Exact Phase Transitions in TMS

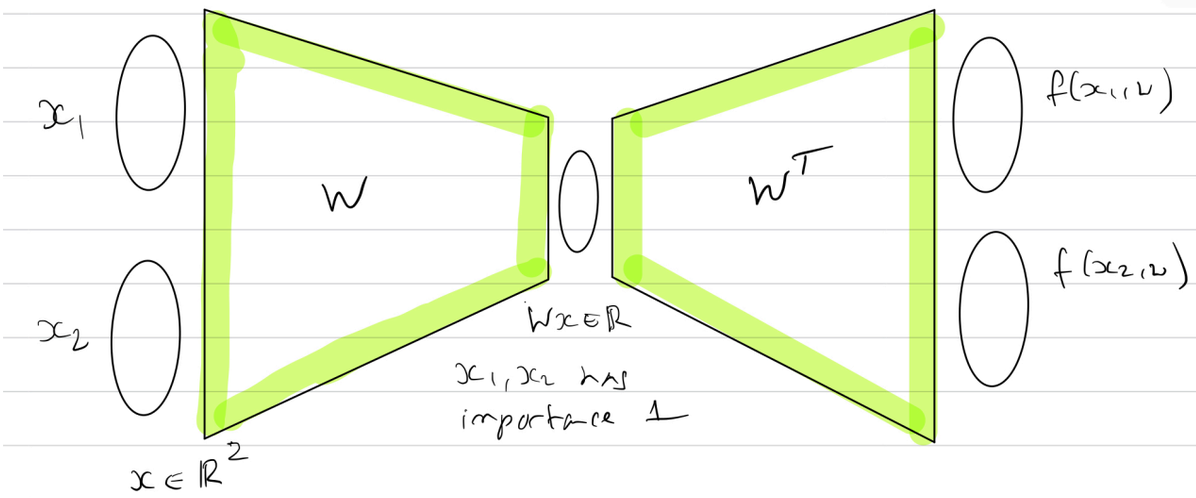
Benjamin Gerraty

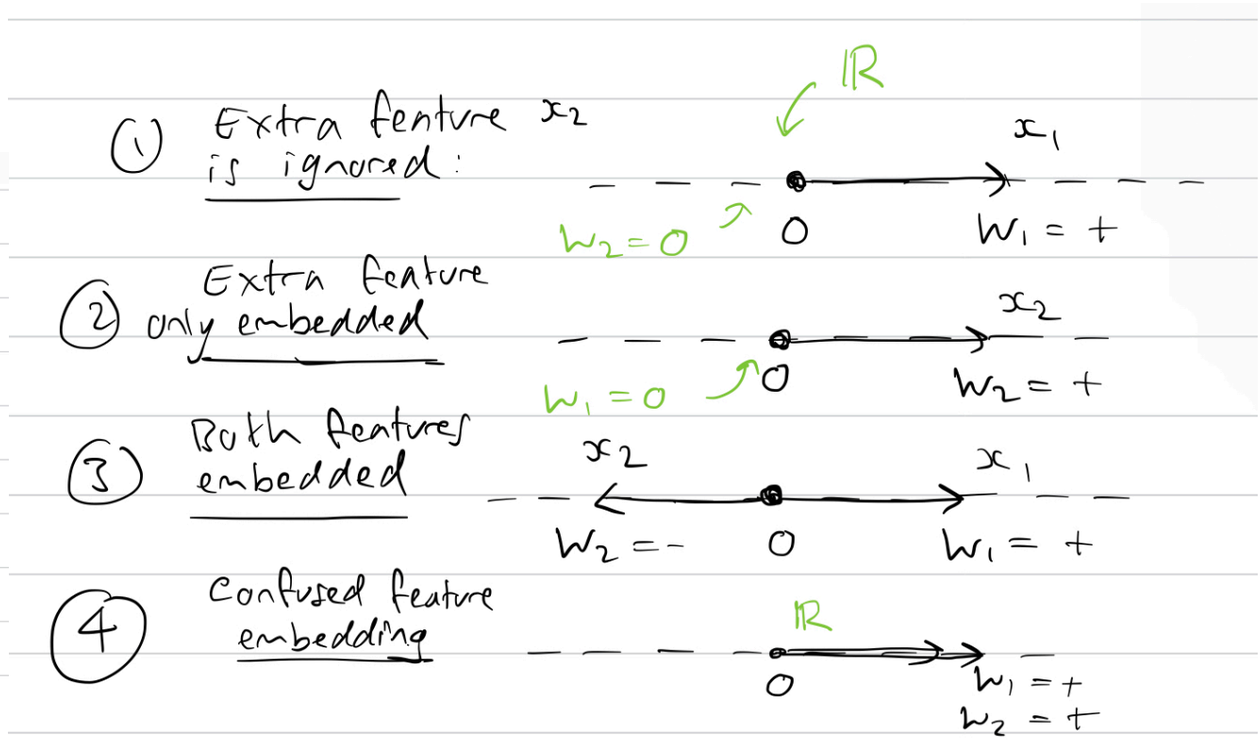
SLT for Alignment Conference
26/06/2023

Set Up

- This short talk will be a presentation of the exact loss landscape for 2 features in 1 dimension, demonstrating the first order phase transition where we see the exact sparsity value the model decides to embed more features than it has dimensions

- Let $f(x, w) = \text{ReLU}(W^T W x)$ be our biasless model. Consider $n = 2$ features $x = (x_1, x_2) \in \mathbb{R}^2$ embedded into $m = 1$ dimensions. The parameter $w \in \mathbb{R}^{1\times 2}$ is $w = [w_1 \quad w_2]$ and importance will be constant so $I = 1$ for both features. We will have the following 4 cases below





Sparsity and Loss

- A feature $x_i \in (x_1, x_2)$ will be set to zero with probability s and otherwise sampled from the uniform distribution $x_i \sim \mathrm{U}[0,1]$. In other words $x_i \sim q(x, s)$ where for $T \subseteq \{1,2\}$ we have

