Distributed and Stochastic Optimization for Machine Learning

Please install the following for the TP:
python3, Pre-install numpy, numba, scikit-learn, ipython notebook / jupyter.

Marco Cuturi

ENSAE ParisTech
École nationale de la statistique et de l'administration économique
université PARIS-SACLAY

Reminders on Convexity: *Sets*

- Line segment between two points in Hilbert space:

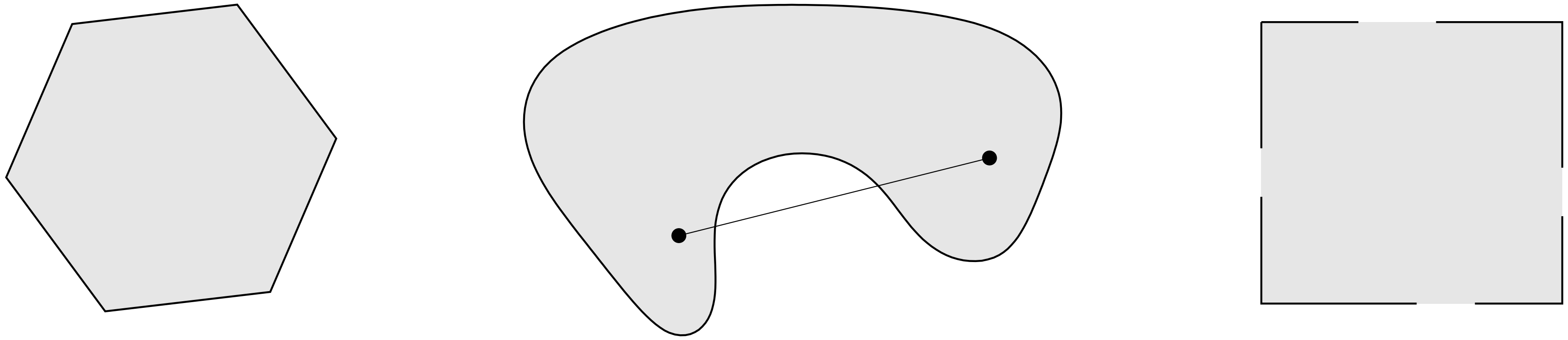
$$\{x = \lambda x_1 + (1-\lambda) x_2, \quad 0 \leq \lambda \leq 1\}$$

- A convex set contains all segments of all its points

Def

C is convex $\Leftrightarrow \forall x_1, x_2 \in C, 0 \leq \lambda \leq 1; \quad \lambda x_1 + (1-\lambda) x_2 \in C$

- Examples

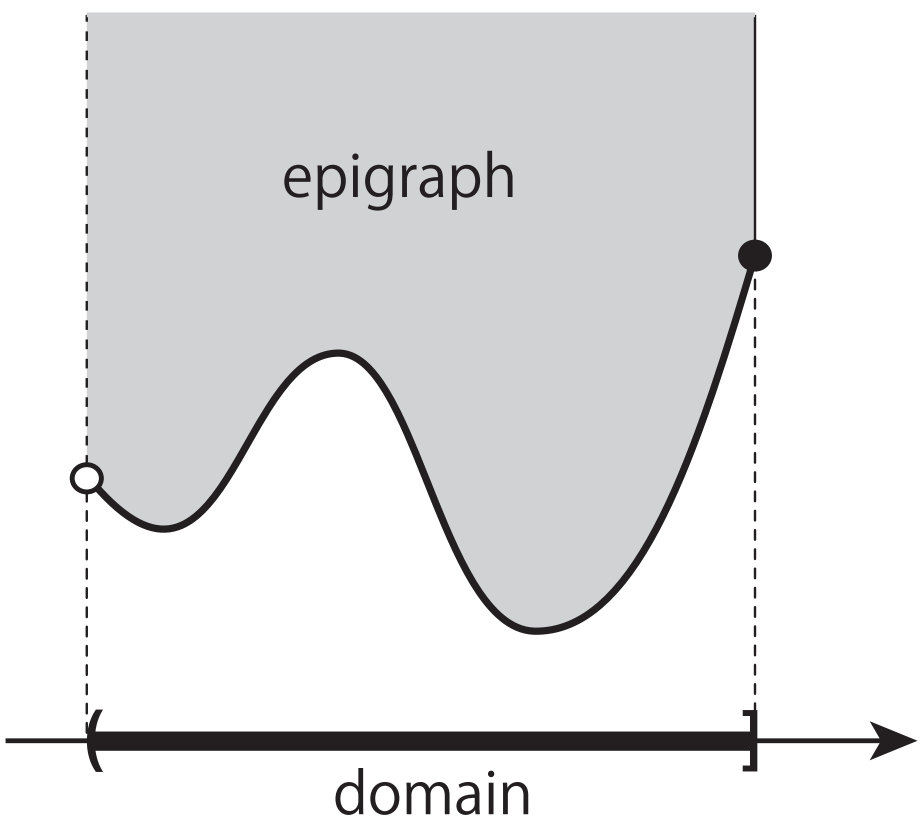


Reminders on Convexity: *Epigraph*

- Epigraphs and domain

Def

$$\mathrm{epi}(f) = \{(x,t) \in \mathbb{R}^p \times \mathbb{R} : f(x) \le t\}$$
$$\mathrm{dom}(f) = \{x \in \mathbb{R}^p : f(x) < \infty\}$$



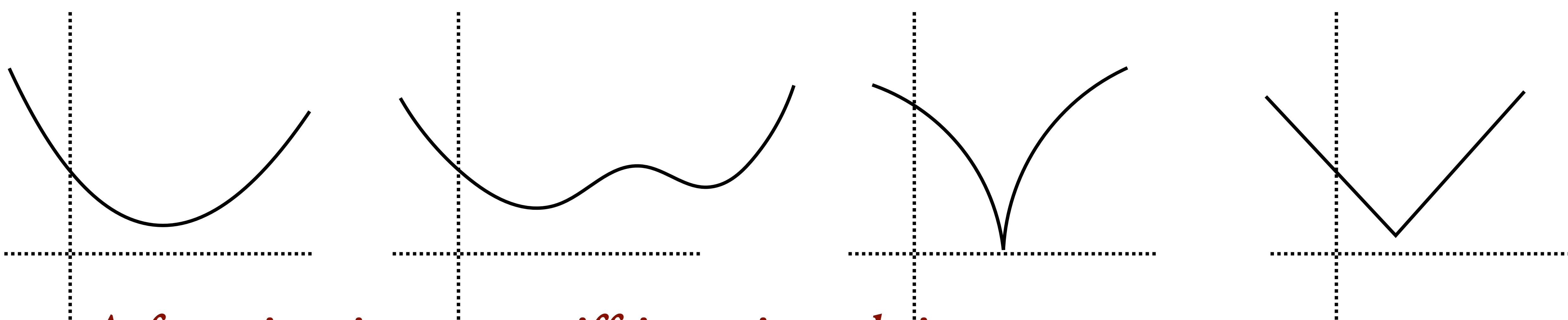
Reminders on Convexity: *Functions*

- Convex function

$$\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$$

Def

$$f : \mathbb{R}^p \to \bar{\mathbb{R}} \text{ convex}$$
$$\Updownarrow$$
$$\forall x_1, x_2 \in \mathbb{R}^p, 0 \leq \lambda \leq 1,$$
$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$



