In *The Relativistic Discriminator: a key element missing from standard GAN* a loss function for a discriminator is proposed but discarded with the claim that it is impractical due to that the formula would require quadratic instead of linear time in the batch size to approximate. This is not the case and I will demonstrate that good approximations can be computed in linear time using amortization and, in a particularly practical approach, using a linear number of evaluations of the function C and $O(n \log n)$ scalar operations.

The loss function in question is

$$L_D = -\mathbb{E}_{x_r \sim \mathbb{P}} \left[\log(\mathbb{E}_{x_f \sim \mathbb{Q}} \left[\sigma \left(C(x_r) - C(x_f) \right) \right] \right] - \mathbb{E}_{x_f \sim \mathbb{Q}} [\log(1 - \mathbb{E}_{x_r \sim \mathbb{P}} [\sigma(C(x_f) - C(x_r))])].$$

It is easy to see that this could be done if $\sigma(x - y)$ could be decomposed as a sum of products of two factors, one that depends only on x and the other which depends only on y, i.e. $\sigma(x - y) = \sum_k A_k(x) B_k(y)$.

Then $\mathbb{E}_{x_f \sim \mathbb{Q}}[\sigma(C(x_r) - C(x_f)] = \mathbb{E}_{x_f \sim \mathbb{Q}}[\sum_k A_k(C(x_r)) B_k(C(x_f))] =$
$= \sum_k A_k(C(x_r)) \mathbb{E}_{x_f \sim \mathbb{Q}}[B_k(C(x_f))]$. The numbers, $\mu_f^{(k)} = \mathbb{E}_{x_f \sim \mathbb{Q}}[B_k(C(x_f))]$ can thus be computed once for each batch and then be used to determine these expectations.

Things can be done similarly for $\mathbb{E}_{x_r \sim \mathbb{P}}[\sigma(C(x_f) - C(x_r))]$ where there would be corresponding numbers $\mu_r^{(k)}$.

Such decompositions of $\sigma(x-y)$ exist. One way to obtain them is from Taylor series, e.g., $\sigma(x-y) = \sum_k c_k (x-y)^k = \sum_k c_k \sum_l \binom{k}{l} x^k (-y)^l = \sum_{k,l} A_{k,l}(x) B_{k,l}(y)$. This can be done in a number of operations independent of the batch size, however Taylor series for, for example, the logistic function require many terms to be accurate for large x and I will showcase a particularly effective approximation which will require sorting of a list with length equal to the batch size.

The logistic function is given by $\sigma(x) = \frac{1}{1+e^{-x}}$. This can, for positive x be written as a geometric series with a remainder term $\sum_{k=0}^{n} (-1)^k e^{-kx} + (-1)^{n+1} e^{-kx} \sigma(x)$. The truncated geometric series is accurate for large x and if the $\sigma(\cdot)$ in the remainder term is approximated by its truncated Taylor series it will be accurate around 0, ensuring that the expression is accurate for all non-negative x even when few terms are used. For negative x $\sigma(x) = 1 - \sigma(-x)$.

Choose the approximation $\sigma(x) \approx \sum_{k=0}^{n} (-1)^n e^{-kx} + (-1)^{n+1} e^{-(n+1)x} \left(\sum_{k=0}^{m} c_k x^k \right)$ and let $x = C(x_r)$ and Y be the probability distribution of $C(x_f)$ where $x_f \sim \mathbb{Q}$.

Then $\mathbb{E}_{x_f \sim \mathbb{Q}} \left[\sigma \left(C(x_r) - C(x_f) \right) \right] = \mathbb{E}_{y \sim Y} \left[\sigma \left(x - y \right) \right] =$
$= \mathbb{E}_{y \sim Y} \left[\sigma \left(x - y \right) | \, x \geq y \right] P_{y \sim Y}(x \geq y) + \mathbb{E}_{y \sim Y} \left[\sigma \left(x - y \right) | \, x < y \right] (1 - P_{y \sim Y}(x \geq y)) \approx$
$\approx F_{n,m}^{+}(x) P_{y \sim Y}(x \geq y) + F_{n,m}^{-}(x)(1 - P_{y \sim Y}(x \geq y))$ where $F_{n,m}^{+}(x) =$
$= \sum_{k=0}^{n} (-1)^k e^{-kx} \mathbb{E}_{y \sim Y}[e^{ky} | x \geq y] + (-1)^{n+1} e^{-(n+1)x} \sum_{k=0}^{m} c_k \sum_{l=0}^{k} \binom{k}{l} x^{k-l} \cdot$
$\cdot \mathbb{E}_{y \sim Y}[(-y)^l e^{(n+1)y} | x \geq y]$ and $F_{n,m}^{-}(x) = 1 - \sum_{k=0}^{n} (-1)^k e^{kx} \cdot$
$\cdot \mathbb{E}_{y \sim Y}[e^{-ky} | x < y] - (-1)^{n+1} e^{(n+1)x} \sum_{k=0}^{m} c_k \sum_{l=0}^{k} \binom{k}{l} (-x)^{k-l} \cdot$
$\cdot \mathbb{E}_{y \sim Y}[y^l e^{-(n+1)y} | x < y]$.