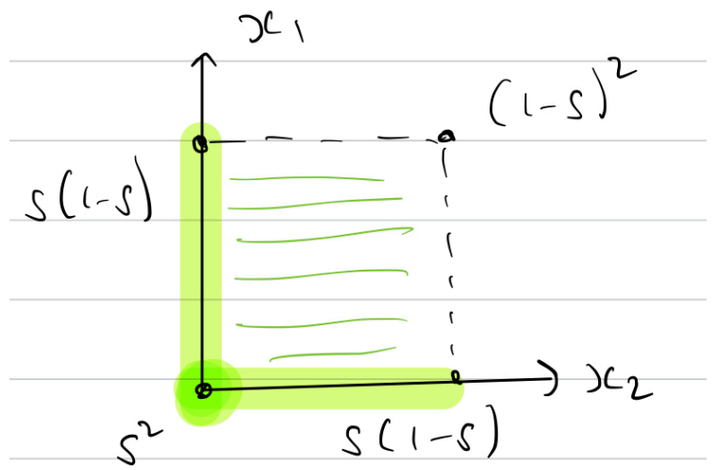
$$q(x, s) = s^2\delta(x_1 = 0, x_2 = 0) + s(1-s)\delta(x_1 = 0) + s(1-s)\delta(x_2 = 0) + (1-s)^2$$

- The *Toy Models Potential* as defined in SLT High 4 is

$$L(w) = \int_{\mathbb{R}^2} q(x, s)\,||x - \mathbf{ReLU}(W^T W x)||^2 dx$$

$$= \int_{\mathbb{R}^2} q(x, s) g(x_1, x_2, w) dx$$

Where $g(x_1, x_2, w) = ||x - \mathrm{ReLU}(W^T W x)||^2$



Lets do the Integral

- Expanding our function $g(x_1, x_2, w)$ we obtain

$$g(x_1, x_2, w) = ||x - \textbf{ReLU}(W^T W x)||^2$$

$$= \left|\left| \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \textbf{ReLU}\left(\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \begin{pmatrix} w_1 & w_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right) \right|\right|^2$$

$$= \left|\left| \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \textbf{ReLU}\begin{pmatrix} w_1^2 x_1 + w_1 w_2 x_2 \\ w_1 w_2 x_1 + w_2^2 x_2 \end{pmatrix} \right|\right|^2$$

- Now our loss function will decompose into four components

$$L(w) = \int_{\mathbb{R}^2} q(x, s) g(x_1, x_2, w) dx$$

$$= s^2 \int g(0,0,w) dx + s(1-s) \int_0^1 g(x_1,0,w) dx_1$$

$$+ (1-s)s \int_0^1 g(0,x_2,w) dx_2 + (1-s)^2 \int_0^1 \int_0^1 g(x_1, x_2, w) dx_1 dx_2$$

Lets do more of the Integral

- Expanding our function $g(x_1, x_2, w)$ we obtain

$$L(w) = s(1-s)\int_0^1 \left|\left| \begin{pmatrix} x_1 \\ 0 \end{pmatrix} - \textbf{ReLU}\begin{pmatrix} w_1^2 x_1 \\ w_1 w_2 x_1 \end{pmatrix} \right|\right|^2 dx_1$$

$$+s(1-s)\int_0^1 \left|\left| \begin{pmatrix} 0 \\ x_2 \end{pmatrix} - \textbf{ReLU}\begin{pmatrix} w_1 w_2 x_2 \\ w_2^2 x_2 \end{pmatrix} \right|\right|^2 dx_2$$

$$+(1-s)^2\int_0^1\int_0^1 \left|\left| \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \textbf{ReLU}\begin{pmatrix} w_1^2 x_1 + w_1 w_2 x_2 \\ w_1 w_2 x_1 + w_2^2 x_2 \end{pmatrix} \right|\right|^2 dx_1 dx_2$$

- Assuming both w_1 and w_2 have the same sign everything inside the ReLU's is positive so the loss is

$$L(w) = \frac{s(1-s)}{3}\left(w_1^4 + 2w_1^2 w_2^2 - 2w_1^2 + w_2^4 - 2w_2^2 + 2\right)$$
$$+\frac{(1-s)^2}{6}\left(2w_1^4 + 3w_1^3 w_2 + 4w_1^2\left(w_2^2 - 1\right) + 3w_1 w_2\left(w_2^2 - 2\right) + 2\left(w_2^4 - 2w_2^2 + 2\right)\right)$$

Critical Points Part 1

- Taking the partial derivatives of the loss and solving $\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial w_2} = 0$ we obtain the critical points

$$(w_1, w_2) = (0,0), \quad (w_1, w_2) = \pm\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$$

- The Hessians of both critical points are

$$\left.\begin{pmatrix} \frac{\partial^2 L}{\partial w_1^2} & \frac{\partial^2 L}{\partial w_1 \partial w_2} \\ \frac{\partial^2 L}{\partial w_2 \partial w_1} & \frac{\partial^2 L}{\partial w_2^2} \end{pmatrix}\right|_{(0,0)} = \begin{pmatrix} \frac{4}{3}(s-1) & -(s-1)^2 \\ -(s-1)^2 & \frac{4}{3}(s-1) \end{pmatrix}$$

$$\lambda_+ = \frac{1}{3}(s-1), \quad \lambda_- = \frac{7}{4}(s-1)$$

Both eigenvalues negative for all s, local maximum

$$\left.\begin{pmatrix} \frac{\partial^2 L}{\partial w_1^2} & \frac{\partial^2 L}{\partial w_1 \partial w_2} \\ \frac{\partial^2 L}{\partial w_2 \partial w_1} & \frac{\partial^2 L}{\partial w_2^2} \end{pmatrix}\right|_{(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})} = \frac{(s-1)}{6}\begin{pmatrix} (-17+9s) & (-11+3s) \\ (-11+3s) & (-17+9s) \end{pmatrix}$$

$$\lambda_+ = (s-1)^2, \quad \lambda_- = \frac{2}{3}(7-3s)(1-s)$$

Both eigenvalues positive for all s, local minimum

Lets just do the Integral again

- When w_1 and w_2 have mixed signs then the $s(1-s)$ terms immediately simplify and the third term finally gives us a case for the ReLU