- *A function is convex iff its epigraph is.*

convex loss functions for regression

- Label is a real number (regression)

$$l(u,y) = \frac{1}{2}(u-y)^2,$$

quadratic

$$l_\tau(u,y) = (1-\tau)\max(u-y,0) + \tau\max(y-u,0), \tau \in [0,1]$$

$$l_\varepsilon(u,y) = \max(|y-u| - \varepsilon), \varepsilon > 0$$

tau-quantile

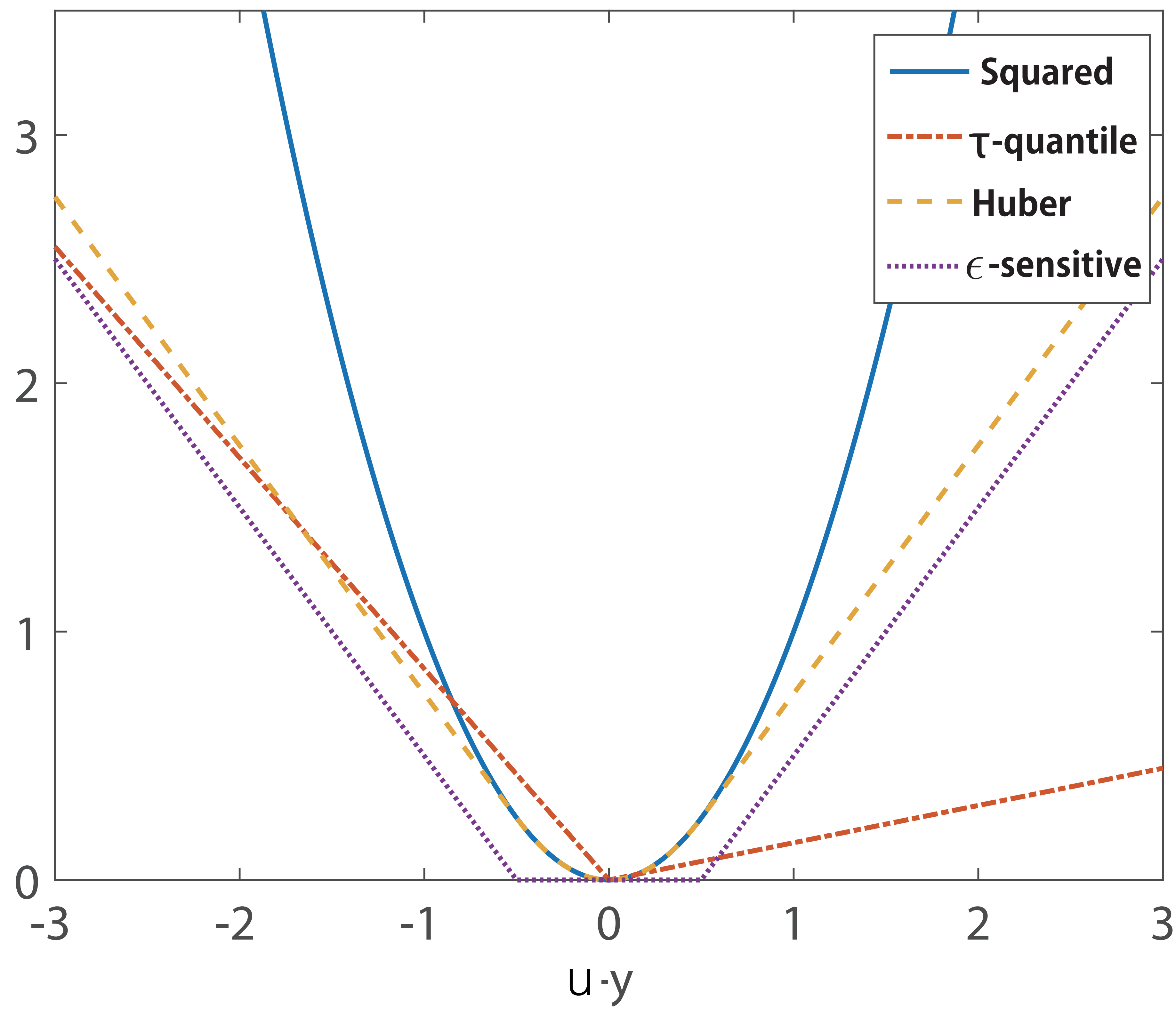
eps-sensitive

$$l_\delta(u,y) = \begin{cases} \frac{1}{2}(y-u)^2 & \text{for } |y-u| \le \delta, \\ \delta\,|y-u| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases}$$

huber

u

convex loss functions for regression



convex loss functions for classification

- Label is a binary, prediction is a number

$$l(u, y) = \log(1 + \exp(-yu)),$$

logistic

$$l(u, y) = |1 - yu|_+ = \max(1 - yu, 0),$$

hinge

$$l(u, y) = \exp(-yu),$$

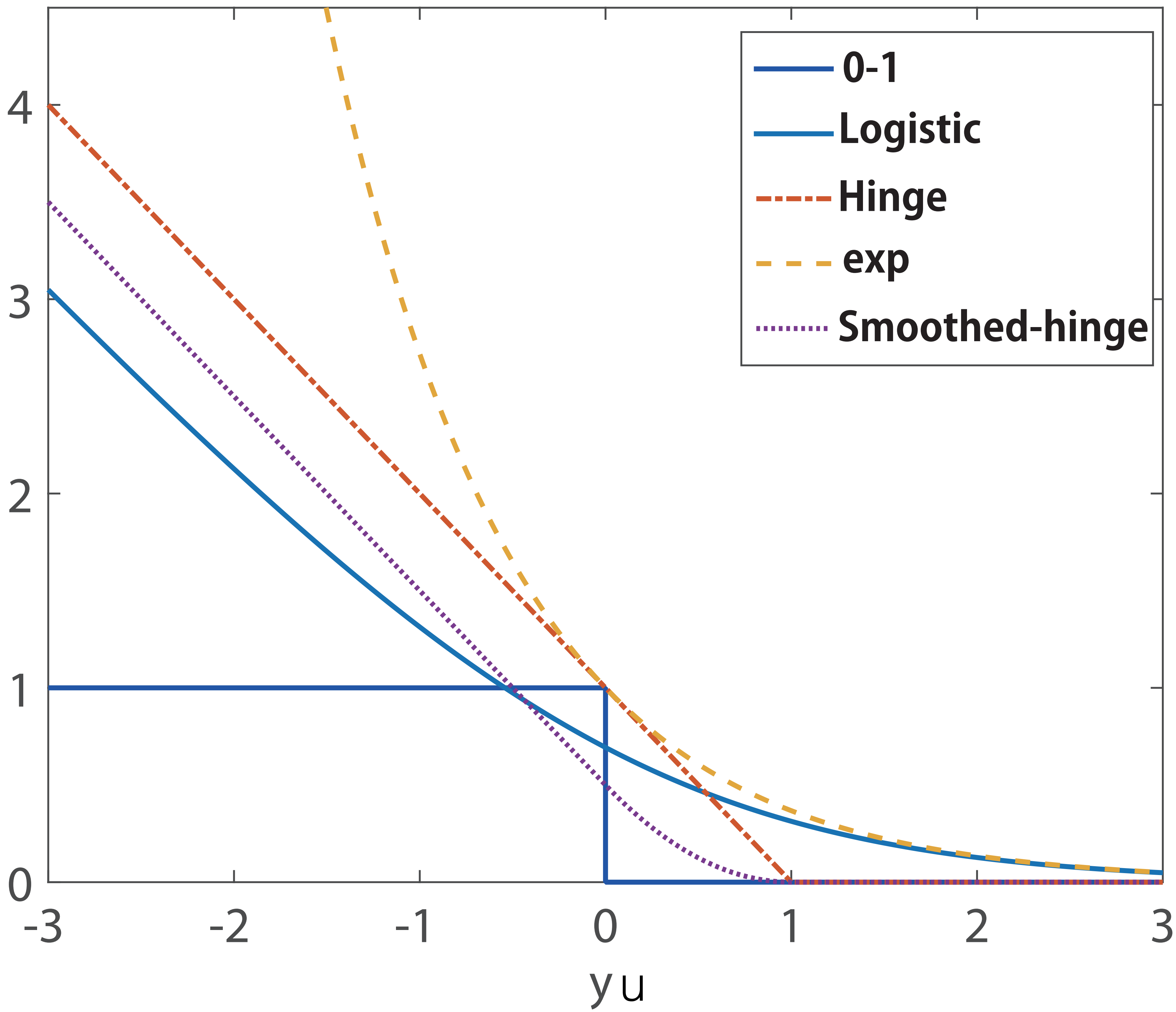
exponential

$$l(u, y) = \begin{cases} 0, & yu \geq 1, \\ \frac{1}{2} - yu, & yu < 0, \\ \frac{1}{2}(1 - yu)^2, & \text{otherwise.} \end{cases}$$

