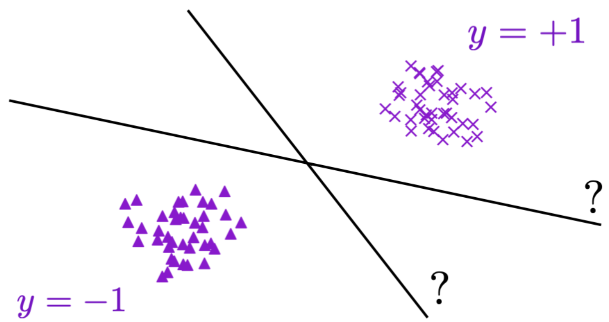
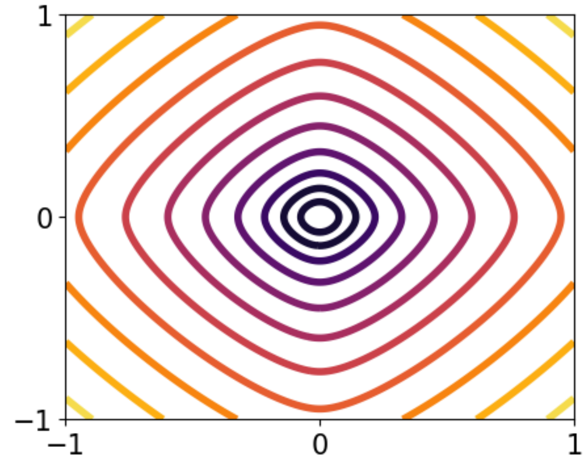
Implicit Bias of Mirror Flow on Separable Data





Radu-Alexandru Dragomir (Télécom Paris)

with Scott Pesme and Nicolas Flammarion (EPFL)

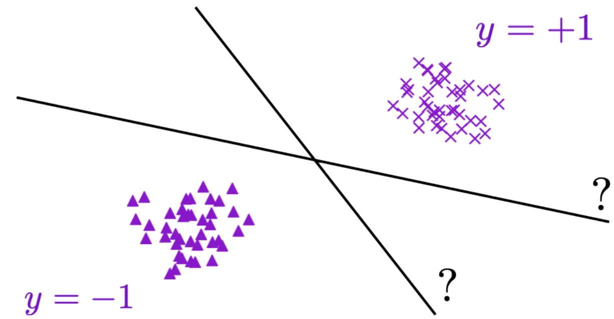
EUROPT 2024, Lund

Setup

Logistic regression:

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

points $x_i \in \mathbb{R}^d$, labels $y_i \in \{-1, +1\}$

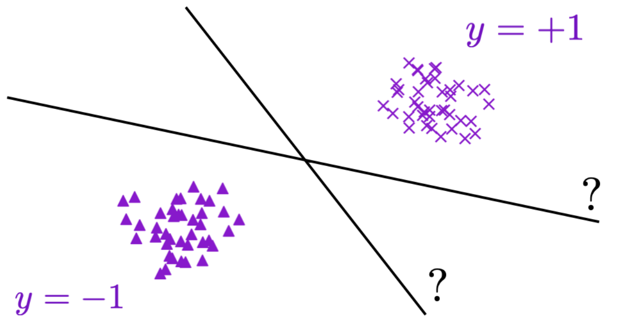


Setup

Logistic regression:

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

points $x_i \in \mathbb{R}^d$, labels $y_i \in \{-1, +1\}$



Linear separability:

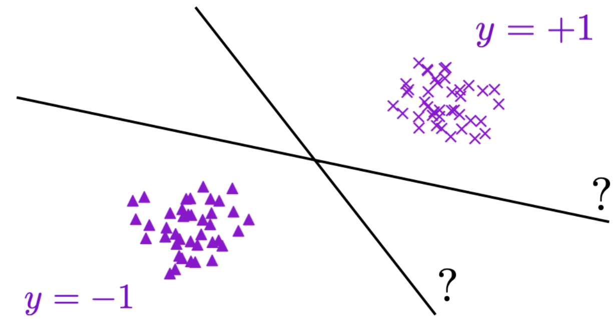
The set $\mathcal{I} = \{\beta^* \,:\, y_i \langle \beta^*, x_i \rangle > 0 \text{ for } i = 1 \ldots n\}$ is nonempty.

Setup

Logistic regression:

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

points $x_i \in \mathbb{R}^d$, labels $y_i \in \{-1, +1\}$



Linear separability:

The set $\mathcal{I} = \{\beta^* \, : \, y_i \langle \beta^*, x_i \rangle > 0$ for $i = 1 \ldots n\}$ is nonempty.

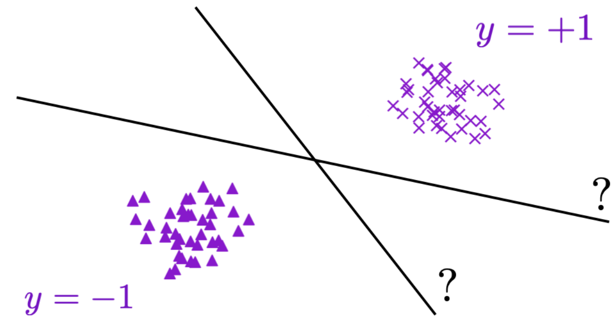
The loss is minimized **at infinity**: $\lim_{s \to \infty} L(s\beta^*) = 0$ for $\beta^* \in \mathcal{I}$

Setup

Logistic regression:

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

points $x_i \in \mathbb{R}^d$, labels $y_i \in \{-1, +1\}$



Linear separability:

The set $\mathcal{I} = \{\beta^* : y_i \langle \beta^*, x_i \rangle > 0$ for $i = 1 \ldots n\}$ is nonempty.

The loss is minimized **at infinity**: $\lim_{s \to \infty} L(s\beta^*) = 0$ for $\beta^* \in \mathcal{I}$

Mirror flow:

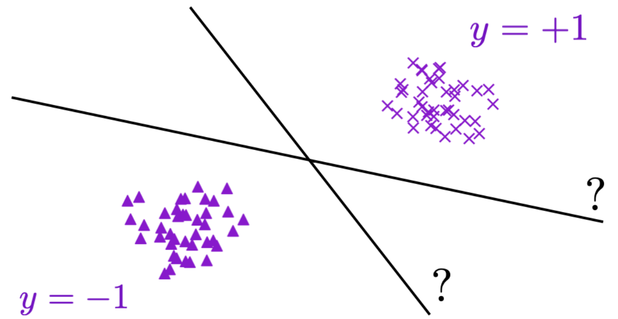
$$\dot{\beta}_t = -\nabla^2 \phi(\beta_t)^{-1} \nabla L(\beta_t)$$

Setup

Logistic regression:

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

points $x_i \in \mathbb{R}^d$, labels $y_i \in \{-1, +1\}$



Linear separability:

The set $\mathcal{I} = \{\beta^* : y_i \langle \beta^*, x_i \rangle > 0$ for $i = 1 \ldots n\}$ is nonempty.

The loss is minimized **at infinity**: $\lim_{s \to \infty} L(s\beta^*) = 0$ for $\beta^* \in \mathcal{I}$

Mirror flow:

$$\dot{\beta}_t = -\nabla^2 \phi(\beta_t)^{-1} \nabla L(\beta_t)$$

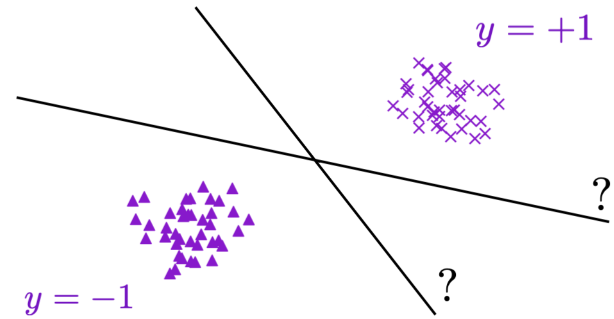
We expect $L(\beta_t) \to 0$ and $\|\beta_t\| \to \infty$.

Setup

Logistic regression:

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

points $x_i \in \mathbb{R}^d$, labels $y_i \in \{-1, +1\}$



Linear separability:

The set $\mathcal{I} = \{\beta^* \,:\, y_i \langle \beta^*, x_i \rangle > 0 \text{ for } i = 1 \dots n\}$ is nonempty.

