This paper talks about the effective usage of adaptive gradient techniques in general adversarial networks. The author states that adaptive gradient algorithms are very popular in training deep neural networks due to their computational efficiency and minimal need for hyperparameter tuning. However, there is not enough evidence showing that adaptive gradient methods converge faster in supervised deep-learning tasks. They perform worse than SGD on image classification tasks. GANs are a popular class of generative models. They consist of a generator and a discriminator, both of which are defined by deep neural networks. They are trained under adversarial cost. Adam is the preferred optimizer for training GANs. GAN is a min-max optimization problem in nature. The main goal of the paper is to design stochastic first-order algorithms with low iteration complexity, low per-iteration cost, and suitable for a general class of non-convex non-concave min-max problems.
The algorithm given for OSD is

Algorithm 1 Optimistic Stochastic Gradient (OSG)

Input: $\mathbf{z}_0 = \mathbf{x}_0 = 0$
for $k = 1, \ldots, N$ **do**
  $\mathbf{z}_k = \Pi_{\mathcal{X}} \left[\mathbf{x}_{k-1} - \eta \cdot \frac{1}{m_{k-1}} \sum_{i=1}^{m_{k-1}} T(\mathbf{z}_{k-1}; \xi_{k-1}^i) \right]$
  $\mathbf{x}_k = \Pi_{\mathcal{X}} \left[\mathbf{x}_{k-1} - \eta \cdot \frac{1}{m_k} \sum_{i=1}^{m_k} T(\mathbf{z}_k; \xi_k^i) \right]$
end for

In the stochastic extra gradient method, the extra call on zk allows the algorithm to anticipate the landscape of T. In contrast, {xk} is an ancillary sequence in OSG and the stochastic gradient is only computed over the sequence of {zk}.
The algorithm for OAdagrad is:

Algorithm 2 Optimistic AdaGrad (OAdagrad)

Input: $\mathbf{z}_0 = \mathbf{x}_0 = 0$, $H_0 = \delta I$
for $k = 1, \ldots, N$ **do**
  $\mathbf{z}_k = \mathbf{x}_{k-1} - \eta H_{k-1}^{-1} \widehat{\mathbf{g}}_{k-1}$
  $\mathbf{x}_k = \mathbf{x}_{k-1} - \eta H_{k-1}^{-1} \widehat{\mathbf{g}}_k$
  Update $\widehat{\mathbf{g}}_{0:k} = [\widehat{\mathbf{g}}_{0:k-1} \ \widehat{\mathbf{g}}_k]$, $s_{k,i} = \|\widehat{\mathbf{g}}_{0:k,i}\|$, $i = 1, \ldots, d$ and set $H_k = \delta I + \text{diag}(s_{k-1})$
end for

Similar to OSG where min variable and max variable are updated simultaneously. ĝk represents the gradient at iteration k. ĝ0k is the i-th row of the matrix obtained by concatenating the sub gradients from iteration 0 through k in the algorithm. OAdagrad updates the discriminator and generator simultaneously. There is no convergence proof for Alternating Adam for non-convex non-concave problem. OAdagrad provides a theoretical justification of a special case of Optimistic Adam as OAdagrad naturally fits into the framework of Optimistic Adam.
In experimental analysis we can conclude that OAdagrad outperforms simultaneous Adam in quantitative metrics (IS and FID) and in sample quality generation. OAdagrad performs better than OSG and Alternating Adam, and OAdagrad results in higher IS on the CIFAR10 benchmark test. The algorithm is proven to enjoy faster adaptive convergence than its non-adaptive counterpart when the gradient is sparse.

References: *Liu, Mingrui, et al. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. ICLR 2020.*