smoothed hinge

u

convex loss functions for regression



convex regularizers

$$\psi(\theta) = \|\theta\|_2^2 = \theta^T\theta,$$ ridge

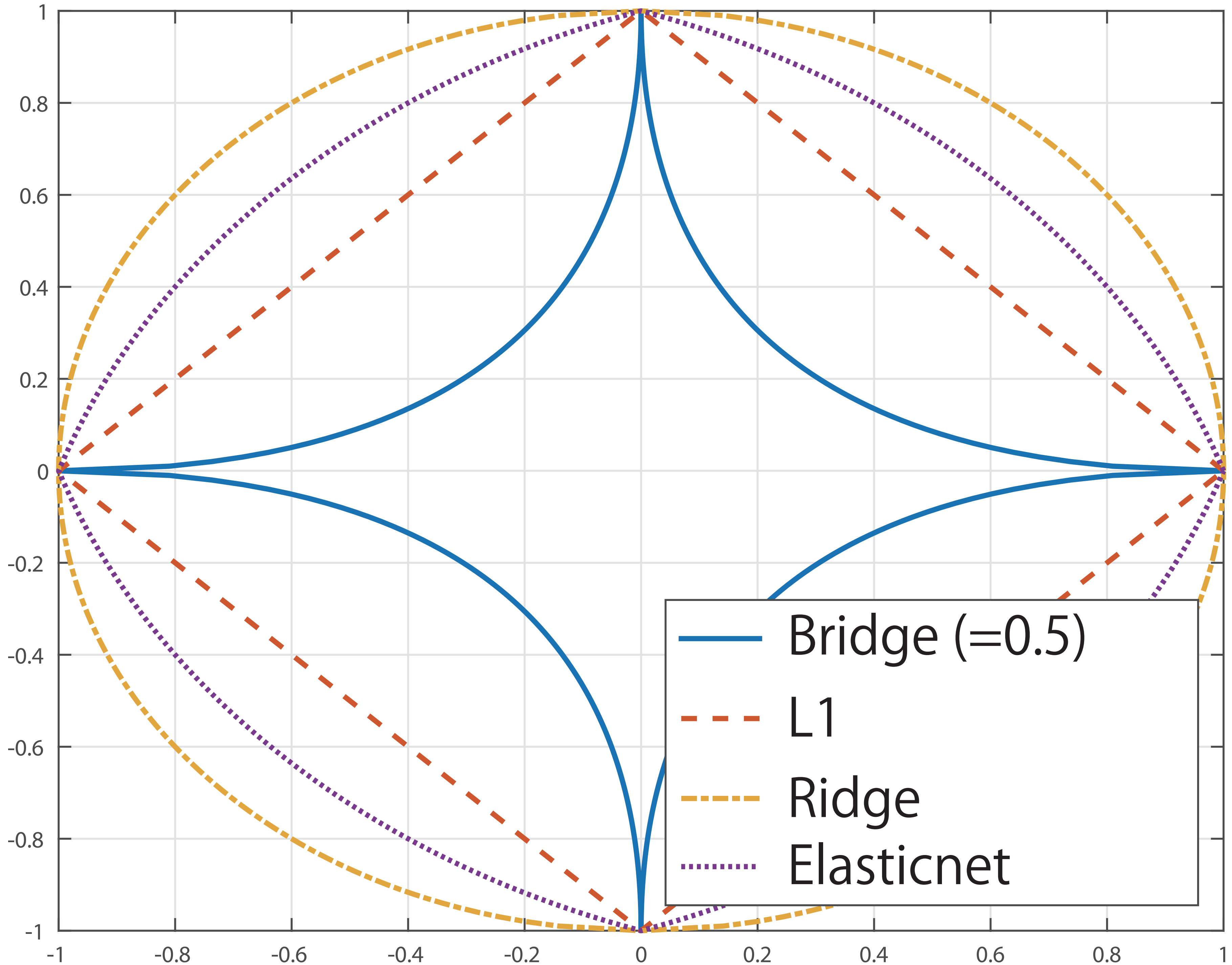
$$\psi(\theta) = \|\theta\|_1 = \sum_i |\theta_i|,$$ L-1

$$\psi(\theta) = a\|\theta\|_1 + b\|\theta\|_2^2,$$ elastic net

$$\psi(\theta) = \|\theta\|_{\mathrm{tr}} = \sum_i^{\min(q,r)} \sigma_j(\theta)$$

trace norm (for matrices)

convex loss functions for regression

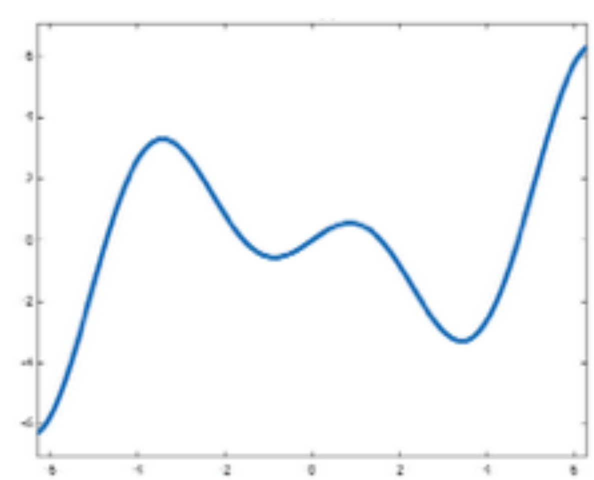


Gradients

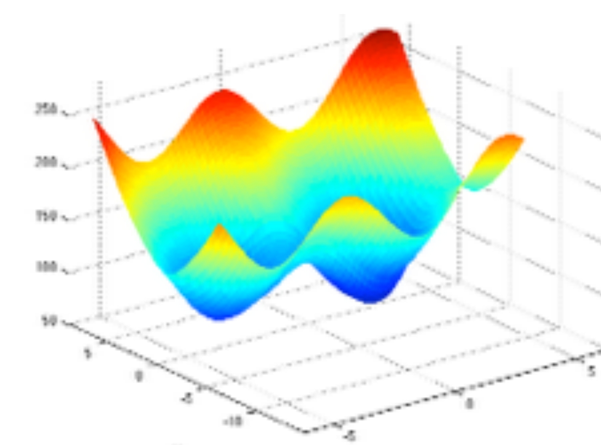
Def For a differentiable function $f : \mathbb{R}^p \to \bar{\mathbb{R}}$, the gradient of f at $x \in \mathrm{dom}(f)$ is

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_p}(x) \end{bmatrix}$$

$$g(x + \varepsilon) = g(x) + g'(x)\varepsilon + o(\varepsilon^2)$$



$$f(x + \varepsilon) = f(x) + \nabla f(x)^T \varepsilon + o(\|\varepsilon\|^2)$$



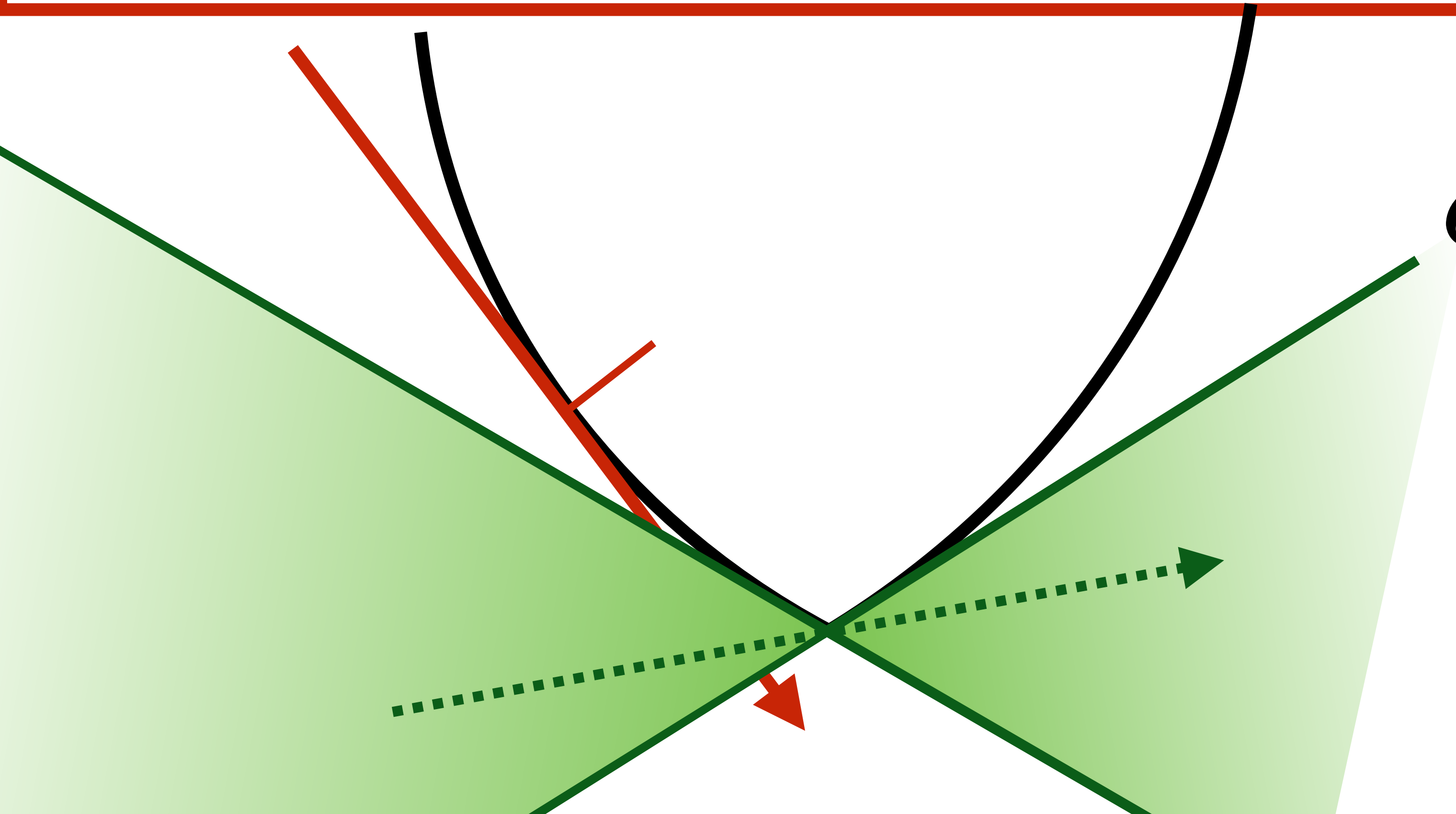
Subgradients

- Subgradients are natural generalization of gradients

Def

For convex function $f : \mathbb{R}^p \to \bar{\mathbb{R}}$, the subdifferential of f at $x \in \mathrm{dom}(f)$ is

