Supplementary Material:
K-Beam Minimax – Efficient Optimization for Deep Adversarial Learning

Jihun Hamm [1] **Yung-Kyun Noh** [2]

1. Simple surfaces

Fig.1 shows the six surfaces $f(u,v)$ and the maximum value function $\phi(u) = \max_{v\in\mathcal{V}} f(u,v)$. From $\phi(u)$ one can check the minima $\arg\min_u \phi(u)$ are:
(a) $u = 0$, (b) $u = 0$, (c) $u = 0$, (d) $u = \pm 0.25$, (e) $u = 0$, and (f) $u = 0$.
The corresponding maxima $R(u) = \arg\max_{v\in\mathcal{V}} f(u,v)$ at the minimum are:
(a) $R(0) = \{0\}$, (b) $R(0) = \{0\}$, (c) $R(0) = [-0.5, 0.5]$, (d) $R(\pm 0.25) = \{-0.25, 0.5\}$, (e) $R(0) = \{-0.5, 0.5\}$, and (f) $R(0) = \{-0.5, 0.5\}$.

Furthermore, $R(\mathcal{U})$ for the whole domain is:
(a) $R(\mathcal{U}) = \{0\}$, (b) $R(\mathcal{U}) = [-0.5, 0.5]$, (c) $R(\mathcal{U}) = \{-0.5, 0.5\}$ except for $R(0) = [-0.5, 0.5]$, (d) $R(\mathcal{U}) = [-0.5, -0.25] \cup \{0.5\}$, (e) $R(\mathcal{U}) = \{-0.5, 0.5\}$, and (f) $R(\mathcal{U}) = \{-0.5, 0.5\}$. These can be verified by solving the minimax problems in closed form.

Note that the origin $(0,0)$ is a critical point for all surfaces. It is also a global saddle point and minimax point for surfaces (a)-(c), but is neither a saddle nor a minimax point for surfaces (d)-(f).

2. Proofs

Lemma 1 (Corollary 4.3.2, Theorem 4.4.2, (Hiriart-Urruty & Lemaréchal, 2001))**.** *Suppose $f(u,v)$ is convex in u for each $v \in A$. Then $\partial\phi_A(u) = \mathrm{co}\{\cup_{v\in A}\nabla_u f(u,v)\}$. Similarly, suppose $f(u,v)$ is convex in u for each $v \in \mathcal{V}$. Then $\partial\phi(u) = \mathrm{co}\{\cup_{v\in\mathcal{V}}\nabla_u f(u,v)\}$.*

Lemma 2 (Chap 3.6, (Dem'yanov & Malozemov, 1974))**.** *A point u is an ϵ-stationary point of $\phi_A(u)$ if and only if $0 \in \mathrm{co}\{\cup_{v\in R_A^\epsilon(u)}\nabla_u f(u,v)\}$.*

Lemma 3. *Suppose $R(u)$ is finite at u. If $d_H(R(u), A) = 0$, then $R(u) = R_A(u)$ and therefore $\partial\phi(u) = \partial\phi_A(u)$.*

[1]The Ohio State University, Columbus, OH, USA. [2]Seoul National University, Seoul, Korea.. Correspondence to: Jihun Hamm <hammj@cse.ohio-state.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Proof. Since $A \subseteq \mathcal{V}$, $\max_{v\in\mathcal{V}} f(u,v) = \max_{v\in R(u)} f(u,v) \geq \max_{v\in A} f(u,v)$. By $d_H(R(u), A) = 0$, we have $R(u) \subseteq A$ and therefore for each $v \in R(u)$, $f(u,v) = \max_{v\in\mathcal{V}} f(u,v) = \max_{v\in A} f(u,v)$, so $v \in R_A(u)$. Conversely, if $v \in R_A(u)$ then $f(u,v) = \max_{v\in A} f(u,v) = \max_{v\in\mathcal{V}} f(u,v)$, so $v \in R(u)$. The remainder of the theorem follows from the definition of subdifferentials. □

Fig. 2 explains several symbols used in the following lemmas.

Lemma 4. *If $d_H(R(u), A) \leq \delta$, then for each $v \in R(u)$ there is one or more $v' \in A$ such that $\phi(u) - f(u,v') \leq l\delta$ and $\|\nabla_u f(u,v) - \nabla_u f(u,v')\| \leq r\delta$.*

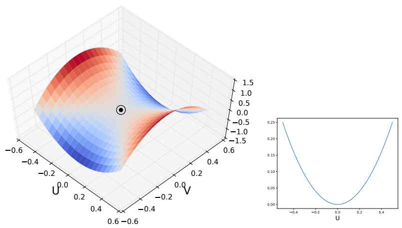
The proof follows directly from the Lipschitz assumptions.

