Nonlinearly Preconditioned Gradient Methods under Generalized Smoothness

Panos Patrinos

panos.patrinos@esat.kuleuven.be

IMT Lucca, O4LC 2025

KU Leuven, ESAT-STADIUS

Goal of the talk

- **fresh view** on adaptive/preconditioned GD methods through **old, forgotten** tool

Φ-convexity

- abstract viewpoint $\Longrightarrow$ deeper understanding
- couples preconditioning with the "right" function class
- leads to new algorithms (**hyperbolic GD**,...)

K. Oikonomidis, J. Quan, E. Laude and P. Patrinos, *Nonlinearly Preconditioned Gradient Methods under Generalized Smoothness.* ICML, **oral**, **2025**

E. Laude and P. Patrinos, *Anisotropic proximal gradient.* Mathematical Programming **12**(4):747-756, **2025**

E. Laude, A. Themelis and P. Patrinos, *Dualities for non-Euclidean smoothness and strong convexity under the light of generalized conjugacy,* SIOPT **33**(4):2721-2749, **2023**

Outline

Is gradient descent all we need?

GD on (nonconvex) $f \in \mathcal{C}^1(\mathbb{R}^n)$

$$x^{k+1} = x^k - \gamma \nabla f(x^k)$$

---

theory

- appropriate class: functions with Lipschitz gradients

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \qquad \forall x, y \in \mathbb{R}^n$$

Is gradient descent all we need?

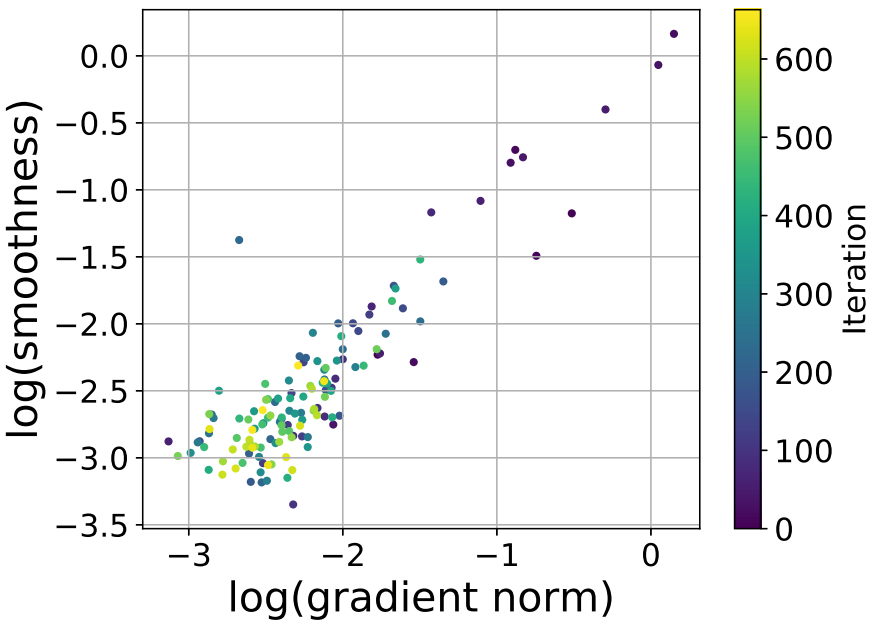
GD on (nonconvex) $f \in \mathcal{C}^1(\mathbb{R}^n)$

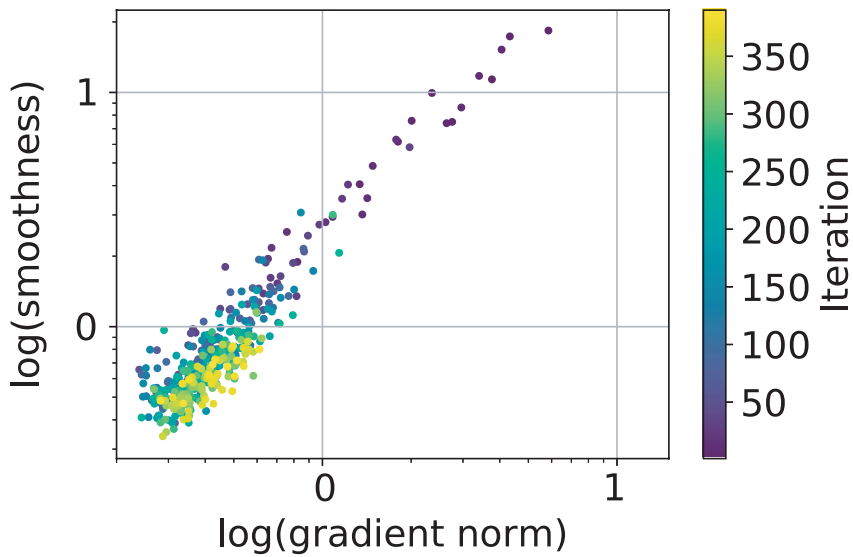
$$x^{k+1} = x^k - \gamma \nabla f(x^k)$$

---

practice

- function class is too narrow





(L_0, L_1)-**smoothness**: $\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|, \qquad \forall x \in \mathbb{R}^n$

- **adaptive methods/nonlinear preconditioning** work better

J. Zhang, H. Tianxing, S. Sra and A. Jadbabaie, *Why gradient clipping accelerates training: A theoretical justification for adaptivity,* ICLR, 2020

Outline

Nonlinearly preconditioned GD

$$x^{k+1} = x^k - \gamma G(\nabla f(x^k))$$

for some **preconditioner** $G : \mathbb{R}^n \to \mathbb{R}^n$

- Normalized GD $\qquad G(y) = \frac{y}{\|y\|+\varepsilon}$

$$x^{k+1} = x^k - \gamma \frac{\nabla f(x^k)}{\|\nabla f(x^k)\| + \varepsilon}$$

- Gradient Clipping $\qquad G(y) = \mathbf{min}\left\{1, \frac{\lambda}{\|y\|}\right\} y$

$$x^{k+1} = x^k - \gamma\, \mathbf{min}\left\{1, \frac{\lambda}{\|\nabla f(x^k)\|}\right\} \nabla f(x^k)$$

- **isotropic preconditioner**: direction does not change, adaptive stepsize

R. Pascanu, T. Mikolov and Y. Bengio, *On the difficulty of training Recurrent Neural Networks,* ICML, 2013

Nonlinearly preconditioned GD

$$x^{k+1} = x^k - \gamma G(\nabla f(x^k))$$

for some **preconditioner** $G : \mathbb{R}^n \to \mathbb{R}^n$

- Adagrad (without memory) $\qquad G_i(y_i) = \frac{y_i}{\sqrt{\varepsilon + y_i^2}}$

$$x_i^{k+1} = x_i^k - \gamma \frac{\nabla_i f(x^k)}{\sqrt{\varepsilon + (\nabla_i f(x^k))^2}} \qquad i = 1, \dots, n$$

- Adam (without memory and momentum) $\qquad G_i(y_i) = \frac{y_i}{|y_i| + \varepsilon}$

$$x_i^{k+1} = x_i^k - \gamma \frac{\nabla_i f(x^k)}{|\nabla_i f(x^k)| + \varepsilon} \qquad i = 1, \dots, n$$

- **separable preconditioner**: direction changes, widely used in practice

---

J. Duchi, E. Hazan and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization,* JMLR, 2011

D.P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization,* ICLR, 2015

Nonlinearly preconditioned GD

$$x^{k+1} = x^k - \gamma G(\nabla f(x^k))$$

for some **preconditioner** $G : \mathbb{R}^n \to \mathbb{R}^n$

desiderata for nonlinear preconditioner

1. $\lim_{y \to 0} G(y) = G(0) = 0$: algorithm stops as soon as $\nabla f(x) = 0$
2. $\lim_{\|y\| \to \infty} \frac{\|G(y)\|}{\|y\|} = 0$: prevents exploding gradients
3. $\lim_{\|y\| \to 0} \frac{\|G(y)\|}{\|y\|} > 0$: prevents vanishing gradients
4. $R = G \circ \nabla f$ is Lipschitz-like: Jacobian of R remains bounded

Anisotropic GD

$$x^{k+1} = x^k - \gamma G(\nabla f(x^k))$$

Anisotropic GD

$$x^{k+1} = x^k - \gamma \nabla \phi^*(\nabla f(x^k))$$

---

key observation

all G seen so far are **gradients of Lipschitz smooth, convex functions**

$$G(y) = \nabla \phi^*(y)$$

$\phi^*(y) = \mathbf{sup}_{x \in \mathbb{R}^n} \{\langle y, x \rangle - \phi(x)\}$ is convex conjugate of $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$

Anisotropic GD

$$x^{k+1} = x^k - \gamma \nabla \phi^*(\nabla f(x^k))$$

---

key observation

all G seen so far are **gradients of Lipschitz smooth, convex functions**

$$G(y) = \nabla \phi^*(y)$$

$h^* : \mathbb{R} \to \mathbb{R}$ cvx, L-smooth, even with

$$(h^*)'(0) = 0$$

- **isotropic**: $\phi^*(y) = h^*(\|y\|)$

- **separable**: $\phi^*(y) = \sum_{i=1}^n (h^*)'(y_i)$

$\phi^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle y, x \rangle - \phi(x)\}$ is convex conjugate of $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$

Anisotropic GD

$$x^{k+1} = x^k - \gamma \nabla \phi^*(\nabla f(x^k))$$

---

key observation

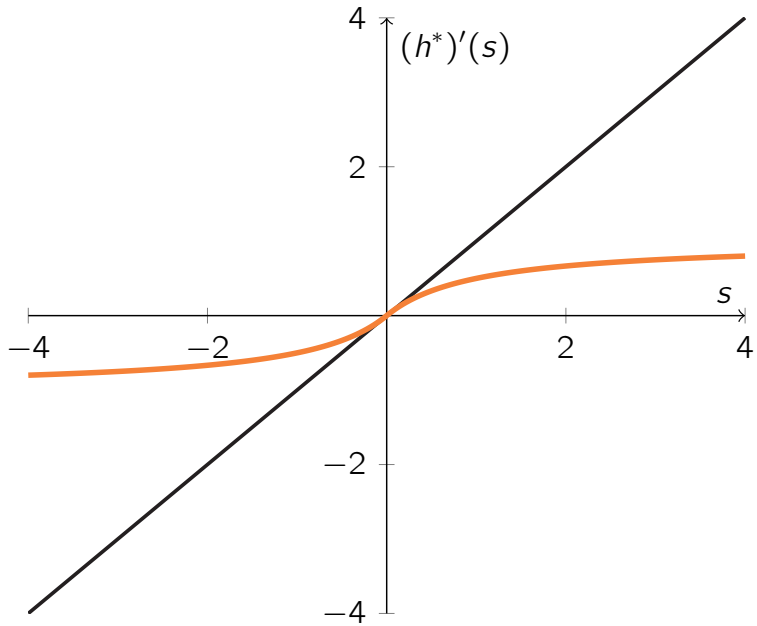
all G seen so far are **gradients of Lipschitz smooth, convex functions**

$$G(y) = \nabla \phi^*(y)$$

$h^* : \mathbb{R} \to \mathbb{R}$ cvx, L-smooth, even with

$$(h^*)'(0) = 0$$

- **isotropic**: $\phi^*(y) = h^*(\|y\|)$
  - normalized GD $(h^*)'(t) = \frac{t}{|t|+\varepsilon}$
- **separable**: $\phi^*(y) = \sum_{i=1}^n (h^*)'(y_i)$



$\phi^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle y, x \rangle - \phi(x)\}$ is convex conjugate of $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$

Anisotropic GD

$$x^{k+1} = x^k - \gamma \nabla \phi^*(\nabla f(x^k))$$

---

key observation

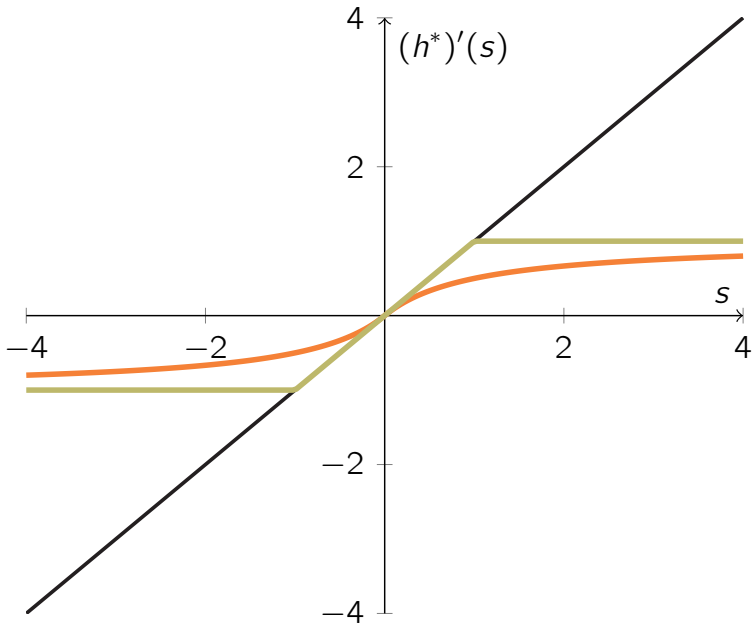
all G seen so far are **gradients of Lipschitz smooth, convex functions**

$$G(y) = \nabla \phi^*(y)$$

$h^* : \mathbb{R} \to \mathbb{R}$ cvx, L-smooth, even with

$$(h^*)'(0) = 0$$

- **isotropic**: $\phi^*(y) = h^*(\|y\|)$
  - normalized GD $(h^*)'(t) = \frac{t}{|t|+\varepsilon}$
  - clipping $(h^*)'(t) = \Pi_{|\cdot| \le \lambda}(t)$
- **separable**: $\phi^*(y) = \sum_{i=1}^n (h^*)'(y_i)$



$\phi^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle y, x \rangle - \phi(x)\}$ is convex conjugate of $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$

Anisotropic GD

$$x^{k+1} = x^k - \gamma \nabla \phi^*(\nabla f(x^k))$$

---

key observation

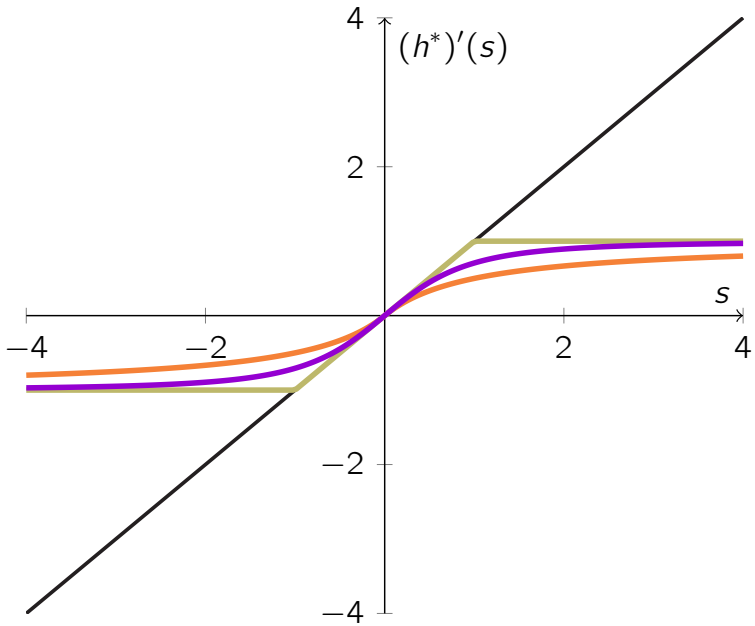
all G seen so far are **gradients of Lipschitz smooth, convex functions**

$$G(y) = \nabla \phi^*(y)$$

$h^* : \mathbb{R} \to \mathbb{R}$ cvx, L-smooth, even with

$$(h^*)'(0) = 0$$

- **isotropic**: $\phi^*(y) = h^*(\|y\|)$
  - normalized GD $(h^*)'(t) = \frac{t}{|t|+\varepsilon}$
  - clipping $(h^*)'(t) = \Pi_{|\cdot| \le \lambda}(t)$
- **separable**: $\phi^*(y) = \sum_{i=1}^n (h^*)'(y_i)$
  - Adagrad $(h^*)'(t) = \frac{t}{\sqrt{t^2+\varepsilon}}$



$\phi^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle y, x \rangle - \phi(x)\}$ is convex conjugate of $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$

Anisotropic GD

$$x^{k+1} = x^k - \gamma \nabla \phi^*(\nabla f(x^k))$$

---

key observation

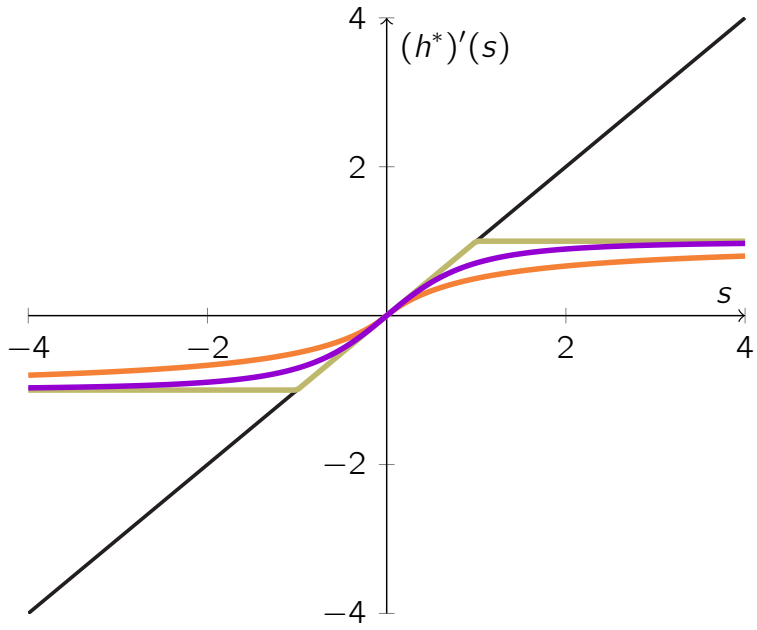
all G seen so far are **gradients of Lipschitz smooth, convex functions**

$$G(y) = \nabla \phi^*(y)$$

$h^* : \mathbb{R} \to \mathbb{R}$ cvx, L-smooth, even with

$$(h^*)'(0) = 0$$

- **isotropic**: $\phi^*(y) = h^*(\|y\|)$
  - normalized GD $(h^*)'(t) = \frac{t}{|t|+\varepsilon}$
  - clipping $(h^*)'(t) = \Pi_{|\cdot| \le \lambda}(t)$
- **separable**: $\phi^*(y) = \sum_{i=1}^n (h^*)'(y_i)$
  - Adagrad $(h^*)'(t) = \frac{t}{\sqrt{t^2+\varepsilon}}$
  - Adam $(h^*)'(t) = \frac{t}{|t|+\varepsilon}$