$$L(w) = s(1-s)\int_0^1 \left|\left|\begin{pmatrix} x_1 \\ 0 \end{pmatrix} - \textbf{ReLU}\begin{pmatrix} w_1^2 x_1 \\ 0 \end{pmatrix}\right|\right|^2 dx_1 + s(1-s)\int_0^1 \left|\left|\begin{pmatrix} 0 \\ x_2 \end{pmatrix} - \textbf{ReLU}\begin{pmatrix} 0 \\ w_2^2 x_2 \end{pmatrix}\right|\right|^2 dx_2$$

$$+(1-s)^2\int_0^1\int_0^1 \left|\left|\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \textbf{ReLU}\begin{pmatrix} w_1^2 x_1 + w_1 w_2 x_2 \\ w_1 w_2 x_1 + w_2^2 x_2 \end{pmatrix}\right|\right|^2 dx_1 dx_2$$

$$= \frac{s(1-s)}{3}\left((w_2^2-1)^2 + (w_1^2-1)^2\right)$$

$$+(1-s)^2\int_0^1\int_0^1 \left(x_1 - \textbf{ReLU}(w_1^2 x_1 + w_1 w_2 x_2)\right)^2 + \left(x_2 - \textbf{ReLU}(w_2^2 x_2 + w_1 w_2 x_1)\right)^2 dx_1 dx_2$$

Lets define $h(x_1, x_2, w_1, w_2) = \int_0^1\int_0^1 \left(x_1 - \text{ReLU}(w_1^2 x_1 + w_1 w_2 x_2)\right)^2 dx_1 dx_2$ so the $(1-s)^2$ term is just $h(x_1, x_2, w_1, w_2) + h(x_2, x_1, w_2, w_1)$

Lets just do the Integral again

- Now the ReLU inside $h(x_1, x_2, w_1, w_2)$ is positive when $x_1 \geq \frac{w_2}{w_1} x_2$, being careful with the cases we arrive at for $w_1 < 0$ and $w_2 > 0$

$$h(x_1, x_2, w_1, w_2) = \begin{cases} \frac{3w_1^3 - w_1^5 + 4w_2}{12w_2} & , w_1 + w_2 > 0 \\ \frac{4w_1 + 4w_1^5 + 6w_1^4 w_2 + w_2^3 + w_1^2 w_2(w_2^2 - 6) + 4w_1^3(w_2^2 - 2)}{12w_1} & , w_1 + w_2 \leq 0 \end{cases}$$

- Our final loss in this region is

$$L(w) = \frac{s(1-s)}{3}\left((w_2^2 - 1)^2 + (w_1^2 - 1)^2\right)$$

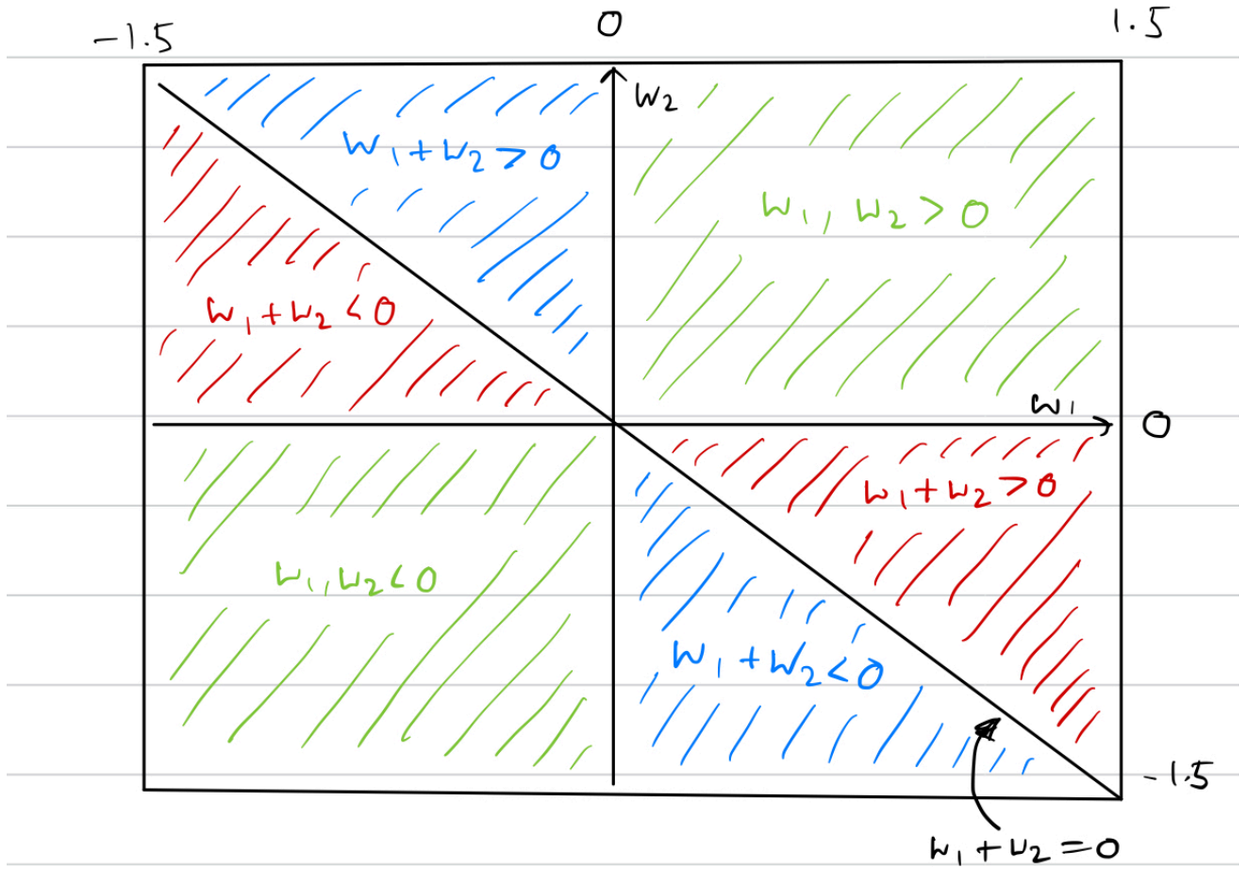
$$+(1-s)^2 \begin{cases} \frac{4w_1^5 + 6w_1^4 w_2 + 4w_1^3(w_2^2 - 2) + w_1^2 w_2(w_2^2 - 6) + 8w_1 - w_2^5 + 4w_2^3}{12w_1} & , w_1 + w_2 \leq 0 \\ \frac{4(w_1^2 - 2)w_2^3 + w_1(w_1^2 - 6)w_2^2 - w_1^3(w_1^2 - 4) + 6w_1 w_2^4 + 4w_2^5 + 8w_2}{12w_2} & , w_1 + w_2 > 0 \end{cases}$$