Lemma 5. *Assume $R(u)$ and $S(u)$ are both finite at u. Let $\zeta = \phi(u) - \max_{v\in S(u)\setminus R(u)} f(u,v)$ be the smallest gap between the global and the non-global maximum values at u. If all local maxima are global maxima, then set $\zeta = \infty$. If $d_H(R(u), A) \leq \delta$ and $d_H(A, S(u)) \leq \delta$ where $\delta < 0.5(\zeta - \epsilon)/l$, then for each $v' \in R_A^\epsilon(u)$, there is $v \in R(u)$ such that $\|v - v'\| \leq \delta$.*

Proof. Let any $v' \in A$ be δ-close to a global maximum, then $f(u,v') \geq \phi(u) - l\delta$. Similarly, let any $v'' \in A$ be δ-close to a non-global maximum, then $f(u,v'') \leq \phi(u) - (\zeta - l\delta)$. Consequently, $f(u,v') \geq f(u,v'') + \zeta - 2l\delta > f(u,v'') + \epsilon$, i.e., any $f(u,v')$ and $f(u,v'')$ are separated by at least ϵ. Therefore, each v' satisfies $v' \in R_A^\epsilon = \{v \in A \mid \phi_A(u) - f(u,v) \leq \epsilon\}$ but no v'' satisfies $v'' \in R_A^\epsilon$. □

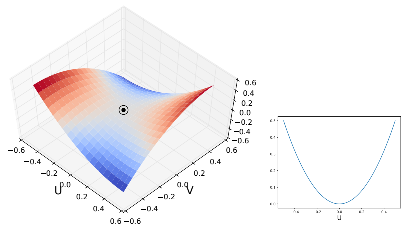
Lemma 6. *Suppose δ is chosen as in Lemma 5 and $\mathcal{U}$ is bounded ($\forall u \in \mathcal{U},\ \|u\| = B < \infty$.) Then any $z' \in \mathrm{co}\{\cup_{v\in R_A^\epsilon}\nabla_u f(u_0,v)\}$ is an $(2r\delta B)$-subgradient of $\phi(u_0)$.*

Proof. From Lemmas 4 and 5, for each $(v^k)' \in R_A^\epsilon$, there is $v^k \in R(u_0)$ such that $\|\nabla_u f(u_0,v^k) - \nabla_u f(u_0,(v^k)')\| \leq r\delta$. Let $z_k = \nabla_u f(u_0,v^k)$ and $z_k' =$



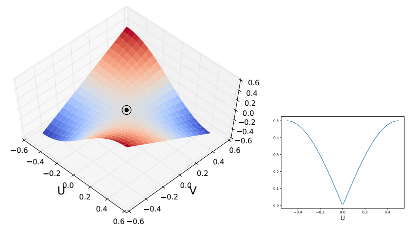
(a) Saddle $(u^2 - v^2)$

critical pts	$\{(0,0)\}$
saddle pts	$\{(0,0)\}$
minimax pts	$\{(0,0)\}$



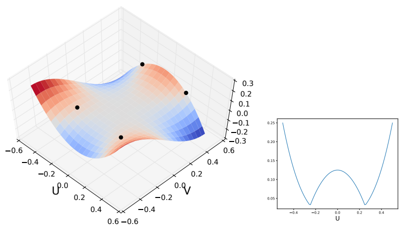
(b) Rotated saddle $(u^2 - v^2 + 2uv)$

critical pts	$\{(0,0)\}$
saddle pts	$\{(0,0)\}$
minimax pts	$\{(0,0)\}$



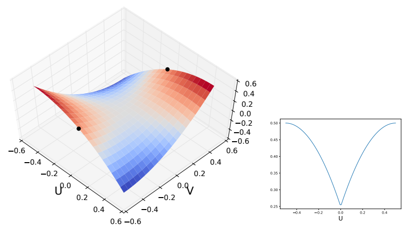
(c) Seesaw $(-v\sin(\pi u))$

critical pts	$\{(0,0)\}$
saddle pts	$\{(0,0)\}$
minimax pts	$\{(0,v)\|v \in [-0.5,0.5]\}$