$\phi^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle y, x \rangle - \phi(x)\}$ is convex conjugate of $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$

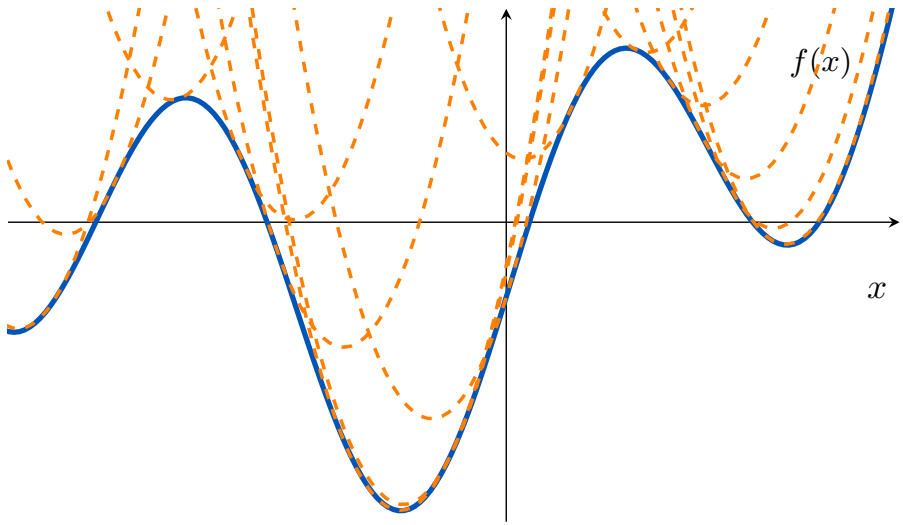
Outline

Euclidean smoothness and GD

$$x^{k+1} = x^k - \gamma \nabla f(x^k) =: T_\gamma(x^k)$$

$f \in \mathcal{C}^1(\mathbb{R}^n)$ is called *L*-**smooth** if **descent lemma** holds

$$f(x) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{L}{2}\|x - \bar{x}\|^2 \qquad \forall \bar{x}, x \in \mathbb{R}^n$$



GD as **majorization-minimization**: minimize rhs of descent lemma

Euclidean smoothness and GD

$$x^{k+1} = x^k - \gamma \nabla f(x^k) =: T_\gamma(x^k)$$

$f \in \mathcal{C}^1(\mathbb{R}^n)$ is called *L*-**smooth** if **descent lemma** holds

$$f(x) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{L}{2}\|x - \bar{x}\|^2 \qquad \forall \bar{x}, x \in \mathbb{R}^n$$

0th, 1st, 2nd order necessary and sufficient conditions, rich calculus

f $L-$smooth

$\Updownarrow$

$\frac{L}{2}\|\cdot\|^2 - f$ convex

$\Updownarrow$

$T_{L^{-1}}$ monotone $\quad \Longleftarrow \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \ \forall x, y \in \mathbb{R}^n$

$\Updownarrow \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \Updownarrow$

$\nabla^2 f(x) \preceq L I_n, \ \forall x \in \mathbb{R}^n \quad \Longleftarrow \quad \|\nabla^2 f(x)\| \leq L, \ \forall x \in \mathbb{R}^n$

Euclidean smoothness and GD

$$x^{k+1} = x^k - \gamma \nabla f(x^k) =: T_\gamma(x^k)$$

$f \in \mathcal{C}^1(\mathbb{R}^n)$ is called **L-smooth** if **descent lemma** holds

$$\begin{aligned} f(x) &\leq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{L}{2}\|x - \bar{x}\|^2 \qquad \forall \bar{x}, x \in \mathbb{R}^n \\ &= f(\bar{x}) + \tfrac{1}{2L}\|L(x - T_{L^{-1}}(\bar{x}))\|^2 - \tfrac{1}{2L}\|L(\bar{x} - T_{L^{-1}}(\bar{x}))\|^2 \end{aligned}$$

Euclidean smoothness and GD

$$x^{k+1} = x^k - \gamma \nabla f(x^k) =: T_\gamma(x^k)$$

$f \in \mathcal{C}^1(\mathbb{R}^n)$ is called L-**smooth** if **descent lemma** holds

$$\begin{aligned} f(x) &\le f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{L}{2}\|x - \bar{x}\|^2 \qquad \forall \bar{x}, x \in \mathbb{R}^n \\ &= f(\bar{x}) + \tfrac{1}{2L}\|L(x - T_{L^{-1}}(\bar{x}))\|^2 - \tfrac{1}{2L}\|L(\bar{x} - T_{L^{-1}}(\bar{x}))\|^2 \\ &= f(\bar{x}) + (L \star \phi)(x - T_{L^{-1}}(\bar{x})) - (L \star \phi)(\bar{x} - T_{L^{-1}}(\bar{x})) \end{aligned}$$

where $\phi = \frac{1}{2}\|\cdot\|^2$, $L \star \phi = L\phi(L^{-1}\cdot)$ (epi-scaling)

Anisotropic smoothness

$$x^+ = x - \gamma \nabla \phi^\star(\nabla f(x)) =: T_\gamma(x)$$

$f \in \mathcal{C}^1(\mathbb{R}^n)$ is called **anisotropically smooth** (aniso-smooth) with respect to a **reference function** $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ if **anisotropic descent lemma** holds

$$f(x) \leq f(\bar{x}) + (L \star \phi)(x - T_{L^{-1}}(\bar{x})) - (L \star \phi)(\bar{x} - T_{L^{-1}}(\bar{x})) \qquad \forall \bar{x}, x \in \mathbb{R}^n$$

Assumption

A1 $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, lsc, strongly convex and even with $\phi(0) = 0$

A2 $\phi \in \mathcal{C}^2(\mathbf{int\,dom}\,\phi)$, $\|\nabla \phi(x^\nu)\| \to \infty$ for $\mathbf{int\,dom}\,\phi \ni x^\nu \to x \in \mathbf{bdry\,dom}\,\phi$

- **A1** is standing assumption
- **A2** implies ϕ^* is $\mathcal{C}^2(\mathbb{R}^n)$ (sometimes needed)
- common choices for ϕ: pick $h : \mathbb{R} \to \overline{\mathbb{R}}$ satisfying **A1** (and **A2**)

isotropic

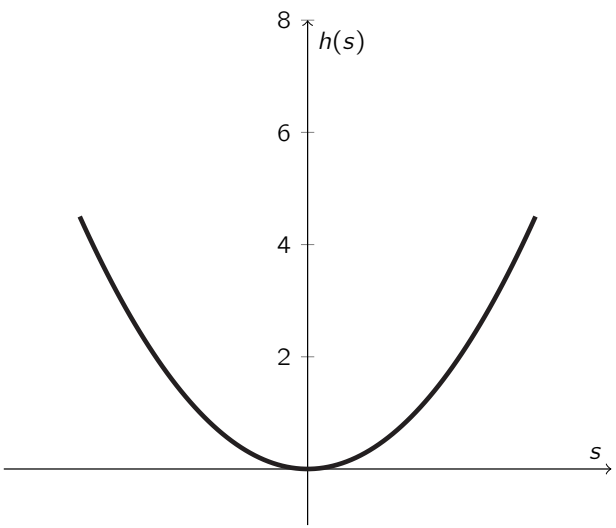
$$\phi(x) = h(\|x\|)$$
$$\nabla \phi^*(y) = (h^*)'(\|y\|)\overline{\mathbf{sign}}(y)$$

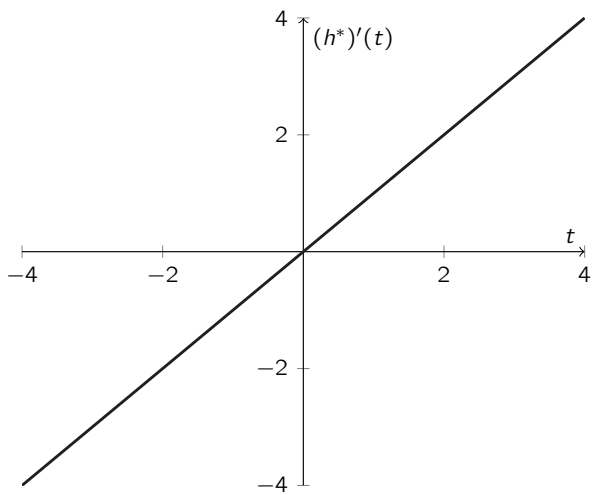
separable

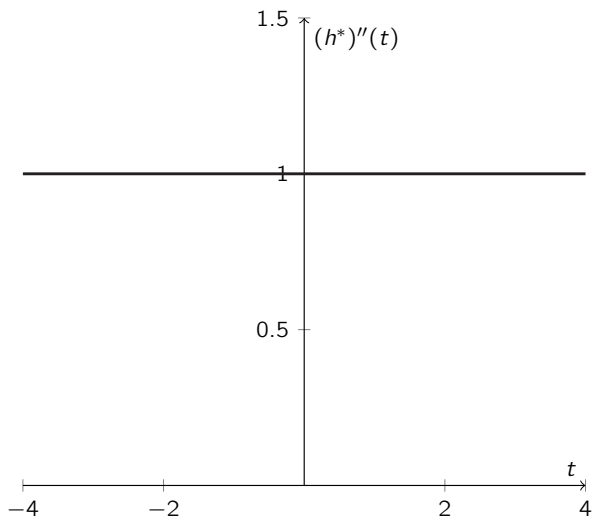
$$\phi(x) = \textstyle\sum_{i=1}^n h(x_i)$$
$$\nabla \phi^*(x) = ((h^*)'(x_1), \ldots, (h^*)'(x_n))$$

Examples of reference functions

$h(s)$	$\mathbf{dom}\, h$	$h^{*\prime}(t)$	$h^{*\prime\prime}(t)$	A1/A2



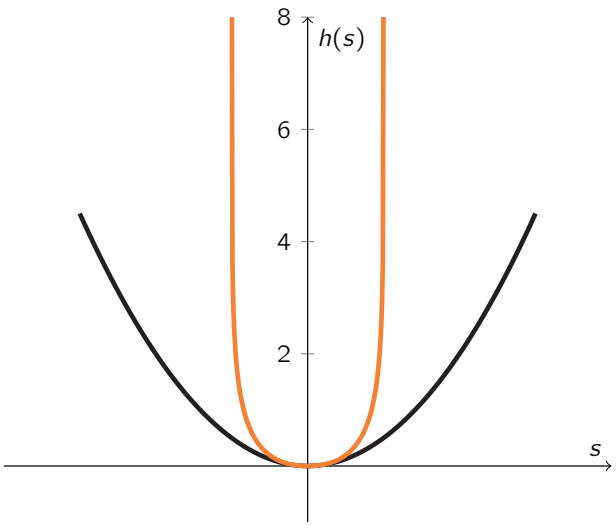


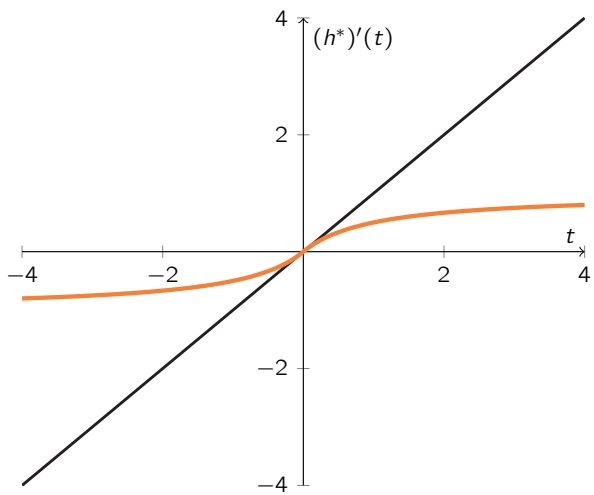


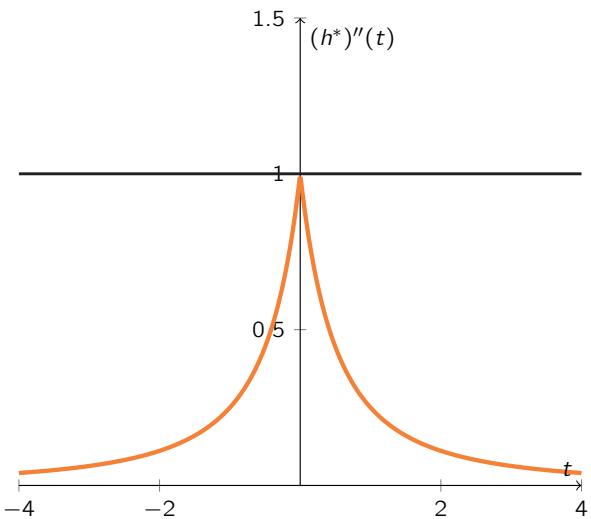
- ϕ grows **faster** than $\|\cdot\|^2 \implies$ **larger function class** than L-smooth

Examples of reference functions

$h(s)$	$\mathbf{dom}\, h$	$h^{*\prime}(t)$	$h^{*\prime\prime}(t)$	A1/A2
$\lambda(-\lvert s\rvert - \mathbf{ln}(1-\lvert s\rvert)$	$(-1,1)$	$\frac{t}{\lvert t\rvert+\lambda}$	$\frac{\lambda}{(\lvert t\rvert+\lambda)^2}$	✓/✓



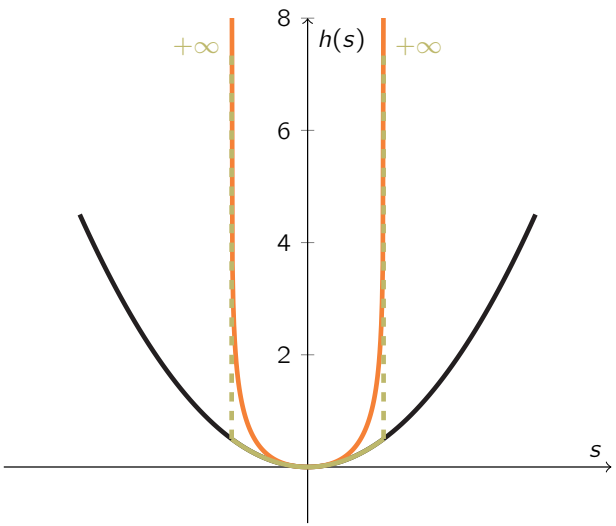


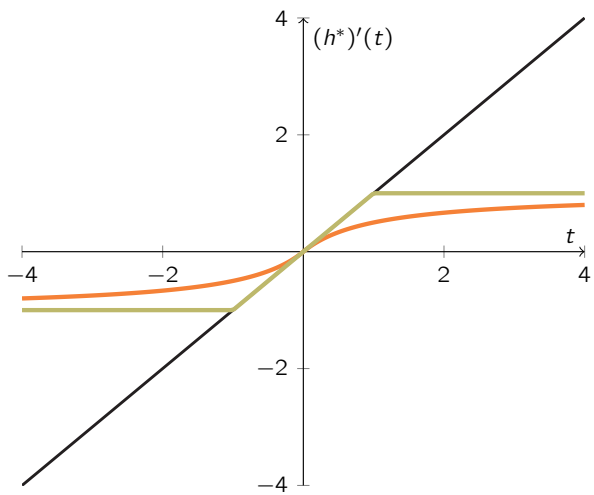


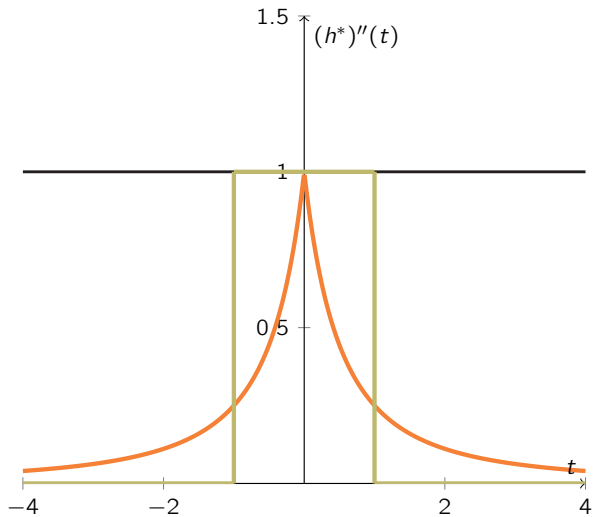
- ϕ grows **faster** than $\|\cdot\|^2 \implies$ **larger function class** than L-smooth

Examples of reference functions

$h(s)$	$\mathbf{dom}\, h$	$h^{*\prime}(t)$	$h^{*\prime\prime}(t)$	A1/A2
$\lambda(-\lvert s\rvert - \mathbf{ln}(1-\lvert s\rvert))$	$(-1,1)$	$\frac{t}{\lvert t\rvert+\lambda}$	$\frac{\lambda}{(\lvert t\rvert+\lambda)^2}$	✓/✓
$\frac{\lambda}{2}s^2 + \delta_{\lvert\cdot\rvert\leq 1}(s)$	$[-1,1]$	$\Pi_{\lvert\cdot\rvert\leq\lambda}(t)$	$\partial_C(\Pi_{\lvert\cdot\rvert\leq\lambda})(t)$	✓/✗