$$\partial f(x) = \{g \in \mathbb{R}^p | \forall y \in \mathbb{R}^p, \langle y - x, g \rangle + f(x) \leq f(y)\}$$



$$\partial f(\boldsymbol{x_0}) = \{\nabla f(\boldsymbol{x_0})\}$$

$$\partial f(\boldsymbol{x_1}) = \{\boldsymbol{g}\}$$

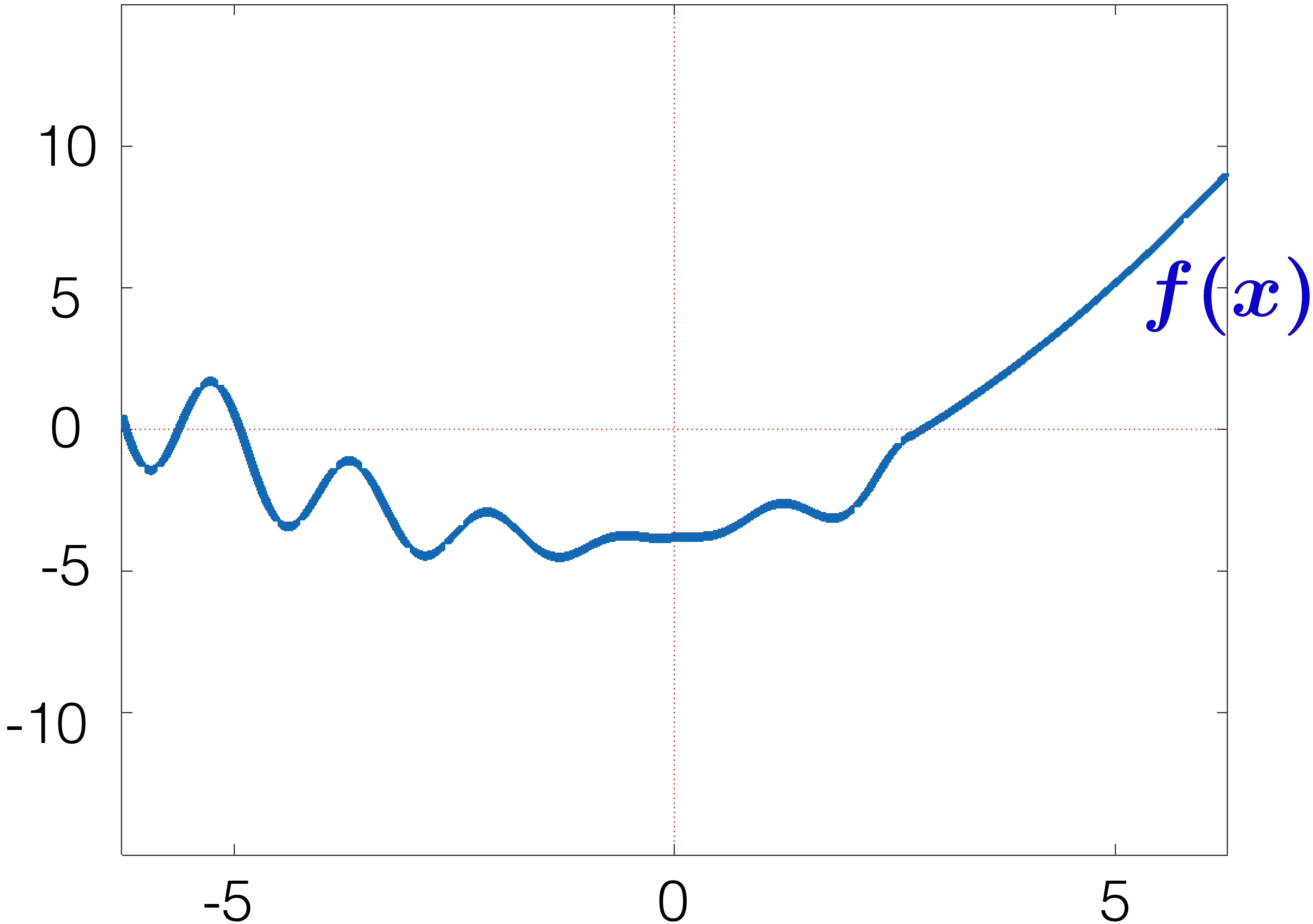
Legendre Transform

Def

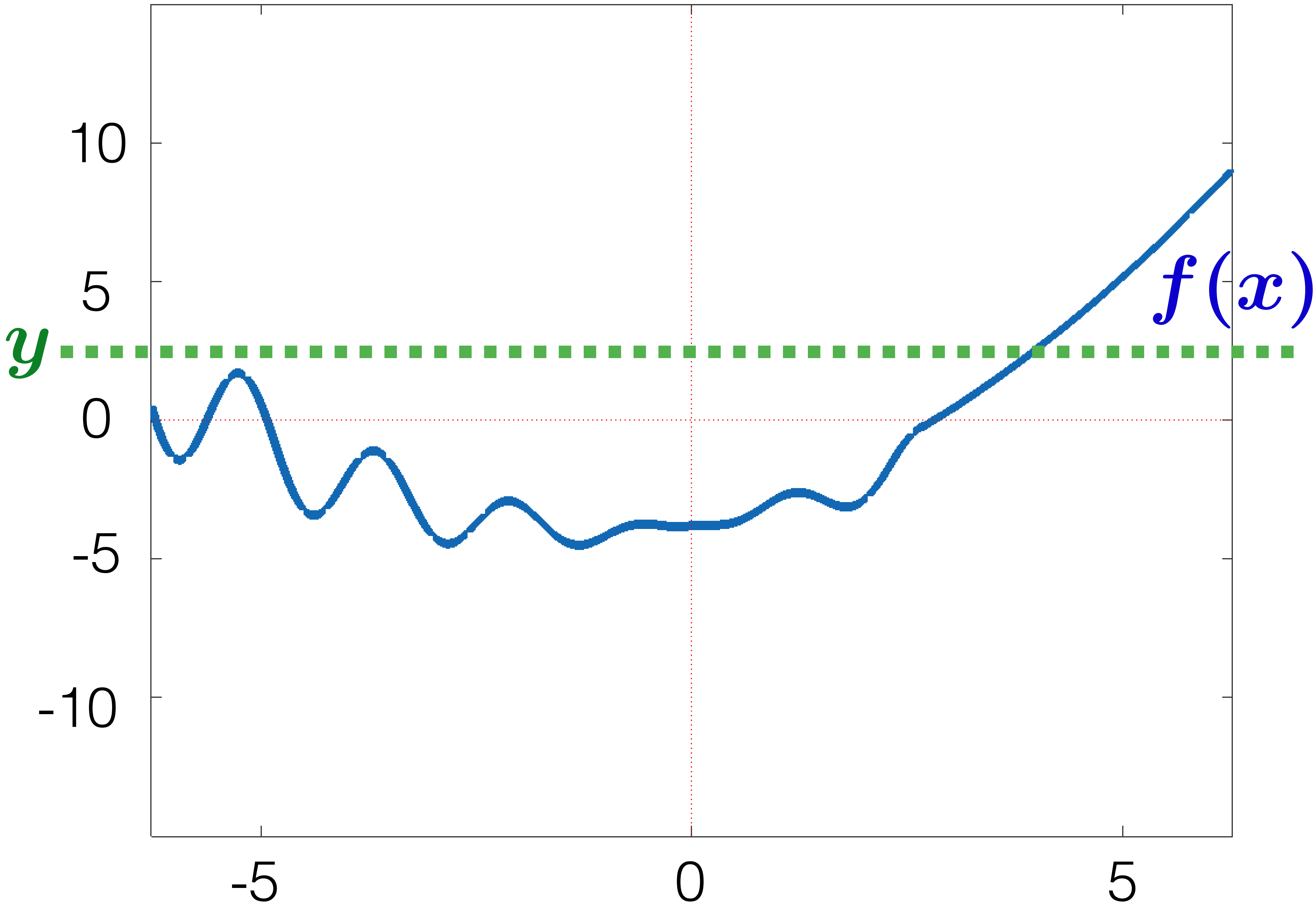
For a (possibly non convex) function $f : \mathbb{R}^p \to \bar{\mathbb{R}}$, the convex conjugate of f is, $\forall \boldsymbol{y} \in \mathbb{R}^p$,

$$f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in \mathbb{R}^p} \langle \boldsymbol{x}, y \rangle - f(\boldsymbol{x})$$