Lets just do the Integral one again again

- The loss for the region $w_1 > 0,\ w_2 < 0$ is the same with regions flipped

$$L(w) = \frac{s(1-s)}{3}\left((w_2^2-1)^2 + (w_1^2-1)^2\right) \quad +(1-s)^2 \begin{cases} \dfrac{4w_1^5 + 6w_1^4 w_2 + 4w_1^3(w_2^2-2) + w_1^2 w_2(w_2^2-6) + 8w_1 - w_2^5 + 4w_2^3}{12w_1} & , w_1 + w_2 \geq 0 \\ \dfrac{4(w_1^2-2)w_2^3 + w_1(w_1^2-6)w_2^2 - w_1^3(w_1^2-4) + 6w_1 w_2^4 + 4w_2^5 + 8w_2}{12w_2} & , w_1 + w_2 < 0 \end{cases}$$

- Why do the mixed signs matter now?



Critical Points Part 2

- The critical points only occur on the line $w_1 + w_2 = 0$ which greatly simplifies solving $\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial w_2} = 0$ giving the critical points

$$(w_1, w_2) = \left(\pm \frac{1}{\sqrt{2}} \sqrt{\frac{3+5s}{1+3s}}, \mp \frac{1}{\sqrt{2}} \sqrt{\frac{3+5s}{1+3s}} \right)$$

- The Hessians of the critical points are

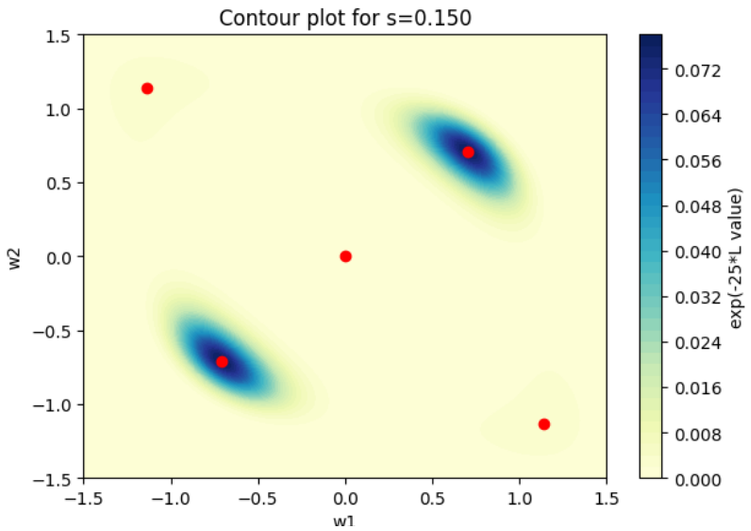
$$\left(\begin{array}{cc} \frac{\partial^2 L}{\partial w_1^2} & \frac{\partial^2 L}{\partial w_1 \partial w_2} \\ \frac{\partial^2 L}{\partial w_2 \partial w_1} & \frac{\partial^2 L}{\partial w_2^2} \end{array} \right) \Bigg|_{\left(-\frac{1}{\sqrt{2}} \sqrt{\frac{3+5s}{1+3s}}, \frac{1}{\sqrt{2}} \sqrt{\frac{3+5s}{1+3s}} \right)}$$

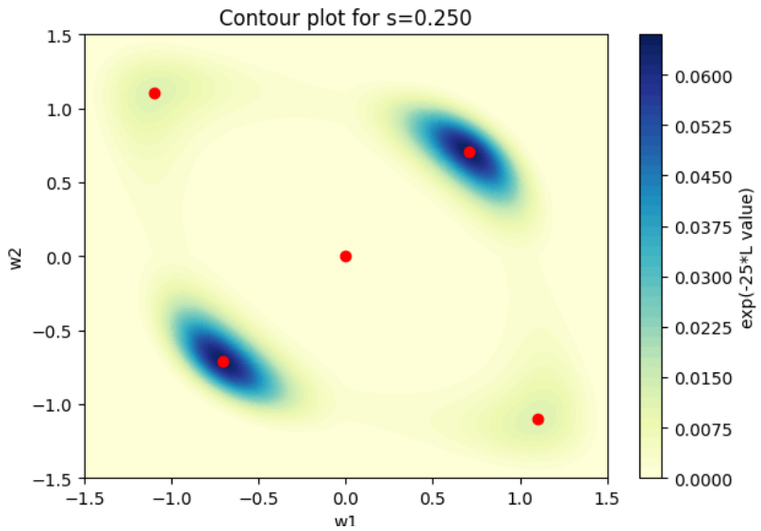
$$= \left(\begin{array}{cc} -\frac{(s-1)(s(89s+30)+9)}{36s+12} & -\frac{(s-1)^2(29s+3)}{36s+12} \\ -\frac{(s-1)^2(29s+3)}{36s+12} & -\frac{(s-1)(s(89s+30)+9)}{36s+12} \end{array} \right)$$

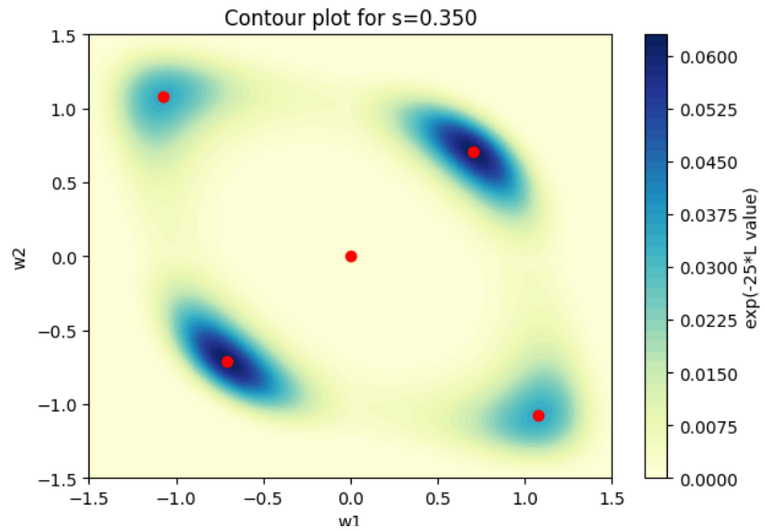
$$\lambda_+ = \frac{3 - s + 57s^2 - 59s^3}{6 + 18s}, \quad \lambda_- = \frac{1}{3}(3 + 2s - 5s^2)$$