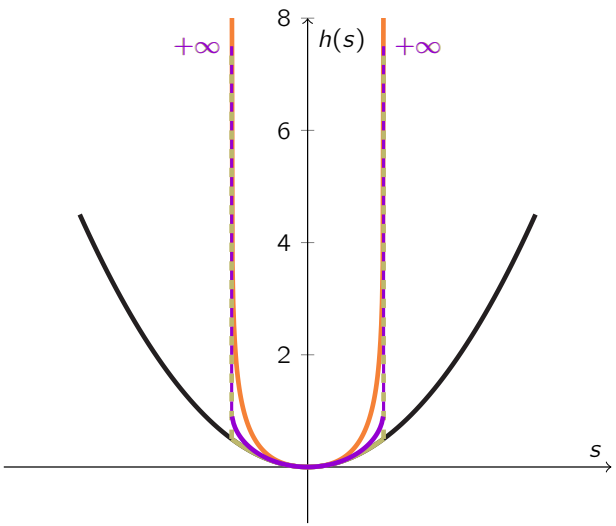


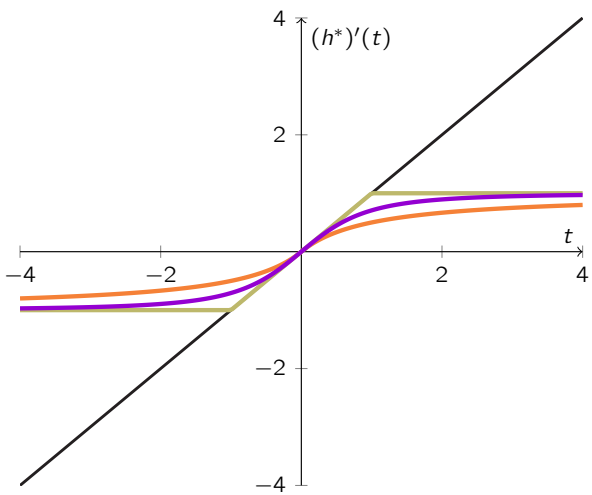


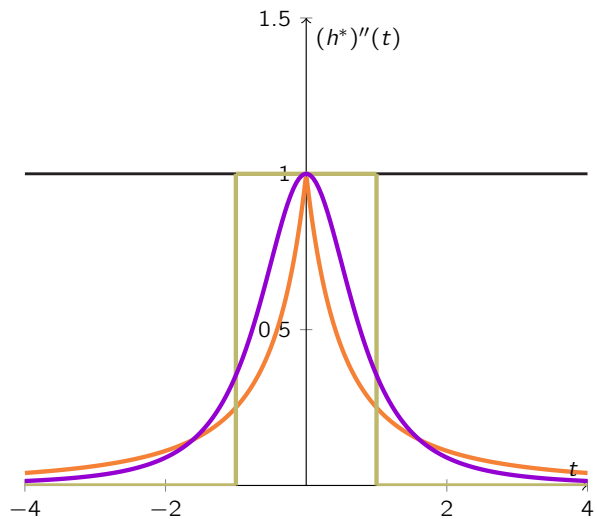
- ϕ grows **faster** than $\|\cdot\|^2 \implies$ **larger function class** than L-smooth

Examples of reference functions

$h(s)$	$\mathbf{dom}\, h$	$h^{*\prime}(t)$	$h^{*\prime\prime}(t)$	A1/A2
$\lambda(-\lvert s\rvert - \ln(1-\lvert s\rvert))$	$(-1,1)$	$\frac{t}{\lvert t\rvert+\lambda}$	$\frac{\lambda}{(\lvert t\rvert+\lambda)^2}$	✓/✓
$\frac{\lambda}{2}s^2 + \delta_{\lvert\cdot\rvert\leq 1}(s)$	$[-1,1]$	$\Pi_{\lvert\cdot\rvert\leq\lambda}(t)$	$\partial_C(\Pi_{\lvert\cdot\rvert\leq\lambda})(t)$	✓/✗
$\lambda(1-\sqrt{1-s^2})$	$[-1,1]$	$\frac{t}{\sqrt{t^2+\lambda^2}}$	$\lambda(t^2+\lambda^2)^{-3/2}$	✓/✓



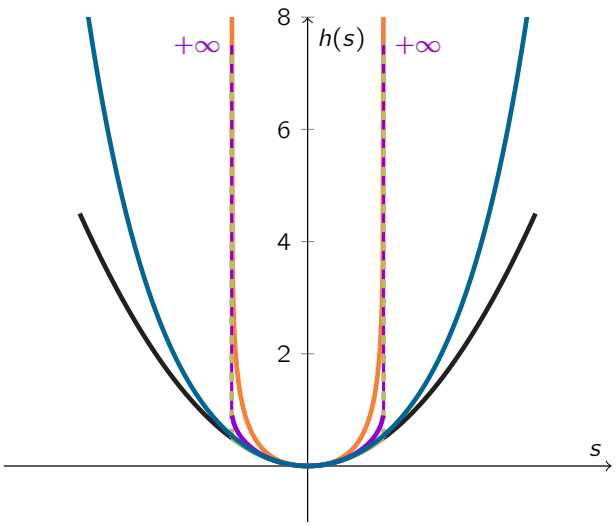


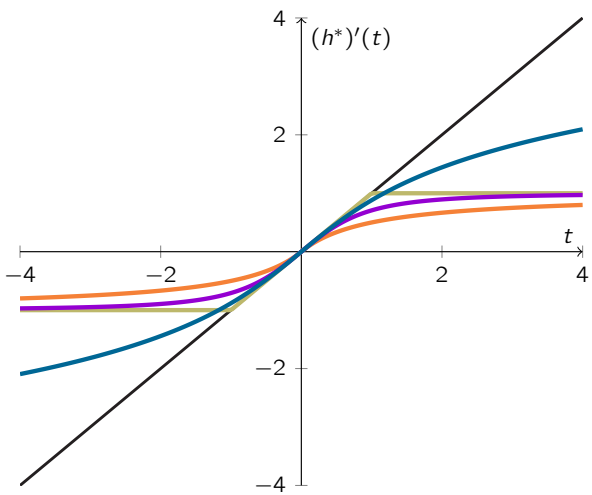


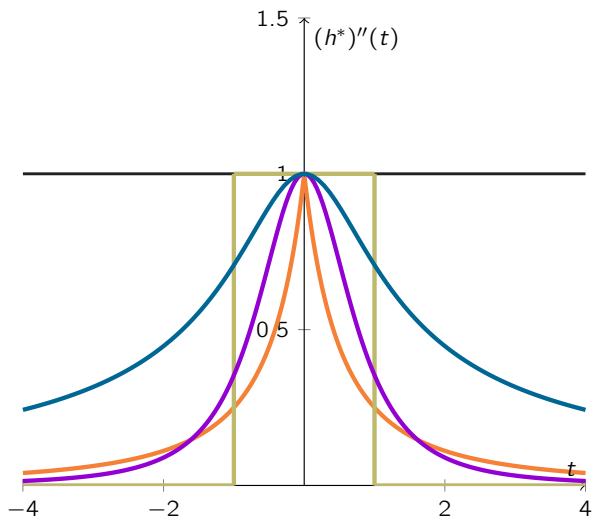
- ϕ grows **faster** than $\|\cdot\|^2 \implies$ **larger function class** than L-smooth

Examples of reference functions

$h(s)$	$\mathbf{dom}\, h$	$h^{*\prime}(t)$	$h^{*\prime\prime}(t)$	A1/A2
$\lambda(-\lvert s\rvert - \mathbf{ln}(1-\lvert s\rvert))$	$(-1,1)$	$\frac{t}{\lvert t\rvert+\lambda}$	$\frac{\lambda}{(\lvert t\rvert+\lambda)^2}$	✓/✓
$\frac{\lambda}{2}s^2 + \delta_{\lvert\cdot\rvert\leq 1}(s)$	$[-1,1]$	$\Pi_{\lvert\cdot\rvert\leq\lambda}(t)$	$\partial_C(\Pi_{\lvert\cdot\rvert\leq\lambda})(t)$	✓/✗
$\lambda(1-\sqrt{1-s^2})$	$[-1,1]$	$\frac{t}{\sqrt{t^2+\lambda^2}}$	$\lambda(t^2+\lambda^2)^{-3/2}$	✓/✓
$\lambda(\mathbf{cosh}(s)-1)$	$\mathbb{R}$	$\mathbf{arcsinh}(\lambda^{-1}t)$	$\frac{1}{\sqrt{t^2+\lambda^2}}$	✓/✓



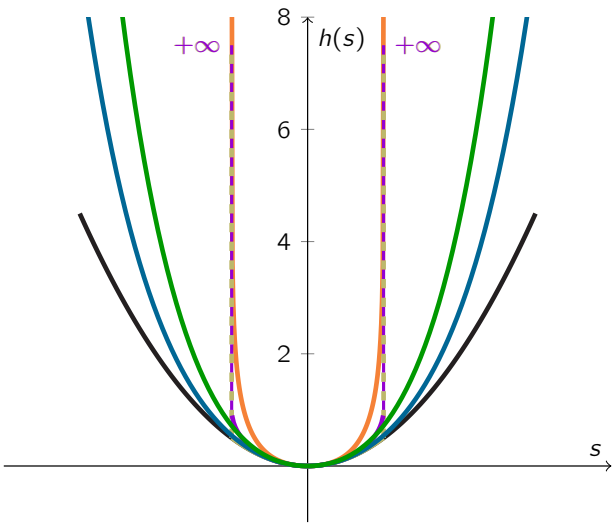


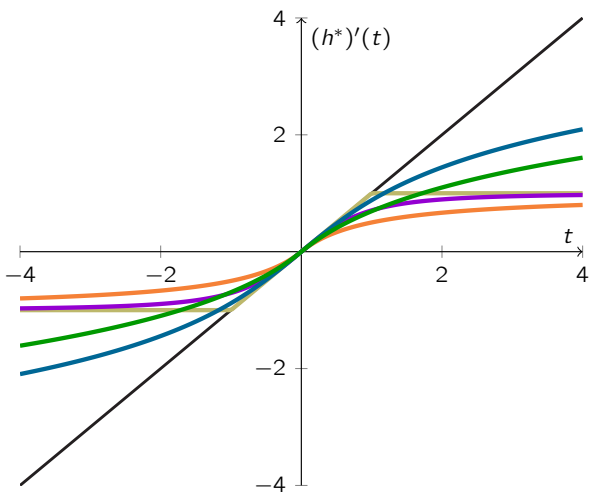


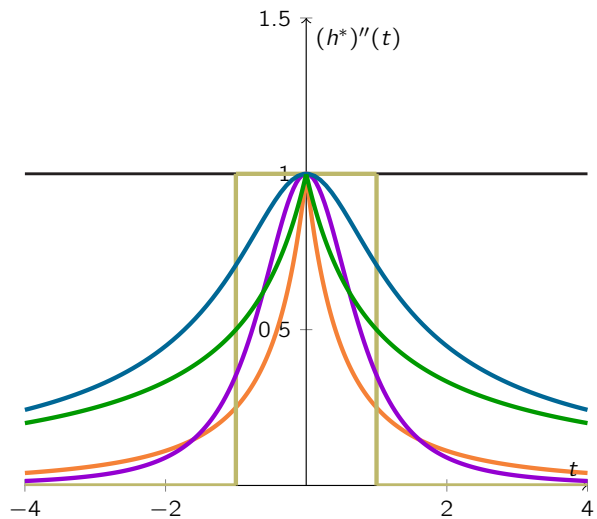
- ϕ grows **faster** than $\|\cdot\|^2 \implies$ **larger function class** than L-smooth

Examples of reference functions

$h(s)$	$\mathbf{dom}\, h$	$h^{*\prime}(t)$	$h^{*\prime\prime}(t)$	A1/A2
$\lambda(-\lvert s\rvert - \mathbf{ln}(1-\lvert s\rvert))$	$(-1,1)$	$\frac{t}{\lvert t\rvert+\lambda}$	$\frac{\lambda}{(\lvert t\rvert+\lambda)^2}$	✓/✓
$\frac{\lambda}{2}s^2 + \delta_{\lvert\cdot\rvert\leq 1}(s)$	$[-1,1]$	$\Pi_{\lvert\cdot\rvert\leq\lambda}(t)$	$\partial_C(\Pi_{\lvert\cdot\rvert\leq\lambda})(t)$	✓/✗
$\lambda(1-\sqrt{1-s^2})$	$[-1,1]$	$\frac{t}{\sqrt{t^2+\lambda^2}}$	$\lambda(t^2+\lambda^2)^{-3/2}$	✓/✓
$\lambda(\mathbf{cosh}(s)-1)$	$\mathbb{R}$	$\mathbf{arcsinh}(\lambda^{-1}t)$	$\frac{1}{\sqrt{t^2+\lambda^2}}$	✓/✓
$\lambda(\mathbf{exp}(\lvert s\rvert)-\lvert s\rvert-1)$	$\mathbb{R}$	$\mathbf{ln}(1+\lambda^{-1}\lvert t\rvert)\overline{\mathbf{sign}}(t)$	$\frac{1}{\lvert t\rvert+\lambda}$	✓/✓



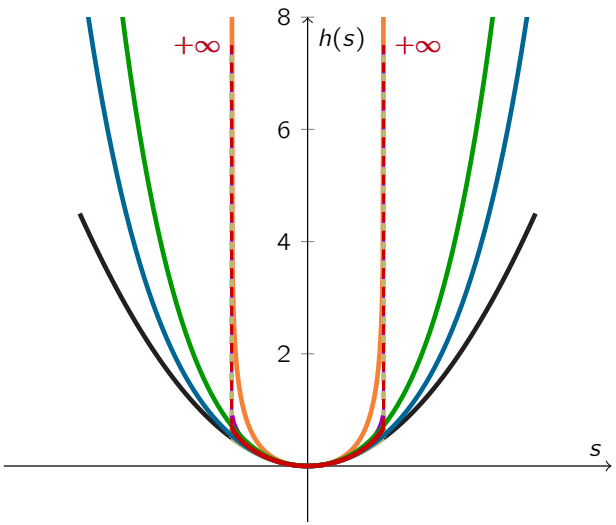


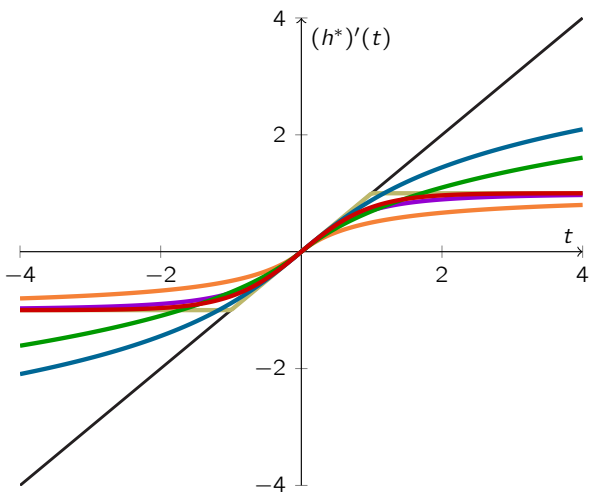


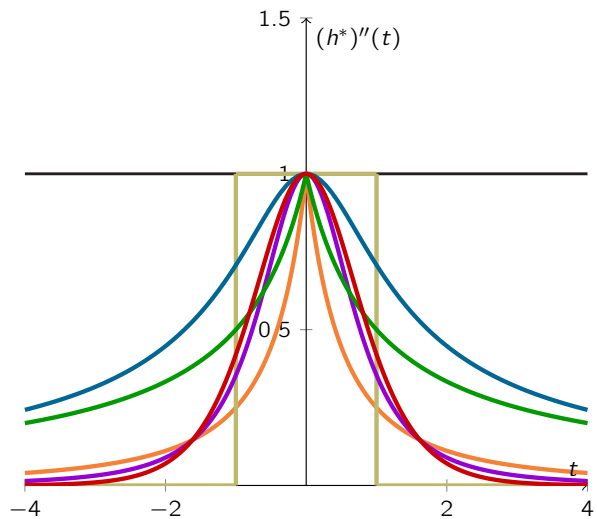
- ϕ grows **faster** than $\|\cdot\|^2 \implies$ **larger function class** than L-smooth

Examples of reference functions

$h(s)$	$\mathbf{dom}\, h$	$h^{*\prime}(t)$	$h^{*\prime\prime}(t)$	A1/A2
$\lambda(-\lvert s\rvert - \mathbf{ln}(1-\lvert s\rvert))$	$(-1,1)$	$\frac{t}{\lvert t\rvert+\lambda}$	$\frac{\lambda}{(\lvert t\rvert+\lambda)^2}$	✓/✓
$\frac{\lambda}{2}s^2 + \delta_{\lvert\cdot\rvert\le 1}(s)$	$[-1,1]$	$\Pi_{\lvert\cdot\rvert\le\lambda}(t)$	$\partial_C(\Pi_{\lvert\cdot\rvert\le\lambda}(t))$	✓/✗
$\lambda(1-\sqrt{1-s^2})$	$[-1,1]$	$\frac{t}{\sqrt{t^2+\lambda^2}}$	$\lambda(t^2+\lambda^2)^{-3/2}$	✓/✓
$\lambda(\mathbf{cosh}(s)-1)$	$\mathbb{R}$	$\mathbf{arcsinh}(\lambda^{-1}t)$	$\frac{1}{\sqrt{t^2+\lambda^2}}$	✓/✓
$\lambda(\mathbf{exp}(\lvert s\rvert)-\lvert s\rvert-1)$	$\mathbb{R}$	$\mathbf{ln}(1+\lambda^{-1}\lvert t\rvert)\overline{\mathbf{sign}}(t)$	$\frac{1}{\lvert t\rvert+\lambda}$	✓/✓
$\lambda\left(s\,\mathbf{arctanh}(s)+\mathbf{ln}\left(\sqrt{1-s^2}\right)\right)$	$(-1,1)$	$\mathbf{tanh}(\lambda^{-1}t)$	$\lambda^{-1}(1-\mathbf{tanh}^2(\lambda^{-1}t))$	✓/✓







- ϕ grows **faster** than $\|\cdot\|^2 \implies$ **larger function class** than L-smooth

Outline

Abstract convexity — some history

INF-CONVOLUTION, SOUS-ADDITIVITÉ, CONVEXITÉ DES FONCTIONS NUMÉRIQUES

PAR JEAN JACQUES MOREAU.

4.c. UN SCHÉMA GÉNÉRAL DE DUALITÉ D'ENVELOPPES. — Soient X et Y deux ensembles quelconques; dans tout ce qui suit, on suppose donnée une *fonction de couplage* (*cf.* Moreau [6], section 14)

$$c : X \times Y \to \overline{\mathbf{R}}.$$

Pour toute fonction $f : X \to \overline{\mathbf{R}}$, on appelle *fonction polaire* de f, la fonction f^c définie sur Y par

$$f^c(y) = \sup_{x \in X} [c(x, y) \dot{+} -f(x)].$$

De même, pour toute $g : Y \to \overline{\mathbf{R}}$, on note g^c, dite fonction polaire de g, la fonction définie sur X par

$$g^c(x) = \sup_{y \in Y} [c(x, y) \dot{+} -g(y)].$$

On appelle *fonction élémentaire* sur X (resp. sur Y) une fonction de la forme $x \mapsto c(x, b) \dot{+} \beta$, avec $b \in Y$ et $\beta \in \overline{\mathbf{R}}$ [resp. une fonction de la forme $y \mapsto c(a, y) \dot{+} \alpha$, avec $a \in X$ et $\alpha \in \overline{\mathbf{R}}$].