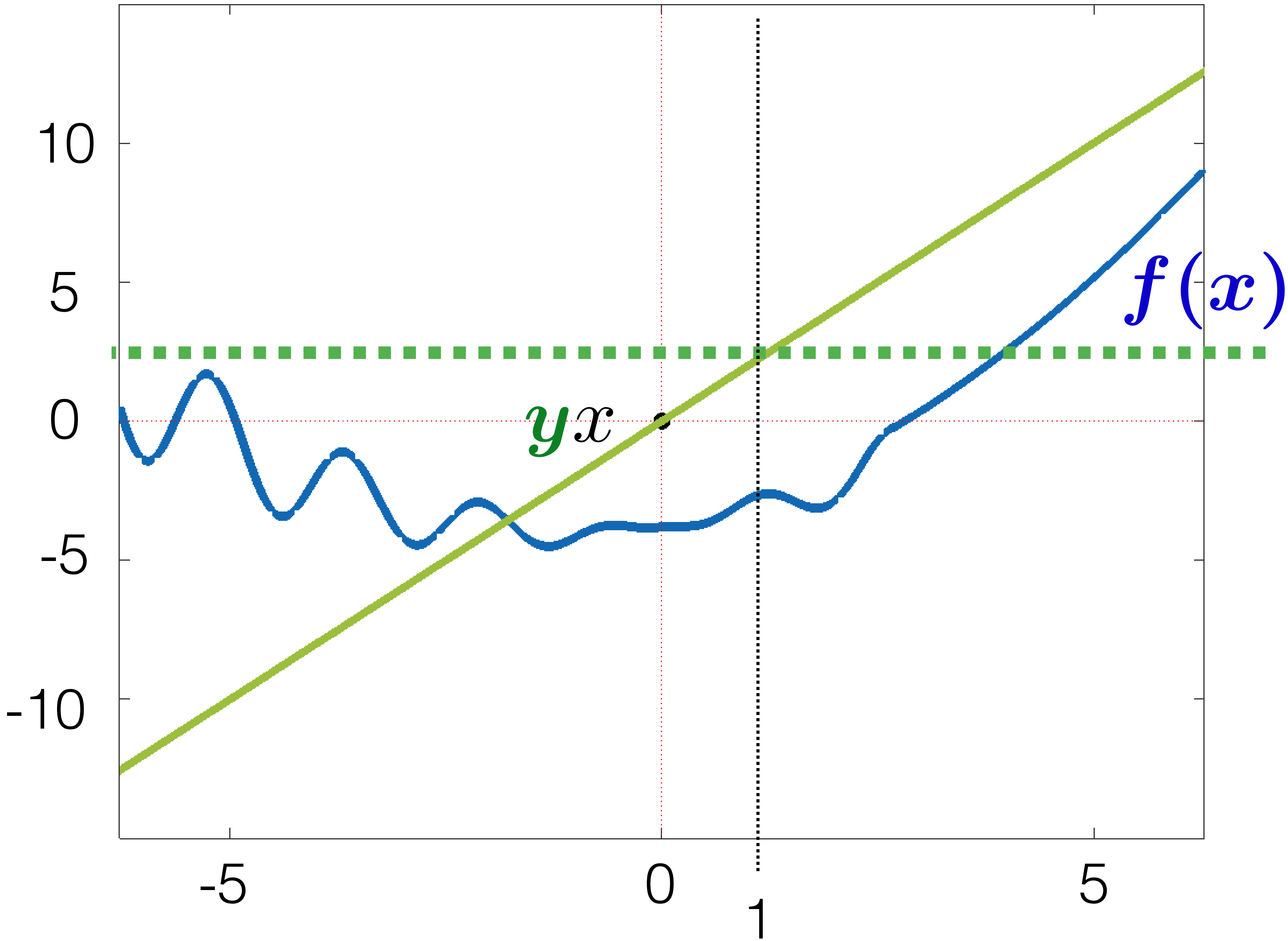
Legendre Transform



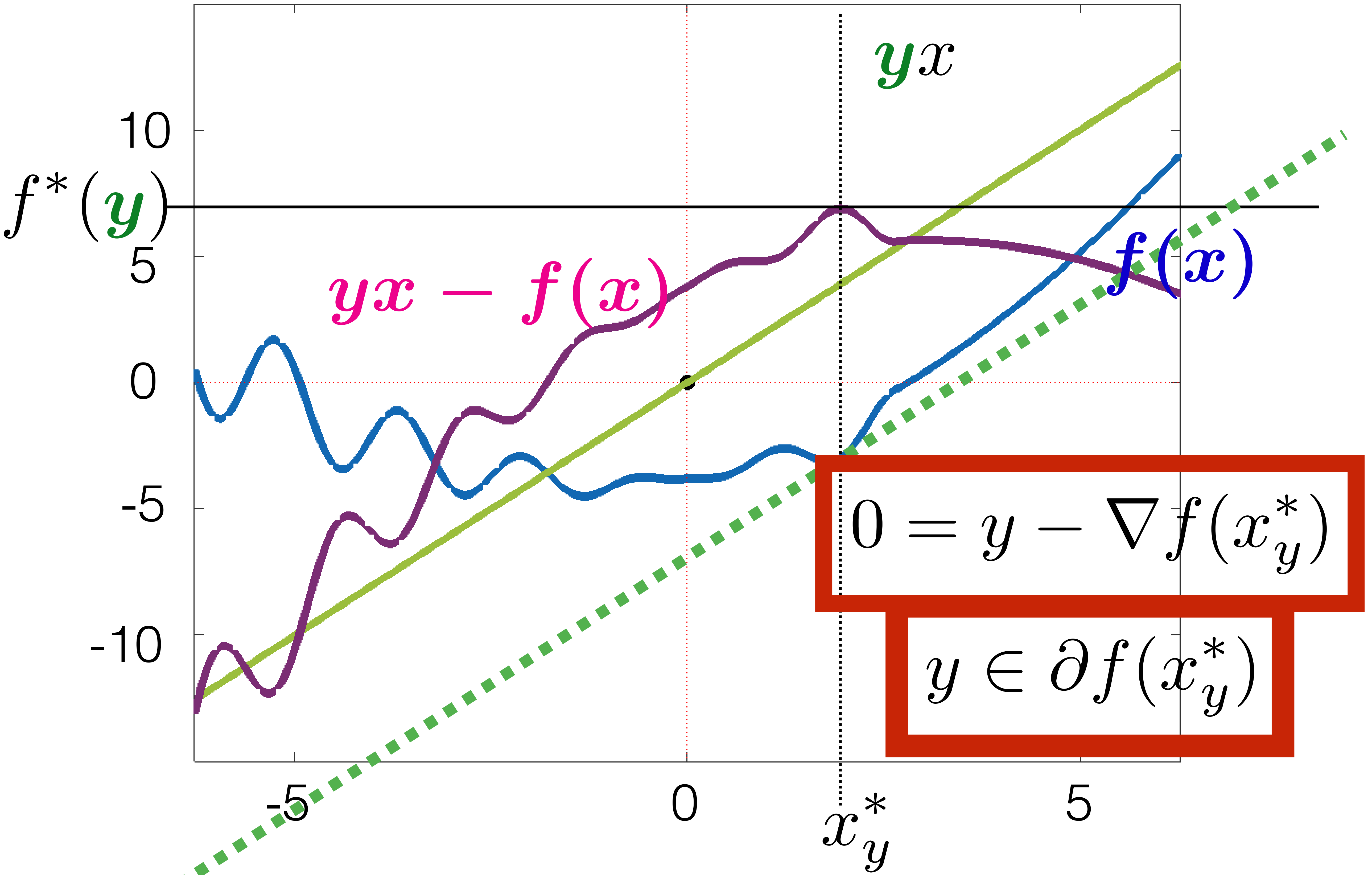
Legendre Transform



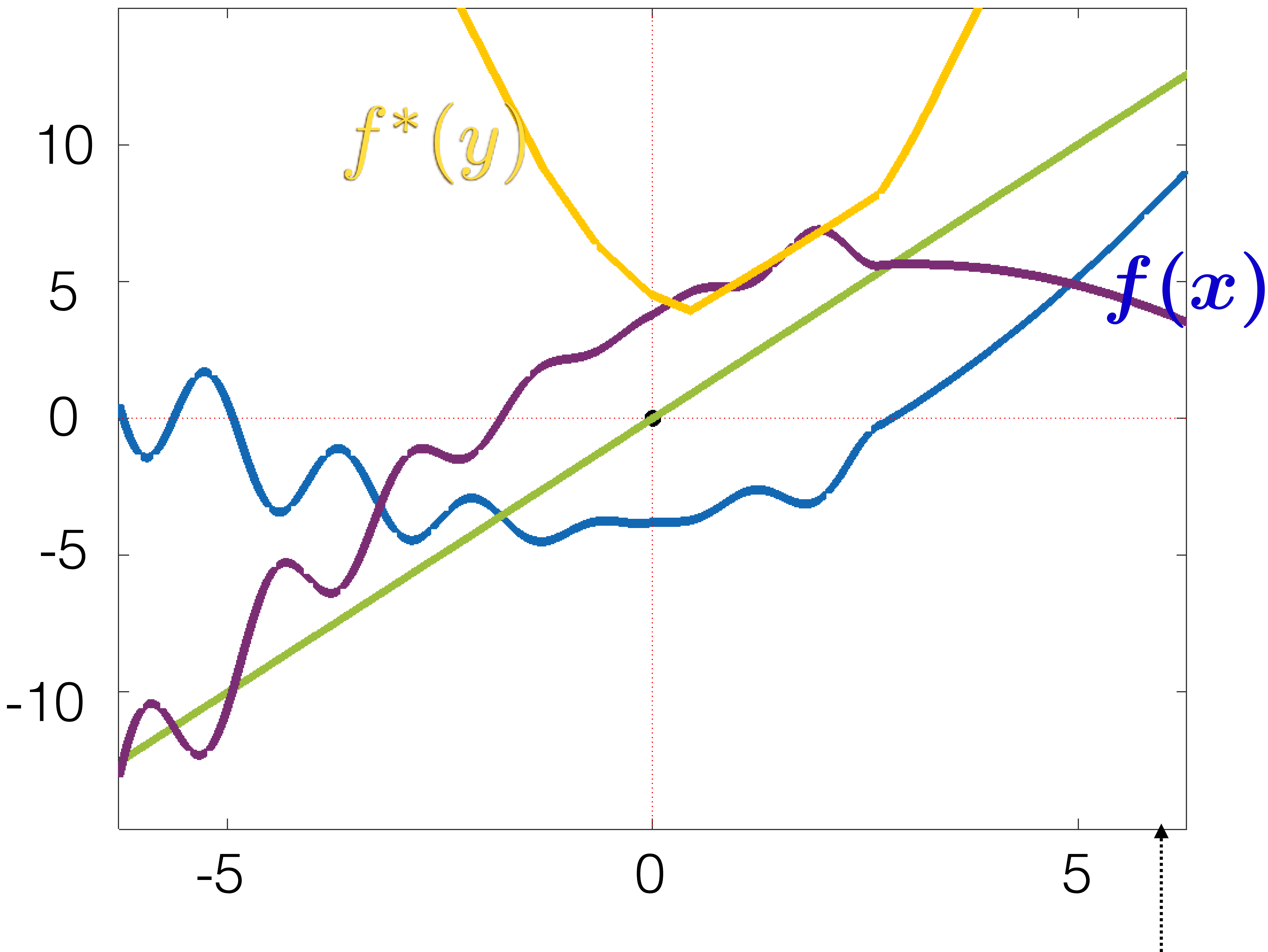
Legendre Transform



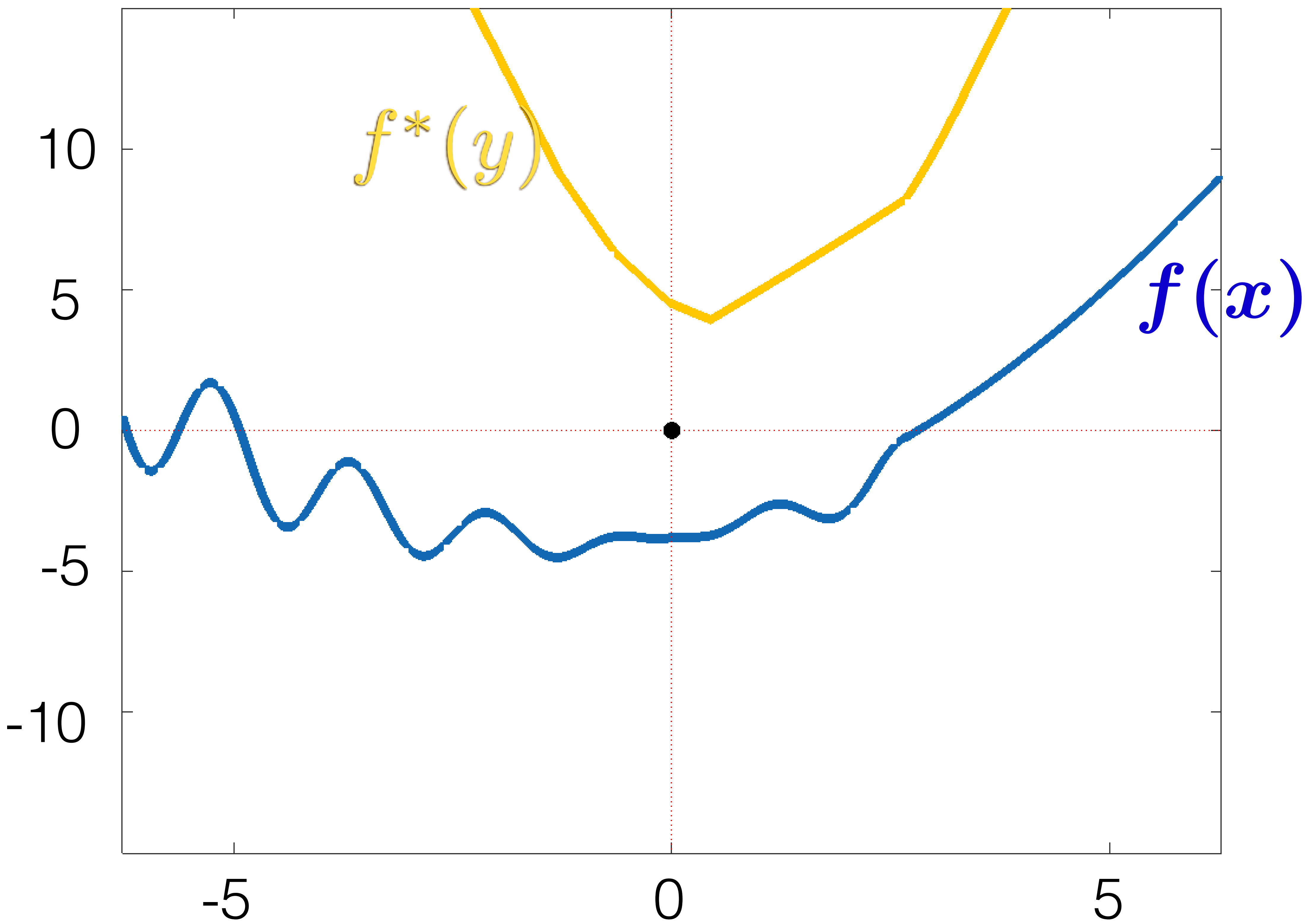
Legendre Transform



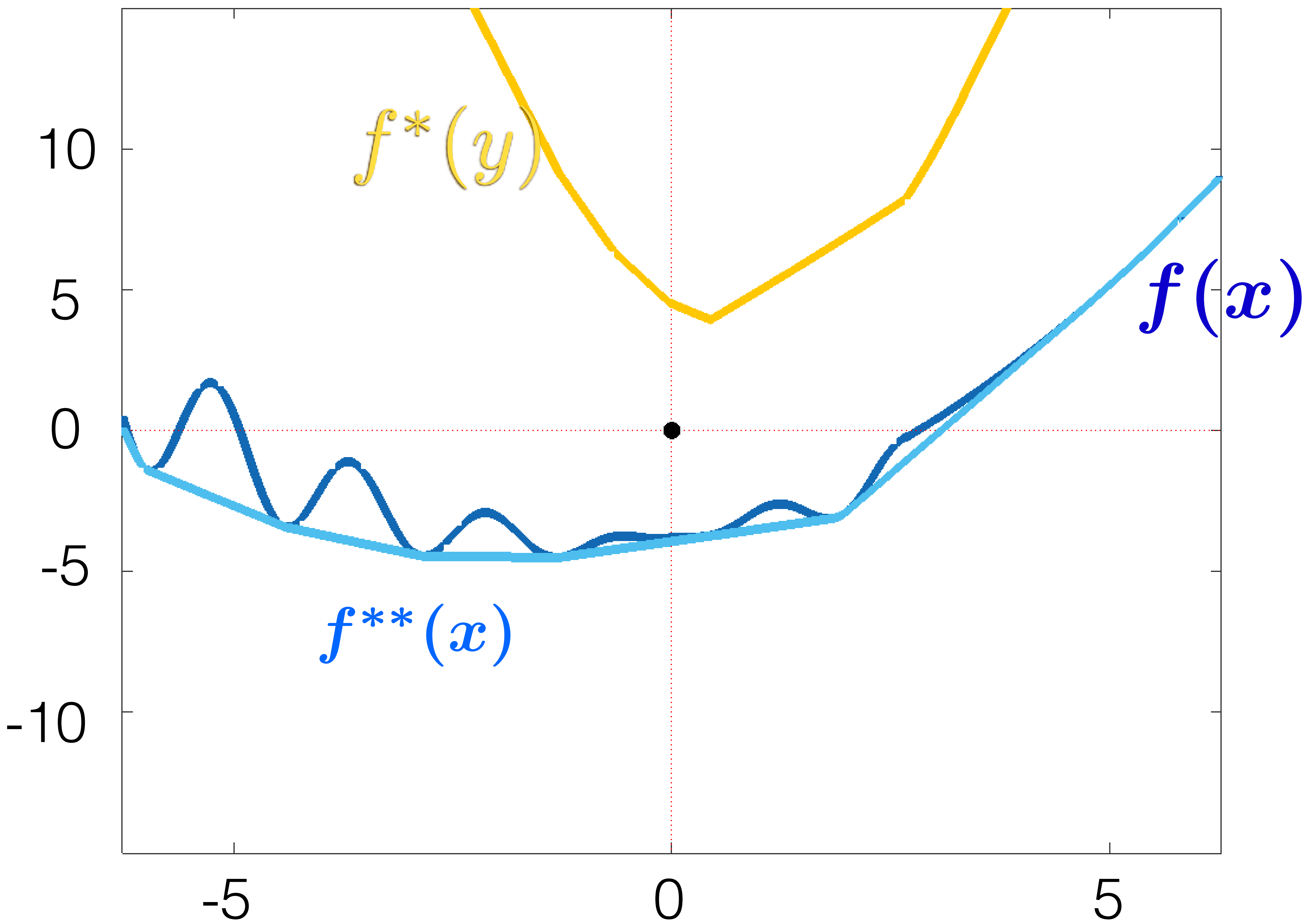
Legendre Transform



Legendre Transform



Legendre Transform



Legendre Transform

Def

For a (possibly non convex) function $f : \mathbb{R}^p \to \bar{\mathbb{R}}$, the convex conjugate of f is $\forall y \in \mathbb{R}^p$,

$$f^*(y) = \sup_{x \in \mathbb{R}^p} \langle x, y \rangle - f(x)$$

	$f(x)$	$f^*(y)$
Squared loss	$\frac{1}{2}x^2$	$\frac{1}{2}y^2$
Hinge loss	$\max\{1-x, 0\}$	$\begin{cases} y & (-1 \leq y \leq 0), \\ \infty & (\text{otherwise}). \end{cases}$
Logistic loss	$\log(1+\exp(-x))$	$\begin{cases} (-y)\log(-y) + (1+y)\log(1+y) & (-1 \leq y \leq 0), \\ \infty & (\text{otherwise}). \end{cases}$
L_1 regularization	$\|x\|_1$	$\begin{cases} 0 & (\max_j \lvert y_j \rvert \leq 1), \\ \infty & (\text{otherwise}). \end{cases}$
L_p regularization $(p > 1)$	$\sum_{j=1}^d \lvert x_j \rvert^p$	$\sum_{j=1}^d \frac{p-1}{p^{\frac{p}{p-1}}} \lvert y_j \rvert^{\frac{p}{p-1}}$

