Adversarial Flows

A gradient flow interpretation of adversarial attacks

Lukas Weigand, Tim Roith, Martin Burger
Friedrich-Alexander Universität Erlangen-Nürnberg, Department Mathematik

May 3, 2023
Adversarial Attacks

Adversarial attack

“panda” 57.7% confidence

+\epsilon

= “gibbon” 99.3% confidence

[GSS14]
Finding adversarial examples

For \( x \in X = \mathbb{R}^n \) and \( y \in Y = \mathbb{R}^m \) and \( \mu \in \mathcal{P}(X \times Y) \) the weights \( \theta \) of a neural network \( f_\theta : X \rightarrow Y \) are trained by minimizing

\[
\min_\theta \mathbb{E}_{(x,y) \sim \mu} l(f_\theta(x), y).
\]

Under the assumption \( y_x \approx y_{x_0} \) in vicinity to \( x_0 \) adversarial examples can be found by solving

\[
\max_{x: \|x-x_0\| \leq \epsilon} \underbrace{l(f_\theta(x), y_{x_0})}_{:=L(x,y)}.
\]
For \( x \in X = \mathbb{R}^n \) and \( y \in Y = \mathbb{R}^m \) and \( \mu \in \mathcal{P}(X \times Y) \) the weights \( \theta \) of a neural network \( f_\theta : X \to Y \) are trained by minimizing

\[
\min_{\theta} \mathbb{E}_{(x,y) \sim \mu} l(f_\theta(x), y)).
\]

Under the assumption \( y_x \approx y_{x_0} \) in vicinity to \( x_0 \) adversarial examples can be found by solving

\[
\max_{x: \|x-x_0\| \leq \epsilon} l(f_\theta(x), y_{x_0}).
\]

(1)

Linearizing the loss in \( x_0 \) we get

\[
l(f_\theta(x), y_{x_0}) \approx l(f_\theta(x_0), y_{x_0}) + \langle x - x_0, \nabla_x l(f_\theta(x_0), y_{x_0}) \rangle.
\]

Thus (1) is roughly equivalent to

\[
\max_{x: \|x-x_0\| \leq \epsilon} \langle x - x_0, \nabla_x l(f_\theta(x), y_{x_0}) \rangle = \max_{x: \|x-x_0\| \leq 1} \left( \frac{x - x_0}{\epsilon}, \nabla_x l(f_\theta(x_0), y_{x_0}) \right).
\]
## Fast Gradient Methods

<table>
<thead>
<tr>
<th></th>
<th>FGM (Fast gradient method)</th>
<th>FGSM (Fast gradient sign method)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Maximization Problem</strong></td>
<td>( \arg \max_{x: |x-x_0|_2 \leq \epsilon} L(x, y) )</td>
<td>( \arg \max_{x: |x-x_0|_\infty \leq \epsilon} L(x, y) )</td>
</tr>
<tr>
<td><strong>Scheme</strong></td>
<td>( x = x_0 + \epsilon \frac{\nabla L(x, y)}{|\nabla L(x, y)|_2} )</td>
<td>( x = x_0 + \epsilon \text{sign}(\nabla L(x, y)) )</td>
</tr>
</tbody>
</table>
| **ODE** | \( \partial_t x(t) = \frac{\nabla L(x, y)}{\|\nabla L(x, y)\|_2} \) | \( \partial_t x(t) \in \text{sign}(\nabla L(x, y)) \)
A smooth energy $\phi : \mathbb{R}^d \to \mathbb{R}$ can decrease along a smooth curve $u : [0, T] \to \mathbb{R}^d$ at most by

$$\frac{d}{dt} \phi(u(t)) = \nabla \phi(u(t)) \cdot \frac{d}{dt} u(t) \geq -\|\nabla \phi(u(t))\| \left\| \frac{d}{dt} u(t) \right\|_* \geq -\frac{1}{q} \|\nabla \phi(u(t))\|^q - \frac{1}{p} \left\| \frac{d}{dt} u(t) \right\|_*^p,$$

where $\frac{1}{p} + \frac{1}{q} = 1$. Curves of maximal slope archive this maximum by satisfying

$$\frac{d}{dt} \phi(u(t)) \leq -\frac{1}{q} \|\nabla \phi(u(t))\|^q - \frac{1}{p} \left\| \frac{d}{dt} u(t) \right\|_*^p.$$
p-curves of maximal slope

