Machine Learning Algorithms - Convex analysis

Caio Corro

Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France

Motivations

Why do we care about convexity?

- Local optimal solutions are also global optimal solutions
- Derive inequality/bounds
  (for example remember that we used $\log u \leq u - 1$)

Applications

- Loss functions
- Regularization functions
- Prediction functions (later)

=> study the properties of these functions to derive algorithms

Convex sets and functions

Sets

Convex set

A set $U \subseteq \mathbb{R}^n$ is convex if and only if:

$$\forall \boldsymbol{u}, \boldsymbol{u}' \in U, \epsilon \in [0,1]: \quad \underbrace{\epsilon \boldsymbol{u} + (1-\epsilon)\boldsymbol{u}'}_{\text{convex combination}} \in U$$

Convex hull

The convex hull of a set U, denoted $\textbf{conv}\, U$,
is the smallest convex set that contains U.

$$\textbf{conv}\, U = \{\epsilon \boldsymbol{u} + (1-\epsilon)\boldsymbol{u}' \mid \boldsymbol{u}, \boldsymbol{u}' \in U \text{ and } \epsilon \in [0,1]\}$$

Example: $\textbf{conv}\, E(k) = \triangle(k)$

Functions

Convex function (synthetic definition)

A function $f : U \to \mathbb{R}$ is convex if and only if:

1. U is a convex set ;
2. $\forall \boldsymbol{u}, \boldsymbol{u}' \in U, \epsilon \in [0, 1]$:

$$f(\underbrace{\epsilon \boldsymbol{u} + (1-\epsilon)\boldsymbol{u}'}_{\text{dom. needs to be conv.}}) \leq \epsilon f(\boldsymbol{u}) + (1-\epsilon) f(\boldsymbol{u}')$$

Concave function

A function f is concave if and only $-f$ is convex.

Functions

Strictly convex function

A function $f : U \to \mathbb{R}$ is strictly convex if and only if:

1. U is a convex set ;
2. $\forall \boldsymbol{u}, \boldsymbol{u}' \in U$ s.t. $\boldsymbol{u} \neq \boldsymbol{u}', \epsilon \in]0, 1[$:

$$f(\underbrace{\epsilon \boldsymbol{u} + (1-\epsilon)\boldsymbol{u}'}_{\text{dom. needs to be conv.}}) < \epsilon f(\boldsymbol{u}) + (1-\epsilon) f(\boldsymbol{u}')$$

Functions

Hessian

Let $f : U \to \mathbb{R}$ be a twice differentiable function, where $U \subseteq R^n$.
The Hessian of f at $\boldsymbol{u} \in U$ is defined as:

$$\nabla^2 f(\boldsymbol{u}) = \begin{bmatrix} \frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1} f(\boldsymbol{u}), & \frac{\partial}{\partial u_1}\frac{\partial}{\partial u_2} f(\boldsymbol{u}), & \dots \\ \frac{\partial}{\partial u_2}\frac{\partial}{\partial u_1} f(\boldsymbol{u}), & \ddots & \\ \vdots & & \ddots \end{bmatrix},$$

that is: $[\nabla^2 f(\boldsymbol{u})]_{i,j} = \frac{\partial}{\partial u_i}\frac{\partial}{\partial u_j} f(\boldsymbol{u})$

Functions

Positive semi-definite matrix

A matrix $\boldsymbol{H} \in \mathbb{R}^{n \times n}$ is pos. semi-def. if and only if:

$$\forall \boldsymbol{u} \in \mathbb{R}^n : \quad \langle \boldsymbol{u}, \boldsymbol{H}\boldsymbol{u} \rangle \geq 0$$

Convex function (analytic definition)

A differentiable function $f : U \to \mathbb{R}$ is convex if and only if:

1. $U \subset \mathbb{R}^n$ is a convex set ;
2. $\forall \boldsymbol{u} \in U : \nabla^2 f(\boldsymbol{u})$ is a pos. semi-def. matrix.