AN EXTENSION OF DUALITY-STABILITY RELATIONS TO NONCONVEX OPTIMIZATION PROBLEMS*

E. J. BALDER†

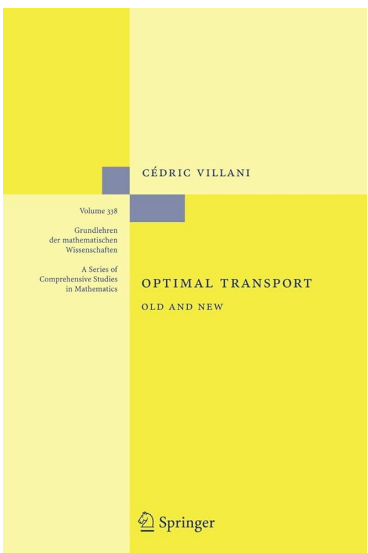
Abstract. By an effective extension of the conjugate function concept a general framework for duality-stability relations in nonconvex optimization problems can be studied. The results obtained show strong correspondences with the duality theory for convex minimization problems. In specializations to mathematical programming problems the canonical Lagrangian of the model appears as the extended Lagrangian considered in exterior penalty function methods.

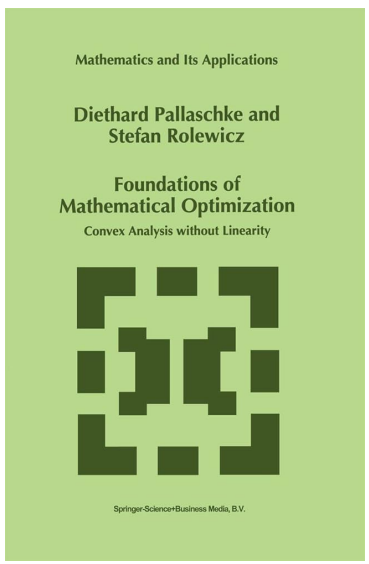
ON Φ-CONVEXITY IN EXTREMAL PROBLEMS*

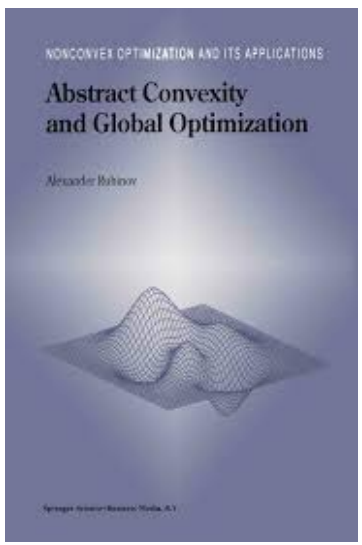
SZYMON DOLECKI† AND STANISŁAW KURCYUSZ‡

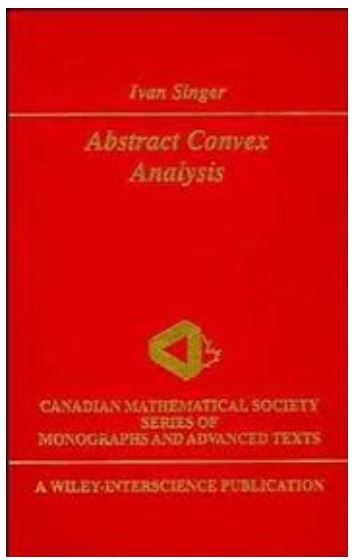
Abstract. For a class of functions Φ on an arbitrary set X, Φ-convex subsets of X and functions on X are defined, the latter being least upper bounds of some functions from Φ. Also the generalized Fenchel transform and Φ-subgradients are determined and their properties investigated.

Φ-convexity and Φ-subdifferentiability of lower-semicontinuous functions on metric spaces are examined with respect to special important families Φ. Among related results, we present a theorem on the existence of minimizing points of nonlinear functions on Banach spaces and extensions of the notion of Hölder continuity. The relevance of the theory to perturbed extremal problems is indicated.



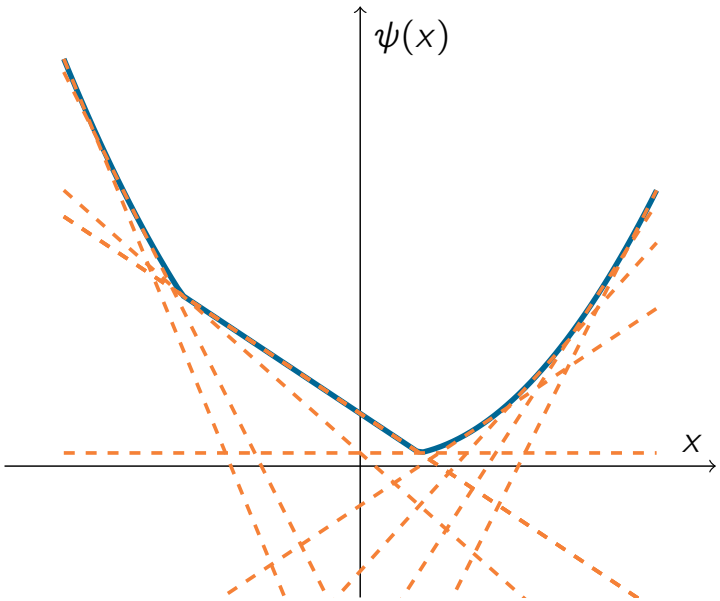






- Moreau replaced inner product with arbitrary coupling in Fenchel conjugate
- early papers: Balder, Dolecki & Kurcyusz (Φ-subdifferential, duality)
- important role in optimal transport theory (Kantorovich duality)

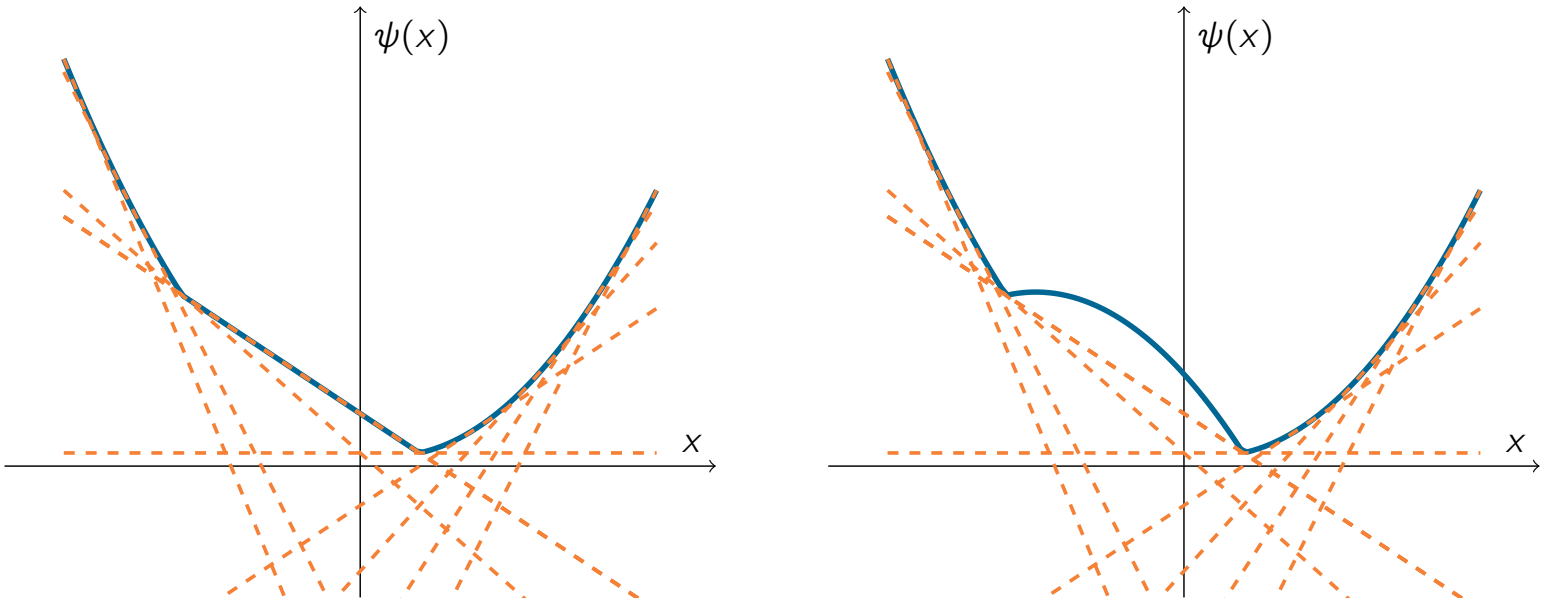
Φ-convexity



proper $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is closed convex if and only if

$$\psi(x) = \sup_{(y,\beta):\ell_{y,\beta}\leq\psi} \ell_{y,\beta}(x), \qquad \text{where} \quad \ell_{y,\beta}(x) = \langle y, x\rangle - \beta$$

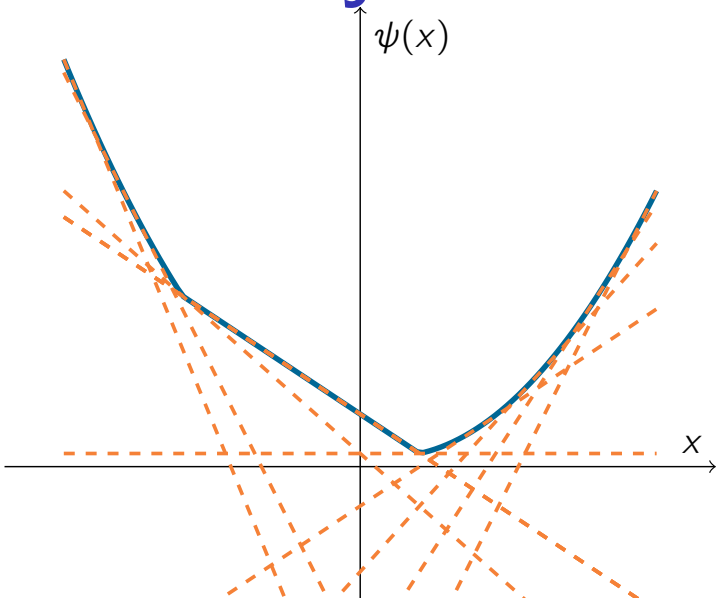
Φ-convexity

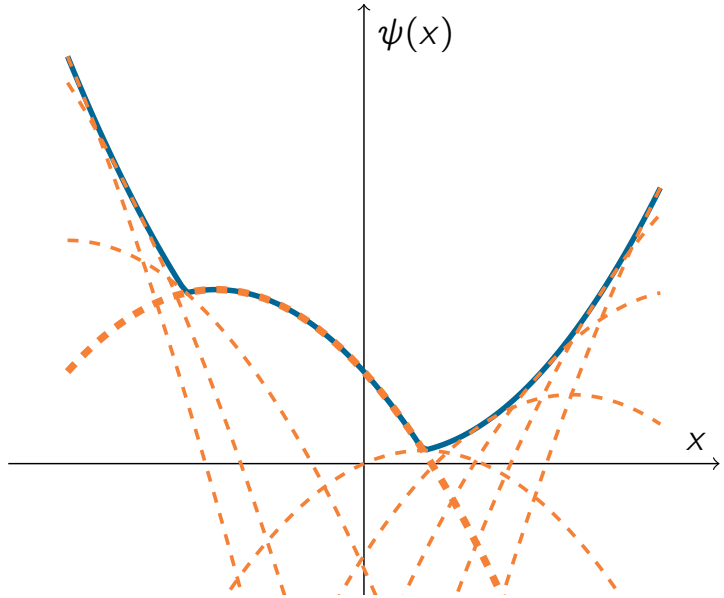


proper $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is closed convex if and only if

$$\psi(x) = \sup_{(y,\beta):\ell_{y,\beta}\leq\psi} \ell_{y,\beta}(x), \qquad \text{where} \quad \ell_{y,\beta}(x) = \langle y, x\rangle - \beta$$

Φ-convexity





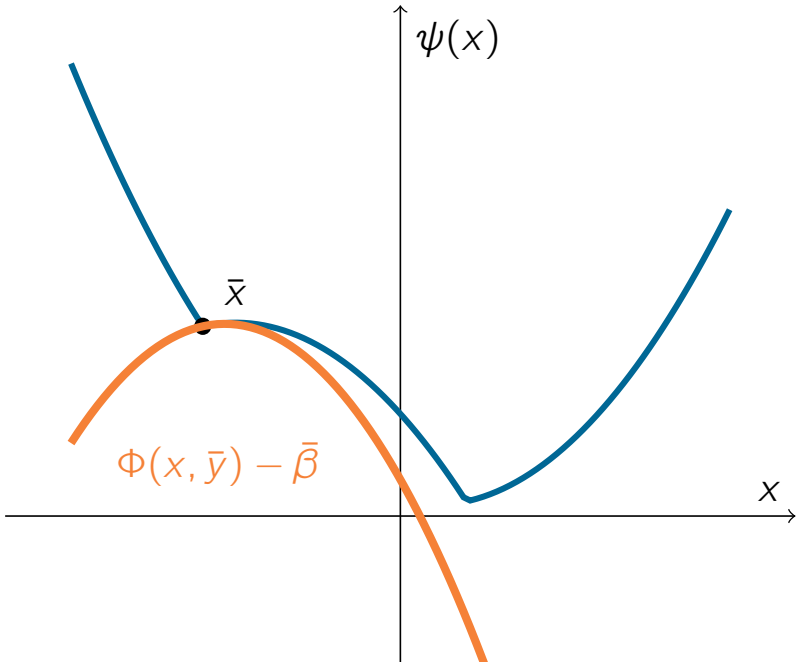
coupling function $\Phi : X \times Y \to \mathbb{R}$ where X, Y nonempty sets

Definition

$\psi : X \to \overline{\mathbb{R}}$ is called **Φ-convex** if it is the pointwise supremum of its **Φ-minorants**

$$\psi(x) = \sup_{(y,\beta):\ell_{y,\beta} \leq \psi} \ell_{y,\beta}(x) \qquad \text{where} \qquad \ell_{y,\beta}(x) = \Phi(x,y) - \beta$$

Φ-convexity



coupling function $\Phi : X \times Y \to \mathbb{R}$ where X, Y nonempty sets

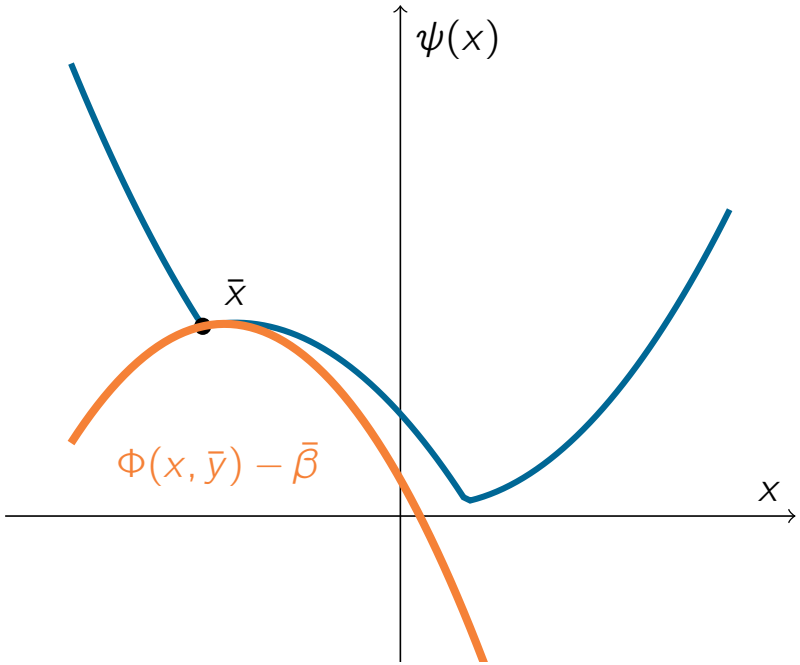
- **task**: given $\bar{x} \in \mathbb{R}^n$, find $\bar{y} \in \mathbb{R}^n$, $\bar{\beta} \in \mathbb{R}$ such that

$$\ell_{\bar{y},\bar{\beta}}(x) = \Phi(x, \bar{y}) - \bar{\beta}$$

minorizes ψ at $\bar{x}$:

$$\psi(\bar{x}) = \ell_{\bar{y},\bar{\beta}}(\bar{x})$$
$$(\forall x \in \mathbb{R}^n) \quad \psi(x) \geq \ell_{\bar{y},\bar{\beta}}(x)$$

Φ-convexity



coupling function $\Phi : X \times Y \to \mathbb{R}$ where X, Y nonempty sets

- **task**: given $\bar{x} \in \mathbb{R}^n$, find $\bar{y} \in \mathbb{R}^n$, $\bar{\beta} \in \mathbb{R}$ such that

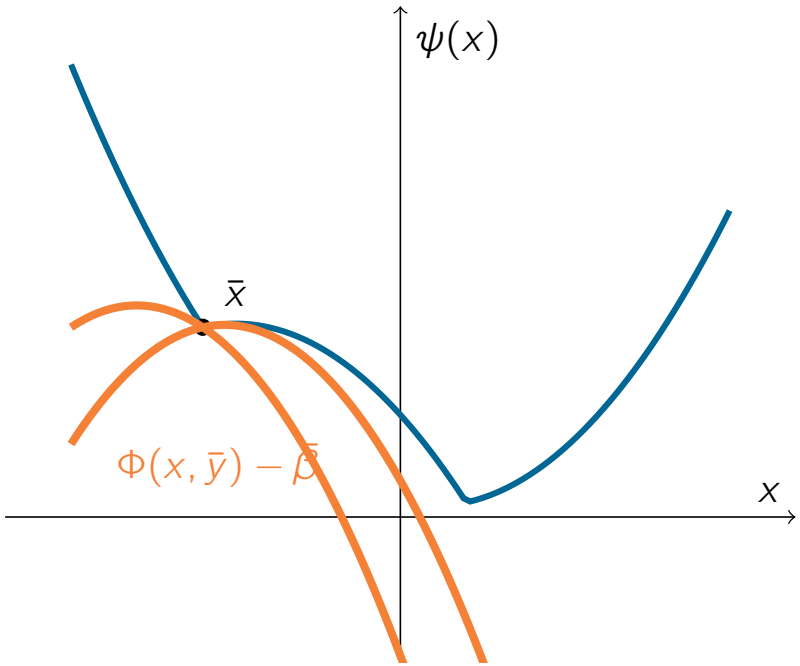
$$\ell_{\bar{y},\bar{\beta}}(x) = \Phi(x,\bar{y}) - \bar{\beta} = \psi(\bar{x}) + \Phi(x,\bar{y}) - \Phi(\bar{x},\bar{y})$$

minorizes ψ at $\bar{x}$:

$$\psi(\bar{x}) = \ell_{\bar{y},\bar{\beta}}(\bar{x}) \iff \bar{\beta} = \Phi(\bar{x},\bar{y}) - \psi(\bar{x})$$

$$(\forall x \in \mathbb{R}^n) \quad \psi(x) \geq \ell_{\bar{y},\bar{\beta}}(x) \iff \psi(x) \geq \psi(\bar{x}) + \Phi(x,\bar{y}) - \Phi(\bar{x},\bar{y})$$