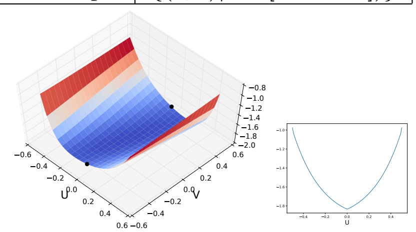
(d) Monkey saddle $(v^3 - 3vu^2)$

critical pts	$\{(0,0)\}$
saddle pts	$\{\}$
minimax pts	$\{(\pm 0.25,-0.25),(\pm 0.25,0.5)\}$



(e) Anti-saddle $(-u^2 + v^2 + 2uv)$

critical pts	$\{(0,0)\}$
saddle pts	$\{\}$
minimax pts	$\{(0,\pm 0.5)\}$



(f) Weapons $(e^{-10(u+.5)e^{-(v+.5)}} + e^{-10(.5-u)e^{v-.5}})$

critical pts	$\{(0,0)\}$
saddle pts	$\{\}$
minimax pts	$\{(0,\pm 0.5)\}$

Figure 1. Examples of saddle point (upper row) and non-saddle point (lower row) problems. The smaller inset after each surface is the max value function $\phi(u) = \max_v f(u,v)$.

$\nabla_u f(u_0, (v^k)')$. Then, for all $k = 1, ..., |R_A^\epsilon|$ and for all u,

$$\begin{aligned} & \phi(u) - \phi(u_0) - \langle z_k',\ u - u_0\rangle \\ = \ & \phi(u) - \phi(u_0) - \langle z_k + z_k' - z_k,\ u - u_0\rangle \\ \geq \ & -\langle z_k' - z_k,\ u - u_0\rangle \\ \geq \ & -\|z_k' - z_k\|\|u - u_0\| \\ \geq \ & -r\delta\|u - u_0\| \geq -2r\delta B. \end{aligned}$$

By taking any convex combination of $\sum_{k=1}^n a_k(\cdot)$ on both sides, we have

$$\phi(u) - \phi(u_0) - \langle \sum_{k=1}^n a_k z_k',\ u - u_0\rangle \geq -2r\delta B,$$

and therefore any $z' \in \mathrm{co}\{\cup_{v \in R_A^\epsilon} \nabla_u f(u_0, v)\}$ is a $(2r\delta B)$-subgradient of $\phi(u_0)$ $\square$

Theorem 7. *Suppose the conditions of Lemmas 4, 5 and 6 hold, and also suppose the max step in Alg.2 is accurate for sufficiently large $i \geq i_0$ for some $i_0 \geq 1$ so that $\max[d_H(R(u_i), A_i), d_H(A_i, S(u_i))] \leq \delta_i$ holds where $\delta_i \leq \min[0.5(\zeta_i - \epsilon_i)/l,\ 0.5\xi_i/(rB)]$ for some non-negative sequence $(\xi_1, \xi_2, ...)$. If the step size satisfies $\rho_i \geq 0, \forall i$, $\sum_{i=1}^\infty \rho_i = \infty$, $\sum_{i=1}^\infty \rho_i^2 < \infty$, and $\sum_{i=1}^\infty \rho_i \xi_i < \infty$, then $\min[\phi(u_1), ..., \phi(u_i)]$ converges to the minimum value ϕ^*.*

Note that a stronger result such as $\liminf_{i\to\infty} \phi(u_i) = \phi^*$ is possible (see, e.g., (Correa & Lemaréchal, 1993)), but we give a simpler proof similar to (Boyd et al., 2003) which assumes $\|\nabla_u f(u,v)\| \leq L$ for some $L > 0$.

Proof. We combine previous lemmas with the standard proof of the ϵ-subgradient descent method. Let $u_{i+1} = u_i - \rho_i g_i$. Then,

$$\begin{aligned} & \|u_{i+1} - u^*\|^2 \\ = \ & \|u_i - u^*\|^2 + \rho_i^2\|g_i\|^2 + 2\rho_i\langle g_i,\ u^* - u_i\rangle \\ \leq \ & \|u_i - u^*\|^2 + \rho_i^2\|g_i\|^2 + 2\rho_i(\phi(u^*) - \phi(u_i) + \xi_i) \end{aligned}$$

from the definition of $\partial_\xi \phi(u)$. Taking $\sum_{i=1}^N(\cdot)$ on both sides gives us

$$\begin{aligned} \|u_{N+1} - u^*\|^2 \ \leq \ & \|u_1 - u^*\|^2 + \sum_{i=1}^N \rho_i^2\|g_i\|^2 \\ & + 2\sum_{i=1}^N \rho_i(\phi(u^*) - \phi(u_i) + \xi_i), \end{aligned}$$

or equivalently,

$$2\sum_{i=1}^N(\rho_i(\phi(u_i) - \phi(u^*) - \xi_i) \leq \|u_1 - u^*\|^2 + \sum_{i=1}^N \rho_i^2\|g_i\|^2.$$

