1. Simple surfaces

Fig. 1 shows the six surfaces \( f(u, v) \) and the maximum value function \( \phi(u) = \max_{v \in V} f(u, v) \). From \( \phi(u) \) one can check the minima \( \arg \min_u \phi(u) \) are:

(a) \( u = 0 \), (b) \( u = \frac{\pi}{2} \), (c) \( u = 0 \), (d) \( u = \pm 0.25 \), (e) \( u = 0 \), and (f) \( u = 0 \).

The corresponding maxima \( R(u) = \arg \max_{v \in V} f(u, v) \) 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, \( R(U) \) for the whole domain is:

(a) \( R(U) = \{0\} \), (b) \( R(U) = [-0.5, 0.5] \), (c) \( R(U) = [-0.5, 0.5] \) except for \( R(0) = [-0.5, 0.5] \), (d) \( R(U) = [-0.5, -0.25] \cup \{0\} \), (e) \( R(U) = [-0.5, 0.5] \), and (f) \( R(U) = (-0.5, 0.5) \). These can be verified by solving the minmax 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 minmax point for surfaces (a)-(c), but is neither a saddle nor a minmax 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 \( \partial \phi(u) = \text{co}\{\nabla_u f(u, v) : v \in V\} \). Similarly, suppose \( f(u, v) \) is convex in \( u \) for each \( v \in V \). Then \( \partial \phi(u) = \text{co}\{\nabla_u f(u, v) : v \in V\} \).

Lemma 2 (Chap 3.6, (Dem’yanov & Malozemov, 1974)). A point \( u \) is an \( \epsilon \)-stationary point of \( \phi(u) \) if and only if \( \epsilon \in \text{co}\{\nabla_u f(u, v) : v \in V\} \).

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

Proof. Since \( A \subseteq V \), \( \max_{v \in V} f(u, v) = \max_{v \in R(u)} f(u, v) = \max_{v \in A} f(u, v) \). By \( d_H(R(u), A) = 0 \), we have \( R(u) \subseteq A \) and therefore for each \( v \in R(u) \), \( f(u, v) = \max_{v \in V} f(u, v) = \max_{v \in A} f(u, v) \), so \( v \in R_A(u) \). Conversely, if \( v \in R_A(u) \) then \( f(u, v) = \max_{v \in V} f(u, v) = \max_{v \in A} f(u, v) \), 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 \( \phi(u) - f(u, v') \leq \delta \) 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 \( v' \in R_A(u) \), there is \( v \in R(u) \) such that \( \|v - v'\| \leq \delta \).

Proof. Let any \( v' \in A \) be \( \delta \)-close to a global maximum, then \( f(u, v') \geq \phi(u) - l\delta \). Similarly, let any \( v'' \in A \) be \( \delta \)-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 \( \epsilon \). Therefore, each \( v' \) satisfies \( v' \in R_A(u) \) but no \( v'' \) satisfies \( v'' \in R_A(u) \).

Lemma 6. Suppose \( \delta \) is chosen as in Lemma 5 and \( U \) is bounded (\( \forall u \in U \), \( \|u\| = B < \infty \)). Then any \( z' \in \text{co}\{\nabla_u f(u_0, v) : v \in V\} \) is an \( (2r\delta B) \)-subgradient of \( \phi(u_0) \).

Proof. From Lemmas 4 and 5, for each \( (v^k)' \in R_A \), 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 \( z_k' = \nabla_u f(u_0, (v^k)') \).
Figure 1. Examples of saddle point (upper row) and non-saddle point (lower row) problems. The smallest inset after each surface is the max value function $\phi(u) = \max_v f(u,v)$.

\[
\nabla_u f(u_0, v^k)'. \text{ Then, for all } k = 1, \ldots, |R^*_k| \text{ and for all } u,
\]

\[
\phi(u) - \phi(u_0) - \langle z_k', u - u_0 \rangle
\]

\[
\geq -\|z_k' - z_k\|_2 \|u - u_0\|_2
\]

\[
\geq -\|z_k' - z_k\|_2 \|u - u_0\|_2
\]

\[
\geq -\delta \|u - u_0\|_2 \geq -2\delta B.
\]

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

\[
\phi(u) - \phi(u_0) - \sum_{k=1}^n a_k z_k' \geq -2\delta B,
\]

and therefore any $z^* \in \text{co}(\cup_{v \in R^*_k} \nabla_v f(u_0, v))$ is a $(2\delta B)$-subgradient of $\phi(u_0)$.

**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 $\delta_i = \min \{0.5|\xi_1 - \epsilon_i|/L, 0.5\xi_i / (\rho_i \|B\|)\}$ for some non-negative sequence $(\xi_1, \xi_2, \ldots)$. 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 $\sum_{i=1}^N \rho_i \xi_i < \infty$, then $\min(\phi(u_1), \ldots, \phi(u_i))$ converges to the minimum value $\phi^*$.

Note that a stronger result such as $\lim_{i \to \infty} \phi(u_i) = \phi^*$ 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 $\epsilon$-subgradient descent method. Let $u_{i+1} = u_i - \rho_i g_i$. Then,

\[
\|u_{i+1} - u^*\|
\]

\[
\|
\|u_i - u^*\|_2^2 + \rho_i^2 \|g_i\|_2^2 + 2\rho_i (\phi(u^*) - \phi(u_i) + \xi_i)
\]

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

\[
\|u_{N+1} - u^*\|_2^2 \leq \|u_1 - u^*\|_2^2 + \sum_{i=1}^N \rho_i^2 \|g_i\|_2^2
\]

or equivalently,

\[
2 \sum_{i=1}^N (\rho_i (\phi(u_i) - \phi(u^*) - \xi_i) \leq \|u_1 - u^*\|_2^2 + \sum_{i=1}^N \rho_i^2 \|g_i\|_2^2.
\]
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 \( \epsilon \)-stationary point of \( \phi(u) \), then \( u_0 \) is also an \( \epsilon' \)-stationary point of \( \phi_A(u) \).

**Proof.** At the \( \epsilon' \)-stationary point of \( \phi_A \), we have \( \max_{v \in R'_A} \langle \nabla_u f(u, v), g \rangle \geq 0 \) for all \( g \) by definition. Since \( R'_A(u) = R'_A(u) + l \delta \), we have \( \max_{v \in R_A} \langle \nabla_u f(u, v), g \rangle \geq \max_{v \in R'_A} \langle \nabla_u f(u, v), g \rangle \geq 0 \) 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>
<thead>
<tr>
<th>Type</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>input dim=10</td>
</tr>
<tr>
<td>Fully connected</td>
<td>hidden nodes=7x7x64</td>
</tr>
<tr>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td>Conv transpose</td>
<td>filter size=5x5x32</td>
</tr>
<tr>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td>Conv transpose</td>
<td>filter size=5x5x1</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>output dim=28x28x1</td>
</tr>
</tbody>
</table>

(a) Generator

<table>
<thead>
<tr>
<th>Type</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>input dim=28x28x1</td>
</tr>
<tr>
<td>Conv</td>
<td>filter size=5x5x16</td>
</tr>
<tr>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td>Max pool</td>
<td>size=2x2, stride=2x2</td>
</tr>
<tr>
<td>Conv</td>
<td>filter size=5x5x32</td>
</tr>
<tr>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td>Max pool</td>
<td>size=2x2, stride=2x2</td>
</tr>
<tr>
<td>Fully connected</td>
<td>hidden nodes=50</td>
</tr>
<tr>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td>Fully connected</td>
<td>output dim=2</td>
</tr>
</tbody>
</table>

(b) Discriminator

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.
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