Let $p_x = P_{y \sim Y}(x \geq y)$, $\mu_{k,x}^{+} = \mathbb{E}_{y \sim Y}[e^{ky} | x \geq y]$, $\mu_{k,x}^{-} = \mathbb{E}_{y \sim Y}[e^{-kx} | x < y]$, $\tilde{\mu}_{l,x}^{+} = \mathbb{E}_{y \sim Y}[(-y)^l e^{(n+1)y} | x \geq y]$ and $\tilde{\mu}_{l,x}^{-} = \mathbb{E}_{y \sim Y}[y^l e^{-(n+1)y} | x < y]$. Once estimates of these numbers have been computed each evaluation of $E_{x_f \sim \mathbb{Q}}[\sigma(C(x_r) - C(x_f))]$ can be performed in time $O(1)$.

In order to determine p_x we maintain a sorted list or tree of $\{y^{(k)}\}_{k=1}^{N}$ where each $y^{(k)}$ has been sampled from Y, structured so that the smallest number greater than some given x can be quickly found. The relevant probability can then be estimated from the place of x would have in the list. This can be done in time $O(n \log n)$.

Since $x \mapsto e^{kx}$ is increasing and $x \mapsto e^{-kx}$ decreasing $\mu_{k,x}^{+} = \mathbb{E}_{y\sim Y}[e^{ky}|x \geq y]$ and $\mu_{k,x}^{-} = \mathbb{E}_{y\sim Y}[e^{-kx}|x < y]$ can be computed without re-sorting in time $O(n)$. $\tilde{\mu}_{l,x}^{+}$ and $\tilde{\mu}_{l,x}^{-}$ each require resorting for each l-term unless it is exploited that $x \mapsto x^l e^{-(n+1)x}$ for positive numbers, first increasing and then decreasing, and $x \mapsto (-x)^l e^{(n+1)x}$, first decreasing and then increasing for negative numbers.

Things can be done similarly for $\mathbb{E}_{x_r\sim\mathbb{P}}[\sigma(C(x_f) - C(x_r))]$.

These methods can also be used to create averaged activation functions, for example, let Y be the probability distribution of the output of a previous layer and x a particular example, then $\mathbb{E}_{y\sim Y}[\max\{x - y, 0\}] = \mathbb{E}_{y\sim Y}[x - y|x - y \geq 0] = x - \mathbb{E}_{y\sim Y}[y|x \geq y]$. This expectation can be computed by sorting samples from Y and for each element in the list keeping the sum of those below it. For x not in the list the location x would have had in the list would have to be determined; and the sum of the elements below it used, that is, the value kept for the element below x.

We can also consider $\mathbb{E}_{y,z,w\sim Y}[\frac{x-y}{z-w}|x \geq y, z \geq w] = (x - \mathbb{E}_{y\sim Y}[y|x \geq y]) \cdot$ $\cdot\mathbb{E}_{z,w\sim Y}[\frac{1}{z-w}|z \geq w]$, but this is not useful, since the reciprocal of 0 is undefined and since it should occur in the sampling when Y is a discrete distribution.

However, $\mathbb{E}_{y,z,w\sim Y}[\frac{\sigma(x-y)}{\sigma(z-w)}] = \mathbb{E}_{y\sim Y}[\sigma(x-y)]\mathbb{E}_{z,w\sim Y}[\frac{1}{\sigma(z-w)}] = \mathbb{E}_{y\sim Y}[\sigma(x - y)]\mathbb{E}_{z,w\sim Y}[1/\frac{1}{1+e^{-z+w}}] = \mathbb{E}_{y\sim Y}[\sigma(x-y)]\,(1 + \mathbb{E}_{z\sim Y}[e^{-z}]\mathbb{E}_{w\sim Y}[e^{w}])$. $\mathbb{E}_{y\sim Y}[\sigma(x-y)]$ can then be approximated using the methods from before.