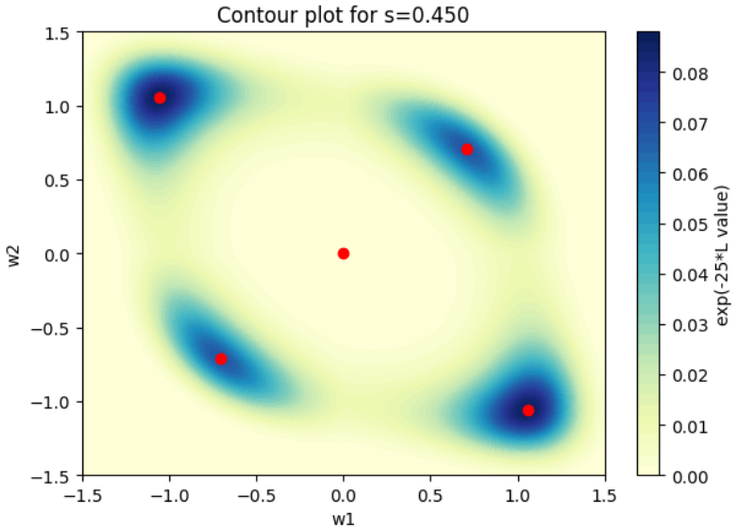
Both eigenvalues positive for all s, local minimum

