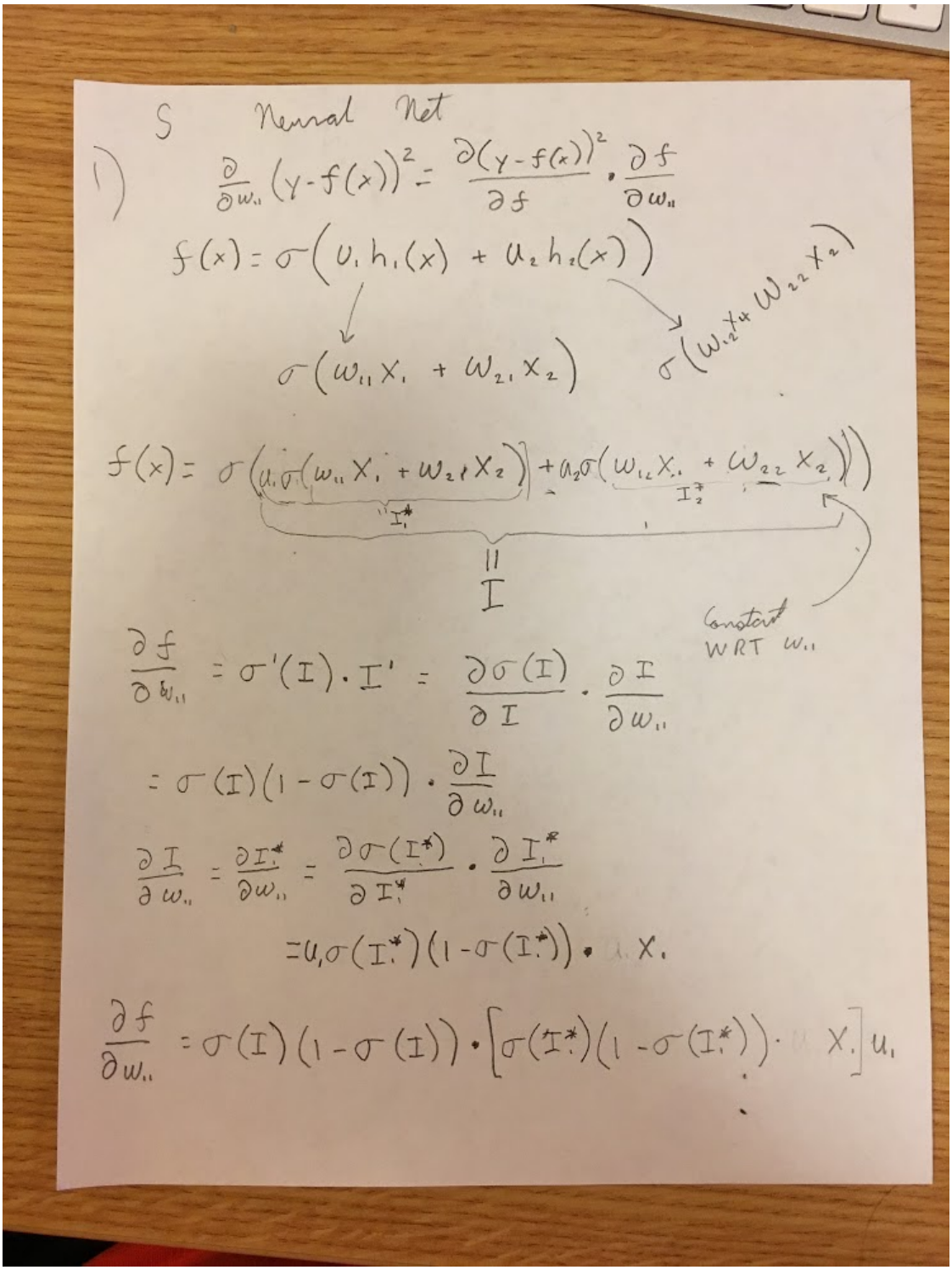
This would imply that we are dealing with the case in which single point and dual point result in the same model