The loss is minimized **at infinity**: $\lim_{s \to \infty} L(s\beta^*) = 0$ for $\beta^* \in \mathcal{I}$

Mirror flow:

$$\dot{\beta}_t = -\nabla^2 \phi(\beta_t)^{-1} \nabla L(\beta_t)$$

We expect $L(\beta_t) \to 0$ and $\|\beta_t\| \to \infty$.

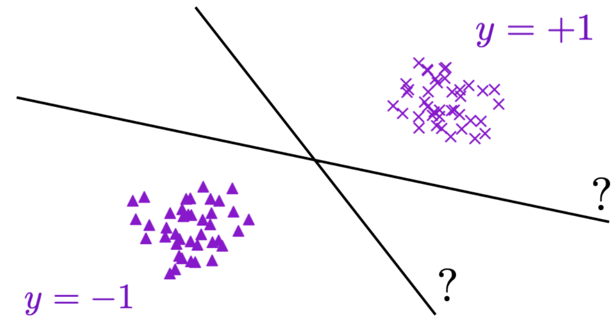
What is the directional limit $\bar{\beta}$ of $\frac{\beta_t}{\|\beta_t\|}$?

Setup

Logistic regression:

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

points $x_i \in \mathbb{R}^d$, labels $y_i \in \{-1, +1\}$



Linear separability:

The set $\mathcal{I} = \{\beta^* \,:\, y_i \langle \beta^*, x_i \rangle > 0$ for $i = 1 \ldots n\}$ is nonempty.

The loss is minimized **at infinity**: $\lim_{s \to \infty} L(s\beta^*) = 0$ for $\beta^* \in \mathcal{I}$

Mirror flow:

$$\dot{\beta}_t = -\nabla^2 \phi(\beta_t)^{-1} \nabla L(\beta_t)$$

We expect $L(\beta_t) \to 0$ and $\|\beta_t\| \to \infty$.

What is the directional limit $\bar{\beta}$ of $\frac{\beta_t}{\|\beta_t\|}$?

Many possible limit directions in $\mathcal{I}$. Which one is **preferred** by the algorithm?

Implicit bias for regression

- **Least squares regression, gradient flow:** [Lemaire 1996]

$$L(\beta) = \|X^T\beta - y\|^2, \quad \dot{\beta}_t = -\nabla L(\beta_t)$$

Implicit bias for regression

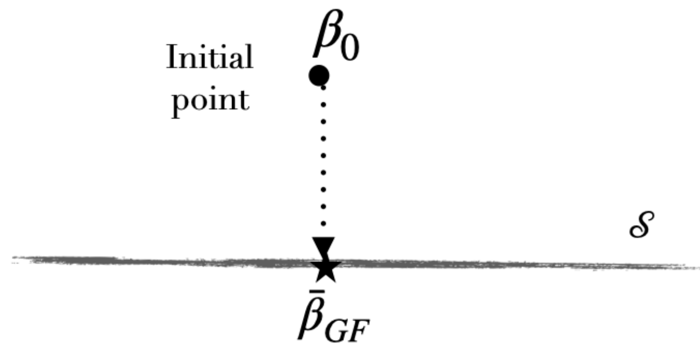
- **Least squares regression, gradient flow:** [Lemaire 1996]

$$L(\beta) = \|X^T\beta - y\|^2, \quad \dot{\beta}_t = -\nabla L(\beta_t)$$

Let $\mathcal{I} = \{\beta \,:\, X^T\beta = y\}$.

Then $\beta_t \to \bar{\beta}_{\mathrm{GF}}$ where

$$\bar{\beta}_{\mathrm{GF}} = \operatorname{argmin}\left\{\|\beta^* - \beta_0\| \,:\, \beta^* \in \mathcal{I}\right\}$$



Implicit bias for regression

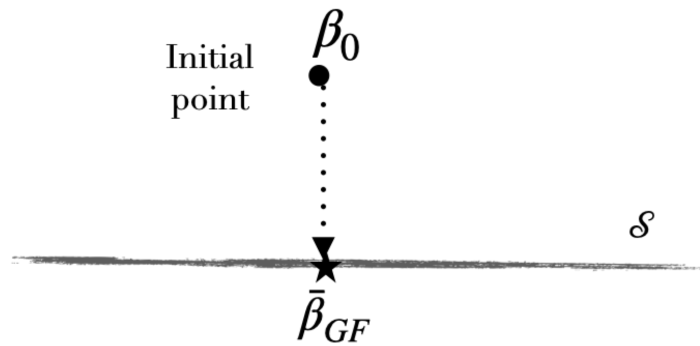
- **Least squares regression, gradient flow:** [Lemaire 1996]

$$L(\beta) = \|X^T\beta - y\|^2, \quad \dot{\beta}_t = -\nabla L(\beta_t)$$

Let $\mathcal{I} = \{\beta \,:\, X^T\beta = y\}$.

Then $\beta_t \to \bar{\beta}_{\mathrm{GF}}$ where

$$\bar{\beta}_{\mathrm{GF}} = \mathrm{argmin}\ \{\|\beta^* - \beta_0\| \,:\, \beta^* \in \mathcal{I}\}$$



- **Least squares regression, mirror flow:** [Gunasekar et al., ICML 2018]

$$L(\beta) = \|X^T\beta - y\|^2, \quad \dot{\beta}_t = -\nabla^2\phi(\beta_t)^{-1}\nabla L(\beta_t)$$

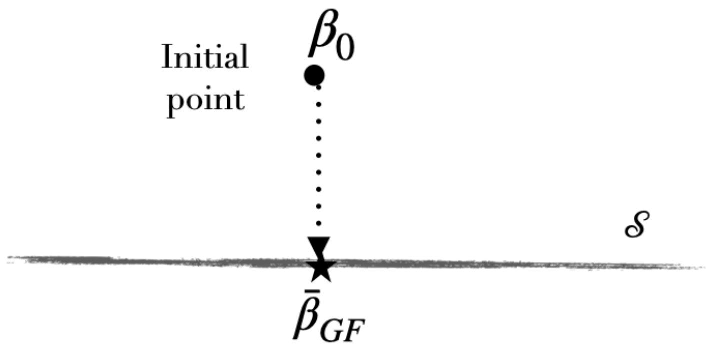
Implicit bias for regression

- **Least squares regression, gradient flow:** [Lemaire 1996]

$$L(\beta) = \|X^T\beta - y\|^2, \quad \dot{\beta}_t = -\nabla L(\beta_t)$$

Let $\mathcal{I} = \{\beta : X^T\beta = y\}$.
Then $\beta_t \to \bar{\beta}_{\mathrm{GF}}$ where

$$\bar{\beta}_{\mathrm{GF}} = \operatorname{argmin}\left\{\|\beta^* - \beta_0\| : \beta^* \in \mathcal{I}\right\}$$



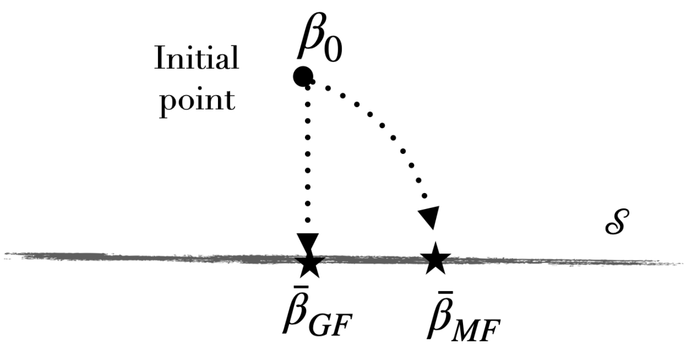
- **Least squares regression, mirror flow:** [Gunasekar et al., ICML 2018]

$$L(\beta) = \|X^T\beta - y\|^2, \quad \dot{\beta}_t = -\nabla^2\phi(\beta_t)^{-1}\nabla L(\beta_t)$$

Then $\beta_t \to \bar{\beta}_{\mathrm{MF}}$ where

$$\bar{\beta}_{\mathrm{MF}} = \operatorname{argmin}\left\{D_\phi(\beta^*, \beta_0) : \beta^* \in \mathcal{I}\right\}$$

(D_ϕ: **Bregman divergence**)



Implicit bias for classification

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

$$\mathcal{I} = \{\beta^* \,:\, y_i \langle \beta^*, x_i \rangle \geq 1, \ \forall i\}$$

The iterates β_t diverge: what is their directional limit?