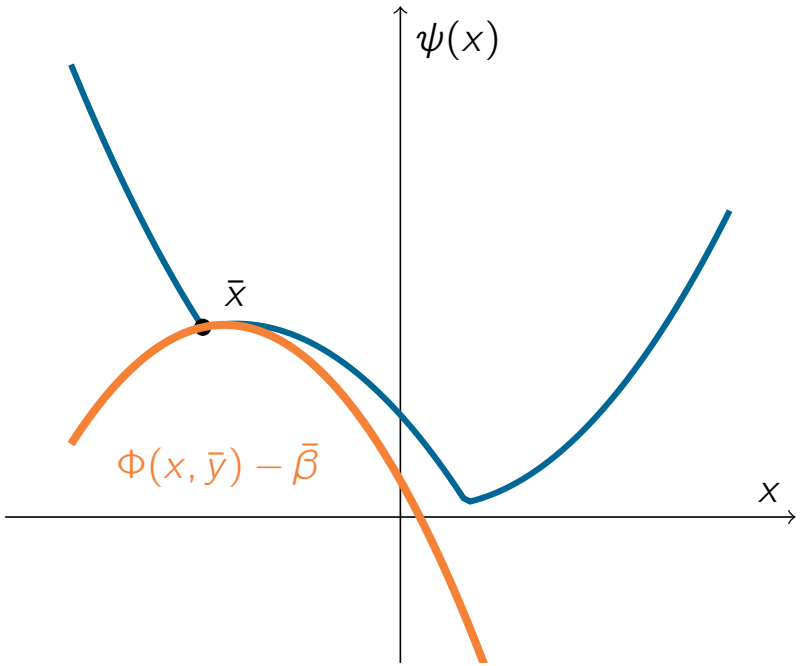
Φ-convexity



- "slopes" $\bar{y}$ of Φ-minorants at $\bar{x}$ = **Φ-subgradients** of ψ at $\bar{x}$

$$\partial_\Phi \psi(\bar{x}) = \{\bar{y} \in \mathbb{R}^n \mid \psi(x) \geq \psi(\bar{x}) + \Phi(x,\bar{y}) - \Phi(\bar{x},\bar{y}),\ \forall x \in \mathbb{R}^n\}$$

Φ-convexity

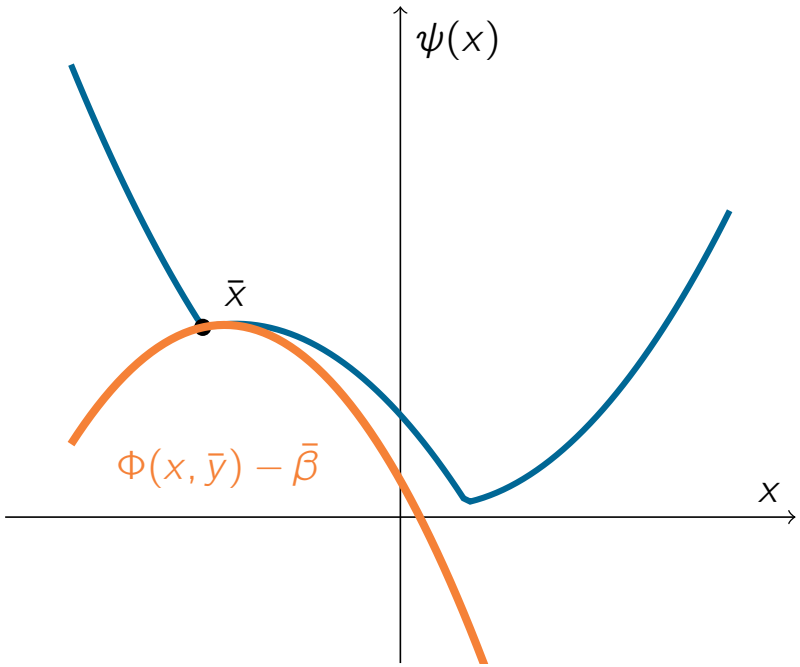


- "slopes" $\bar{y}$ of Φ-minorants at $\bar{x} =$ **Φ-subgradients** of ψ at $\bar{x}$

$$\partial_\Phi \psi(\bar{x}) = \{\bar{y} \in \mathbb{R}^n \mid \psi(x) \geq \psi(\bar{x}) + \Phi(x, \bar{y}) - \Phi(\bar{x}, \bar{y}),\ \forall x \in \mathbb{R}^n\}$$

- $\bar{y} \in \partial_\Phi \psi(\bar{x}) \iff \bar{x} \in \mathbf{argmin}_{x \in X}\, \psi(x) - \Phi(x, \bar{y})$

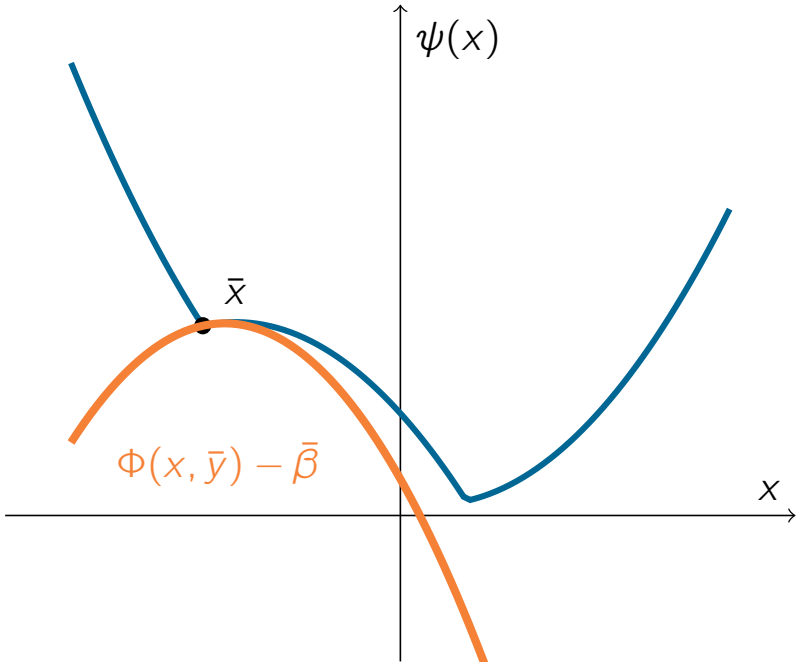
Φ-convexity

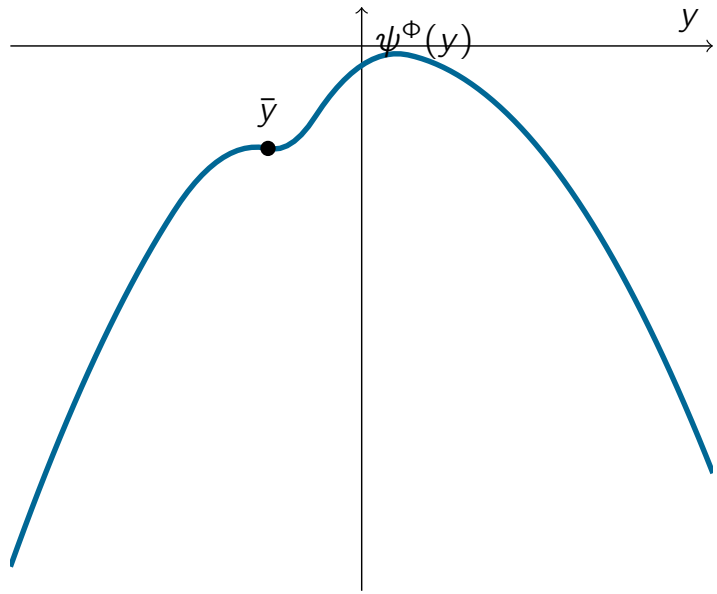


- $-$vertical translation $\bar{\beta}$ of Φ-minorant with slope $\bar{y} =$ **Φ-conjugate** at $\bar{y}$

$$\bar{\beta} = \Phi(\bar{x}, \bar{y}) - \psi(\bar{x})$$

Φ-convexity





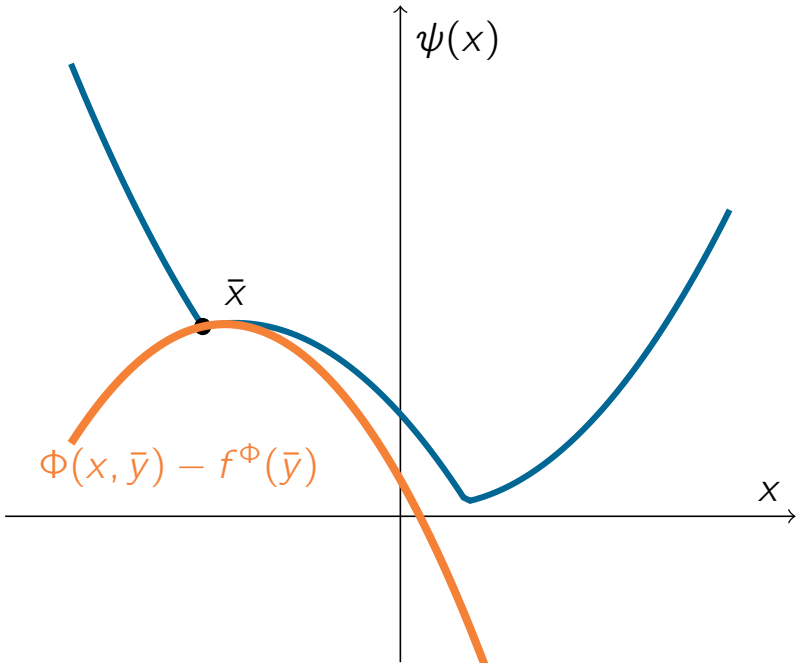
$$\psi^\Phi(y) = \sup_{x\in X}\{\Phi(x,y)-\psi(x)\}$$

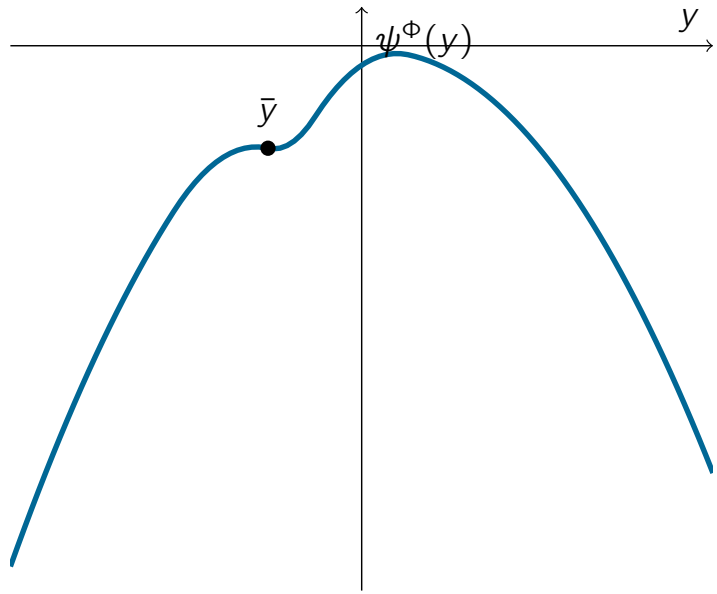
- $-$vertical translation $\bar{\beta}$ of Φ-minorant with slope $\bar{y}$ = **Φ-conjugate** at $\bar{y}$

$$\bar{\beta} = \Phi(\bar{x},\bar{y})-\psi(\bar{x}) = \sup_{x\in X}\{\Phi(x,\bar{y})-\psi(x)\} =: \psi^\Phi(\bar{y})$$

- ψ^Φ is Φ-convex (even if ψ is not)

Φ-convexity





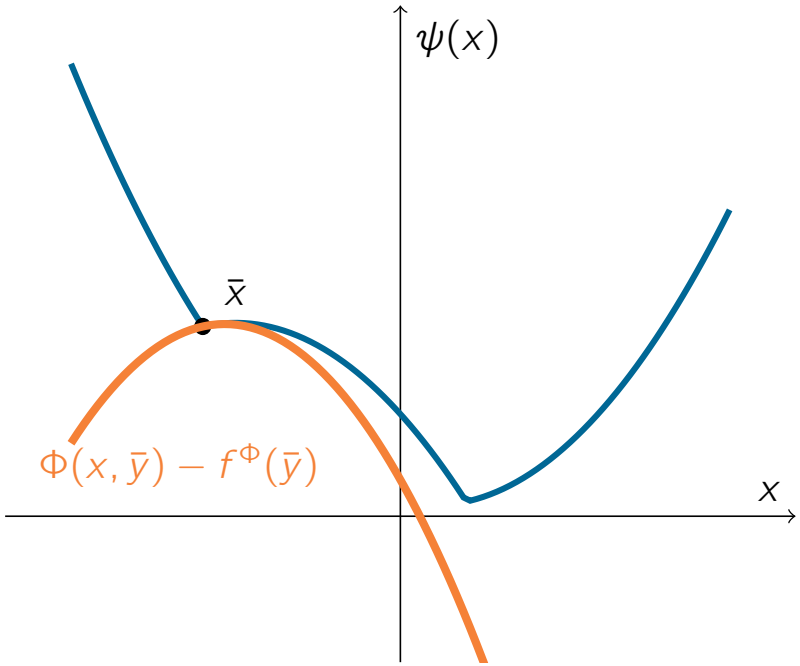
$$\psi^{\Phi}(y) = \sup_{x \in X} \{\Phi(x,y) - \psi(x)\}$$

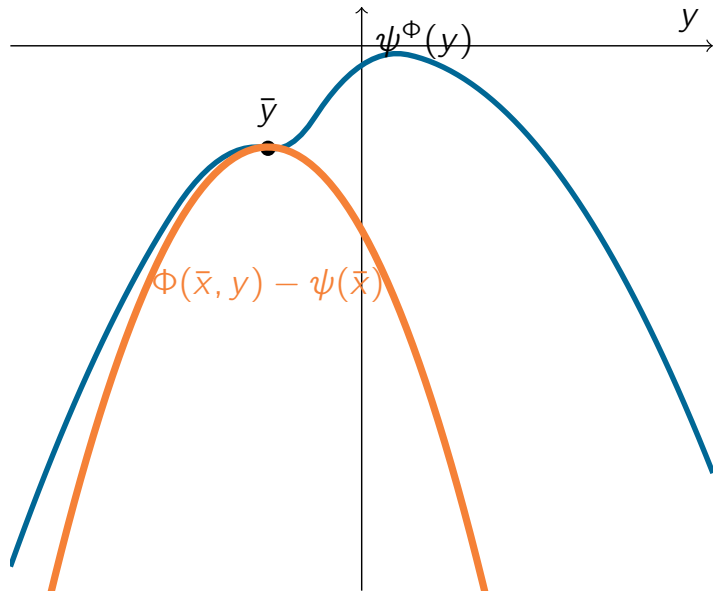
- **Fenchel-Young inequality**

$$\psi(x) + \psi^{\Phi}(y) \geq \Phi(x,y) \qquad \forall x, y \in \mathbb{R}^n$$

- $\bar{y} \in \partial_{\Phi}\psi(\bar{x}) \iff \psi(\bar{x}) + \psi^{\Phi}(\bar{y}) = \Phi(\bar{x},\bar{y})$

Φ-convexity





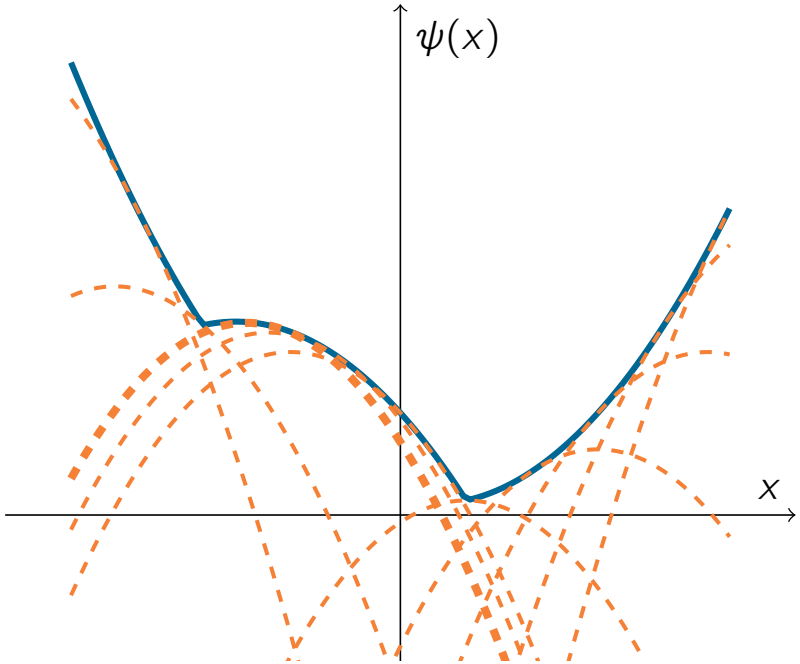
$$\psi^{\Phi}(y) = \sup_{x \in X} \{\Phi(x,y) - \psi(x)\}$$

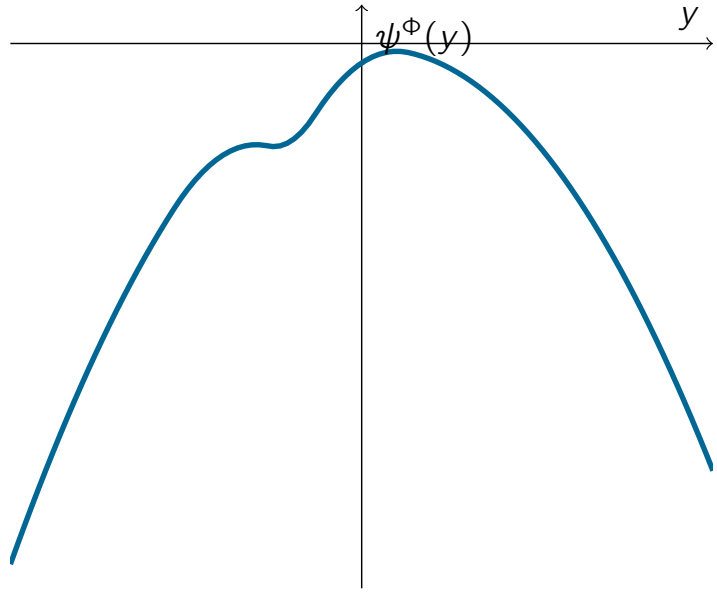
- **Fenchel-Young inequality**

$$\psi(x) + \psi^{\Phi}(y) \geq \Phi(x,y) \qquad \forall x, y \in \mathbb{R}^n$$

- $\bar{y} \in \partial_{\Phi}\psi(\bar{x}) \iff \psi(\bar{x}) + \psi^{\Phi}(\bar{y}) = \Phi(\bar{x},\bar{y}) \implies \bar{x} \in \partial_{\Phi}\psi^{\Phi}(\bar{y})$