$$\Rightarrow U = V = X$$

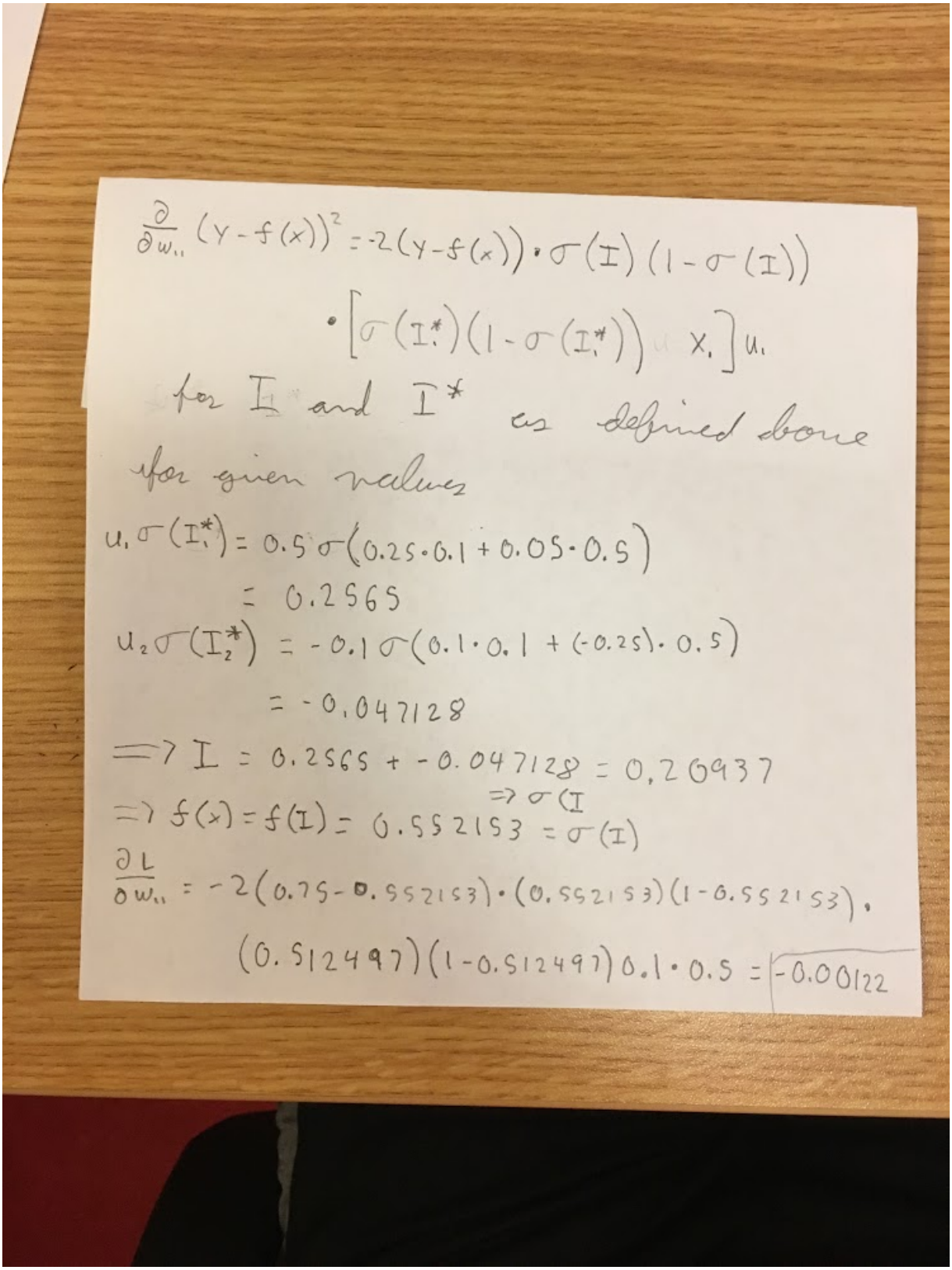
Eli Pinkus
155 Final

5 Neural Net

Parts 1 and 2



Eli Pinkus
155 Final



$$\frac{\partial}{\partial w_{11}}(y - f(x))^2 = -2(y - f(x)) \cdot \sigma(I)(1 - \sigma(I)) \cdot \left[\sigma(I_1^*)(1 - \sigma(I_1^*))\, x_1\right] u_1$$

for I and I^* as defined above for given values

$$u_1 \sigma(I_1^*) = 0.5\,\sigma(0.25 \cdot 0.1 + 0.05 \cdot 0.5) = 0.2565$$

$$u_2 \sigma(I_2^*) = -0.1\,\sigma(0.1 \cdot 0.1 + (-0.25) \cdot 0.5) = -0.047128$$

$$\Rightarrow I = 0.2565 + -0.047128 = 0.20937$$

$$\Rightarrow f(x) = f(I) = 0.552153 = \sigma(I) \quad (\Rightarrow \sigma(I))$$

$$\frac{\partial L}{\partial w_{11}} = -2(0.75 - 0.552153) \cdot (0.552153)(1 - 0.552153) \cdot (0.512497)(1 - 0.512497)\, 0.1 \cdot 0.5 = -0.00122$$

Eli Pinkus
155 Final

Part 3

We have:

$$\frac{\partial}{\partial w_{11}} L\big(y, f(x)\big) = \frac{\partial}{\partial w_{11}} \big(y - f(x)\big)^2 = \frac{\partial}{\partial f} \big(y - f(x)\big)^2 \cdot \frac{\partial f}{\partial w_{11}}$$

We can clearly see that the term $\frac{\partial f}{\partial w_{11}}$ is the culprit behind vanishing gradients. Calculating this gradient involves calculating, and multiplying by through chain rule, gradients of composite functions from all previous layers, so we have a lot of multiplied terms that can cause the result gradient to vanish. Additionally, we are using

$$\sigma(s) = \frac{e^s}{1 + e^s}$$

as our non-linear activation function

and we have that $\max_s \sigma'(s) = \frac{1}{4} < 1$, so in the process of calculating $\frac{\partial f}{\partial w_{11}}$, we must also calculate and multiply by many such small values which is exacerbated by having more layers and thus more derivatives of activation functions.