8 IGR-M in 2D model with non-linear (sigmoid) activation

Beyond the linear case in Section-4.1 of the manuscript, now we consider a 2D nonlinear model that has a Sigmoid activation function to explore the effect of IGR-M. The loss function E is minimized using two learnable parameters (w_1, w_2) but with a sigmoid layer in-between. Here the optimization problem is as follows:

$$(\hat{w_1}, \hat{w_2}) = \arg\min_{w_1,w_2} \frac{1}{2}(y - w_1\sigma(w_2 x))^2 \equiv \arg\min_{w_1,w_2} \frac{1}{2}\left(y - \frac{w_1}{1+e^{-w_2 x}}\right)^2 := E(w_1, w_2)$$

where σ is the Sigmoid activation function. The norm of the gradient has the following expression in this case:

$$\|\nabla E\|^2 = \left|\frac{\partial E}{\partial w_1}\right|^2 + \left|\frac{\partial E}{\partial w_2}\right|^2 = \left(\frac{1}{(1+e^{-w_2 x})^2} + \frac{w_1^2 x^2 e^{-2w_2 x}}{(1+e^{-w_2 x})^4}\right)\left(y - \frac{w_1}{1+e^{-w_2 x}}\right)^2.$$

The dashed black curve plots global minima given by the equation $w_2 = -\frac{\log(\frac{w_1}{y}-1)}{x}$. Unlike the linear case, (where the IGR was proportional to the norm of the weights w_1 and w_2), here the IGR $\|\nabla E\|^2$ has a more complicated level set (Figure 3. So, to help understand the effect of IGR-M, we plot two reference curves, one is the dark blue curve that represents the gradient flow for the original loss function, given as

$$\mathbf{x}'(t) = -\nabla E(\mathbf{x}(t)).$$

where $\mathbf{x} = [w_1, w_2]^T$. The other is the solid black curve that shows the gradient flow for implicit regularizer $\|\nabla E\|^2$ given as:

$$\mathbf{x}'(t) = -\nabla\|\nabla E(\mathbf{x}(t))\|_2^2.$$

A method with a stronger IGR would have a trajectory closer to the solid black curve. So, we plot the trajectories of (GD) ($\beta = 0$) and (GD+M) with $\beta = 0.5, 0.8$, and 0.9, with the same initialization $(w_1 = 6, w_2 = 2)$. The effective learning-rate is kept the same for all the trajectories which equals the learning rate for GD, i.e., $\frac{h}{1-\beta} = 0.01$. We see how the trajectory for (GD+M) is closer to the gradient flow for implicit regularizer (the solid black curve) than that of (GD). More explicitly, we observe that all the trajectories converge to the curve of global minima. However, with a larger β, the trajectory becomes closer to the gradient flow minimizing $\|\nabla E\|^2$ (the solid black curve). This observation agrees with our theorem which states the modified loss is a weighted combination of the original loss E and implicit regularizer $\|\nabla E\|^2$, and larger β leads to a larger weight for the regularizer $\|\nabla E\|^2$, hence making it closer to the solid black curve.

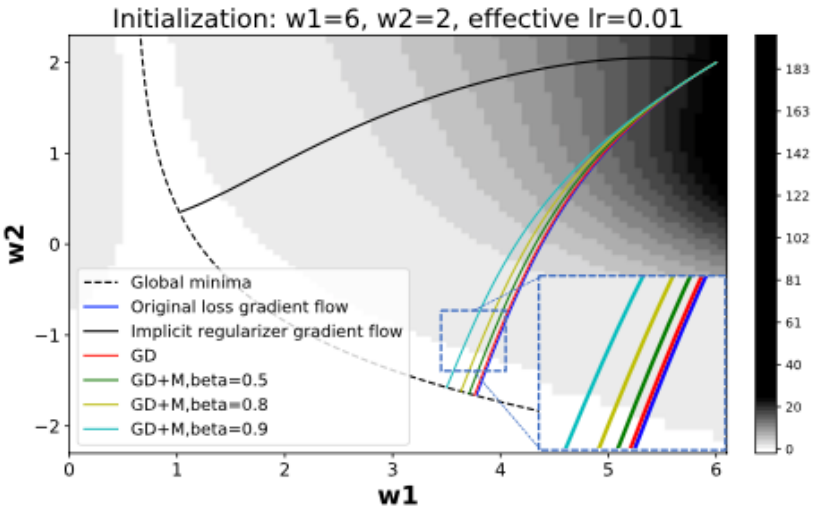

Figure 3: Trajectories for (GD) and (GD+M) for various β but with the same effective learning rate $\frac{h}{(1-\beta)}$. With increasing β, the trajectory becomes closer to the gradient flow of the implicit regularizer (solid back line), hence supporting our theory. The background color denotes the magnitude $\|\nabla E\|_2^2$