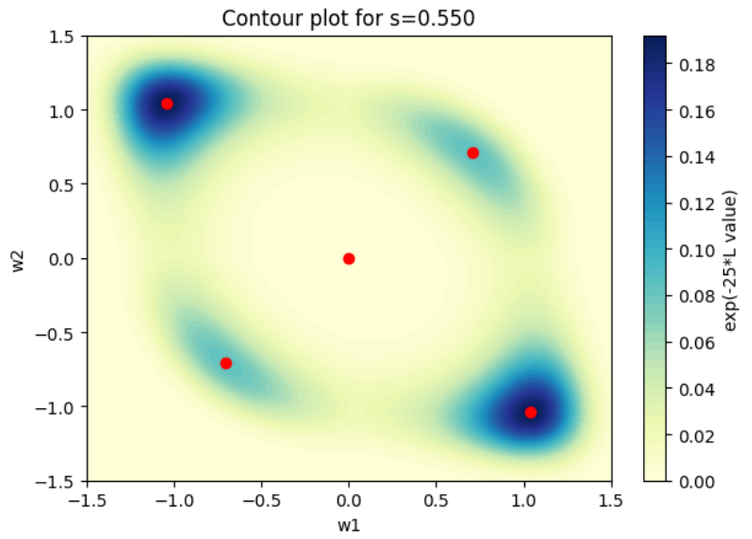
Contour Plots

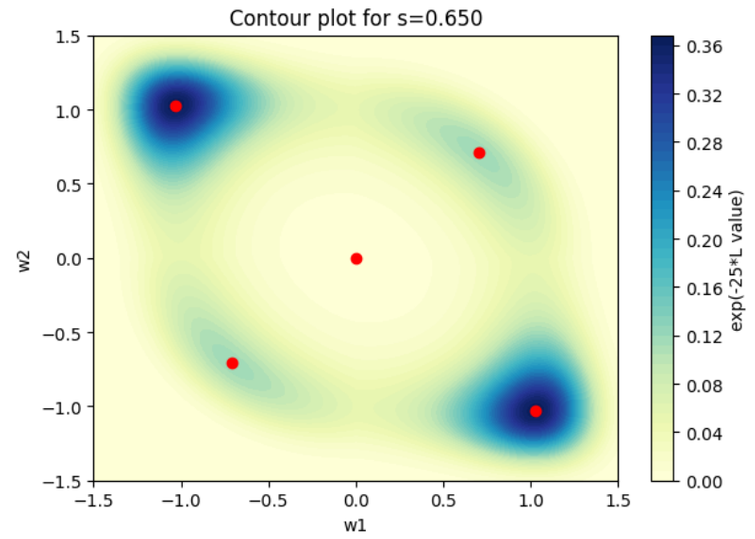


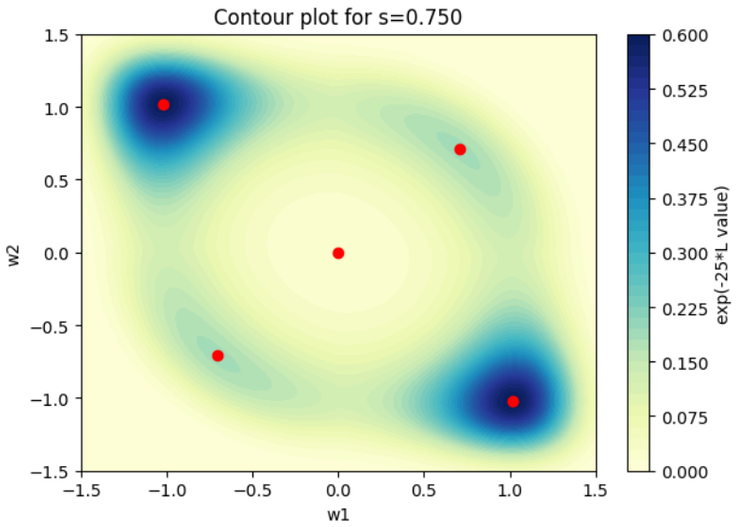


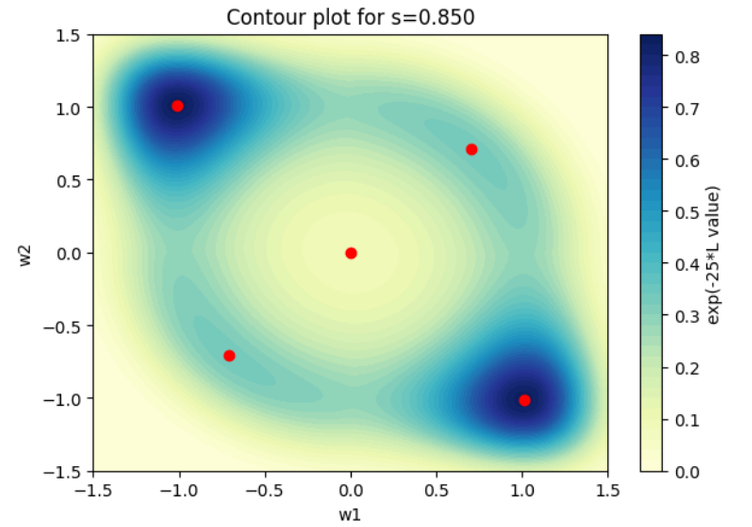


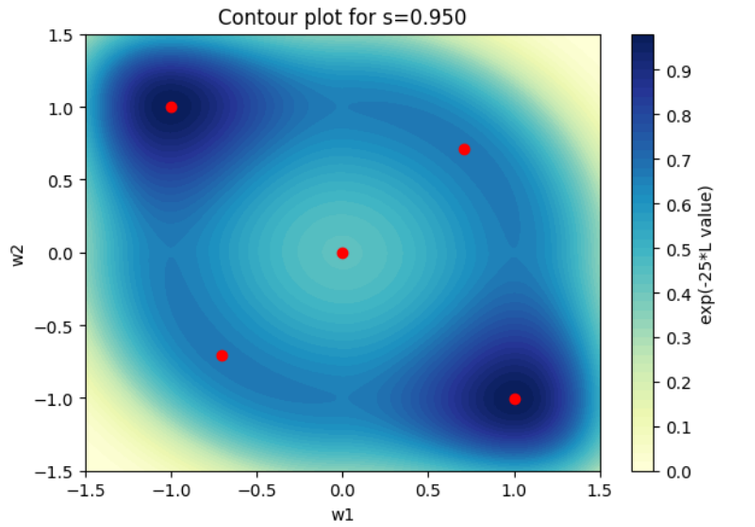




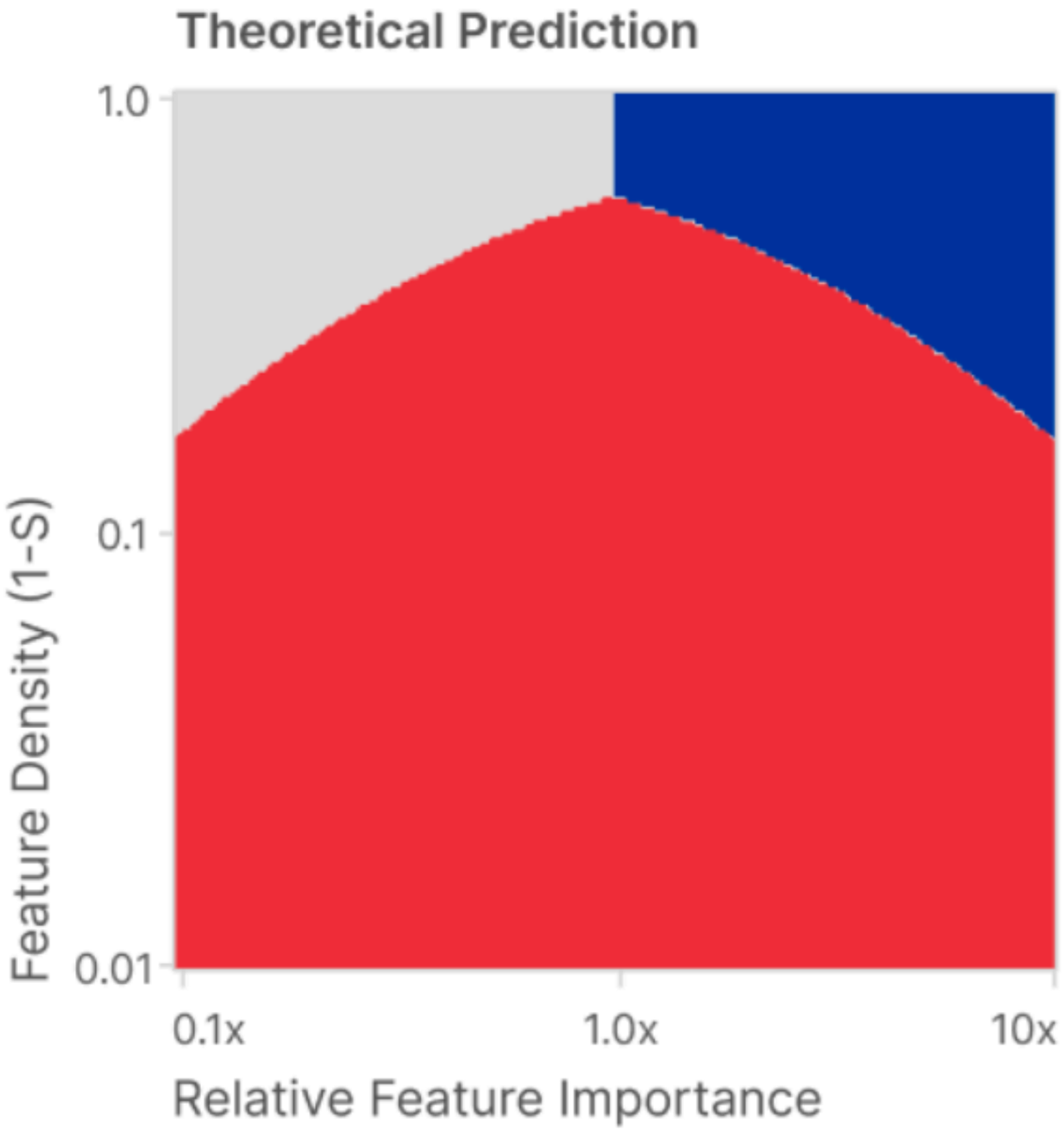








Something we missed last time



Not Represented
(Extra Feature is 0)

