## A WHY GRADIENT DESCENT IMPLICITLY REGULARIZES

This is a sketch of why gradient descent implicitly regularizes. Suppose we have a model $Xw$ for a vector of data $y \in \mathbb{R}^n$ and want to minimize the norm of the error,

$$L(w) = \|Xw - y\|_2^2 = \|e\|_2^2$$

where we introduce some short-hand notation. We use the gradient learning rule,

$$w(t+1) = w(t) - \eta X^T e(t)$$
$$\Rightarrow e(t+1) = e(t) - \eta X X^T e(t)$$
$$\Rightarrow e(t+1) = (I - \eta X X^T) e(t)$$

Each matrix satisfies $X \in \mathbb{R}^{n \times d_1}$ where $n$ is the number of samples and $d_1$ is the dimension of each sample. In the overparameterized setting we have $d_1 > n$ and so $XX^T$ will generically have full-rank and the error will go to zero.

This lies in the difference between $XX^T$ which appears here in the error analysis and $X^T X$ which appears in the solution. So we can have $XX^T \in \mathbb{R}^{n \times n}$ generically full-rank only if we have more parameters than there is data. On the other hand, we only have $X^T X$ full-rank if also it's satisfied that there is more data than parameters. This is important because in this case we can compute the pseudo-inverse easily. Generically, we can show that if we use gradient descent we have something like the following,

$$\underbrace{(X^T X)^{-1} X}_{\text{left inverse}} \quad \underbrace{X^{-1}}_{\text{inverse}} \quad \underbrace{X^T (X X^T)^{-1}}_{\text{right inverse}}$$

for the cases where we are under-parameterized, minimally parameterized, or over-parameterized to model the data.

So under gradient flow we'll suppose the parameters update according to,

$$\dot{w} = -\eta X^T e$$
$$w(0) = 0$$

Observe that the gradient $\dot{w}$ is invariantly in the span of $X^T$ so we may conclude that $w(t)$ is always in the span of $X^T$. Generically, any solution in the over-parameterized setting is a global optimizer such that $Xw = y$. This means that the limit of the flow can be written as $w^* = X^T \alpha$ for some coefficient vector with the constraint that $Xw^* = y$. After some manipulations we find that,

$$y = Xw^* = XX^T \alpha$$
$$\Rightarrow \alpha = (XX^T)^{-1} y$$
$$\Rightarrow w^* = X^T (XX^T)^{-1} y = X^+ y$$

This means that the solution $X^+$ picked from gradient flow is the Moore-Penrose psuedoinverse. This can be defined as the matrix,

$$X^+ = \lim_{\lambda \to 0^+} X^T (XX^T + \lambda I)^{-1}$$

Also observe that there is a unique minimizer for the regularized problem,

$$\min_w L(w) + \lambda \|w\|_2^2$$

with value $w_\lambda = X^T (XX^T + \lambda I)^{-1} y$. Perhaps, $Xw = y$ has a set of solutions, but it is clear this set is convex so there is a unique minimum norm solution. On the other hand, each $w_\lambda$ corresponds to a best solution with norm below the minimum. However, we have $w^* = \lim_{\lambda \to 0^+} w_\lambda$ from continuity. Since $w^*$ is an exact solution it can't have less than the minimum-norm and it is clear $w^*$ can't have above the minimum-norm either since this is not the case for any of the $w_\lambda$. We conclude that gradient descent does indeed find the minimum norm solution.