Optimization for Data Science

Lec 02': Projection

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

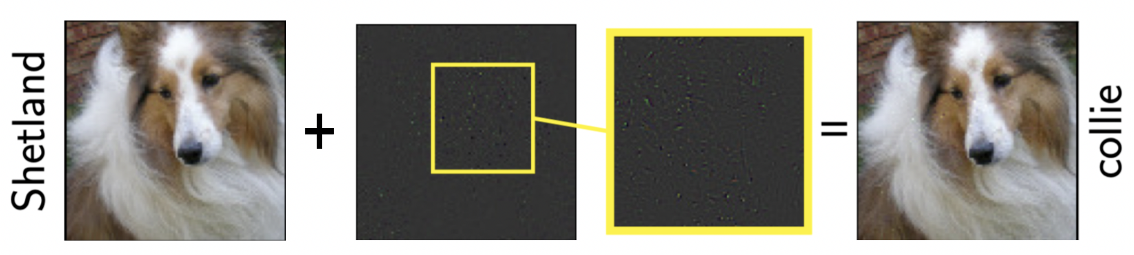
Problem

Constrained smooth minimization:

$$f_\star = \inf_{\mathbf{w} \in C} f(\mathbf{w}).$$

- Constraint on the domain: closed set $C \subseteq \mathbb{R}^d$
- $f : \mathbb{R}^d \to \mathbb{R}$ is smooth, e.g. continuously differentiable
- f can be convex or nonconvex; C can be convex or nonconvex
- Minimizer may or may not be attained
- Maximization is just negation

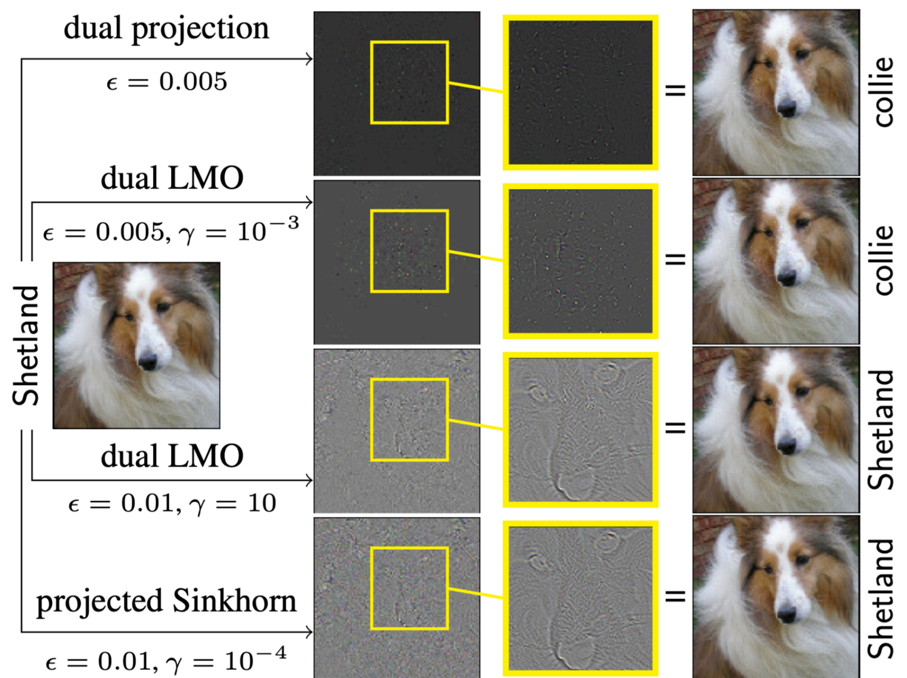
White-box Adversarial Attacks



- Mathematically, a neural network is a function $f(\mathbf{w}; \mathbf{x})$
- Typically, input $\mathbf{x}$ is given and network weights $\mathbf{w}$ optimized
- Could also freeze weights $\mathbf{w}$ and optimize $\mathbf{x}$, adversarially!

$$\min_{\boldsymbol{\delta}} \ \text{size}(\boldsymbol{\delta}) \quad \text{s.t.} \quad \text{pred}[f(\mathbf{w}; \mathbf{x} + \boldsymbol{\delta})] \neq \mathsf{y}$$

- More generally: $\max_{\boldsymbol{\delta}} \ \ell(\mathbf{w}; \mathbf{x} + \boldsymbol{\delta}, \mathsf{y}) \quad \text{s.t.} \quad \text{size}(\boldsymbol{\delta}) \leq \epsilon \ \text{ and } \ \mathbf{0} \leq \mathbf{x} + \boldsymbol{\delta} \leq \mathbf{1}$



dual projection
$\epsilon = 0.005$
dual LMO
$\epsilon = 0.005, \gamma = 10^{-3}$
Shetland
dual LMO
$\epsilon = 0.01, \gamma = 10$
projected Sinkhorn
$\epsilon = 0.01, \gamma = 10^{-4}$
=
collie
=
collie
=
Shetland
=
Shetland

Convexity

A point set $C \subseteq \mathbb{R}^d$ is convex iff for any $\mathbf{w}, \mathbf{z} \in C$, the line segment $[\mathbf{w}, \mathbf{z}] \subseteq C$.