If $n = 1$, the second definition simplifies to $\forall \boldsymbol{u} \in U : f''(\boldsymbol{u}) \geq 0$

Other important properties

Proper function

A $f : U \to \mathbb{R} \cup \{-\infty, +\infty\}$ is proper if and only if:

1. $\forall \boldsymbol{u} \in U : f(\boldsymbol{u}) \neq -\infty$,
2. $\exists \boldsymbol{u} \in U$ s.t. $f(\boldsymbol{u}) \neq +\infty$.

Closed function

A function is closed if and only if its epigraph is closed. This property is equivalent to lower semi-continuity.
=> You can simply ignore this for this course.

Extended real value extension

Let $f : U \to \mathbb{R}$, $U \subseteq \mathbb{R}^n$ be a function.
The e.r.v. extension of f is the function $\widetilde{f} : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ defined as follows:

$$\widetilde{f}(\boldsymbol{u}) = \begin{cases} f(\boldsymbol{u}) & \text{if } \boldsymbol{u} \in U, \text{or equivalently } \boldsymbol{u} \in \mathbf{dom}\, f, \\ +\infty & \text{otherwise.} \end{cases}$$

We define $\mathbf{dom}\, \widetilde{f} = \{\boldsymbol{u} \in \mathbb{R}^n \mid \widetilde{f}(\boldsymbol{u}) \neq \infty\}$.

Property

If f is convex, then $\tilde{f}$ is also convex (to prove).

Notation

In general, we just assume we directly manipulate the e.r.v. extension, i.e. $f = \widetilde{f}$.

Extended real value extension

Indicator function

Let S be a set. The indicator function of S is defined as:

$$\delta_S(\boldsymbol{s}) = \begin{cases} 0 & \text{if } \boldsymbol{s} \in S, \\ +\infty & \text{otherwise.} \end{cases}$$

Application

Transform a constrained optimization problem into an "unconstrained" problem:

$$\begin{aligned} \min_{\boldsymbol{u} \in \mathbb{R}^n} \; & f(\boldsymbol{u}) \\ \text{s.t. } & \boldsymbol{u} \in S \end{aligned} \qquad = \qquad \min_{\boldsymbol{u} \in \mathbb{R}^n} \; f(\boldsymbol{u}) + \delta_S(\boldsymbol{u})$$

Operations preserving convexity and closedness

Operations on set of functions

Weighted sum of functions (Beck, th. 2.7 & 2.16)

Let $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}, i \in \{1, ..., m\}$, be a set of convex (closed) functions, and $\alpha_1, ..., \alpha_m \geq 0$.

Then, the function $f(\boldsymbol{u}) = \sum_{i=1}^{m} \alpha_i f_i(\boldsymbol{u})$ is convex (closed).

Operations on set of functions

Weighted sum of functions (Beck, th. 2.7 & 2.16)

Let $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}, i \in \{1, ..., m\}$, be a set of convex (closed) functions, and $\alpha_1, ..., \alpha_m \geq 0$.

Then, the function $f(\boldsymbol{u}) = \sum_{i=1}^{m} \alpha_i f_i(\boldsymbol{u})$ is convex (closed).

Maximum of functions (Beck, th. 2.7 & 2.16)

Let $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}, i \in \{1, ..., m\}$, be a set of convex (closed) functions. Then, the function $f(\boldsymbol{u}) = \max(f_1(\boldsymbol{u}), ..., f_m(\boldsymbol{u}))$ is convex (closed).

Example: maximum of affine functions.

Linear transformation (Beck, th. 2.7 & 2.16)

Let:

- $\boldsymbol{A} \in \mathbb{R}^{m \times n}$,
- $\boldsymbol{b} \in \mathbb{R}^{m}$,
- $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ be a convex (closed) function.

Then, the function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ defined as follows:

$$h(\boldsymbol{u}) = f(\boldsymbol{A}\boldsymbol{u} + \boldsymbol{b})$$

is convex.