$W = \begin{bmatrix} 1 & 0 \end{bmatrix}$
$W \perp \begin{bmatrix} 0 & 1 \end{bmatrix}$

Dedicated Dimension
(Other Not Represented)

$W = \begin{bmatrix} 0 & 1 \end{bmatrix}$
$W \perp \begin{bmatrix} 1 & 0 \end{bmatrix}$

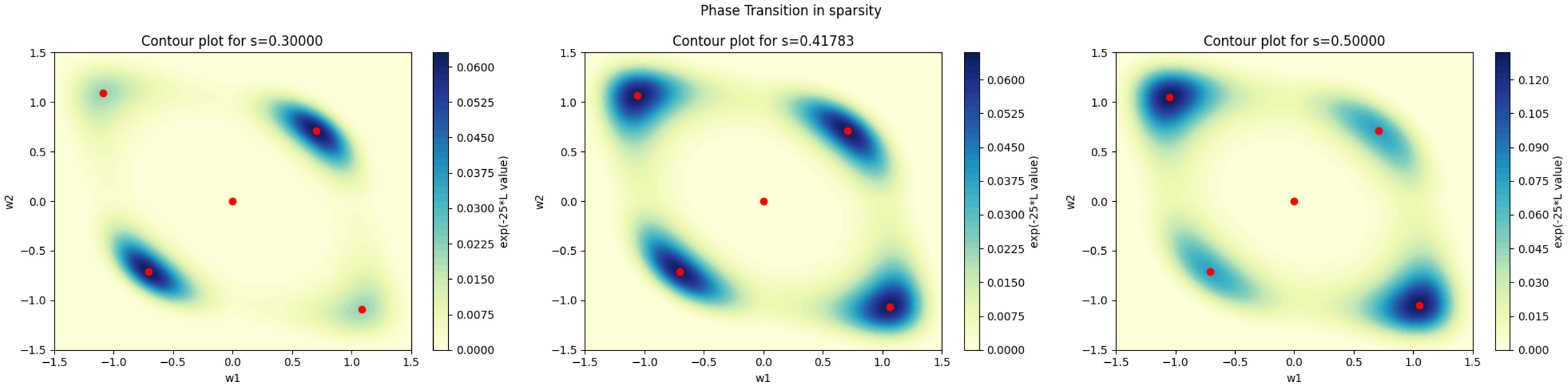
Superposition
(Antipodal Pair)

$W = \begin{bmatrix} 1 & -1 \end{bmatrix}$
$W \perp \begin{bmatrix} 1 & 1 \end{bmatrix}$

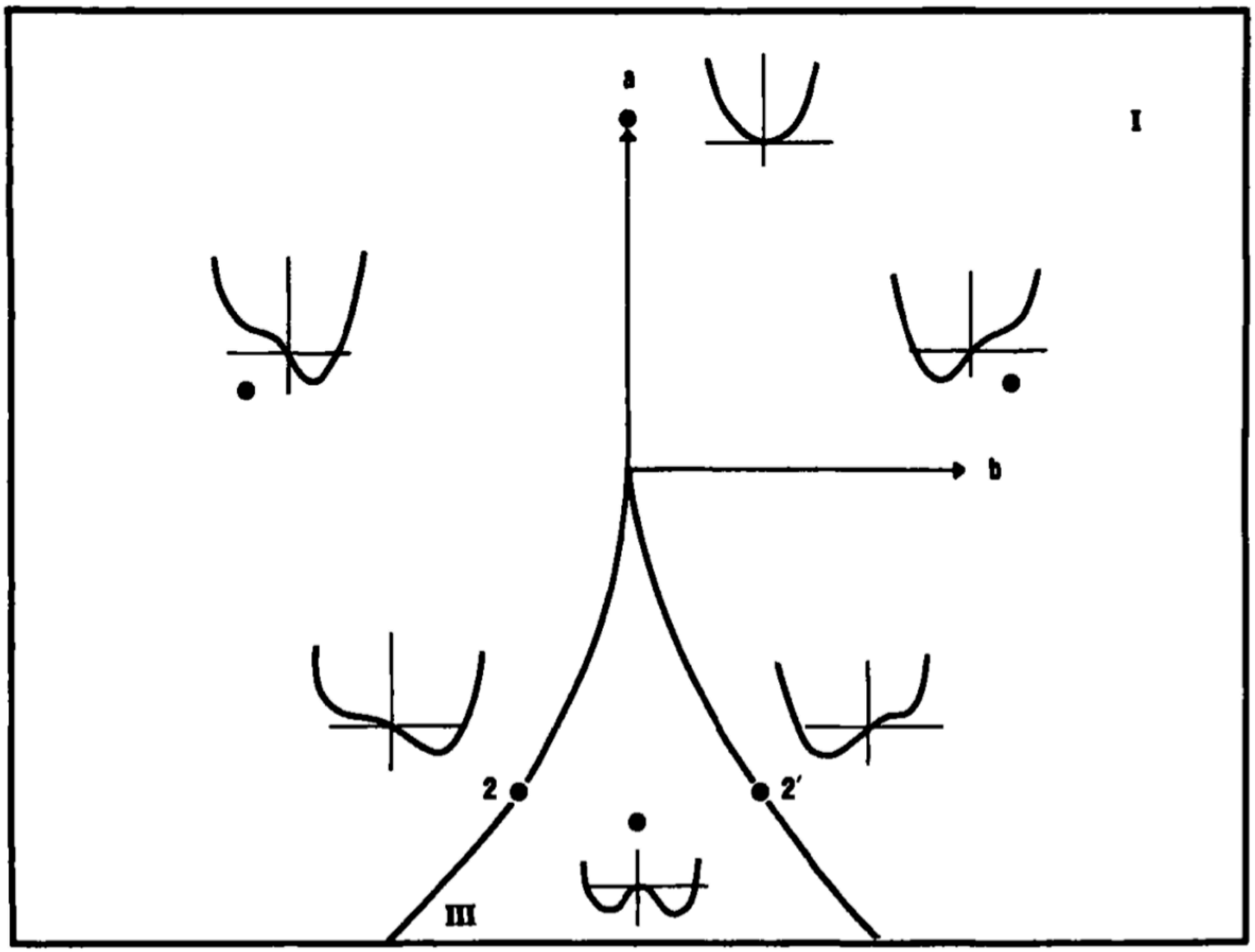
Phase Transition

- Note there is a clear $w = (w_1, w_2) = -(w_1, w_2)$ symmetry from the $O(1)$ group so it suffices to solve $L\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = L\left(-\frac{1}{\sqrt{2}}\sqrt{\frac{3+5s}{1+3s}}, \frac{1}{\sqrt{2}}\sqrt{\frac{3+5s}{1+3s}}\right)$ which gives