Φ-convexity





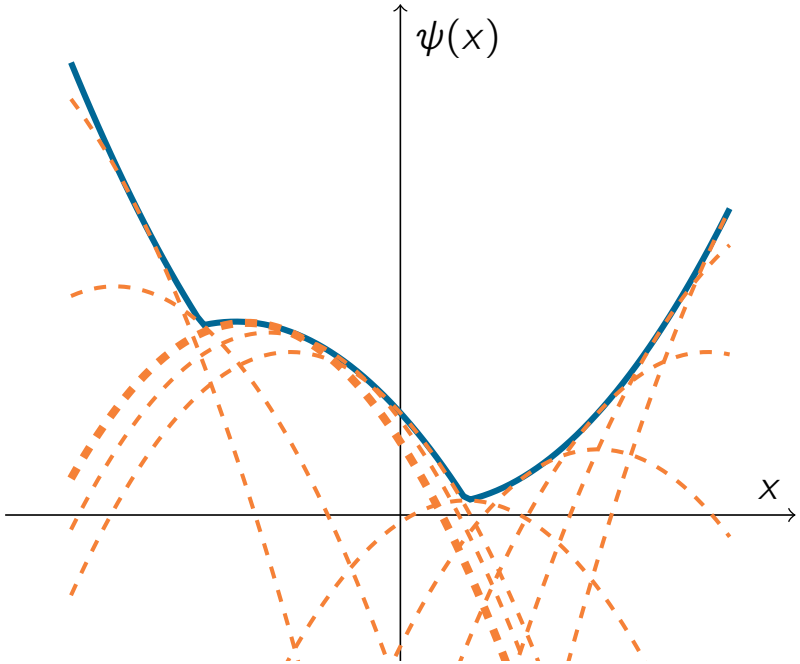
$$\psi^\Phi(y) = \sup_{x \in X} \{\Phi(x,y) - \psi(x)\}$$

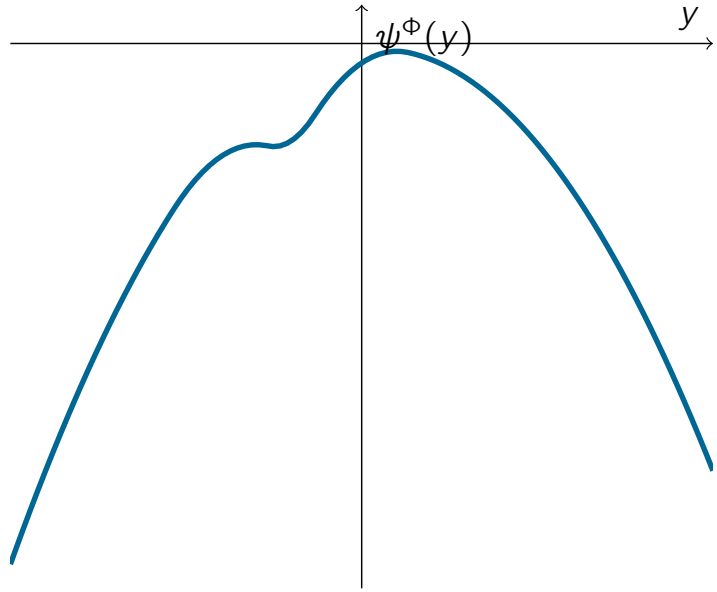
- biconjugate

$$\psi^{\Phi\Phi}(x) = \sup_{y \in Y} \{\Phi(x,y) - \psi^\Phi(y)\} = \sup_{(y,\beta):\ell_{y,\beta} \le \psi} \ell_{y,\beta}(x)$$

- $\psi^{\Phi\Phi}$ largest Φ-convex function minorizing f
- $\bar{y} \in \partial_\Phi \psi(\bar{x}) \implies \psi(\bar{x}) = \psi^{\Phi\Phi}(\bar{x})$
- $\psi : X \to \overline{\mathbb{R}}$ Φ-convex $\iff \psi = \psi^{\Phi\Phi}$

Φ-convexity





$$\psi^\Phi(y) = \sup_{x\in X}\{\Phi(x,y) - \psi(x)\}$$

summary

for any proper $\psi : X \to \overline{\mathbb{R}}$

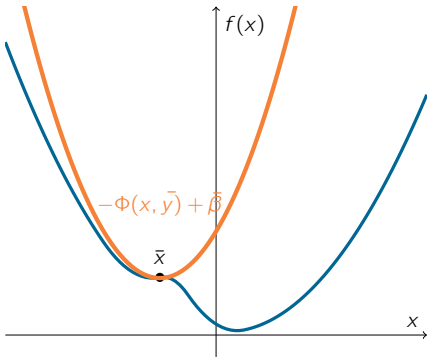
$$\bar{y} \in \partial_\Phi \psi(\bar{x}) \iff \bar{x} \in \operatorname*{argmin}_{x\in X}\{\psi(x) - \Phi(x,\bar{y})\} \iff \psi(\bar{x}) + \psi^\Phi(\bar{y}) = \Phi(\bar{x},\bar{y})$$

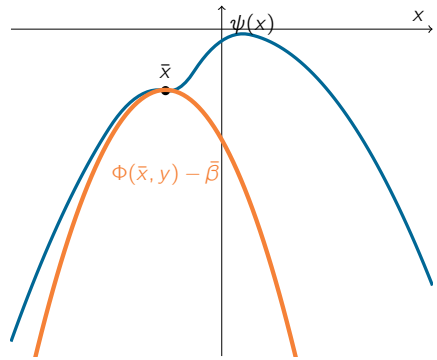
$$\Longrightarrow \begin{cases} \bar{x} \in \partial\psi^\Phi(\bar{y}) \\ \psi(\bar{x}) = \psi^{\Phi\Phi}(\bar{x}) \end{cases}$$

Φ-convexity and anisotropic smoothness

$$X = Y = \mathbb{R}^n \qquad \Phi(x, y) = -(L^{-1} \star \phi)(x - y)$$

- $-\Phi$-**majorants** for $f \in \mathcal{C}^1(\mathbb{R}^n) \iff$ Φ-**minorants** for $\psi = -f$





- slopes of $-\Phi$-**majorants** of $f \iff$ Φ-**subgradients** of $\psi = -f$

$$\bar{y} \in \partial_\Phi(-f)(\bar{x}) \iff \bar{x} \in \operatorname*{\mathbf{argmin}}_{x \in \mathbb{R}^n} (L^{-1} \star \phi)(x - \bar{y}) - f(x)$$

$$\text{Fermat} \implies \nabla\phi(L(\bar{x} - \bar{y})) = \nabla f(\bar{x})$$

$$\iff \bar{y} = \bar{x} - \tfrac{1}{L}\nabla\phi^\star(\nabla f(\bar{x})) =: T_{L^{-1}}(\bar{x})$$

- **anisotropic descent** for $f \iff$ Φ-**subgradient** inequality for $\psi = -f$

$$f(x) \le f(\bar{x}) + (L^{-1} \star \phi)(x - \bar{y}) - (L^{-1} \star \phi)(\bar{x} - \bar{y}), \qquad \forall x \in \mathbb{R}^n$$

Outline

Characterization of anisotropic smoothness

$$f(x) \leq f(\bar{x}) + (L^{-1} \star \phi)(x - \bar{y}) - (L^{-1} \star \phi)(\bar{x} - \bar{y}), \qquad \forall x \in \mathbb{R}^n$$

where $\bar{y} = \bar{x} - \frac{1}{L}\nabla\varphi^\star(\nabla f(\bar{x})) =: T_{L^{-1}}(\bar{x})$

---

f $L-$anisosmooth $\quad \Longleftrightarrow \quad$ $f = \xi \,\square\, L^{-1} \star \phi$

Characterization of anisotropic smoothness

$$f(x) \leq f(\bar{x}) + (L^{-1} \star \phi)(x - \bar{y}) - (L^{-1} \star \phi)(\bar{x} - \bar{y}), \qquad \forall x \in \mathbb{R}^n$$

where $\bar{y} = \bar{x} - \frac{1}{L}\nabla\varphi^\star(\nabla f(\bar{x})) =: T_{L^{-1}}(\bar{x})$

---

f $L-$anisosmooth $\quad \Longleftrightarrow \quad$ $f = \xi \,\square\, L^{-1} \star \phi$

proof: (only $\Longrightarrow$)

$$f \text{ } L\text{-anisosmooth} \iff \partial_\Phi(-f)(x) \neq \emptyset, \forall x \in \mathbb{R}^n$$

Characterization of anisotropic smoothness

$$f(x) \leq f(\bar{x}) + (L^{-1} \star \phi)(x - \bar{y}) - (L^{-1} \star \phi)(\bar{x} - \bar{y}), \qquad \forall x \in \mathbb{R}^n$$

where $\bar{y} = \bar{x} - \frac{1}{L}\nabla\varphi^\star(\nabla f(\bar{x})) =: T_{L^{-1}}(\bar{x})$

---

f $L-$anisosmooth $\quad \Longleftrightarrow \quad$ $f = \xi \,\square\, L^{-1} \star \phi$

proof: (only $\Longrightarrow$)

$$f \text{ } L\text{-anisosmooth} \iff \partial_\Phi(-f)(x) \neq \emptyset, \forall x \in \mathbb{R}^n$$
$$\Longrightarrow \; -f \text{ } \Phi\text{-convex} \iff -f = (-f)^{\Phi\Phi}$$

Characterization of anisotropic smoothness

$$f(x) \leq f(\bar{x}) + (L^{-1} \star \phi)(x - \bar{y}) - (L^{-1} \star \phi)(\bar{x} - \bar{y}), \qquad \forall x \in \mathbb{R}^n$$

where $\bar{y} = \bar{x} - \frac{1}{L}\nabla\varphi^\star(\nabla f(\bar{x})) =: T_{L^{-1}}(\bar{x})$

---

f $L-$anisosmooth $\quad\Longleftrightarrow\quad$ $f = \xi \,\square\, L^{-1} \star \phi$

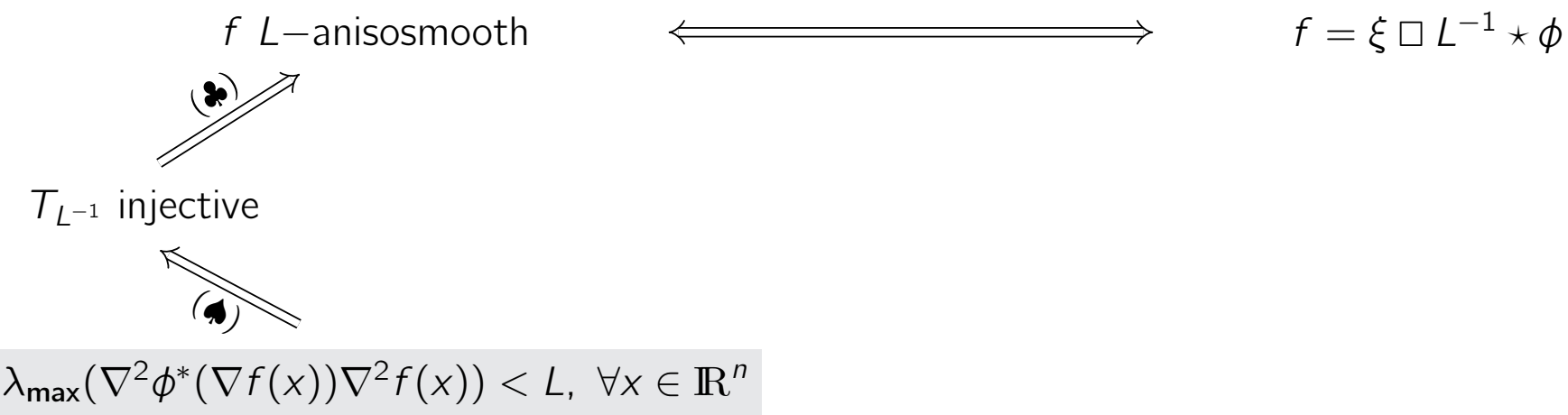
proof: (only $\Longrightarrow$)

$$\begin{aligned} f \ L\text{-anisosmooth} &\iff \partial_\Phi(-f)(x) \neq \emptyset, \forall x \in \mathbb{R}^n \\ &\Longrightarrow -f \ \Phi\text{-convex} \iff -f = (-f)^{\Phi\Phi} \\ &\iff f(x) = -\sup_{y \in \mathbb{R}^n} -(L^{-1} \star \phi)(x - y) - (-f)^\Phi(y) = (-f)^\Phi \,\square\, (L^{-1} \star \phi) \end{aligned}$$

Characterization of anisotropic smoothness

$$f(x) \leq f(\bar{x}) + (L^{-1} \star \phi)(x - \bar{y}) - (L^{-1} \star \phi)(\bar{x} - \bar{y}), \qquad \forall x \in \mathbb{R}^n$$

where $\bar{y} = \bar{x} - \frac{1}{L}\nabla\varphi^\star(\nabla f(\bar{x})) =: T_{L^{-1}}(\bar{x})$



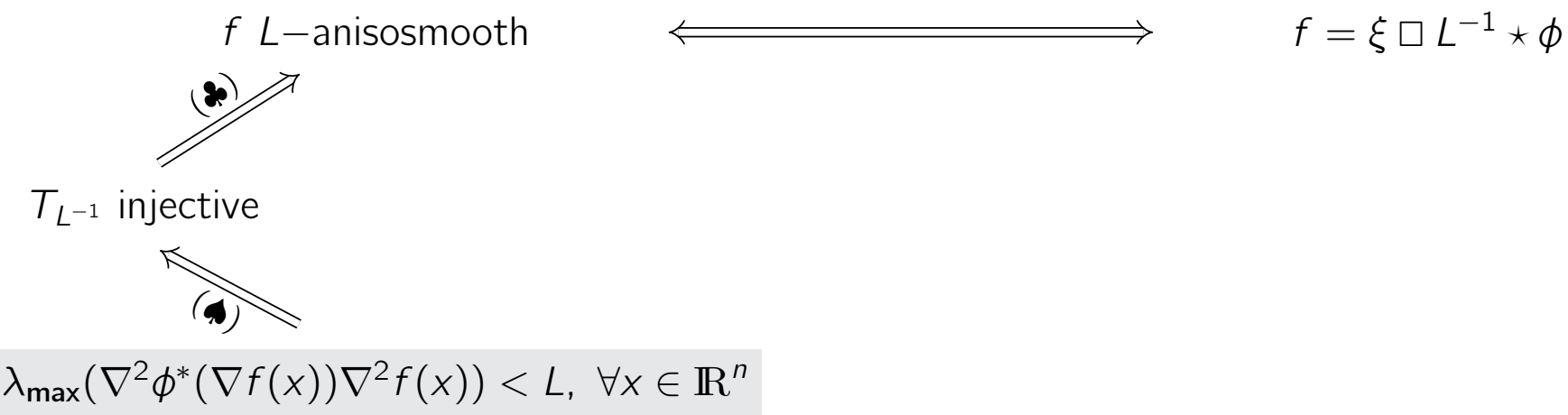
$$\lambda_{\max}(\nabla^2\phi^*(\nabla f(x))\nabla^2 f(x)) < L, \ \forall x \in \mathbb{R}^n$$

- f L-anisosmooth at $\bar{x} \in \mathbb{R}^n \iff \bar{x} \in \mathbf{argmin}_{x \in \mathbb{R}^n}(L^{-1} \star \phi)(x - \bar{y}) - f(x)$

Characterization of anisotropic smoothness

$$f(x) \leq f(\bar{x}) + (L^{-1} \star \phi)(x - \bar{y}) - (L^{-1} \star \phi)(\bar{x} - \bar{y}), \qquad \forall x \in \mathbb{R}^n$$

where $\bar{y} = \bar{x} - \frac{1}{L}\nabla\varphi^\star(\nabla f(\bar{x})) =: T_{L^{-1}}(\bar{x})$



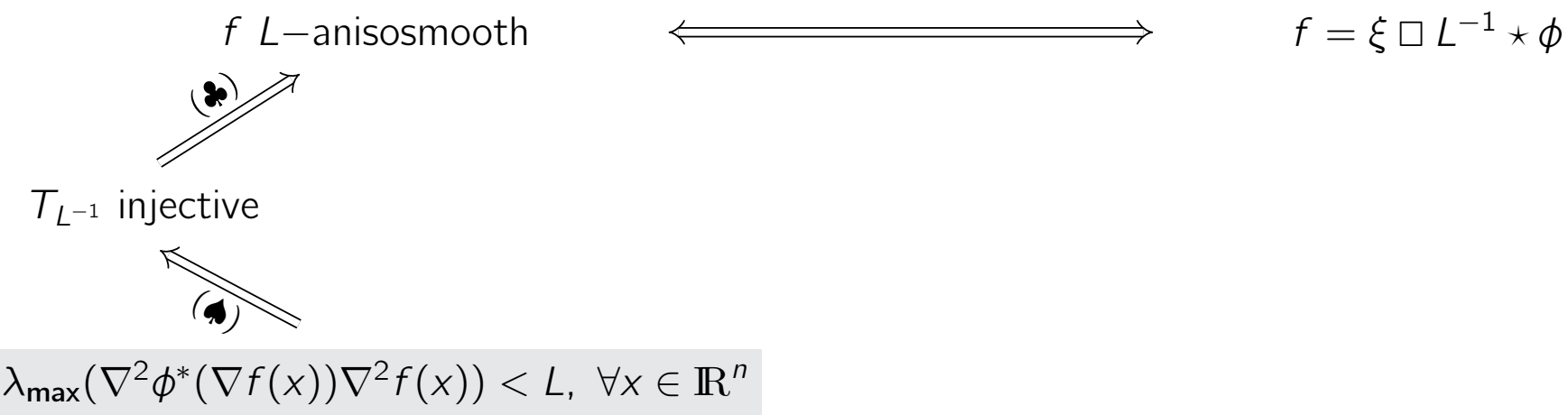
$$\lambda_{\max}(\nabla^2\phi^*(\nabla f(x))\nabla^2 f(x)) < L, \ \forall x \in \mathbb{R}^n$$

