Vanishing generator gradients in original GAN

The original value function suggested in the GAN paper is

$$\min_{\theta_G} \max_{\theta_D} \left(\mathbb{E}_{x \sim p_{data}} \left[\log D_{\theta_D}(x) \right] + \mathbb{E}_{z \sim p(z)} \left[\log(1 - D_{\theta_D}(G_{\theta_G}(z))) \right] \right), \tag{1}$$

where the dependence on discriminator, generator parameters θ_D, θ_G have been explicitly included. Since the discriminator wants to maximize this as a function of θ_D, a good discriminator will output 1 on real samples and 0 on fake samples. Remember that D is a probability and the output of a sigmoid, so $D_{\theta_D}(s) = \sigma_{\theta_D}(s)$, where s is the discriminator input ($s = G_{\theta_G}(z)$); the vanishing gradient/saturation referred to when the discriminator is good originate from the sigmoid. From the chain rule,

$$\nabla_{\theta_G} \mathbb{E}_{z \sim p(z)} \left[\log(1 - D_{\theta_D}(G_{\theta_G}(z))) \right] = \mathbb{E}_{z \sim p(z)} \left[\frac{1}{1 - \sigma_{\theta_D}(s)} (-\sigma'_{\theta_D}(s)) \Big|_{s = G_{\theta_G}(z)} \nabla_{\theta_G} G_{\theta_G}(z) \right]. \tag{2}$$

The difference in behavior comes from the product of the first two terms in the two cases. In the case above, $\sigma'(s) \to 0$ as $\sigma \to 0$, while the fraction remains finite. For a generator function $\log(D_{\theta_D}(G_{\theta_G}(z)))$, the fraction diverges and the two balance to give nonvanishing gradient. Figure 1 plots $\log(1 - \sigma(s))$ vs. $\log(\sigma(s))$, *as a function of* s, and we see the vanishing gradient/saturation in the region $s \to -\infty$ ($\sigma \to 0$).

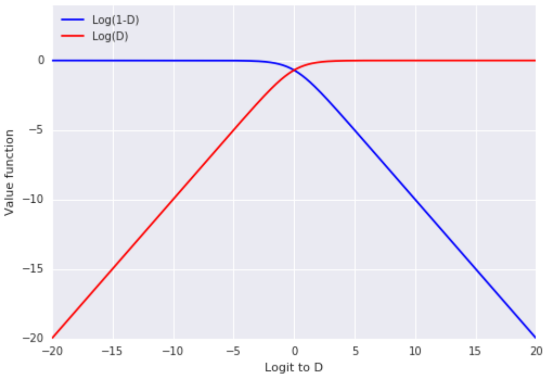

Figure 1: Comparing two generator value functions, $\log(1 - \sigma(s))$ and $\log(\sigma(s))$, as a function of the logit value s to discriminator.