Legendre Transform

Def

For a (possibly non convex) function $f : \mathbb{R}^p \to \bar{\mathbb{R}}$, the convex conjugate of f is $\forall y \in \mathbb{R}^p$,

$$f^*(y) = \sup_{x \in \mathbb{R}^p} \langle x, y \rangle - f(x)$$

f^* is convex, even if f is not.

$$y \in \partial f(x) \Leftrightarrow f(x) + f^*(y) = \langle x, y \rangle \Leftrightarrow x \in \partial f^*(y)$$

$$\forall x, y, f(x) + f^*(y) \geq \langle x, y \rangle$$

Fenchel Duality Theorem

Theorem

Let $f : \mathbb{R}^p \to \bar{R}$ and $g : \mathbb{R}^q \to \bar{R}$ be closed convex, and $A \in \mathbb{R}^{q \times p}$ a linear map. Suppose that either condition (a) or (b) is satisfied. Then

$$\inf_{x \in \mathbb{R}^p} f(x) + g(Ax) = \sup_{y \in \mathbb{R}^q} -f^*(A^T y) - g^*(-y)$$

$(a) \exists x \in \mathbb{R}^p$ s.t. $x \in \mathrm{ri}(\mathrm{dom}(f))$ and $Ax \in \mathrm{ri}(\mathrm{dom}(g))$

$(b) \exists y \in \mathbb{R}^q$ s.t. $A^T y \in \mathrm{ri}(\mathrm{dom}(f^*))$ and $-y \in \mathrm{ri}(\mathrm{dom}(g^*))$

Fenchel Duality Theorem

Theorem