The epigraph of a function $f : \mathbb{R}^d \to (-\infty, \infty]$ is defined as the set

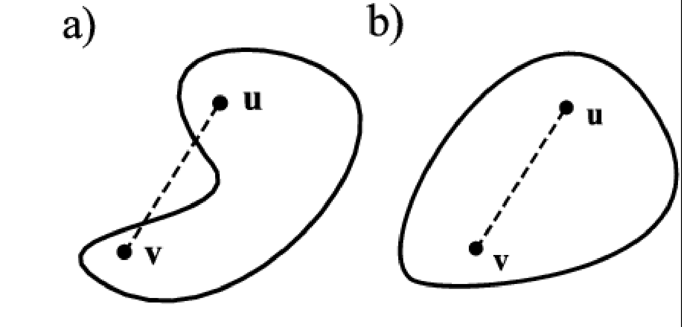
$$\text{epi}\, f := \{(\mathbf{w}, t) \in \mathbb{R}^{d+1} : f(\mathbf{w}) \leq t\}$$

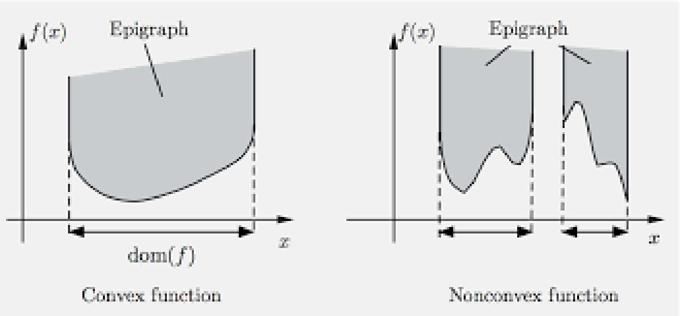
A function $f : \mathbb{R}^d \to (-\infty, \infty]$ is convex iff its epigraph is a convex set, or equivalently

$$\forall \mathbf{w}, \forall \mathbf{z}, \forall \lambda \in [0, 1], \quad f(\lambda \mathbf{w} + (1 - \lambda)\mathbf{z}) \leq \lambda f(\mathbf{w}) + (1 - \lambda) f(\mathbf{z})$$

Theorem: second-order test for convexity

f is convex iff $\nabla^2 f$ is positive semidefinite.


a)
u
v
b)
u
v


f(x)
Epigraph
x
dom(f)
Convex function
f(x)
Epigraph
x
Nonconvex function

Calculus of Convexity

- f, g convex $\implies \alpha \cdot f + \beta \cdot g$ is convex for any $\alpha, \beta \geq 0$
- f convex $\implies f(A\mathbf{w})$ is convex
- f convex increasing and g convex $\implies f \circ g$ is convex
- f convex $\implies (\mathbf{w}, t > 0) \mapsto t f(\mathbf{w}/t)$ is convex
- f_t convex $\implies f = \sup_t f_t$ is convex
- $f(\mathbf{w}, \mathbf{z})$ convex $\implies g = \min_{\mathbf{z}} f(\mathbf{w}, \mathbf{z})$ is convex
- Is $\log(\sum_j \exp(w_j))$ convex?

A Nice Univariate Result

Theorem: constrained univariate convex minimization

For any univariate convex function f and convex interval $C = [a, b]$, we have

$$\mathrm{P}_C\left(\operatorname*{argmin}_{w \in \mathbb{R}} f(w)\right) \subseteq \operatorname*{argmin}_{w \in C} f(w),$$

where $\mathrm{P}_C(w) = \mathrm{P}_{[a,b]}(w) = (a \vee w) \wedge b$ is the closest point in C to w.

- Not true if C is not an interval (i.e. not convex)
- Not true if f is not convex
- Not true when dimension $d \geq 2$, even when both f and C are convex
- Except when $\operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}) \subseteq C$

An Algorithm that does NOT work

$$\eta \leftarrow \operatorname*{argmin}_{\eta \geq 0} f(\mathbf{w}_\eta), \quad \text{s.t.} \quad \mathbf{w}_\eta := \mathbf{w} - \eta \cdot \nabla f(\mathbf{w}) \in C$$

$$\mathbf{w} \leftarrow \mathbf{w}_\eta$$

- Does NOT work
  - $f(\mathbf{w}) := \frac{1}{2}(w_1^2 + w_2^2)$
  - $C = \{\mathbf{w} \geq \mathbf{0} : w_1 + w_2 = 1\}$
  - stuck at $\mathbf{w} = (1, 0)$ while minimum is at $\mathbf{w}_\star = (\frac{1}{2}, \frac{1}{2})$
- Important to leave the constraint set C

(Euclidean) Projection

Let $C \subseteq \mathbb{R}^d$ be a closed set. The Euclidean projection of a point $\mathbf{w} \in \mathbb{R}^d$ to C is:

$$\mathrm{P}_C(\mathbf{w}) := \operatorname*{argmin}_{\mathbf{z} \in C} \|\mathbf{z} - \mathbf{w}\|_2,$$

i.e. the point(s) in C that are closest to the given point $\mathbf{w}$.