A smooth energy \( \phi : \mathbb{R}^d \to \mathbb{R} \) can decrease along a smooth curve \( u : [0, T] \to \mathbb{R}^d \) at most by

\[
\frac{d}{dt} \phi(u(t)) = \nabla \phi(u(t)) \cdot \frac{d}{dt} u(t) \geq -\|\nabla \phi(u(t))\| \left\| \frac{d}{dt} u(t) \right\|_* \geq -\frac{1}{q} \|\nabla \phi(u(t))\|^q - \frac{1}{p} \left\| \frac{d}{dt} u \right\|_*^p,
\]

where \( \frac{1}{p} + \frac{1}{q} = 1 \). Curves of maximal slope archive this maximum by satisfying

\[
\frac{d}{dt} \phi(u(t)) \leq -\frac{1}{q} \|\nabla \phi(u(t))\|^q - \frac{1}{p} \left\| \frac{d}{dt} u(t) \right\|_*^p.
\]

Generalization to Metric Spaces \((S, d(\cdot, \cdot))\) in [AGS05]:

<table>
<thead>
<tr>
<th>Metric Slope and Velocity</th>
</tr>
</thead>
<tbody>
<tr>
<td>(</td>
</tr>
<tr>
<td>(</td>
</tr>
</tbody>
</table>
Definition

A Lipschitz curve \( u(t) : [0, T] \to S \) is an \( \infty \)-curve of maximal slope for some energy \( \phi : S \to (0, \infty) \), if \( \phi \circ u \) is equivalent to a nonincreasing map \( \varphi \) and

\[
\varphi'(t) \leq -\frac{\partial \phi(t)}{|u(t)|} \quad \text{for a.e. } t \in [0, T]
\]

\[|u(t)| \leq 1\]
Definition

A Lipschitz curve $u(t) : [0, T] \to S$ is a $\infty$-curve of maximal slope for some energy $\phi : S \to (0, \infty]$, if $\phi \circ u$ is equivalent to a nonincreasing map $\varphi$ and

$$\varphi'(t) \leq -|\partial \phi(t)| \quad \text{for a.e. } t \in [0, T]$$

Minimizing Movement

$$U_{n+1} \in \arg\min_u \frac{1}{p \tau^{p-1}} d^p(u, U_n) + \phi(u) \quad \rightarrow U_{n+1} \in \arg\min_{u : d(u, U_n) \leq \tau} \phi(u)$$
\(\infty\)-curves of maximal slope in reflexive Banach spaces

The application of the chain rule to \(\infty\)-curves of maximum slope we obtain:

\[
\frac{d}{dt} \phi(u(t)) = \langle D\phi(u(t)), u'(t) \rangle = -\|D\phi(u)\|_\ast \|u'(t)\| \leq 1 \quad \text{for a.e. } t \in [0, T]
\]
$\infty$-curves of maximal slope in reflexive Banach spaces

The application of the chain rule to $\infty$-curves of maximum slope we obtain:

$$\frac{d}{dt} \phi(u(t)) = \langle D\phi(u(t)), u'(t) \rangle = -\|D\phi(u)\|_*$$

for a.e. $t \in [0, T]$\n
$$\|u'(t)\| \leq 1$$

By the characterisation of subdifferential of convex function this is equivalent to [RMS08]:

$\infty$-curves of maximum slope in reflexive Banach spaces

$$u'(t) \in \partial \| \cdot \|_\star (-D\phi(u(t)))$$

for a.e. $t \in [0, T]$.

For $(\mathbb{R}^d, \| \cdot \|_2)$ and $(\mathbb{R}^d, \| \cdot \|_\infty)$ with $\phi(x) = -L(x, y)$ we recover FGM and FGSM:

$$\frac{d}{dt} u(t) \in \partial \| \cdot \|_\star (-\nabla \phi(u(t))) \quad \Rightarrow \quad \begin{cases} \partial_t x(t) = \frac{\nabla L(x,y)}{\|\nabla L(x,y)\|_2} \\ \partial_t x(t) \in \text{sign}(\nabla L(x,y)) \end{cases}$$
Adversarial training

To approximate the data distribution $\mu \in \mathcal{P}(X \times Y)$ by a robust neural network its weights $\theta$ are trained by solving

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mu} \left[ \max_{\tilde{x} \in B_{\epsilon}(x)} l(f_{\theta}(\tilde{x}), y) \right].$$

In the smooth case this problem can be rewritten as

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mu} \left[ G(\mu, \mu^0) \leq \epsilon \right],$$

where $G$ is the optimal transport distance of the form

$$G(\mu, \tilde{\mu}) := \inf_{\gamma \in \Gamma(\mu, \tilde{\mu})} \gamma - \text{ess sup}_{(x,y), (\tilde{x}, \tilde{y})} c((x,y), (\tilde{x}, \tilde{y})).$$

with $c((x,y), (\tilde{x}, \tilde{y})) := \|x - \tilde{x}\|_2$ if $y = \tilde{y} + \infty$ if $y \neq \tilde{y}$.
Adversarial training

To approximate the data distribution $\mu \in \mathcal{P}(X \times Y)$ by a robust neural network its weights $\theta$ are trained by solving