$$s = 0, \quad s = \frac{1}{43}\left(3 + 4\sqrt{14}\right) \approx 0.41783$$

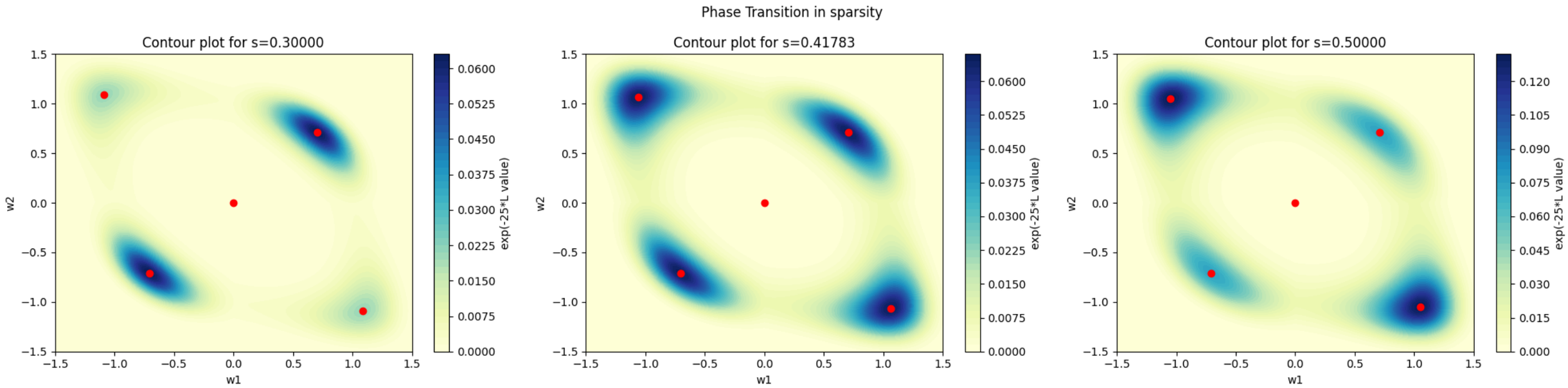


3 Critical Points region



Takeaways

- If you need a Toy visible loss landscape for your experiments/theory this week take this one!



First Order Phase Transition

Bonus Content - Non Constant Importance

- What about non-constant importance? Here $I_1 = 1$ and $I_2 = I$ is a new parameter

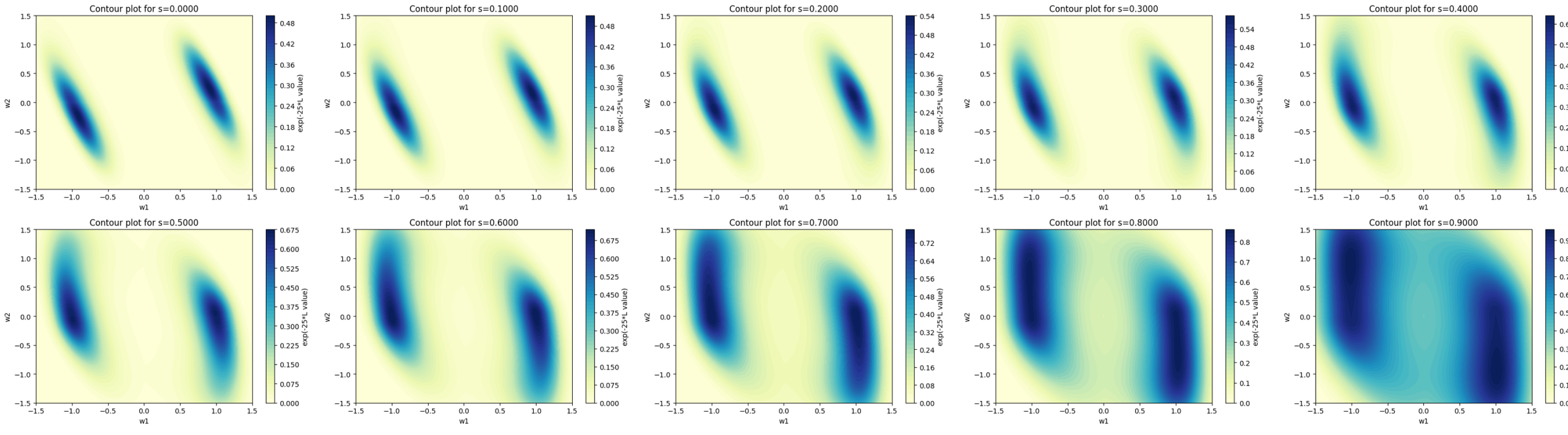
$$L(w) = \int q(x,s)\,||\mathbf{I}(x - \mathbf{ReLU}(W^T W x))||^2 dx$$

$$= s(1-s)\left(\int_0^1 \left(x_1 - w_1^2 x_1\right)^2 dx_1 + I \int_0^1 \left(x_2 - w_2^2 x_2\right)^2 dx_2 \right)$$

$$+(1-s)^2 \int_0^1 \int_0^1 \left(x_1 - \mathbf{ReLU}(w_1^2 x_1 + w_1 w_2 x_2)\right)^2 + I\left(x_2 - \mathbf{ReLU}(w_2^2 x_2 + w_1 w_2 x_1)\right)^2 dx_1 dx_2$$

Contour Plots

Phase Transition in sparsity for l = 0.1



Phase Transition in sparsity for l = 0.4