Implicit bias for classification

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

$$\mathcal{I} = \{\beta^* \,:\, y_i \langle \beta^*, x_i \rangle \geq 1, \ \ \forall i\}$$

The iterates β_t diverge: what is their directional limit?

- **Gradient flow** [Soudry et al., JMLR 2018]: $\frac{\beta_t}{\|\beta_t\|} \to \bar{\beta}_{\mathrm{GF}}$ where

$$\bar{\beta}_{\mathrm{GF}} \propto \operatorname{argmin}\left\{ \|\beta^*\|_2 \,:\, \beta^* \in \mathcal{I} \right\}$$

Implicit bias for classification

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

$$\mathcal{I} = \{\beta^* \,:\, y_i \langle \beta^*, x_i \rangle \geq 1, \ \ \forall i\}$$

The iterates β_t diverge: what is their directional limit?

- **Gradient flow** [Soudry et al., JMLR 2018]: $\frac{\beta_t}{||\beta_t||} \to \bar{\beta}_{\mathrm{GF}}$ where

$$\bar{\beta}_{\mathrm{GF}} \propto \operatorname{argmin} \left\{ \, ||\beta^*||_2 \,:\, \beta^* \in \mathcal{I} \right\}$$

$\to$ max-margin classifier (SVM)

Implicit bias for classification

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right) \qquad \mathcal{I} = \{\beta^* \ : \ y_i \langle \beta^*, x_i \rangle \geq 1, \ \ \forall i\}$$

The iterates β_t diverge: what is their directional limit?

- **Gradient flow** [Soudry et al., JMLR 2018]: $\frac{\beta_t}{\|\beta_t\|} \to \bar{\beta}_{\mathrm{GF}}$ where

$$\bar{\beta}_{\mathrm{GF}} \propto \operatorname{argmin} \left\{ \|\beta^*\|_2 \ : \ \beta^* \in \mathcal{I} \right\}$$

$\rightarrow$ max-margin classifier (SVM)

- **Mirror flow:** our work. $\frac{\beta_t}{\|\beta_t\|} \to \bar{\beta}_{\mathrm{MF}}$ where

$$\bar{\beta}_{\mathrm{MF}} \propto \operatorname{argmin} \left\{ \phi_\infty(\beta^*) \ : \ \beta^* \in \mathcal{I} \right\}$$

$\rightarrow$ ϕ_∞-max margin classifier

ϕ_∞: **horizon function** of ϕ (limit of ϕ "at infinity")

Mirror flow: why and how

$$\dot{\beta}_t = -\nabla^2\phi(\beta_t)^{-1}\nabla L(\beta_t)$$

Mirror flow: why and how

$$\dot{\beta}_t = -\nabla^2\phi(\beta_t)^{-1}\nabla L(\beta_t)$$

Potential function ϕ is

- **strictly convex** and C^2 on $\mathbb{R}^d$ (**full domain**),
- **coercive** and has **coercive gradients**.

Mirror flow: why and how

$$\dot{\beta}_t = -\nabla^2\phi(\beta_t)^{-1}\nabla L(\beta_t)$$

Potential function ϕ is

- **strictly convex** and C^2 on $\mathbb{R}^d$ (**full domain**),
- **coercive** and has **coercive gradients**.

Motivation: reparametrized problems $\beta = F(\theta)$

Gradient flow on $\theta \mapsto L(F(\theta))$ $\quad\Leftrightarrow\quad$ **Mirror flow** on $\beta \mapsto L(\beta)$

Mirror flow: why and how

$$\dot{\beta}_t = -\nabla^2\phi(\beta_t)^{-1}\nabla L(\beta_t)$$

Potential function ϕ is

- **strictly convex** and C^2 on $\mathbb{R}^d$ (**full domain**),
- **coercive** and has **coercive gradients**.

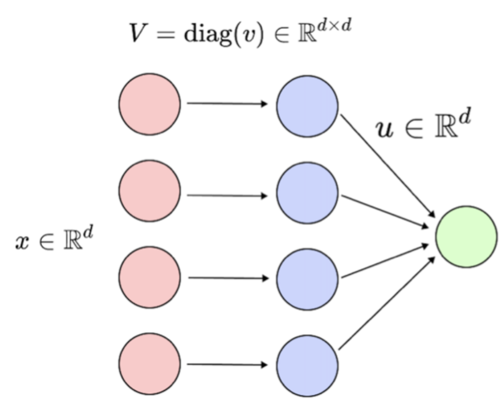
Motivation: reparametrized problems $\beta = F(\theta)$

Gradient flow on $\theta \mapsto L(F(\theta))$ $\Leftrightarrow$ **Mirror flow** on $\beta \mapsto L(\beta)$

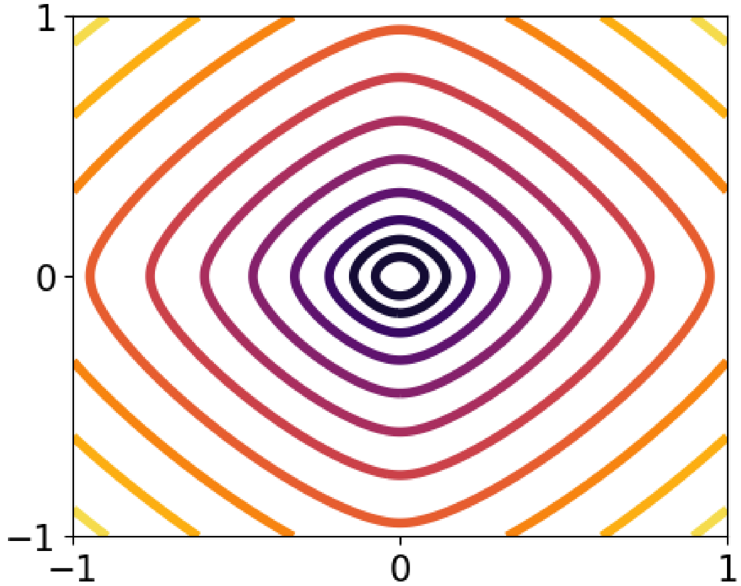
Example: $\beta = u \odot v$ ("diagonal neural networks")

Gradient flow on $L(u \odot v) \Leftrightarrow$ mirror flow on $L(\beta)$
with hyperbolic potential

$$\phi(\beta) = \sum_{i=1}^{d} \left(\beta_i \text{arcsinh}(\beta_i) - \sqrt{\beta_i^2 + 1} \right)$$



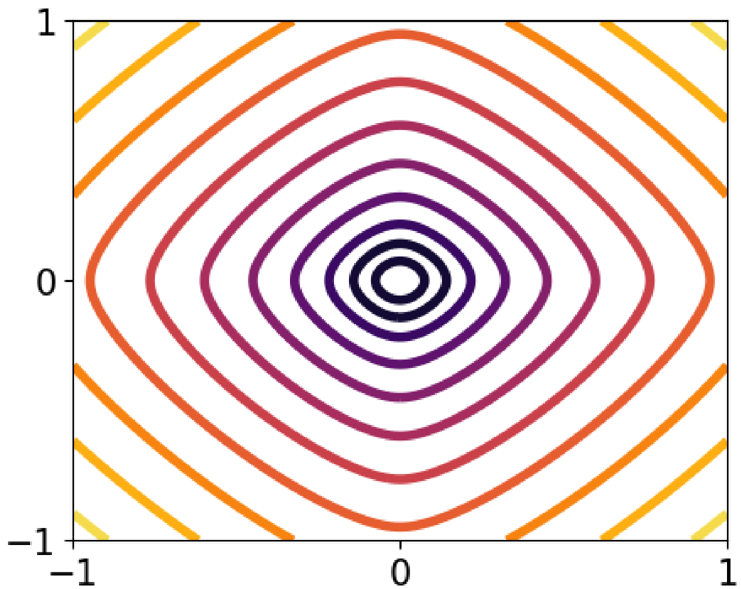
Homogenous or not?