Gradients

Scalar input

Derivative

Let $f : \mathbb{R} \to \mathbb{R}$ be a function and $u, w \in \mathbb{R}$ be variables such that:

$$w = f(u).$$

For a given u, how does an infinitesimal change of u impact w?

Scalar input

Derivative

Let $f : \mathbb{R} \to \mathbb{R}$ be a function and $u, w \in \mathbb{R}$ be variables such that:

$$w = f(u).$$

For a given u, how does an infinitesimal change of u impact w?

$$\frac{\partial w}{\partial u} = f'(u) = \lim_{\epsilon \to 0} \frac{f(u+\epsilon) - f(u)}{\epsilon}$$

Scalar input

Derivative

Let $f : \mathbb{R} \to \mathbb{R}$ be a function and $u, w \in \mathbb{R}$ be variables such that:

$$w = f(u).$$

For a given u, how does an infinitesimal change of u impact w?

$$\frac{\partial w}{\partial u} = f'(u) = \lim_{\epsilon \to 0} \frac{f(u+\epsilon) - f(u)}{\epsilon}$$

Linear approximation

Let $h : \mathbb{R} \to \mathbb{R}$ be function parameterized by $a \in \mathbb{R}$ defined as follows:

$$h_a(u) = f(a) + f'(a) \cdot (u - a)$$

Then, h_a is an approximation of f for u "close to" a.

Scalar input

Example

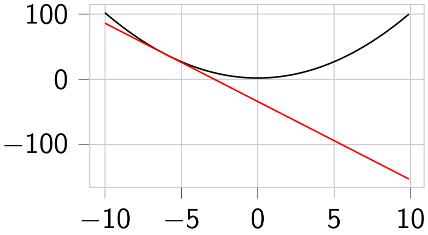
$$
\begin{aligned}
f(u) &= u^2 + 2 \\
f'(u) &= 2u \\
h_a(u) &= f(a) + f'(a) \cdot (u - a) \\
&= a^2 + 2 + 2a(u - a) \\
&= 2au + 2 - a^2
\end{aligned}
$$

Scalar input

Example

$$
\begin{aligned}
f(u) &= u^2 + 2 \\
f'(u) &= 2u \\
h_a(u) &= f(a) + f'(a) \cdot (u - a) \\
&= a^2 + 2 + 2a(u - a) \\
&= 2au + 2 - a^2
\end{aligned}
$$

Intuition: the sign of $f'(u)$ gives the slope of the approximation, we could use this information to move closer to the minimum of $f(u)$.

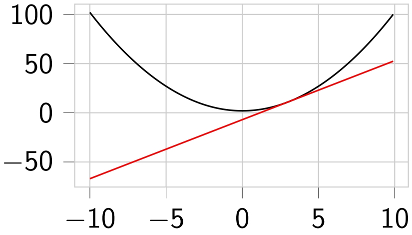


- $a = -6$
- Black: $f(u)$
- Red: $h_{-6}(u)$

Scalar input

Example

$$f(u) = u^2 + 2$$
$$f'(u) = 2u$$
$$h_a(u) = f(a) + f'(a) \cdot (u - a)$$
$$= a^2 + 2 + 2a(u - a)$$
$$= 2au + 2 - a^2$$



- $a = 3$
- Black: $f(a)$
- Red: $h_3(u)$

Intuition: the sign of $f'(u)$ gives the slope of the approximation, we could use this information to move closer to the minimum of $f(u)$.

Scalar input

Chain rule

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ and $h : \mathbb{R} \rightarrow \mathbb{R}$ be two functions and u, v, w be variables such that:

$$\begin{aligned} v &= f(u), \\ w &= h(v) \qquad\qquad \text{i.e. } w = h(f(u)) = h \circ f(u). \end{aligned}$$

For a given u, how does an infinitesimal change of u impact w?