Let $f : \mathbb{R}^p \to \bar{R}$ and $g : \mathbb{R}^q \to \bar{R}$ be closed convex, and $A \in \mathbb{R}^{q \times p}$ a linear map. Suppose that either condition (a) or (b) is satisfied. Then

$$\inf_{x \in \mathbb{R}^p} f(x) + g(Ax) = \sup_{y \in \mathbb{R}^q} -f^*(A^T y) - g^*(-y)$$

$(a) \exists x \in \mathbb{R}^p$ s.t. $x \in \mathrm{ri}(\mathrm{dom}(f))$ and $Ax \in \mathrm{ri}(\mathrm{dom}(g))$

$(b) \exists y \in \mathbb{R}^q$ s.t. $A^T y \in \mathrm{ri}(\mathrm{dom}(f^*))$ and $-y \in \mathrm{ri}(\mathrm{dom}(g^*))$

Fenchel Duality and ERM

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l_{\boldsymbol{\theta}}(z_i) + \psi(\boldsymbol{\theta})$$

$$l_{\boldsymbol{\theta}}(z_i) = l(y_i, x_i^T \boldsymbol{\theta})$$

$$\frac{1}{n} \sum_i l_{\boldsymbol{\theta}}(z_i) = \mathbf{l}(\mathbf{y}, X\boldsymbol{\theta}) = g(X\boldsymbol{\theta})$$

$$X \in \mathbb{R}^{n \times p}$$

$$\sup_{\boldsymbol{y} \in \mathbb{R}^{\boldsymbol{n}}} -\psi^*(-X^T y) - g^*(y) = -\inf_{\boldsymbol{y} \in \mathbb{R}^{\boldsymbol{n}}} g^*(y) + \psi^*(-X^T y)$$

$$\sup_{\boldsymbol{y} \in \mathbb{R}^{\boldsymbol{n}}} \sum_i l_i^*(y_i) + \psi^*(-X^T y)$$

