To recap Algorithm 1, in each iteration step 1 samples points from p_k and p_{-k}, step 2 solves the inner maximization by moving samples of p_{-k} to locations where D has the maximum outputs, and step 3 solves the outer minimization by increasing D outputs on p_k samples and decreasing outputs on p_{-k} samples. Step 2 is implemented by performing gradient ascent on D, using initial samples of p_{-k} as starting points. Given that D could be a non-concave function during the course of Algorithm 1 execution, samples of p_{-k} could be stuck in local maxima points in $\mathcal{X} \setminus \text{Supp}(p_k)$. We now show that due to this gradient-based search method used by step 2, Algorithm 1 has the following convergence property:

Proposition 1. *When p_{-k} is a uniform distribution over the space $\mathcal{X} \setminus Supp(p_k)$, algorithm 1 converges to a D solution with no local maxima and global maxima at support of p_k.*

Proof. We assume that D has enough capacity such that step 3's update of D in $\text{Supp}(p_k)$ does not affect D's outputs in $\mathcal{X} \setminus \text{Supp}(p_k)$. We assume that the environment in which Algorithm 1 is simulated has a numeric limit of ϵ (e.g., $\epsilon = 10^{-12}$) such that $\mathcal{X} \setminus \text{Supp}(p_k)$ is a finite set. (This assumption is valid when the algorithm runs on a computer.) Since we are analyzing an algorithm's behavior, p_k and p_{-k} are both empirical distributions.

We first prove that when p_{-k} is a uniform distribution, any local maximum point in $\mathcal{X} \setminus \text{Supp}(p_k)$ can be eliminated by running Algorithm 1 for a finite number of iterations.

To proceed, we first state the condition under which a local maximum point will be eliminated: a local maximum point q in $\mathcal{X} \setminus \text{Supp}(p_k)$ will be eliminated if via one or more iterations of the algorithm a sufficient number of p_{-k} samples reach q. When this condition is satisfied, the cumulative effects of step 3 cause the local maximum to disappear by decreasing $D(q)$ to a sufficiently small value.

We next show that the above condition is always satisfied when p_{-k} is a uniform distribution and Algorithm 1 runs for a finite number of iterations. Let U be the set of points in $\mathcal{X} \setminus \text{Supp}(p_k)$ that reach q when performing gradient ascent on D in step 2. U is non-empty when a sufficiently small step size is used for performing the gradient ascent, because it at least contains the point q itself when step size approaches 0. For U is non-empty, a sufficient number of p_{-k} samples could fall on U and subsequently reach q if enough samplings of the uniform distribution are done via step 1.

For a given D, the set of local maxima points in $\mathcal{X} \setminus \text{Supp}(p_k)$ is a finite set. However, as new local maxima are constantly being created due to D's update in each iteration, it is possible that this set will never be empty. We now prove that in a finite iterations of the algorithm this set actually goes to empty. Let $\mathcal{Q}_t$ be the set of local maxima points of D in $\mathcal{X} \setminus \text{Supp}(p_k)$ in iteration t. We have shown in the first proof that all the elements of $\mathcal{Q}_t$ are going to be reached by p_{-k} samples in a finite number of iterations, hence if $\mathcal{Q}_t$ is non-empty as $t \to \infty$, D values in $\mathcal{X} \setminus \text{Supp}(p_k)$ will be decreased by an $\geq \epsilon$ amount for an infinite number of times, which contradicts our assumption that D is a finite function in the finite set of $\mathcal{X} \setminus \text{Supp}(p_k)$.

□

Proposition 2. *When p_{-k} is sampled from a uniform distribution over the space $\mathcal{X} \setminus Supp(p_k)$, algorithm 1 converges to a D solution with global maxima at support of p_k and no local maxima outside support of p_k.*

Proof. We assume that D has enough capacity such that step 3's update of D in $\text{Supp}(p_k)$ does not affect D's outputs in $\mathcal{X} \setminus \text{Supp}(p_k)$.

We assume that the environment in which Algorithm 1 is simulated has a numeric limit of ϵ (e.g., $\epsilon = 10^{-12}$) such that $\mathcal{X} \setminus \text{Supp}(p_k)$ is a finite set. (This assumption is valid when the algorithm runs on a computer.)Since p_{-k} is sampled from a uniform distribution, the probability of any point in $mathcalX$ lies in the support of p_{-k} is positive.

We first prove that when p_{-k} is a uniform distribution, any local maximum point in $\mathcal{X} \setminus \text{Supp}(p_k)$ can be eliminated by running Algorithm 1 for a finite number of iterations.

To proceed, we first state the condition under which a local maximum point will be eliminated: a local maximum point q in $\mathcal{X} \setminus \text{Supp}(p_k)$ will be eliminated if via one or more iterations of the algorithm a sufficient number of p_{-k} samples reach q. When this condition is satisfied, the cumulative effects of step 3 cause the local maximum to disappear by decreasing $D(q)$ to a sufficiently small value.

We next show that the above condition is always satisfied when p_{-k} is a uniform distribution and Algorithm 1 runs for a finite number of iterations. Let U be the set of points in $\mathcal{X} \setminus \text{Supp}(p_k)$ that reach q when performing gradient ascent on D in step 2. U is non-empty when a sufficiently small step size is used for performing the gradient ascent, because it at least contains the point q itself when step size approaches 0. For U is non-empty, a sufficient number of p_{-k} samples could fall on U and subsequently reach q if enough samplings of the uniform distribution are done via step 1.

For a given D, the set of local maxima points in $\mathcal{X} \setminus \text{Supp}(p_k)$ is a finite set. However, as new local maxima are constantly being created due to D's update in each iteration, it is possible that this set will never be empty. We now prove that in a finite iterations of the algorithm this set actually goes to empty. Let $\mathcal{Q}_t$ be the set of local maxima points of D in $\mathcal{X} \setminus \text{Supp}(p_k)$ in iteration t. We have shown in the first proof that all the elements of $\mathcal{Q}_t$ are going to be reached by p_{-k} samples in a finite number of iterations, hence if $\mathcal{Q}_t$ is non-empty as $t \to \infty$, D values in $\mathcal{X} \setminus \text{Supp}(p_k)$ will be decreased by an $\geq \epsilon$ amount for an infinite number of times, which contradicts our assumption that D is a finite function in the finite set of $\mathcal{X} \setminus \text{Supp}(p_k)$.

□

We note that the above convergence property holds for any random initialization of D. However, uniform distribution is not a necessary condition; any p_{-k} distribution that has non-zero density everywhere in the data space will suffice. For a particular or particular type of initialization of D, its local maxima points could follow some distribution, and hence the assumption on p_{-k} could be relaxed. However, in practise whether or not a p_{-k} distribution is sufficient for a given D can be difficult to measure.