- f L-anisosmooth at $\bar{x} \in \mathbb{R}^n \iff \bar{x} \in \mathbf{argmin}_{x \in \mathbb{R}^n}(L^{-1} \star \phi)(x - \bar{y}) - f(x)$
- $\bar{x}$ is stationary point of $(L^{-1} \star \phi)(\cdot - \bar{y}) - f \iff \bar{y} = T_{L^{-1}}(\bar{x})$

Characterization of anisotropic smoothness

$$f(x) \leq f(\bar{x}) + (L^{-1} \star \phi)(x - \bar{y}) - (L^{-1} \star \phi)(\bar{x} - \bar{y}), \qquad \forall x \in \mathbb{R}^n$$

where $\bar{y} = \bar{x} - \frac{1}{L}\nabla\varphi^\star(\nabla f(\bar{x})) =: T_{L^{-1}}(\bar{x})$



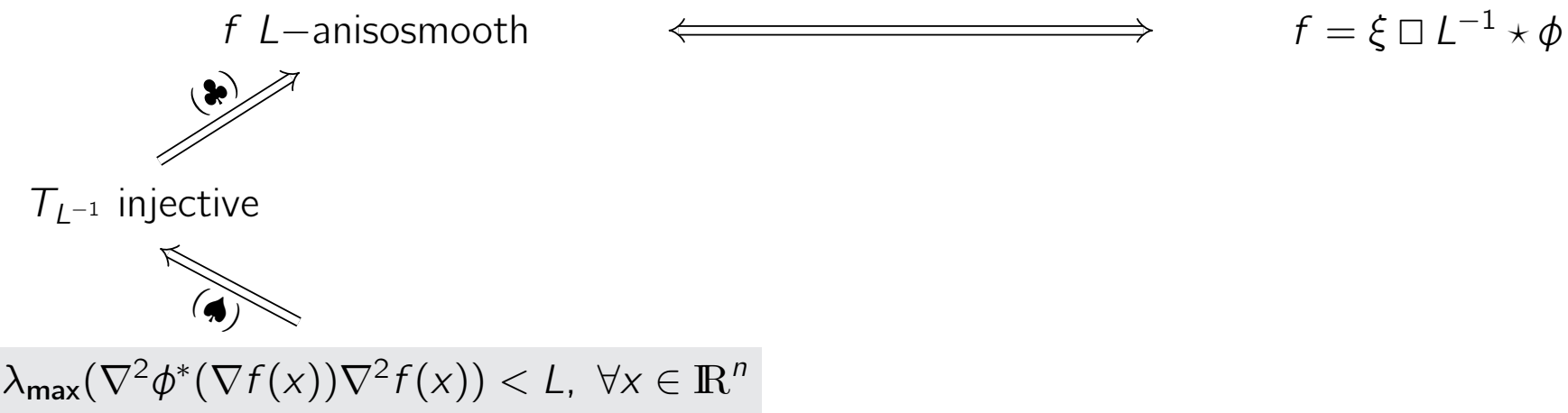
$$\lambda_{\max}(\nabla^2\phi^*(\nabla f(x))\nabla^2 f(x)) < L, \ \forall x \in \mathbb{R}^n$$

- f L-anisosmooth at $\bar{x} \in \mathbb{R}^n \iff \bar{x} \in \mathbf{argmin}_{x \in \mathbb{R}^n}(L^{-1} \star \phi)(x - \bar{y}) - f(x)$
- $\bar{x}$ is stationary point of $(L^{-1} \star \phi)(\cdot - \bar{y}) - f \iff \bar{y} = T_{L^{-1}}(\bar{x})$
- LMI holds $\implies$ $\bar{x}$ is (strong) **local minimizer**

Characterization of anisotropic smoothness

$$f(x) \leq f(\bar{x}) + (L^{-1} \star \phi)(x - \bar{y}) - (L^{-1} \star \phi)(\bar{x} - \bar{y}), \qquad \forall x \in \mathbb{R}^n$$

where $\bar{y} = \bar{x} - \frac{1}{L}\nabla\varphi^\star(\nabla f(\bar{x})) =: T_{L^{-1}}(\bar{x})$



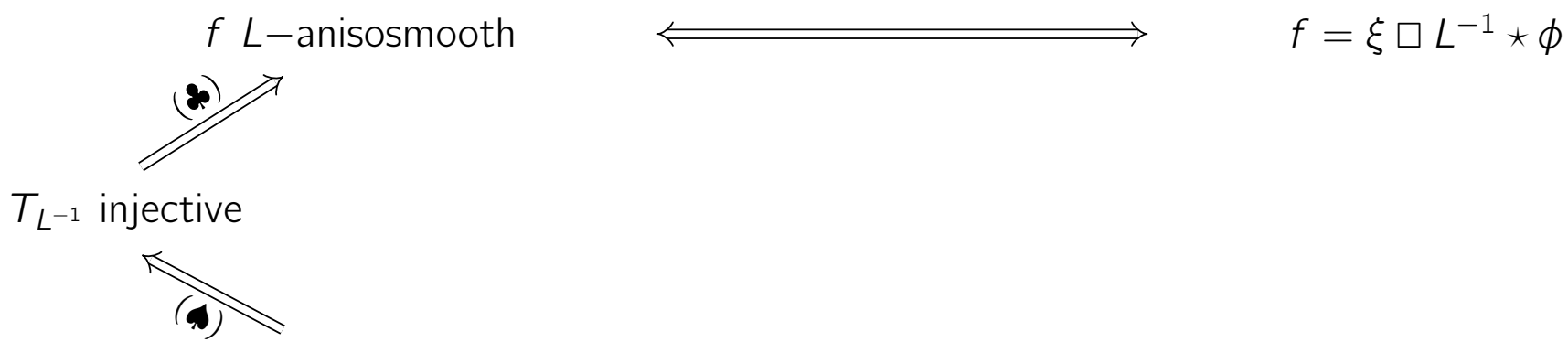
$$\lambda_{\max}(\nabla^2\phi^*(\nabla f(x))\nabla^2 f(x)) < L, \ \forall x \in \mathbb{R}^n$$

- f L-anisosmooth at $\bar{x} \in \mathbb{R}^n \iff \bar{x} \in \mathbf{argmin}_{x \in \mathbb{R}^n}(L^{-1} \star \phi)(x - \bar{y}) - f(x)$
- $\bar{x}$ is stationary point of $(L^{-1} \star \phi)(\cdot - \bar{y}) - f \iff \bar{y} = T_{L^{-1}}(\bar{x})$
- LMI holds $\implies$ $\bar{x}$ is (strong) **local minimizer**
- is $\bar{x}$ a **global minimizer**? yes, if

Characterization of anisotropic smoothness

$$f(x) \le f(\bar{x}) + (L^{-1} \star \phi)(x - \bar{y}) - (L^{-1} \star \phi)(\bar{x} - \bar{y}), \qquad \forall x \in \mathbb{R}^n$$

where $\bar{y} = \bar{x} - \frac{1}{L}\nabla\varphi^\star(\nabla f(\bar{x})) =: T_{L^{-1}}(\bar{x})$



$$\lambda_{\max}(\nabla^2\phi^*(\nabla f(x))\nabla^2 f(x)) < L, \ \forall x \in \mathbb{R}^n$$

- f L-anisosmooth at $\bar{x} \in \mathbb{R}^n \iff \bar{x} \in \mathbf{argmin}_{x\in\mathbb{R}^n}(L^{-1} \star \phi)(x - \bar{y}) - f(x)$
- $\bar{x}$ is stationary point of $(L^{-1} \star \phi)(\cdot - \bar{y}) - f \iff \bar{y} = T_{L^{-1}}(\bar{x})$
- LMI holds $\implies$ $\bar{x}$ is (strong) local minimizer
- is $\bar{x}$ a global minimizer? yes, if
  1. $(L^{-1} \star \phi)(\cdot - \bar{y}) - f$ attains its minimum $\Longleftarrow$ (♣)

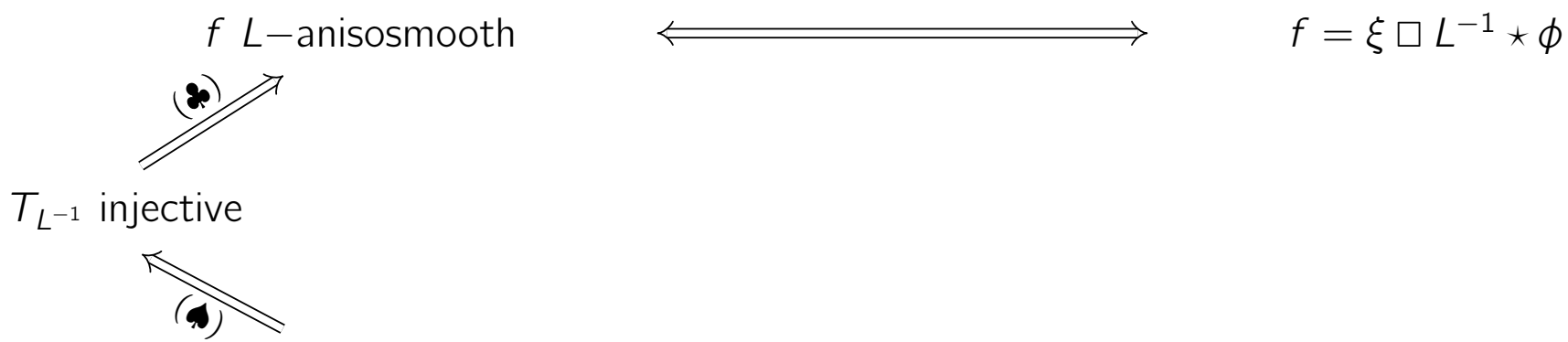
(♣) either $\mathbf{dom}\,\phi$ bounded or a very mild growth condition on f: there exist $r \in (0, L)$, $\beta \in \mathbb{R}$:

$$f(x) \le (r^{-1} \star \phi)(x) - \beta, \qquad \forall x \in \mathbb{R}^n$$

Characterization of anisotropic smoothness

$$f(x) \le f(\bar{x}) + (L^{-1} \star \phi)(x - \bar{y}) - (L^{-1} \star \phi)(\bar{x} - \bar{y}), \qquad \forall x \in \mathbb{R}^n$$

where $\bar{y} = \bar{x} - \frac{1}{L}\nabla\varphi^\star(\nabla f(\bar{x})) =: T_{L^{-1}}(\bar{x})$



$\lambda_{\max}(\nabla^2\phi^*(\nabla f(x))\nabla^2 f(x)) < L, \ \forall x \in \mathbb{R}^n$

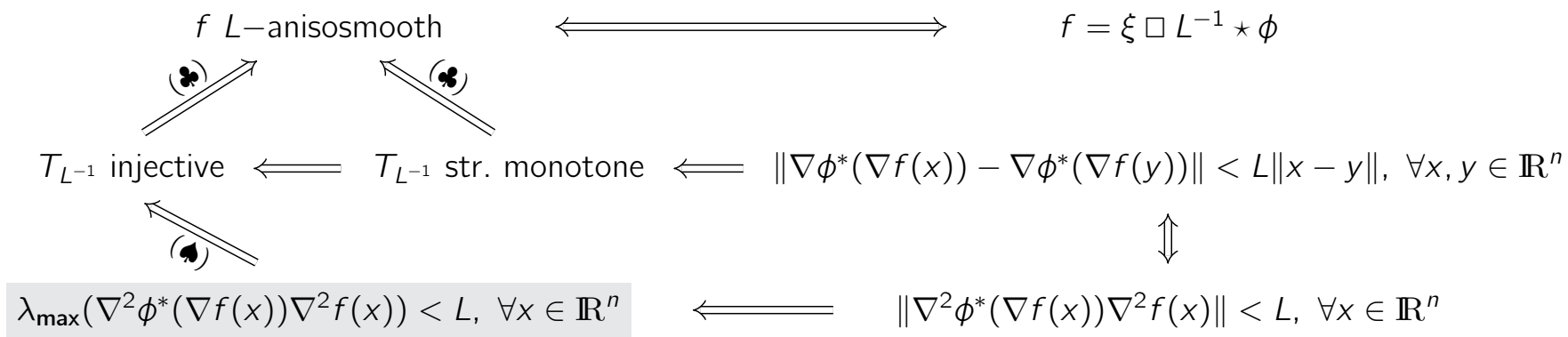
- f L-anisosmooth at $\bar{x} \in \mathbb{R}^n \iff \bar{x} \in \mathbf{argmin}_{x\in\mathbb{R}^n}(L^{-1} \star \phi)(x - \bar{y}) - f(x)$
- $\bar{x}$ is stationary point of $(L^{-1} \star \phi)(\cdot - \bar{y}) - f \iff \bar{y} = T_{L^{-1}}(\bar{x})$
- LMI holds $\implies$ $\bar{x}$ is (strong) **local minimizer**
- is $\bar{x}$ a **global minimizer**? yes, if
  1. $(L^{-1} \star \phi)(\cdot - \bar{y}) - f$ attains its minimum $\impliedby$ (♣)
  2. $\bar{x}$ is unique stationary point $\iff$ $T_{L^{-1}}$ injective $\impliedby$ (♠)+ LMI

(♠) either $\mathbf{dom}\,\phi$ bounded or very mild growth condition on ∇f: $\limsup_{\|x\|\to\infty} \frac{\|\nabla\phi^*(\nabla f(x))\|}{\|x\|} < L$

Characterization of anisotropic smoothness

$$f(x) \le f(\bar{x}) + (L^{-1} \star \phi)(x - \bar{y}) - (L^{-1} \star \phi)(\bar{x} - \bar{y}), \qquad \forall x \in \mathbb{R}^n$$

where $\bar{y} = \bar{x} - \frac{1}{L}\nabla\varphi^\star(\nabla f(\bar{x})) =: T_{L^{-1}}(\bar{x})$



- f L-anisosmooth at $\bar{x} \in \mathbb{R}^n \iff \bar{x} \in \mathbf{argmin}_{x \in \mathbb{R}^n}(L^{-1} \star \phi)(x - \bar{y}) - f(x)$
- $\bar{x}$ is stationary point of $(L^{-1} \star \phi)(\cdot - \bar{y}) - f \iff \bar{y} = T_{L^{-1}}(\bar{x})$
- LMI holds $\implies$ $\bar{x}$ is (strong) **local minimizer**
- is $\bar{x}$ a **global minimizer**? yes, if
  1. $(L^{-1} \star \phi)(\cdot - \bar{y}) - f$ attains its minimum $\impliedby$ ($\clubsuit$)
  2. $\bar{x}$ is unique stationary point $\iff$ $T_{L^{-1}}$ injective $\impliedby$ $T_{L^{-1}}$ strict. monotone

2nd-order characterization: Isotropic case

$$\nabla^2 f(x) \prec L \nabla^2 \phi^*(\nabla f(x))^{-1} \qquad \text{with } \phi(x) = h(\|x\|)$$

$$\nabla^2 \phi^*(y)^{-1} = \begin{cases} \alpha(\|y\|)\,\mathbf{P}_{\|} + \beta(\|y\|)\,\mathbf{P}_{\perp} & y \neq 0 \\ \alpha(\|y\|)\, I_n & y = 0 \end{cases}$$

where $\alpha(t) = \frac{1}{(h^*)''(t)}$ $\beta(t) = \frac{t}{(h^*)'(t)}$ $\mathbf{P}_{\|} = \frac{yy^\top}{\|y\|^2}$ $\mathbf{P}_{\perp} = I_n - \mathbf{P}_{\|}$

$h(s)$	$(h^*)'(t)$	$\alpha(t)$	$\beta(t)$
$\lambda(\mathbf{cosh}(s) - 1)$	$\mathbf{arcsinh}(\lambda^{-1} t)$	$\sqrt{t^2 + \lambda^2}$	$\frac{t}{\mathbf{arcsinh}(\lambda^{-1} t)}$
$\lambda(\mathbf{exp}(\lvert s\rvert) - \lvert s\rvert - 1)$	$\mathbf{ln}(1 + \lambda^{-1}\lvert t\rvert)\overline{\mathbf{sign}}(t)$	$\lvert t\rvert + \lambda$	$\frac{t}{\mathbf{ln}(1+\lambda^{-1}\lvert t\rvert)}$
$\lambda(-\lvert s\rvert - \mathbf{ln}(1 - \lvert s\rvert))$	$\frac{t}{\lvert t\rvert+\lambda}$	$\frac{(\lvert t\rvert+\lambda)^2}{\lambda}$	$\lvert t\rvert + \lambda$
$\lambda(1 - \sqrt{1 - s^2})$	$\frac{t}{\sqrt{t^2+\lambda^2}}$	$\lambda^{-2}(t^2 + \lambda^2)^{3/2}$	$\sqrt{t^2 + \lambda^2}$
$\lambda\left(s\,\mathbf{arctanh}(s) + \mathbf{ln}\left(\sqrt{1 - s^2}\right)\right)$	$\mathbf{tanh}(\lambda^{-1} t)$	$\lambda\,\mathbf{cosh}^2(\lambda^{-1} t)$	$\frac{t}{\mathbf{tanh}(\lambda^{-1} t)}$

sufficient condition

$$\lambda_{\mathbf{max}}(\nabla^2 f(x)) \leq L\,\mathbf{min}\{\alpha(\|\nabla f(x)\|), \beta(\|\nabla f(x)\|)\}$$
$$(= L\beta(\|\nabla f(x)\|) \text{ for all examples})$$

Connection with (L_0, L_1)-smoothness

$h(s)$	$(h^*)'(t)$	$\alpha(t)$	$\beta(t)$
$\lambda(-\lvert s\rvert - \mathbf{ln}(1-\lvert s\rvert))$	$\frac{t}{\lvert t\rvert+\lambda}$	$\frac{(\lvert t\rvert+\lambda)^2}{\lambda}$	$\lvert t\rvert + \lambda$