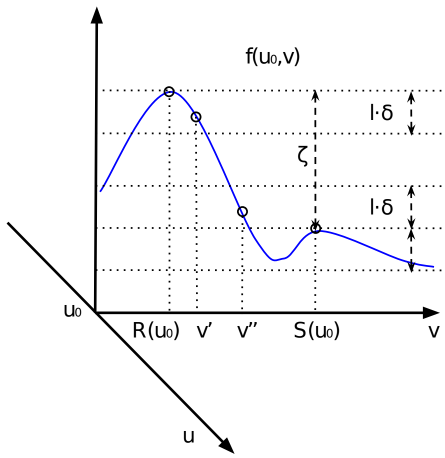


Figure 2. Consider a slice of $f(u,v)$ at $u = u_0$. ζ: smallest gap between the f values of global maxima $R(u)$ and non-global maxima $S(u) \setminus R(u)$. v' is no farther than δ to a point in $R(u_0)$ and v'' is no farther than δ to a point in $S(u_0) \setminus R(u_0)$. By choosing $\epsilon < \zeta - 2l\delta$, we have $v' \in R_A^\epsilon(u_0)$ and $v'' \notin R_A^\epsilon(u_0)$. See Lemma 5.

If we define $\underline{\phi}(u_i) := \min[\phi(u_1), ..., \phi(u_i)]$, then $\sum_{i=1}^N \rho_i(\phi(u_i) - \phi^*) \geq (\sum_{i=1}^N \rho_i)(\underline{\phi}(u_i) - \phi^*)$. Combining the two inequalities, we have

$$\begin{aligned} 0 \;\leq\; \underline{\phi}(u_i) - \phi^* &\leq \frac{\sum_{i=1}^N \rho_i(\phi(u_i) - \phi^*)}{\sum_{i=1}^N \rho_i} \\ &\leq \frac{\|u_1 - u^*\|^2 + \sum_{i=1}^N \rho_i^2 \|g_i\|^2 + 2\sum_{i=1}^N \rho_i \xi_i}{2\sum_{i=1}^N \rho_i} \\ &\leq \frac{\|u_1 - u^*\|^2 + \sum_{i=1}^N \rho_i^2 L^2 + 2\sum_{i=1}^N \rho_i \xi_i}{2\sum_{i=1}^N \rho_i}. \end{aligned}$$

With $\sum_{i=1}^\infty \rho_i = \infty$, $\sum_{i=1}^\infty \rho_i^2 < \infty$, and $\sum_{i=1}^\infty \rho_i \xi_i < \infty$, we get $\underline{\phi}(u_i) \to \phi^*$. □

Lemma 8. *For any $\epsilon > 0$, one can choose a fixed $A = (v^1, ..., v^k)$ such that $\phi(u) - \phi_A(u) \leq \epsilon$ holds for all u. Furthermore, if $\hat{u} = \arg\min_u \phi_A(u)$ is the minimizer of the approximation, then $\phi(\hat{u}) - \phi(u^*) \leq \epsilon$.*

Proof. Since $\mathcal{V}$ is compact and f is continuous, we can find a finite grid A such as a uniform ϵ/l-grid for l-Lipschitz f so that $\phi(u) - \phi_A(u) \leq \epsilon$ for all u. Furthermore, we have

$$\begin{aligned} \phi(\hat{u}) - \phi(u^*) &= \phi(\hat{u}) - \phi_A(\hat{u}) + \phi_A(\hat{u}) - \phi(u^*) \\ &\leq \phi(\hat{u}) - \phi_A(\hat{u}) + \phi_A(u^*) - \phi(u^*) \\ &\leq \phi(\hat{u}) - \phi_A(\hat{u}) \leq \epsilon, \end{aligned}$$

since $\phi_A(u) = \max_{v \in A} f(u,v) \leq \max_{v \in \mathcal{V}} f(u,v) = \phi(u)$ for all u. □

Lemma 9. *Let $\epsilon = \epsilon' + l\delta$ $(\epsilon, \epsilon' \geq 0)$ where l is the Lipschitz coefficient of $f(u,v)$ in v. If u_0 is an ϵ-stationary point of $\phi(u)$, then u_0 is also an ϵ'-stationary point of $\phi_A(u)$.*