Hyperbolic potential: **not homogenous**

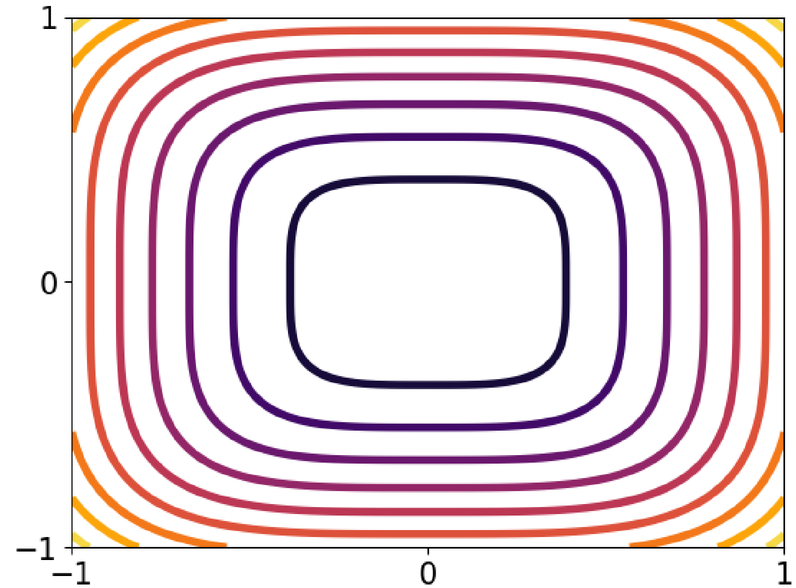
$$\phi(\beta) = \sum_{i=1}^{d} \left(\beta_i \mathsf{arcsinh}(\beta_i) - \sqrt{\beta_i^2 + 1} \right)$$

Homogenous or not?



Hyperbolic potential: **not homogenous**

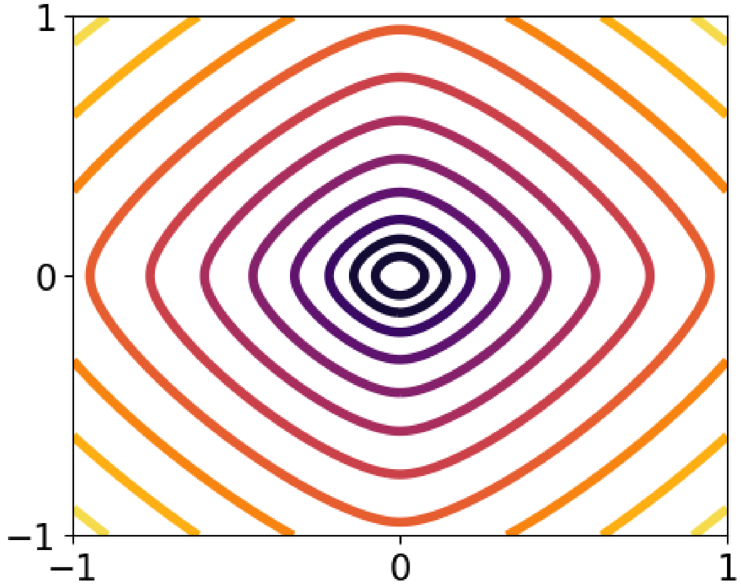
$$\phi(\beta) = \sum_{i=1}^{d} \left(\beta_i \text{arcsinh}(\beta_i) - \sqrt{\beta_i^2 + 1} \right)$$



Homogenous potential

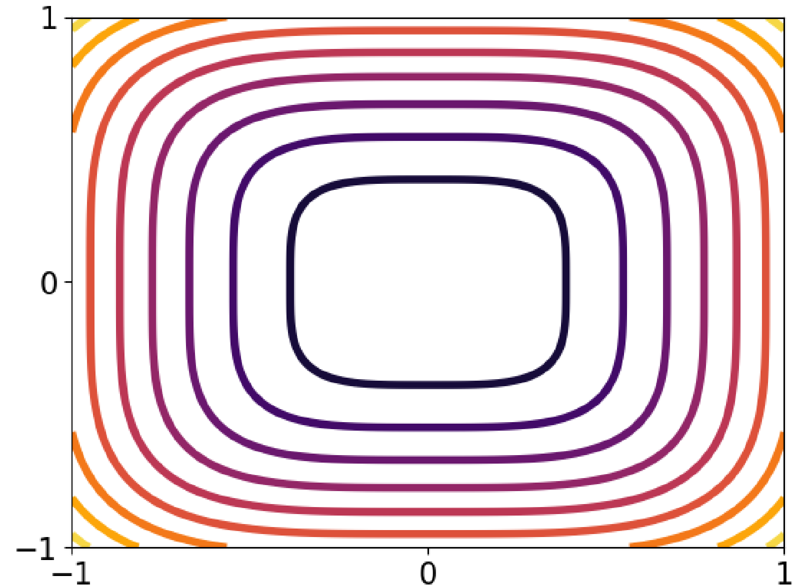
$$\phi(\beta) = \sum_{i=1}^{d} \beta_i^4$$

Homogenous or not?



Hyperbolic potential: **not homogenous**

$$\phi(\beta) = \sum_{i=1}^{d} \left(\beta_i \text{arcsinh}(\beta_i) - \sqrt{\beta_i^2 + 1} \right)$$

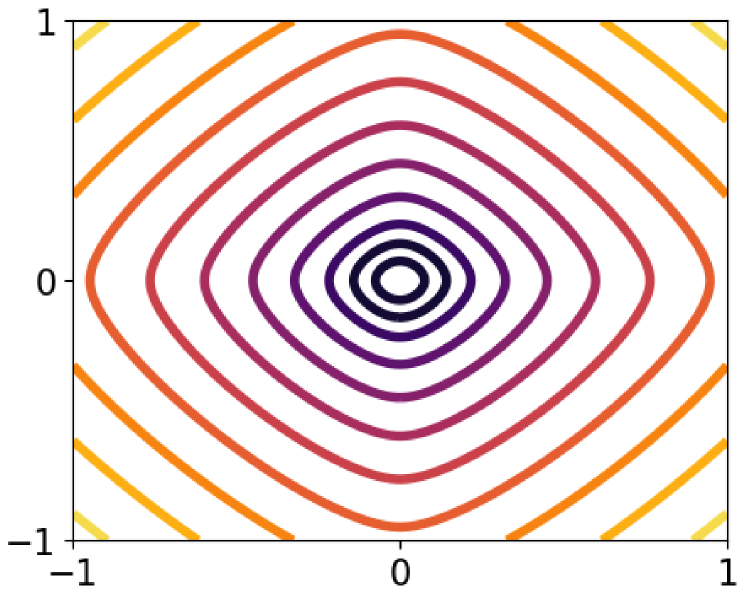


Homogenous potential

$$\phi(\beta) = \sum_{i=1}^{d} \beta_i^4$$

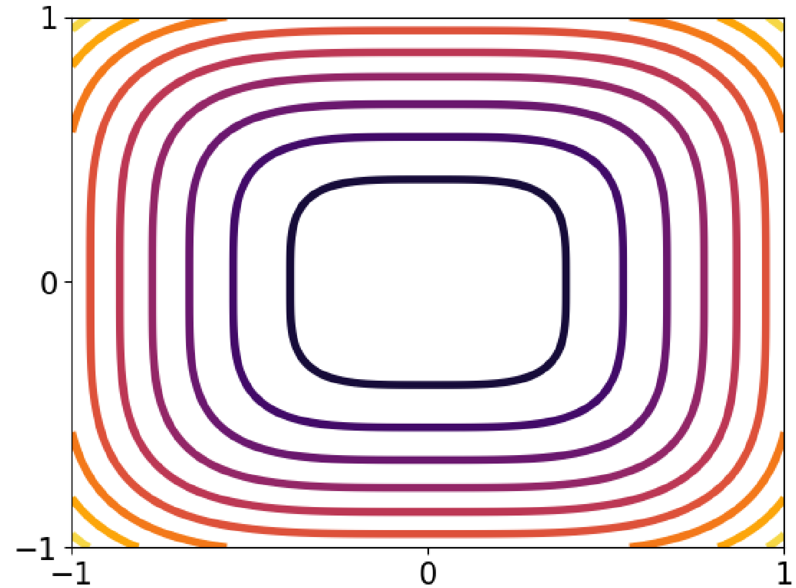
We want to define the shape of the potential ϕ **"at infinity"**

Homogenous or not?