Scalar input

Chain rule

Let $f : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ be two functions and u, v, w be variables such that:

$$v = f(u),$$
$$w = h(v) \qquad \text{i.e. } w = h(f(u)) = h \circ f(u).$$

For a given u, how does an infinitesimal change of u impact w?

$$\frac{\partial w}{\partial u} = \frac{\partial w}{\partial v} \cdot \frac{\partial v}{\partial u}$$

Scalar input

Example: explicit differentiation

$$f(u) = (2u+1)^2 = 4u^2 + 4u + 1$$
$$f'(u) = 8u + 4$$

Example: differentiation using the chain rule

$$v = 2u + 1 \qquad \frac{\partial v}{\partial u} = 2$$
$$w = v^2 = f(u) \qquad \frac{\partial w}{\partial v} = 2z$$
$$\frac{\partial w}{\partial u} = \frac{\partial w}{\partial v} \cdot \frac{\partial v}{\partial u} = 2v * 2 = 4(2u+1) = 8u + 4 = f'(u)$$

Vector input

Let $f : \mathbb{R}^m \rightarrow \mathbb{R}$ be a function and $\boldsymbol{u} \in \mathbb{R}^m, w \in \mathbb{R}$ be variables such that:

$$w = f(\boldsymbol{u}).$$

Partial derivative

For a given $\boldsymbol{u}$, how does an infinitesimal change of u_i impact w?

$$\frac{\partial w}{\partial u_i}$$

i.e. each input $u_j, j \neq i$ is considered as a constant.

Gradient

For a given $\boldsymbol{u}$, how does an infinitesimal change of $\boldsymbol{u}$ impact w?

$$\nabla_{\boldsymbol{u}} f(\boldsymbol{u}) = \begin{bmatrix} \frac{\partial}{\partial u_1} f(\boldsymbol{u}) \\ \frac{\partial}{\partial u_2} f(\boldsymbol{u}) \\ \vdots \end{bmatrix}$$

Vector input

Chain rule

Let $f : \mathbb{R}^m \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}$ be two functions and $\boldsymbol{u}^m, \boldsymbol{v}^n, w$ be variables such that:

$$\boldsymbol{v} = f(\boldsymbol{u}),$$
$$w = h(\boldsymbol{v})$$

For a given u_i, how does an infinitesimal change of u_i impact w?

$$\frac{\partial w}{\partial u_i} = \sum_j \frac{\partial w}{\partial v_j} \cdot \frac{\partial v_j}{\partial u_i}$$

Vector example

$$\boldsymbol{v} = \boldsymbol{W}\boldsymbol{u} + b \quad \text{or} \quad v_j = \sum_i W_{j,i} u_i + b_j \qquad \frac{\partial v_j}{\partial u_i} = W_{j,i}$$

$$w = \sum_j v_j \qquad \frac{\partial w}{\partial v_j} = 1$$

$$\frac{\partial w}{\partial u_i} = \sum_j \frac{\partial w}{\partial v_j} \cdot \frac{\partial v_j}{\partial u_i} = \sum_j 1 * W_{j,i}$$

Subgradients

Subgradient

Given a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, a subgradient at $\boldsymbol{u} \in U$ is a vector $\boldsymbol{g} \in \mathbb{R}^n$ such that:

$$\forall \boldsymbol{u}' \in \mathbb{R}^n : \quad f(\boldsymbol{u}') \geq f(\boldsymbol{u}) + \langle \boldsymbol{g}, \boldsymbol{u}' - \boldsymbol{u} \rangle$$

The set of subgradients at point $\boldsymbol{u}$ is called the subdifferential and is denoted $\partial f(\boldsymbol{u})$.