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mu} \left[ \max_{\tilde{x} \in B_\epsilon(x)} l(f_{\theta}(\tilde{x}), y) \right].$$

In the smooth case this problem can be rewritten as

$$\min_{\theta} \max_{\mu : G(\mu, \mu_0) \leq \epsilon} \mathbb{E}_{(x,y) \sim \mu} L(x, y)$$

where $G$ is the optimal transport distance of the form

$$G(\mu, \tilde{\mu}) := \inf_{\gamma \in \Gamma(\mu, \tilde{\mu})} \gamma - \text{ess sup} \ c((x, y), (\tilde{x}, \tilde{y}))$$

with

$$c((x, y), (\tilde{x}, \tilde{y})) := \begin{cases} \|x - \tilde{x}\| & \text{if } y = \tilde{y} \\ +\infty & \text{if } y \neq \tilde{y}. \end{cases}$$
Let \((X, \| \cdot \|)\) be a separable and reflexive Banach space and \(\mu, \tilde{\mu} \in P_{\infty}(X)\) then

\[
W_{\infty}(\mu, \tilde{\mu}) := \inf_{\gamma \in \Gamma(\mu, \tilde{\mu})} \gamma - \text{ess sup} \| x - y \|.
\]
Let $(X, \| \cdot \|)$ be a separable and reflexive Banach space and $\mu, \tilde{\mu} \in \mathcal{P}_\infty(X)$ then

$$W_\infty(\mu, \tilde{\mu}) := \inf_{\gamma \in \Gamma(\mu, \tilde{\mu})} \gamma - \text{ess sup} \|x - y\|.$$ 

Since $G$ does not allow movement of mass along the $y$ direction the inner maximization problem is equivalent to

**Distributional adversary**

$$\max_{\tilde{\mu} : W_\infty(\mu, \tilde{\mu}) \leq \epsilon} \int l(f_\theta(x), y) d\mu_y \quad \text{for } \nu \text{ a.e. } y \in Y \text{ and } d\mu = d\mu_y d\nu$$
Lipschitz curves in $W_\infty$

**Lemma**

If in addition $X$ satisfies the satisfying the bounded approximation property, i.e. there exists a sequence of finite rank linear operators $T_n : X \to X$ such that

$$\lim_{n \to \infty} \|T_n(x) - x\| = 0 \quad \forall x \in X.$$ 

and $\mu \in AC^\infty([0, T], W_\infty)$ then for $L^1$-a.e. $t \in [0, T]$ there exists a velocity field $v_t \in L^\infty(\mu_t, X)$ such that:

$$\frac{d}{dt} \int_X \varphi d\mu_t = \int_X \langle D\varphi, v_t \rangle d\mu_t \quad \forall \varphi \in C^1_b(x)$$

holds in the sense of distributions and

$$\|v_t\|_{L^p(\mu_t; X)} = |\mu|'(t) \text{ for } L^1 \text{- a.e. } t \in [0, T].$$

[Lis07]
Properties of potential energies

For $C^1$ smooth potential the slope of the potential energy

$$\phi(\mu) := \int V(x) d\mu(x)$$

- Slop: $|\partial \phi|(\mu) = \int |DV(x)| d\mu(x)$
- Chain rule: $\frac{d}{dt} \phi(\mu_t) = \int \langle DV(x), v_t(x) \rangle d\mu_t(x)$
$\infty$-curve of maximal slope for potential energies

$\infty$-curves of maximal slope satisfy

$$
\frac{d}{dt} \phi(\mu_t) = \int \langle DV(x), v_t(x) \rangle d\mu_t(x) = -\|v_t\|_{L(\mu_t; X)} \quad \text{for a.e. } t \in [0, T]
$$

$\iff$

$$
v_t(x) \in - \arg \max_{v:\|v\|=1} \langle DV(x), v_t(x) \rangle \quad \text{for } \mu_t \text{ a.e. } x \in X \text{ and a.e. } t \in [0, T]
$$

$\iff$

$$
v_t(x) \in \partial \| \cdot \|_\ast(-DV(x)) \quad \text{for } \mu_t \text{ a.e. } x \in X \text{ and a.e. } t \in [0, T]
$$

Curve of maximal slope

$$
\partial_t u_t \in \text{div}(\partial \| \cdot \|_\ast(-DV) u_t)
$$
Conclusion

- FGM and FGSM are essentially explicit Euler discretisation of $\infty$-curve of maximum slope in $(\mathbb{R}^d, \| \cdot \|)$.
- If $\partial\| \cdot \|_\ast$ has a closed form we can compute adversarial attacks w.r.t. any norm $\| \cdot \|$.
- Computing the corresponding ODE for each data point $x_i$ of the empirical distribution $\mu_y = \sum_{i=1}^{n} \delta_{x_i}$ yields the curve of maximum slope in $W_\infty$.
References