Hyperbolic potential: **not homogenous**

$$\phi(\beta) = \sum_{i=1}^{d} \left(\beta_i \mathrm{arcsinh}(\beta_i) - \sqrt{\beta_i^2 + 1} \right)$$



Homogenous potential

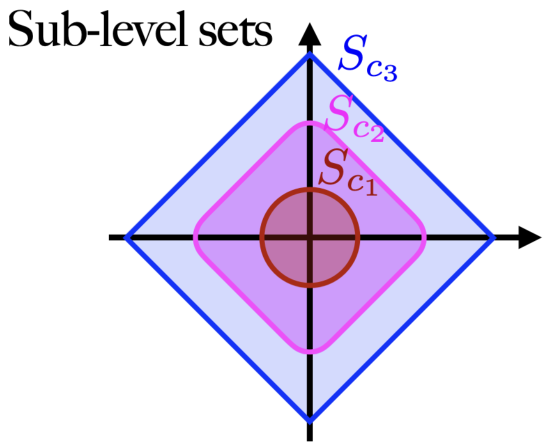
$$\phi(\beta) = \sum_{i=1}^{d} \beta_i^4$$

We want to define the shape of the potential ϕ **"at infinity"**

$\rightarrow$ horizon function ϕ_∞

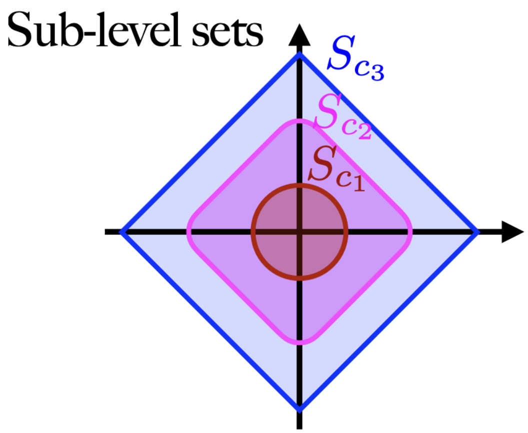
Horizon function: geometric construction

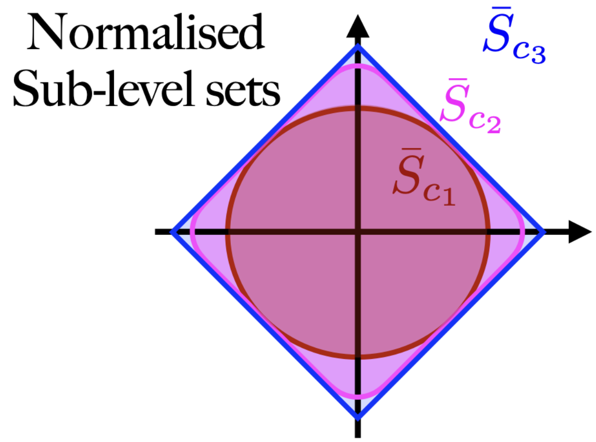
Horizon function: geometric construction



$$S_c = \{\beta \, : \, \phi(\beta) \leq c\}$$

Horizon function: geometric construction

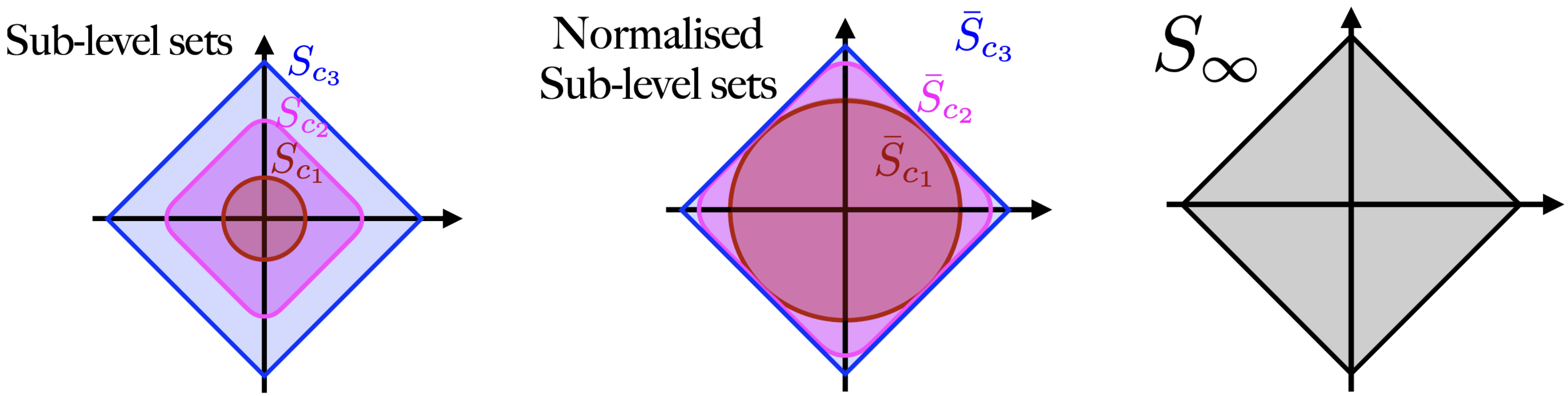




$$S_c = \{\beta \,:\, \phi(\beta) \leq c\}$$

$$\bar{S}_c = S_c / \max_{\beta \in S_c} \|\beta\|$$

Horizon function: geometric construction

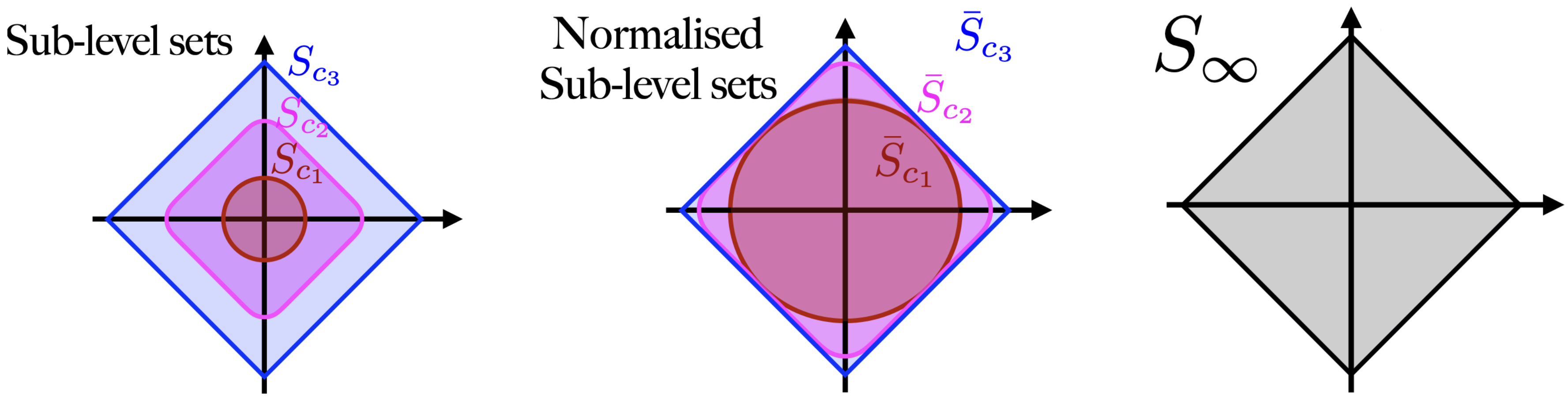


$S_c = \{\beta \, : \, \phi(\beta) \leq c\}$

$\bar{S}_c = S_c / \max_{\beta \in S_c} \|\beta\|$

We say that ϕ admits a **horizon** if $\bar{S}_c$ converges to a set S_∞ as $c \to \infty$.