Subgradient

Given a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, a subgradient at $\boldsymbol{u} \in U$ is a vector $\boldsymbol{g} \in \mathbb{R}^n$ such that:

$$\forall \boldsymbol{u}' \in \mathbb{R}^n : \quad f(\boldsymbol{u}') \geq f(\boldsymbol{u}) + \langle \boldsymbol{g}, \boldsymbol{u}' - \boldsymbol{u} \rangle$$

The set of subgradients at point $\boldsymbol{u}$ is called the subdifferential and is denoted $\partial f(\boldsymbol{u})$.

Properties

If f is convex, then:

- if f is differentiable at $\boldsymbol{u}$, then $\partial f(\boldsymbol{u}) = \{\nabla f(\boldsymbol{x})\}$
  (i.e. the gradient is the single subgradient)
- the function $h(\boldsymbol{u}') = f(\boldsymbol{u}) + \langle \boldsymbol{g}, \boldsymbol{u}' - \boldsymbol{u} \rangle$ is a linear sub-estimator of f
- a similar definition for concave function is the super-gradient

Subgradient

Given a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, a subgradient at $\boldsymbol{u} \in U$ is a vector $\boldsymbol{g} \in \mathbb{R}^n$ such that:

$$\forall \boldsymbol{u}' \in \mathbb{R}^n : \quad f(\boldsymbol{u}') \geq f(\boldsymbol{u}) + \langle \boldsymbol{g}, \boldsymbol{u}' - \boldsymbol{u} \rangle$$

The set of subgradients at point $\boldsymbol{u}$ is called the subdifferential and is denoted $\partial f(\boldsymbol{u})$.

Properties

If f is convex, then:

- if f is differentiable at $\boldsymbol{u}$, then $\partial f(\boldsymbol{u}) = \{\nabla f(\boldsymbol{x})\}$
  (i.e. the gradient is the single subgradient)
- the function $h(\boldsymbol{u}') = f(\boldsymbol{u}) + \langle \boldsymbol{g}, \boldsymbol{u}' - \boldsymbol{u} \rangle$ is a linear sub-estimator of f
- a similar definition for concave function is the super-gradient

Existence of the subgradient (Beck, th. 3.14)

Let f be a proper convex function.
Then, $\forall \boldsymbol{u} \in int(\textbf{dom}\, f)$, the subdifferential $\partial f(\boldsymbol{u})$ is non-empty.

Computing subgradients 1/2

There are "rules" that allows to compute the subgradient of a function at a given point (see Beck, Section 2.4).

- Strong subgradient result: the subdifferential set at a given point is known
- Weak subgradient result: one or several subgradients at a given point are known, but not all

Computing subgradients 1/2

There are "rules" that allows to compute the subgradient of a function at a given point (see Beck, Section 2.4).

- Strong subgradient result: the subdifferential set at a given point is known
- Weak subgradient result: one or several subgradients at a given point are known, but not all

Multiplication by a positive scalar (Beck, th. 3.35)

Let $f : \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$ be a proper function, $h(\boldsymbol{u}) = \alpha f(\boldsymbol{u})$ with $\alpha > 0$. Then:

$$\forall \boldsymbol{u} \in \mathbf{dom}\, f, \boldsymbol{g} \in \mathbb{R}^k : \alpha \boldsymbol{u} \in \partial h(\boldsymbol{u}) \quad \text{if and only if} \quad \boldsymbol{g} \in \partial f(\boldsymbol{u})$$

Computing subgradients 2/2

Summation (Beck, th. 3.36)

Let $f_1 : \mathbb{R}^k \to \; \cup\{\infty\} R$ and $f_2 : \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$ be proper functions and $h(\boldsymbol{u}) = f_1(\boldsymbol{u}) + f_2(\boldsymbol{u})$. Then, $\forall \boldsymbol{u} \in \textbf{dom}\, h, \boldsymbol{g} \in \mathbb{R}^k$, we have $\boldsymbol{g} \in \partial h(\boldsymbol{u})$ if and only if:

$$\boldsymbol{g} = \boldsymbol{g}^{(1)} + \boldsymbol{g}^{(2)} \quad \text{such that} \quad \boldsymbol{g}^{(1)} \in \partial f_1(\boldsymbol{u}) \text{ and } \boldsymbol{g}^{(2)} \in \partial f_2(\boldsymbol{u})$$