Proof. At the ϵ'-stationary point of ϕ_A, we have $\max_{v \in R_A^{\epsilon'}} \langle \nabla_u f(u,v),\ g\rangle \geq 0$ for all g by definition. Since $R^\epsilon(u) = R^{\epsilon'+l\delta}(u) \supseteq R_A^{\epsilon'}(u)$, we have $\max_{v \in R^\epsilon} \langle \nabla_u f(u,v),\ g\rangle \geq \max_{v \in R_A^{\epsilon'}} \langle \nabla_u f(u,v),\ g\rangle \geq 0$ for all g. □

3. GAN training for MNIST

We also trained GANs to generate MNIST images with the K-beam method. The objective function is the same as the MoG experiments, but the generator G and the discriminator networks D are more complex as shown in Table 1.

Table 1. Generator and discriminator networks for GAN-MNIST

(a) Generator

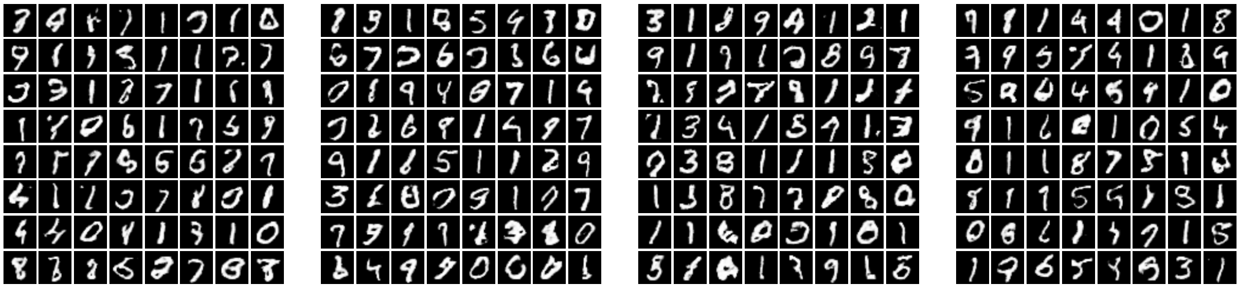
Type	Size
Input	input dim=10
Fully connected	hidden nodes=7x7x64
ReLU	·
Conv transpose	filter size=5x5x32
ReLU	·
Conv transpose	filter size=5x5x1
Sigmoid	output dim=28x28x1

(b) Discriminator

Type	Size
Input	input dim=28x28x1
Conv	filter size=5x5x16
ReLU	·
Max pool	size=2x2, stride=2x2
Conv	filter size=5x5x32
ReLU	·
Max pool	size=2x2, stride=2x2
Fully connected	hidden nodes=50
ReLU	· · ·
Fully connected	output dim=2

The networks are trained with the batch size of 128 using the Adam optimizer with the learning rate of 10^{-3}.

Fig. 3 shows typical training results for $K = 1, 2, 5, 10$ and $J = 1$. Images generated with a larger K look slightly more natural than those with a smaller K. However, an important difference is that GAN training often fails to converge to a good solution due to "mode collapsing" (Nagarajan & Kolter, 2017) when K is small, as observed by an abrupt change in the cost function during optimization.



(a) $K = 1$ (b) $K = 2$ (c) $K = 5$ (d) $K = 10$

Figure 3. MNIST images generated using GAN after 10000 iterations, trained with $K = 1, 2, 5, 10$.

The mode collapsing rarely happens with larger K's such as K=10 with GAN-MNIST. This difference in stability is not directly observable by qualitatively comparing the best generated images from each setting, but it can be measured objectively by average convergence and variance as shown in the figures of the main paper.

References

Boyd, Stephen, Xiao, Lin, and Mutapcic, Almir. Subgradient methods. *lecture notes of EE392o, Stanford University, Autumn Quarter*, 2003.

Correa, Rafael and Lemaréchal, Claude. Convergence of some algorithms for convex minimization. *Mathematical Programming*, 62(1):261–275, 1993.

Dem'yanov, Vladimir Fedorovich and Malozemov, Vassili Nikolaevich. *Introduction to minimax*. John Wiley & Sons, 1974.

Hiriart-Urruty, Jean-Baptiste and Lemaréchal, Claude. *Fundamentals of convex analysis*. Springer, 2001.

Nagarajan, Vaishnavh and Kolter, J Zico. Gradient descent gan optimization is locally stable. In *Advances in Neural Information Processing Systems*, pp. 5591–5600, 2017.