Horizon function: geometric construction



$S_c = \{\beta \,:\, \phi(\beta) \leq c\}$

$\bar{S}_c = S_c / \max_{\beta \in S_c} \|\beta\|$

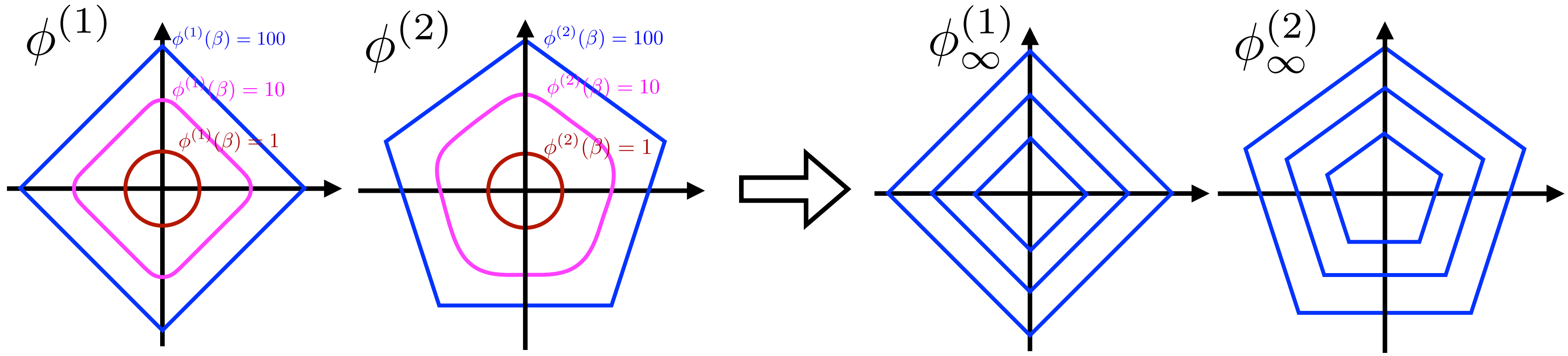
We say that ϕ admits a **horizon** if $\bar{S}_c$ converges to a set S_∞ as $c \to \infty$.

Horizon function: *Minkowski gauge* of S_∞

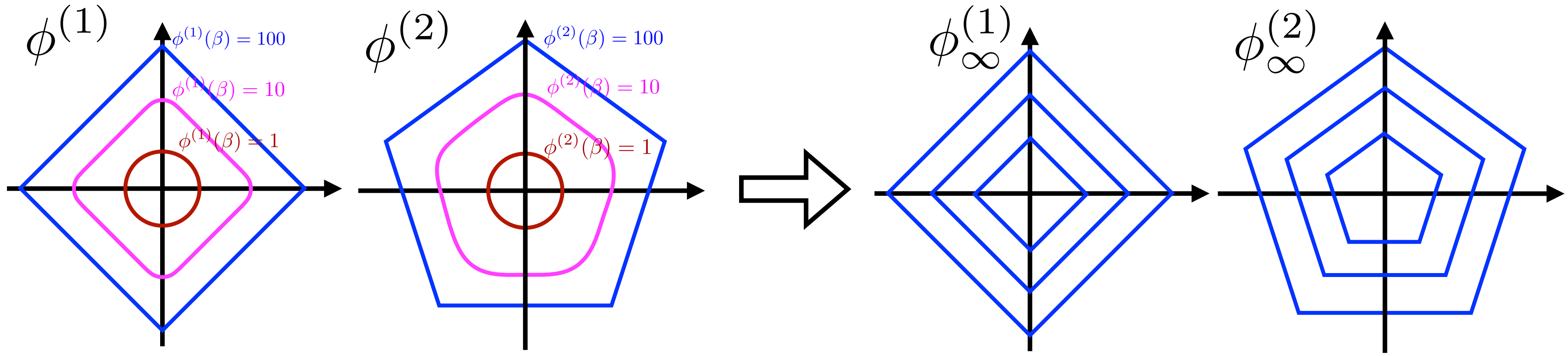
$$\phi_\infty(\beta) = \inf\{r > 0 \,:\, \frac{\beta}{r} \in S_\infty\}$$

ϕ_∞ is **1-homogenous** and its level sets are λS_∞ for $\lambda > 0$.

Horizon function: illustration and conditions for existence

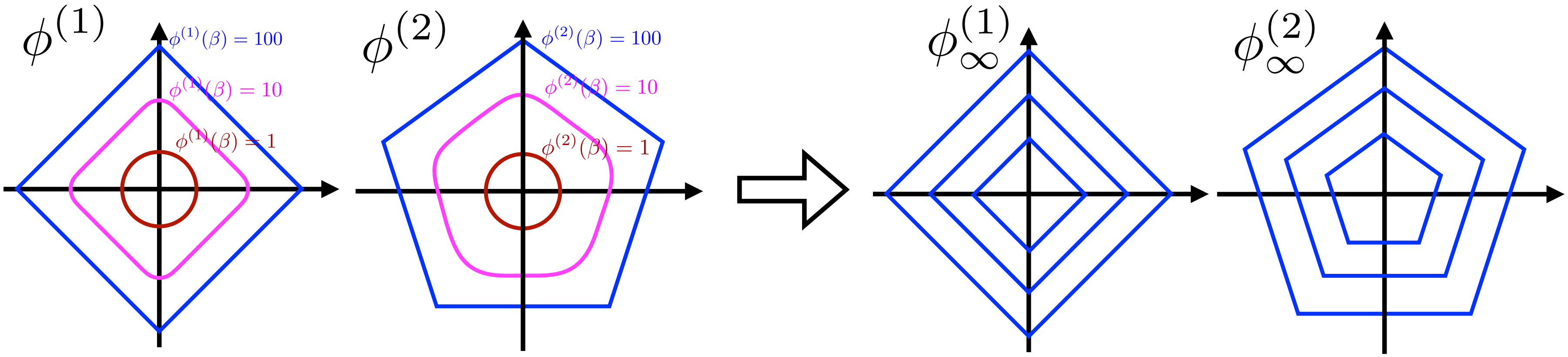


Horizon function: illustration and conditions for existence



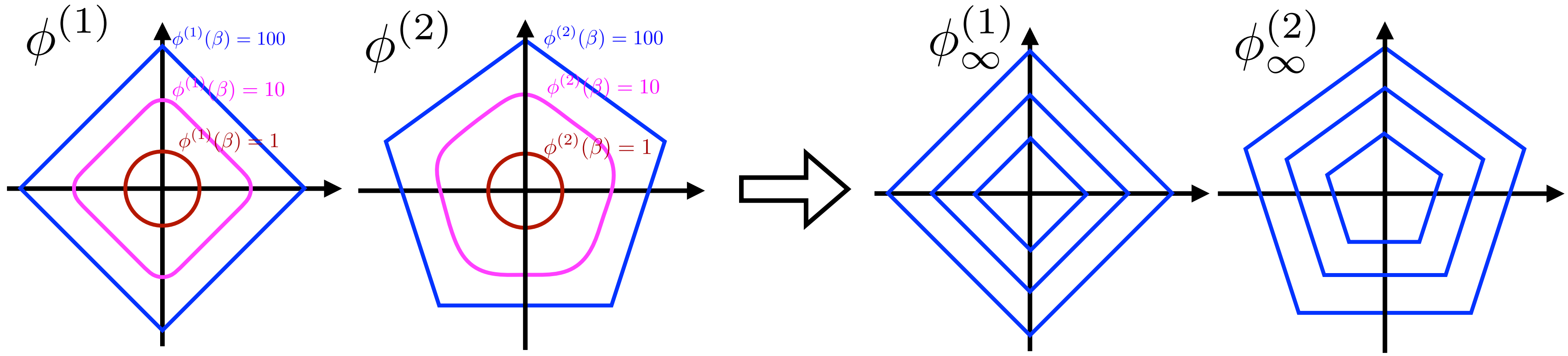
Does ϕ always admit a horizon?

Horizon function: illustration and conditions for existence



Does ϕ always admit a horizon? Yes, for all *reasonable* functions (or *tame*).

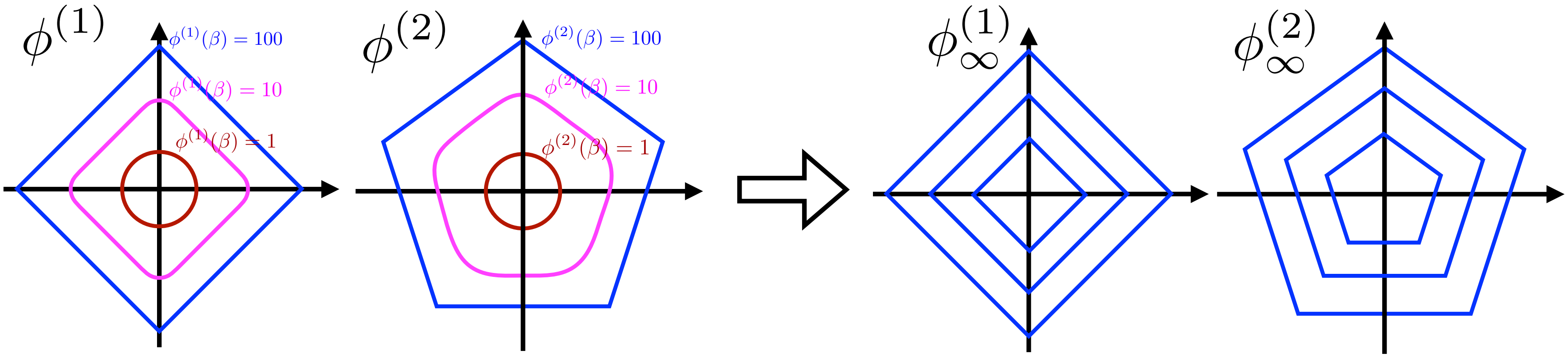
Horizon function: illustration and conditions for existence



Does ϕ always admit a horizon? Yes, for all *reasonable* functions (or *tame*).