Computing subgradients 2/2

Summation (Beck, th. 3.36)

Let $f_1 : \mathbb{R}^k \rightarrow \cup\{\infty\}R$ and $f_2 : \mathbb{R}^k \rightarrow \mathbb{R} \cup \{\infty\}$ be proper functions and $h(\boldsymbol{u}) = f_1(\boldsymbol{u}) + f_2(\boldsymbol{u})$. Then, $\forall \boldsymbol{u} \in \mathbf{dom}\, h, \boldsymbol{g} \in \mathbb{R}^k$, we have $\boldsymbol{g} \in \partial h(\boldsymbol{u})$ if and only if:

$$\boldsymbol{g} = \boldsymbol{g}^{(1)} + \boldsymbol{g}^{(2)} \quad \text{such that} \quad \boldsymbol{g}^{(1)} \in \partial f_1(\boldsymbol{u}) \text{ and } \boldsymbol{g}^{(2)} \in \partial f_2(\boldsymbol{u})$$

Maximization (Beck, th. 3.50)

Let $f_1 ... f_n : \mathbb{R}^k \rightarrow \mathbb{R} \cup \{\infty\}$ be a set of proper functions and

$$h(\boldsymbol{u}) = \max(f_1(\boldsymbol{u}), ..., f_n(\boldsymbol{u}))$$

Let $\boldsymbol{u} \in \mathbb{R}^k$ and $\mathrm{I}(\boldsymbol{u}) = \{i \in \{1...n\} | f_i(\boldsymbol{u}) = h(\boldsymbol{u})\}$.

If $\boldsymbol{g} \in \partial f_i(\boldsymbol{u})$ for any $i \in \mathrm{I}(\boldsymbol{u})$, then $\boldsymbol{g} \in \partial h(\boldsymbol{u})$.
(Note: we could get stronger result than this)

Optimality conditions

Unconstrained optimization problem (Fermat's theorem)

Let $f : \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$ be a proper convex function and $\hat{\boldsymbol{u}} \in \textbf{dom}\, f$.
If $0 \in \partial f(\hat{\boldsymbol{u}})$,
then $f(\hat{\boldsymbol{u}})$ is the global minimum of f.

Proof

By the subgradient definition:

$$\forall \boldsymbol{u} \in \mathbb{R}^k, \boldsymbol{g} \in \partial f(\hat{\boldsymbol{u}}) : \quad f(\boldsymbol{u}) \geq f(\hat{\boldsymbol{u}}) + \langle \boldsymbol{g}, \boldsymbol{u} - \hat{\boldsymbol{u}} \rangle$$

In particular, we know that $0 \in \partial f(\hat{\boldsymbol{u}})$, therefore:

$$f(\boldsymbol{u}) \geq f(\hat{\boldsymbol{u}})$$

Fenchel conjugates

Definition

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a function.
The Fenchel conjugate of f is the function $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ defined as follows:

$$f^*(\boldsymbol{t}) = \sup_{\boldsymbol{u} \in \mathbf{dom}\, f} \langle \boldsymbol{t}, \boldsymbol{u} \rangle - f(\boldsymbol{u})$$

Definition

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a function.
The Fenchel conjugate of f is the function $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ defined as follows:

$$f^*(\boldsymbol{t}) = \sup_{\boldsymbol{u} \in \mathbf{dom}\, f} \langle \boldsymbol{t}, \boldsymbol{u} \rangle - f(\boldsymbol{u})$$

The biconjugate of f is the function $f^{**} : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ defined as follows:

$$f^{**}(\boldsymbol{u}) = \sup_{\boldsymbol{t} \in \mathbf{dom}\, f^*} \langle \boldsymbol{u}, \boldsymbol{t} \rangle - f^*(\boldsymbol{t})$$