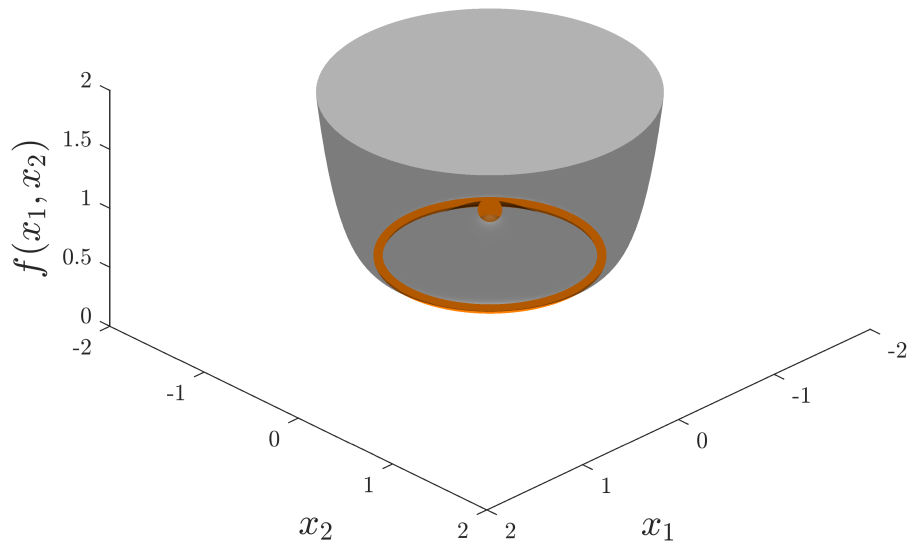
- NGD becomes
$$x^+ = x - \frac{\gamma}{\|\nabla f(x)\| + \lambda}\nabla f(x)$$
- 2nd-order condition $\nabla^2 f(x) \prec L\nabla^2\phi^*(\nabla f(x))^{-1}$:
$$\nabla^2 f(x) \prec L(\lambda + \|\nabla f(x)\|)\left(I_n + (\lambda\|\nabla f(x)\|)^{-1}\nabla f(x)\nabla f(x)^\top\right)$$
- sufficient condition $\lambda_{\mathbf{max}}(\nabla^2 f(x)) \le L\beta(\|\nabla f(x)\|)$:
$$\begin{aligned}\lambda_{\mathbf{max}}(\nabla^2 f(x)) &\le L(\lambda + \|\nabla f(x)\|)\\ &= L_0 + L_1\|\nabla f(x)\| \qquad (L_0 = L\lambda,\ L_1 = L)\end{aligned}$$
- "one-sided" version of (L_0, L_1) smoothness (less strict)
- class of anisosmooth functions strictly larger than (L_0, L_1)-smooth

Example: anisosmooth $\supset$ (L_0, L_1)-smooth

- in 1d the condition becomes $f''(x) \leq \lambda^{-1} L(\lambda + |f'(x)|)^2$
  - a^x, $a > 1$, a^{b^x}, $a, b > 1$, strongly convex self-concordant functions are all anisosmooth
- n-d example

$$f(x) = \mathbf{exp}(\|x\|^2) - 2\|x\|^2$$

  - f is not (L_0, L_1)-smooth but is anisosmooth for $\lambda = 1$, $L = 10$



Anisotropic smoothness & clipping

- 2nd-order condition becomes

$$\begin{cases} \lambda_{\max}(\nabla^2 f(x)) \le \lambda L, & \|\nabla f(x)\| \le \lambda \\ \lambda_{\max}(\mathbf{P}_\perp \nabla^2 f(x) \mathbf{P}_\perp) \le L\|\nabla f(x)\|, & \|\nabla f(x)\| > \lambda \end{cases}$$

  where $\mathbf{P}_\perp = I_n - \frac{\nabla f(x) \nabla f(x)^\top}{\|\nabla f(x)\|^2}$ projection on $\mathbf{ker}\, \nabla f(x)$
- in 1d: $f''(x) \le L\lambda$ if $|f'(x)| \le \lambda$
- if $\{x \mid |f'(x)| \le \lambda\}$ is bounded for some λ then any function with locally Lipschitz second derivative is anisosmooth

Outline

Nonconvex functions

$$x^{k+1} = x^k - \gamma \nabla \phi^*(\nabla f(x^k))$$

where $f \in \mathcal{C}^1(\mathbb{R}^n)$ is anisosmooth with respect to ϕ

Theorem

If $\gamma L \in (0, 2)$, *then*

$$\min_{0 \le k \le K} \phi(\nabla \phi^*(\nabla f(x^k))) \le \frac{L}{\min\{2 - \gamma L, \gamma L\}} \frac{f(x^0) - f^\star}{K + 1}.$$

Nonconvex functions: momentum

choose $x^0 \in \mathbb{R}^n$, $\gamma, \beta > 0$, set $v^{-1} = 0$ and iterate:

$$v^k = \beta v^{k-1} + (1-\beta)\nabla\phi^*(\nabla f(x^k))$$
$$x^{k+1} = x^k - \gamma v^k$$

Theorem

If $\gamma L \in (0,1)$, $\beta \in (0, 1/2)$, then

$$\min_{0\le k\le K} \phi(\nabla\phi^*(\nabla f(x^k))) \le \frac{f(x^0) - f^\star}{\gamma(1-2\beta)(K+1)}.$$

Moreover if f satisfies the (slightly) stronger condition

$$\|\nabla\phi^*(\nabla f(x)) - \nabla\phi^*(\nabla f(\bar{x}))\| \le L\|x - \bar{x}\|$$

and $\beta \in (0,1)$ and $\gamma = \text{`}(1-\beta)^2/L$ then

$$\min_{1\le k\le K} \phi(\nabla\phi^*(\nabla f(x^k))) \le \frac{1}{K}\left(\frac{f(x^0) - f_\star}{\beta\gamma} + \frac{1}{1-\beta}\phi(\nabla\phi^*(\nabla f(x^0)))\right)$$

Nonconvex functions: anisotropic PL

choose $x^0 \in \mathbb{R}^n$, $\gamma, \beta > 0$, set $v^{-1} = 0$ and iterate:

$$v^k = \beta v^{k-1} + (1-\beta)\nabla\phi^*(\nabla f(x^k))$$
$$x^{k+1} = x^k - \gamma v^k$$

Definition

We say that f satisfies the anisotropic gradient dominance condition relative to ϕ with constant $\mu > 0$ if for all $x \in \mathbb{R}^n$

$$\phi(\nabla\phi^*(\nabla f(x))) \geq \mu(f(x) - f_\star).$$

Theorem

If $\gamma L \in (0,1)$, $\beta \in (0, 1/2)$, *then*

$$f(x^k) - f_\star \leq \alpha^k (f(x^0) - f_\star),$$

where $\alpha = \mathbf{max}\{1 - \gamma\beta(1-2\beta)\mu, \beta(1+2\beta)\}$.

Convex functions: isotropic reference functions

$$x^{k+1} = x^k - \tfrac{1}{L}\nabla\phi^*(\nabla f(x^k))$$

Theorem

Let f be convex and $\phi = h \circ \|\cdot\|$ with $h : \mathbb{R} \to \overline{\mathbb{R}}$ proper, lsc, strongly convex, even with $h(0) = 0$. The following holds for $k \in \mathbb{N}$:

1. $\|x^{k+1} - x^\star\| \leq \|x^k - x^\star\|$, *where* $x^\star \in \mathbf{argmin}\, f$,
2. $\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|$,
3. *if* $(h^*)'(t)/t$ *is nonincreasing on* $\mathbb{R}_+$

$$f(x^k) - f_\star \leq \frac{L\|\nabla f(x^0)\|\|x^0 - x^\star\|^2}{(h^*)'(\|\nabla f(x^0)\|)(k+1)}$$

Convex functions: general reference functions

$$x^{k+1} = x^k - \tfrac{1}{L}\nabla\phi^*(\nabla f(x^k))$$

Theorem

Let f be convex and $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ proper, lsc, strongly convex, even with $\phi(0) = 0$ and 2-subhomogeneous:

$$\phi(\theta x) \leq \theta^2\phi(x) \text{ for all } \theta \in [0, 1]$$

Then

$$f(x^k) - f_\star \leq \frac{4D_0}{k}$$

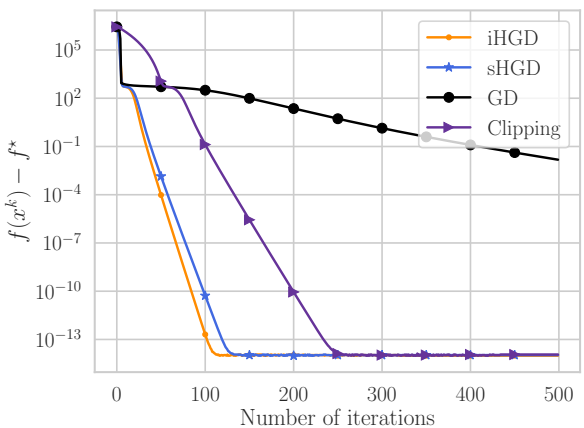
where $D_0 = \mathbf{sup}\{(L^{-1} \star \phi)(x - x^\star) \mid f(x) \leq f(x^0)\}$.

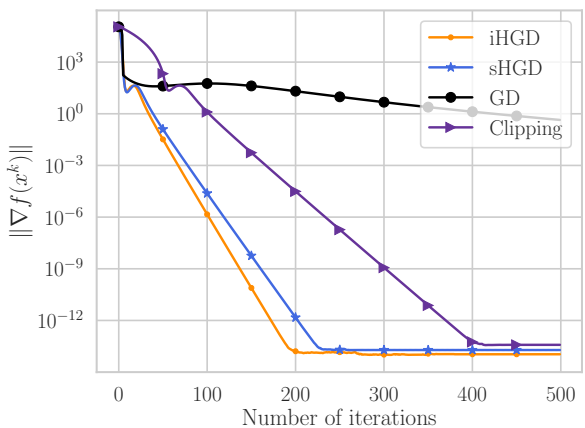
Outline

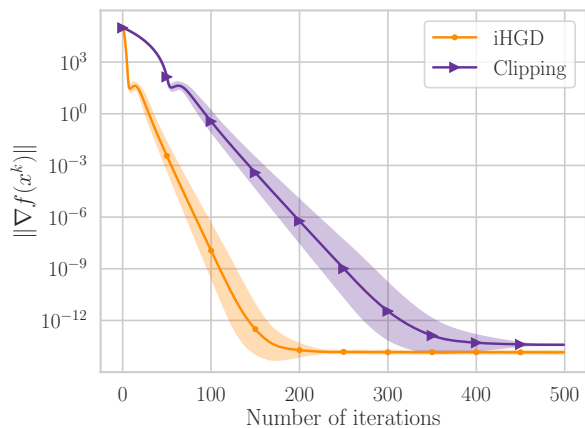
Phase retrieval

$$\underset{x\in\mathbb{R}^n}{\textbf{minimize}}\, f(x) = \frac{1}{2m}\sum_{i=1}^{m}(y_i - (a_i^\top x)^2)^2$$

- $n = 100,\ m = 3000$

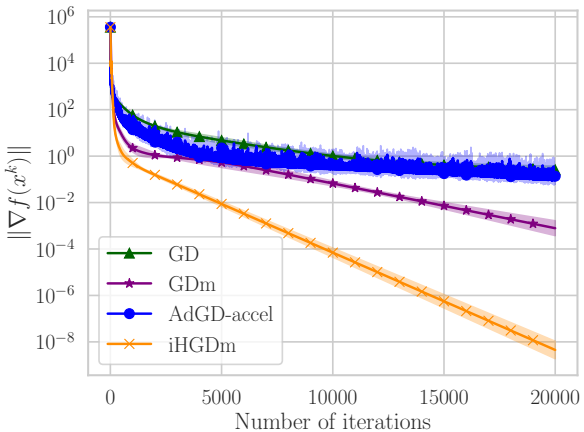


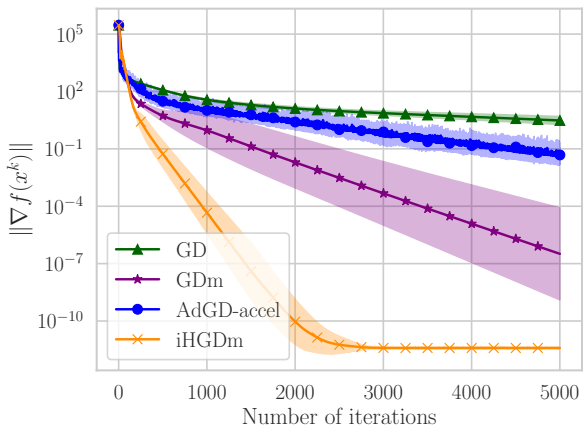


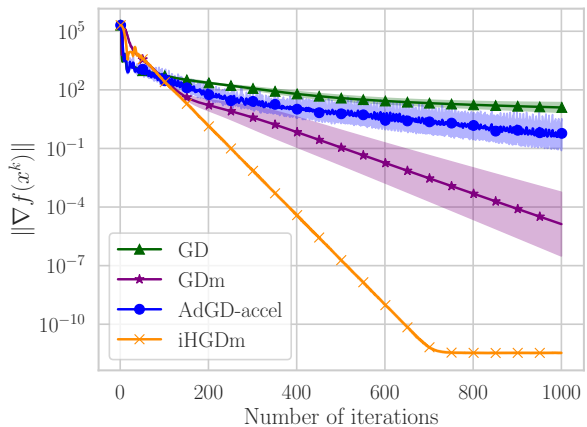


Matrix factorization

$$\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}} f(U, V) = \frac{1}{2} \|UV^\top - A\|_F^2$$

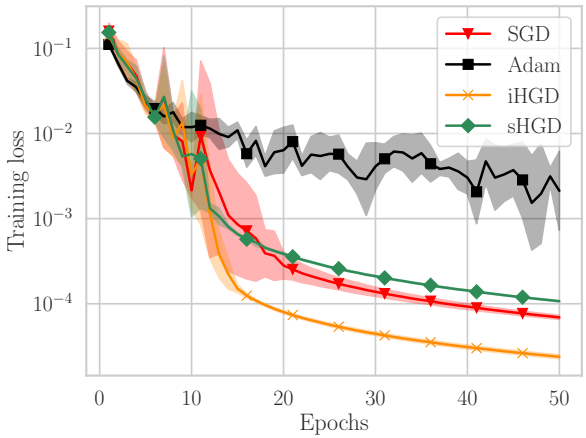


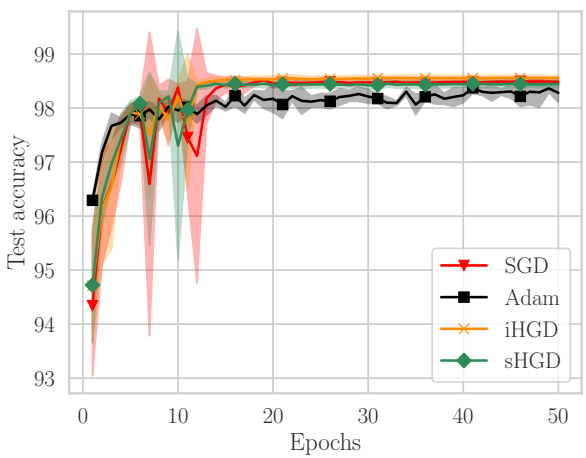




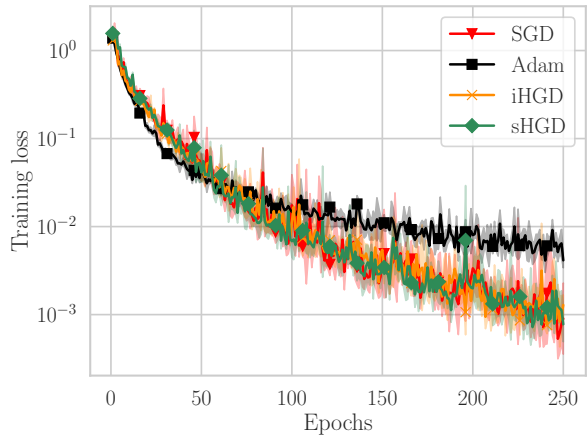
- MovieLens100K data set, rank $r = \{10, 20, 30\}$

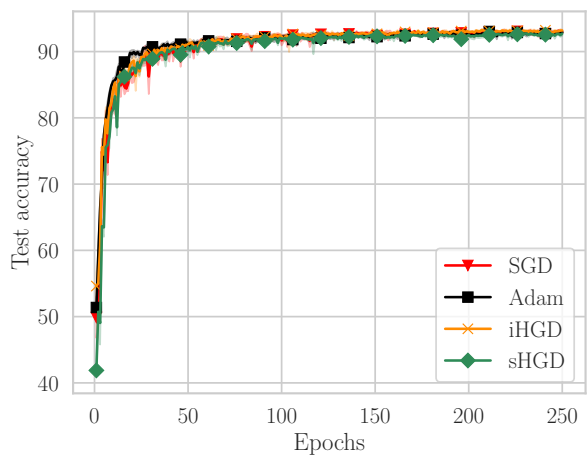
Neural network training



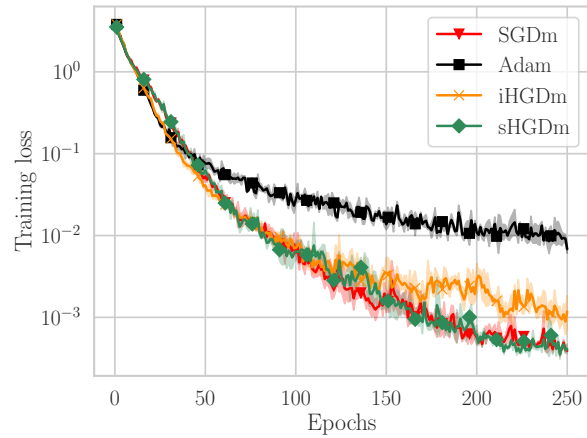


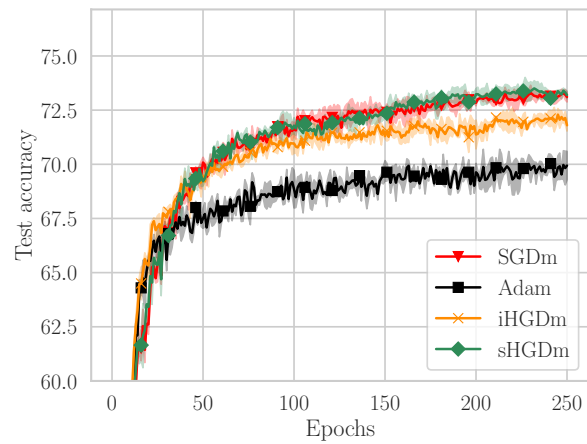
MNIST MLP
without momentum





Cifar 10 Resnet-18
without momentum





Cifar 100 Resnet-34
with momentum

The end

- joint work with Kostas Oikonomidis, Jan Quan, Manu Laude



Questions?

K. Oikonomidis, J. Quan, E. Laude and P. Patrinos, *Nonlinearly Preconditioned Gradient Methods under Generalized Smoothness.* ICML, **2025**

E. Laude and P. Patrinos, *Anisotropic proximal gradient.* Mathematical Programming **12**(4):747-756, **2025**