If ϕ is **definable in a o-minimal structure**, then it admits a horizon.

Horizon function: illustration and conditions for existence

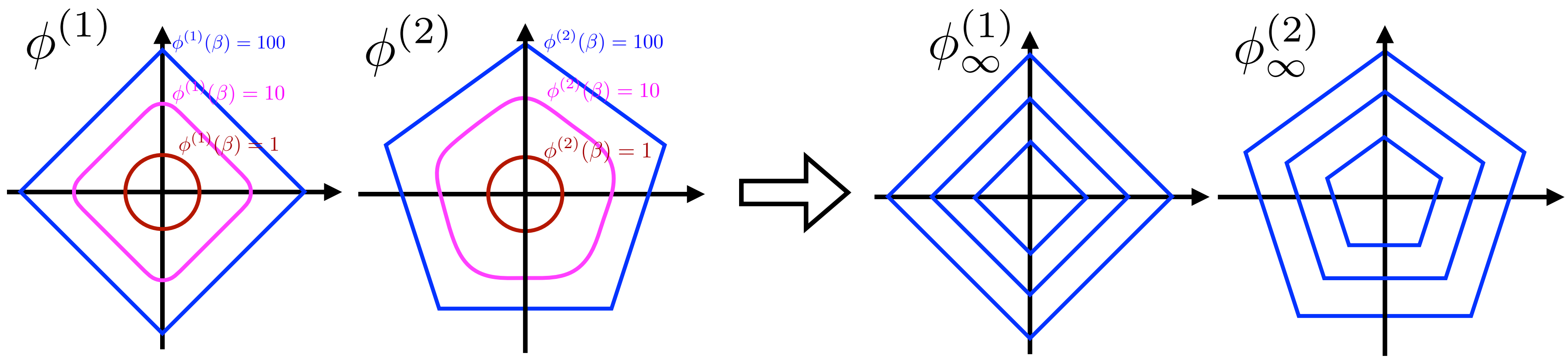


Does ϕ always admit a horizon? Yes, for all *reasonable* functions (or *tame*).

If ϕ is **definable in a o-minimal structure**, then it admits a horizon.

E.g. semianalytic, globally subanalytic, log-exp. This includes polynomials, power functions, exp, log, and **reasonable** combinations of those...

Horizon function: illustration and conditions for existence



Does ϕ always admit a horizon? Yes, for all *reasonable* functions (or *tame*).

If ϕ is **definable in a o-minimal structure**, then it admits a horizon.

E.g. semianalytic, globally subanalytic, log-exp. This includes polynomials, power functions, exp, log, and **reasonable** combinations of those...

Explicit formula for separable potentials

If $\phi(\beta) = \sum_{i=1}^{d} h(\beta_i)$ with $h : \mathbb{R} \to \mathbb{R}$ **tame** and even,

$$\phi_\infty(\beta) \propto \lim_{s\to\infty} \frac{1}{s} h^{-1}\left[\phi(s\beta)\right]$$

Main result

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

$$\mathcal{I} = \{\beta^* \,:\, y_i \langle \beta^*, x_i \rangle \geq 1 \ \forall i\}$$

Main result

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

$$\mathcal{I} = \{\beta^* \, : \, y_i \langle \beta^*, x_i \rangle \geq 1 \; \forall i\}$$

Theorem

The mirror flow iterates converge **in direction** towards $\bar{\beta}$ satisfying

$$\bar{\beta} \propto \operatorname{argmin} \left\{ \phi_\infty(\beta^*) \, : \, \beta^* \in \mathcal{I} \right\}$$

Main result

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right)$$

$$\mathcal{I} = \{\beta^* \ : \ y_i \langle \beta^*, x_i \rangle \geq 1 \ \forall i\}$$

Theorem

The mirror flow iterates converge **in direction** towards $\bar{\beta}$ satisfying

$$\bar{\beta} \propto \operatorname{argmin} \left\{ \phi_\infty(\beta^*) \ : \ \beta^* \in \mathcal{I} \right\}$$

(provided that ϕ admits a horizon and that the argmin is nonempty)

Main result

$$\min_{\beta \in \mathbb{R}^d} L(\beta) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right) \qquad \mathcal{I} = \{\beta^* : y_i \langle \beta^*, x_i \rangle \geq 1 \ \forall i\}$$

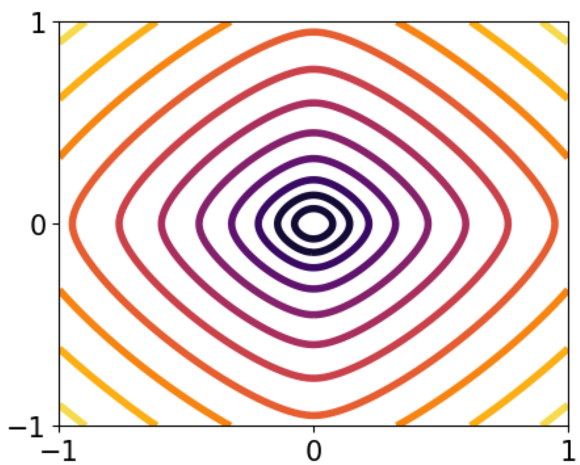
Theorem

The mirror flow iterates converge **in direction** towards $\bar{\beta}$ satisfying

$$\bar{\beta} \propto \operatorname{argmin}\left\{\phi_\infty(\beta^*) : \beta^* \in \mathcal{I}\right\}$$

(provided that ϕ admits a horizon and that the argmin is nonempty)

Application: hyperbolic potential



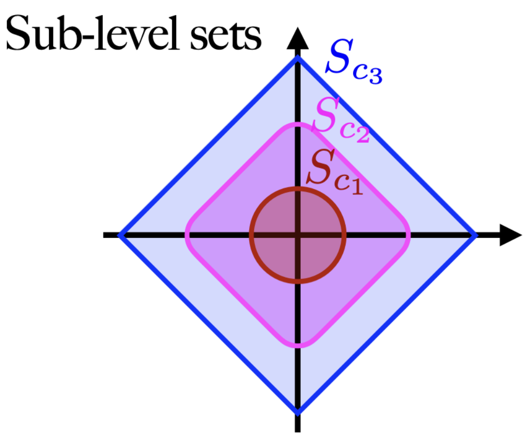
$$\phi(\beta) = \sum_{i=1}^{d} \left(\beta_i \operatorname{arcsinh}(\beta_i) - \sqrt{\beta_i^2 + 1}\right)$$

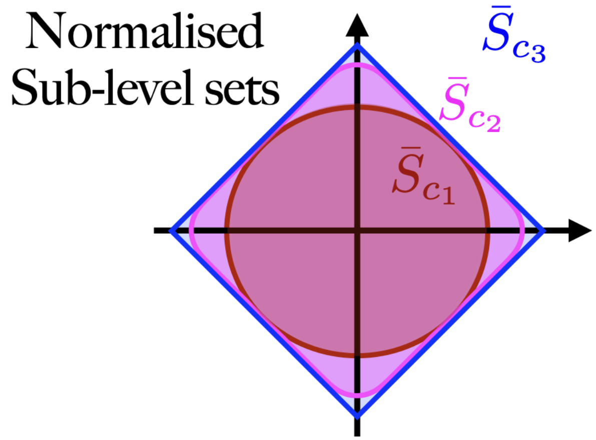
$$\phi_\infty(\beta) \propto \|\beta\|_1$$