If f is proper, closed and convex, then $f^{**} = f$
$=>$ important property is often used to build variational formulation of functions

One small theorem

Fenchel-Young inequality

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function, $\boldsymbol{u} \in \textbf{dom}\, f$ and $\boldsymbol{t} \in \textbf{dom}\, f^*$:

$$f(\boldsymbol{u}) + \; f^*(\boldsymbol{t}) \geq \langle \boldsymbol{u}, \boldsymbol{t} \rangle$$

Proof

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function, $\boldsymbol{u} \in \textbf{dom}\, f$ and $\boldsymbol{t} \in \textbf{dom}\, f^*$:

$$\begin{aligned} \langle \boldsymbol{u}, \boldsymbol{t} \rangle - f(\boldsymbol{u}) &\leq \sup_{\boldsymbol{u}' \in \textbf{dom}\, f} \boldsymbol{u}'^\top \boldsymbol{y} - f(\boldsymbol{u}') \\ &= f^*(\boldsymbol{t}) \end{aligned}$$

By re-arranging terms we get the expected inequality.

Subdifferential of a Fenchel conjugate

Let $f : \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$ be function. Let $\boldsymbol{t} \in \textbf{dom}\, f^*$ and

$$\hat{\boldsymbol{u}} = \operatorname*{arg\,max}_{\boldsymbol{u} \in \textbf{dom}\, f} \langle \boldsymbol{u}, \boldsymbol{t} \rangle - f(\boldsymbol{u})$$

Then, $\hat{\boldsymbol{u}}$ is a subgradient of f^* at $\boldsymbol{t}$, i.e. $\hat{\boldsymbol{u}} \in \partial f^*(\boldsymbol{t})$.

Subdifferential of a Fenchel conjugate

Let $f : \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$ be function. Let $\boldsymbol{t} \in \textbf{dom}\, f^*$ and

$$\hat{\boldsymbol{u}} = \operatorname*{arg\,max}_{\boldsymbol{u} \in \textbf{dom}\, f} \langle \boldsymbol{u}, \boldsymbol{t} \rangle - f(\boldsymbol{u})$$

Then, $\hat{\boldsymbol{u}}$ is a subgradient of f^* at $\boldsymbol{t}$, i.e. $\hat{\boldsymbol{u}} \in \partial f^*(\boldsymbol{t})$.

Proof

Although this can be proved via Danskin's theorem, here is a simpler proof.
We have $f^*(\boldsymbol{t}) = \max_{\boldsymbol{u} \in \textbf{dom}\, f} \langle \boldsymbol{u}, \boldsymbol{t} \rangle - f(\boldsymbol{u}) = \langle \hat{\boldsymbol{u}}, \boldsymbol{t} \rangle - f(\hat{\boldsymbol{u}})$.
For all $\boldsymbol{t}' \in \textbf{dom}\, f^*$ we have:

$$\begin{aligned} f^*(\boldsymbol{t}) + \langle \hat{\boldsymbol{u}}, \boldsymbol{t}' - \boldsymbol{t} \rangle &= \langle \hat{\boldsymbol{u}}, \boldsymbol{t} \rangle - f(\hat{\boldsymbol{u}}) + \langle \hat{\boldsymbol{u}}, \boldsymbol{t}' \rangle - \langle \hat{\boldsymbol{u}}, \boldsymbol{t} \rangle \\ &= \langle \hat{\boldsymbol{u}}, \boldsymbol{t}' \rangle - f(\hat{\boldsymbol{u}}) \\ &\leq \max_{\boldsymbol{u} \in \textbf{dom}\, f} \boldsymbol{u}^\top \boldsymbol{t}' - f(\boldsymbol{u}) \\ &= f^*(\boldsymbol{t}') \end{aligned}$$

Hence $\hat{\boldsymbol{u}}$ is a subgradient of f^* at $\boldsymbol{t}$.