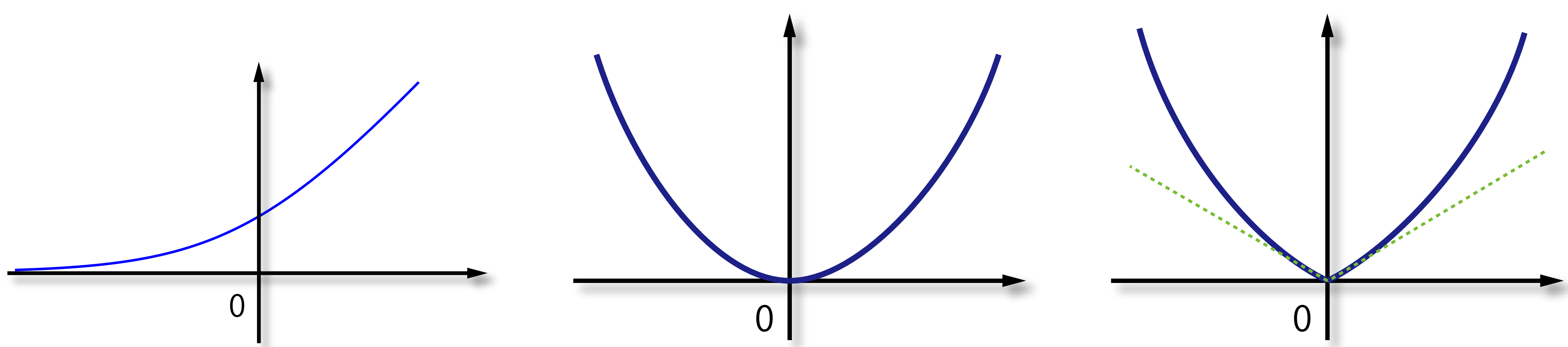
Smoothness of Functions

Def

smoothness: gradient is Lipschitz continuous

$$\|\nabla f(x) - \nabla f(x')\| \leq \boldsymbol{L}\|x - x'\|$$

strong convexity: f is $\boldsymbol{\mu}$-strongly convex if $x \to f(x) - \frac{\boldsymbol{\mu}}{2}\|x\|^2$ is convex.



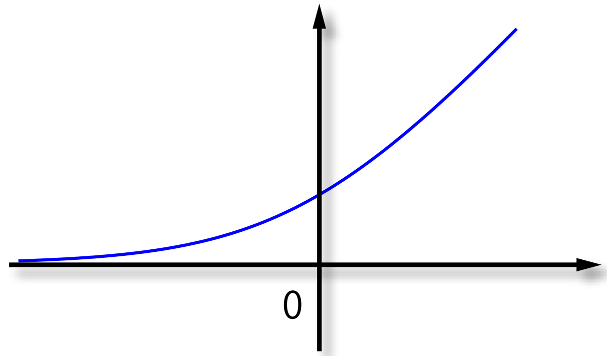
Smoothness of Functions

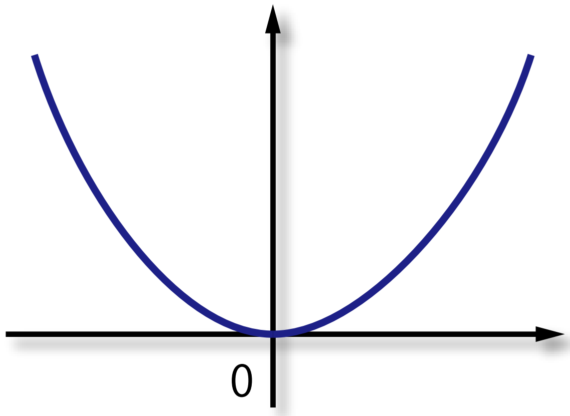
Def

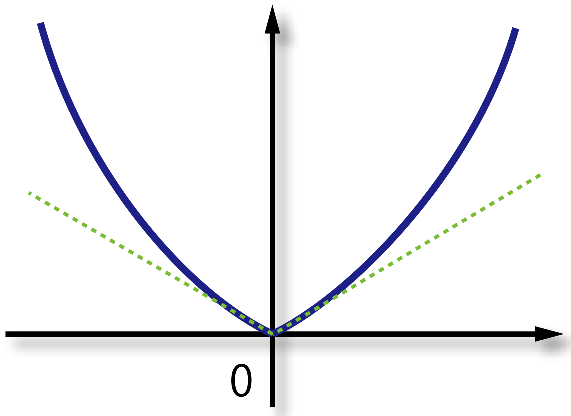
smoothness: gradient is Lipschitz continuous

$$\|\nabla f(x) - \nabla f(x')\| \leq \boldsymbol{L}\|x - x'\|$$

strong convexity: f is $\boldsymbol{\mu}$-strongly convex if $x \to f(x) - \frac{\boldsymbol{\mu}}{2}\|x\|^2$ is convex.



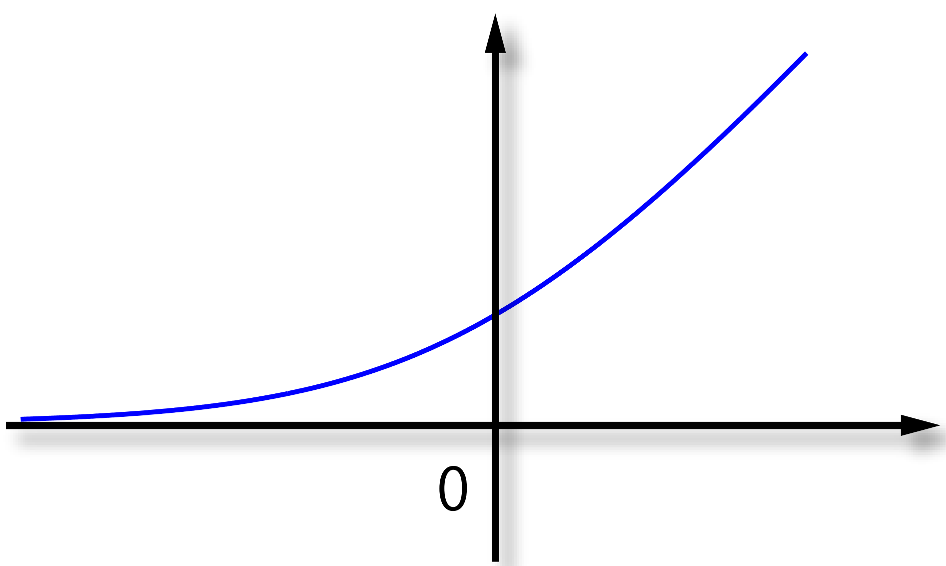




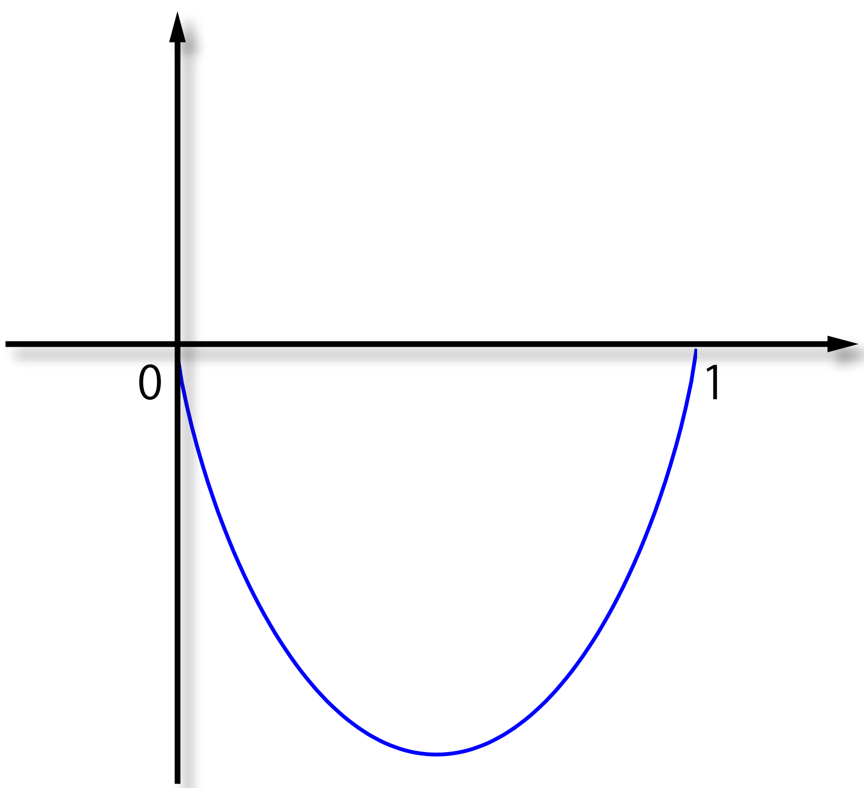
Smoothness of Functions

Theorem

f is $\boldsymbol{L}$ -smooth $\Leftrightarrow$ f^* is $\frac{1}{L}$ - strongly convex.



Logistic: loss is smooth, not strongly convex



Dual Logistic: strongly convex, not smooth