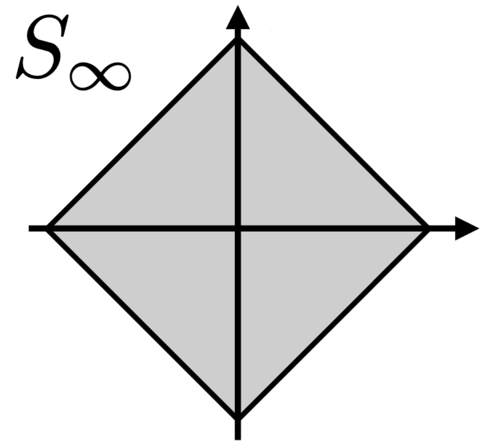
Implicit bias towards **sparsity** in diagonal neural nets

(known result, different proof)

Conclusions



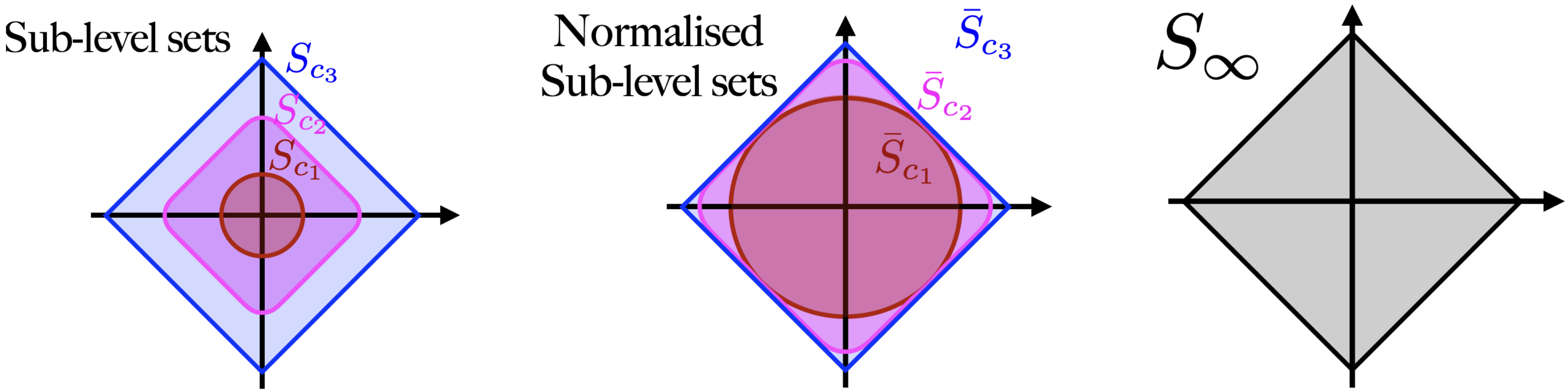




$$S_c = \{\beta \, : \, \phi(\beta) \leq c\} \qquad \bar{S}_c = S_c / \max_{\beta \in S_c} \|\beta\|$$

- Building an understanding of optimization **at infinity** through **horizon function**.

Conclusions

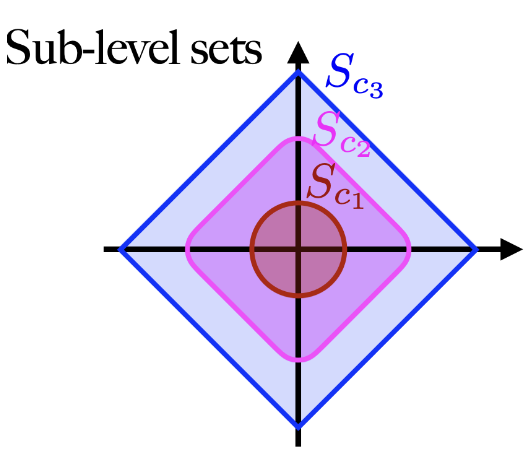


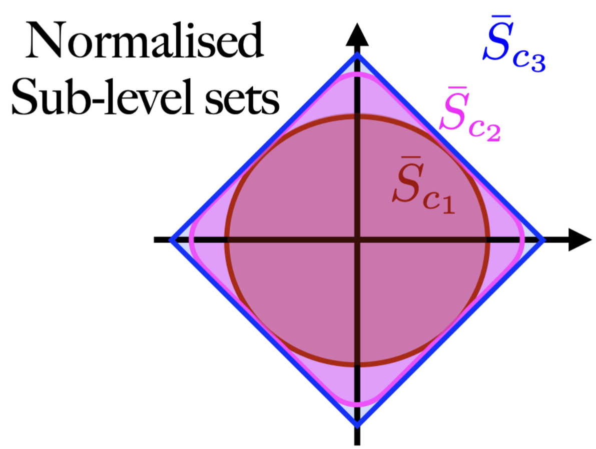
$$S_c = \{\beta \,:\, \phi(\beta) \leq c\}$$

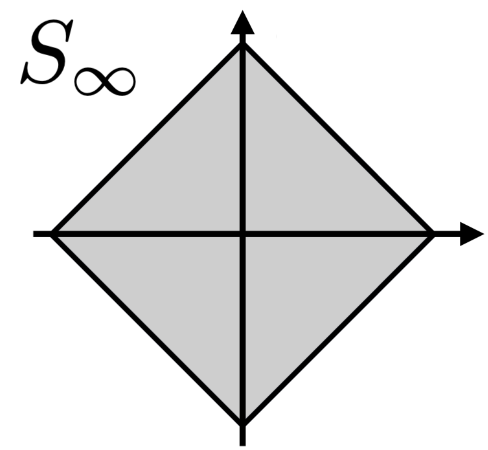
$$\bar{S}_c = S_c / \max_{\beta \in S_c} \|\beta\|$$

- Building an understanding of optimization **at infinity** through **horizon function**.
- Convergence **rates**? Degenerate case?

Conclusions





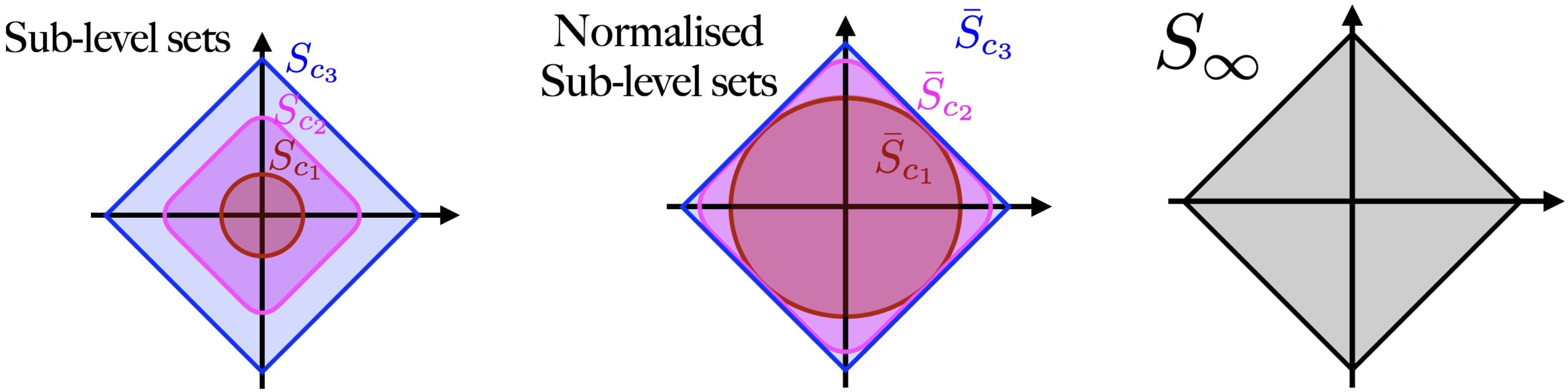


$$S_c = \{\beta \,:\, \phi(\beta) \leq c\}$$

$$\bar{S}_c = S_c / \max_{\beta \in S_c} \|\beta\|$$

- Building an understanding of optimization **at infinity** through **horizon function**.
- Convergence **rates**? Degenerate case?
- **Strong assumptions:** ϕ is defined **everywhere** and **coercive** (excludes $-\log(\beta), \beta\log(\beta), -\sqrt{\beta}...$)

Conclusions



$S_c = \{\beta \,:\, \phi(\beta) \le c\}$ $\qquad$ $\bar{S}_c = S_c / \max_{\beta \in S_c} \|\beta\|$

- Building an understanding of optimization **at infinity** through **horizon function**.
- Convergence **rates**? Degenerate case?
- **Strong assumptions:** ϕ is defined **everywhere** and **coercive** (excludes $-\log(\beta), \beta \log(\beta), -\sqrt{\beta}$...)

Thank you ! (paper out on arXiv:2406.12763)