- We always have $\mathrm{P}_C(\mathbf{w}) \neq \emptyset$ and compact
- $\mathrm{P}_C(\mathbf{w}) = \mathbf{w}$ iff $\mathbf{w} \in C$
- $\mathrm{P}_C(\mathbf{w}) = \mathrm{bd}\, C$ if $\mathbf{w} \notin C$
- In $\mathbb{R}^d$, P_C is unique iff C is convex

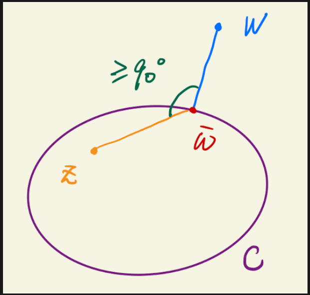
Geometrically

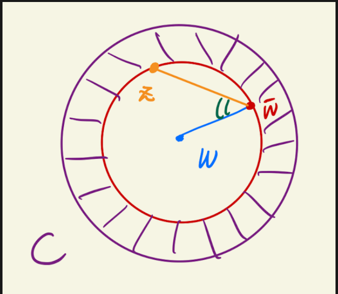
Theorem:

If C is convex, then $\bar{\mathbf{w}} = \mathrm{P}_C(\mathbf{w})$ iff for all $\mathbf{z} \in C$

$$\langle \mathbf{z} - \bar{\mathbf{w}}, \mathbf{w} - \bar{\mathbf{w}} \rangle \leq 0,$$

or equivalently, $\frac{1}{2}\|\mathbf{z} - \mathbf{w}\|_2^2 \geq \frac{1}{2}\|\mathbf{z} - \bar{\mathbf{w}}\|_2^2 + \frac{1}{2}\|\bar{\mathbf{w}} - \mathbf{w}\|_2^2$.





Example: Projection to the hypercube

$$\min_{\mathbf{a} \le \boldsymbol{\delta} \le \mathbf{b}} \|\boldsymbol{\delta} - \boldsymbol{\gamma}\|_2 \quad = \quad \min_{\mathbf{a} \le \boldsymbol{\delta} \le \mathbf{b}} \|\boldsymbol{\delta} - \boldsymbol{\gamma}\|_2^2$$

- Problem is separable: reduce to each dimension separately
- Apply the nice univariate result $\boldsymbol{\delta}_\star = (\boldsymbol{\gamma} \vee \mathbf{a}) \wedge \mathbf{b}$

Example: Projection to the ball

$$\min_{\|\mathbf{z}\|_2 \le \lambda} \|\mathbf{w} - \mathbf{z}\|_2 \quad = \quad \min_{\|\mathbf{z}\|_2 \le \lambda} \|\mathbf{w} - \mathbf{z}\|_2^2$$

- Decompose $\mathbf{z} = r \cdot \bar{\mathbf{z}}$, where $r \ge 0, \|\bar{\mathbf{z}}\|_2 = 1$
- Apply the nice univariate result $\mathbf{w}_\star = \left(\frac{\lambda}{\|\mathbf{w}\|_2} \wedge 1 \right) \cdot \mathbf{w}$

Algorithm 1: Projected gradient descent for constrained smooth minimization

Input: $\mathbf{w}_0 \in \mathbb{R}^d$, constraint $C \subseteq \mathbb{R}^d$, smooth function $f : \mathbb{R}^d \to \mathbb{R}$

for $t = 0, 1, \ldots$ **do**

  $\mathbf{g}_t \leftarrow \nabla f(\mathbf{w}_t)$ // compute the gradient

  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \mathbf{g}_t$ // η_t is the step size

  $\mathbf{w}_{t+1} \leftarrow \mathrm{P}_C(\mathbf{w}_{t+1})$ // project back to the constraint

- $C = \mathbb{R}^d$: reduces to gradient descent
- Motivation from L-smoothness:

$$
\begin{aligned}
f(\mathbf{w}) &\le f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, \nabla f(\mathbf{w}_t) \rangle + \tfrac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2 \\
&= \tfrac{1}{2\eta_t} \|\mathbf{w} - (\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t))\|_2^2 + f(\mathbf{w}_t) - \tfrac{\eta_t}{2} \|\nabla f(\mathbf{w}_t)\|_2^2
\end{aligned}
$$

A. A. Goldstein. "Convex programming in Hilbert space". *Bulletin of the American Mathematical Society*, vol. 70, no. 5 (1964), pp. 709–710, E. S. Levitin and B. T. Polyak. "Constrained Minimization Methods". *USSR Computational Mathematics and Mathematical Physics*, vol. 6, no. 5 (1966), pp. 1–50. [English translation in *Zh. Vȳchisl. Mat. mat. Fiz.* vol. 6, no. 5, pp. 787–823, 1965].