Proximal operators and proximal gradient methods

Pierre Ablin

The Training Problem

Solving the *training problem*: $$\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

Reference method: Gradient descent

$$\nabla\left(\frac{1}{n}\sum_{i=1}^{n} f_i(w)\right) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(w)$$

Gradient Descent Algorithm

Set $w^1 = 0$, choose $\alpha > 0$.

for $t = 1, 2, 3, \ldots, T$

$$w^{t+1} = w^t - \frac{\alpha}{n}\sum_{i=1}^{n}\nabla f_i(w^t)$$

Output w^{T+1}

Convergence GD I

Theorem

Let f be convex and L-smooth.

$$f(w^T) - f(w^*) \leq \frac{2L||w^0 - w^*||_2^2}{T} = O\left(\frac{1}{T}\right).$$

Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

$\Rightarrow$ for $\dfrac{f(w^T) - f(w^*)}{||w^0 - w^*||_2^2} \leq \epsilon$ we need $T \geq \dfrac{2L}{\epsilon} = O\left(\dfrac{1}{\epsilon}\right)$

Convergence GD I

Theorem

Let f be convex and L-smooth.

$$f(w^T) - f(w^*) \leq \frac{2L||w^0 - w^*||_2^2}{T} = O\left(\frac{1}{T}\right).$$

Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

Is f always differentiable?

$$\Rightarrow \text{for } \frac{f(w^T) - f(w^*)}{||w^0 - w^*||_2^2} \leq \epsilon \text{ we need } T \geq \frac{2L}{\epsilon} = O\left(\frac{1}{\epsilon}\right)$$

Convergence GD I

Not true for many problems

Theorem

Let f be convex and L-smooth.

$$f(w^T) - f(w^*) \leq \frac{2L||w^0 - w^*||_2^2}{T} = O\left(\frac{1}{T}\right).$$

Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

Is f always differentiable?

$$\Rightarrow \text{for } \frac{f(w^T) - f(w^*)}{||w^0 - w^*||_2^2} \leq \epsilon \text{ we need } T \geq \frac{2L}{\epsilon} = O\left(\frac{1}{\epsilon}\right)$$

Change notation: Keep loss and regularizer separate

Data fit function

$$F(w) := \frac{1}{n}\sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)$$

The Training problem

$$\min_w F(w) + \lambda R(w)$$

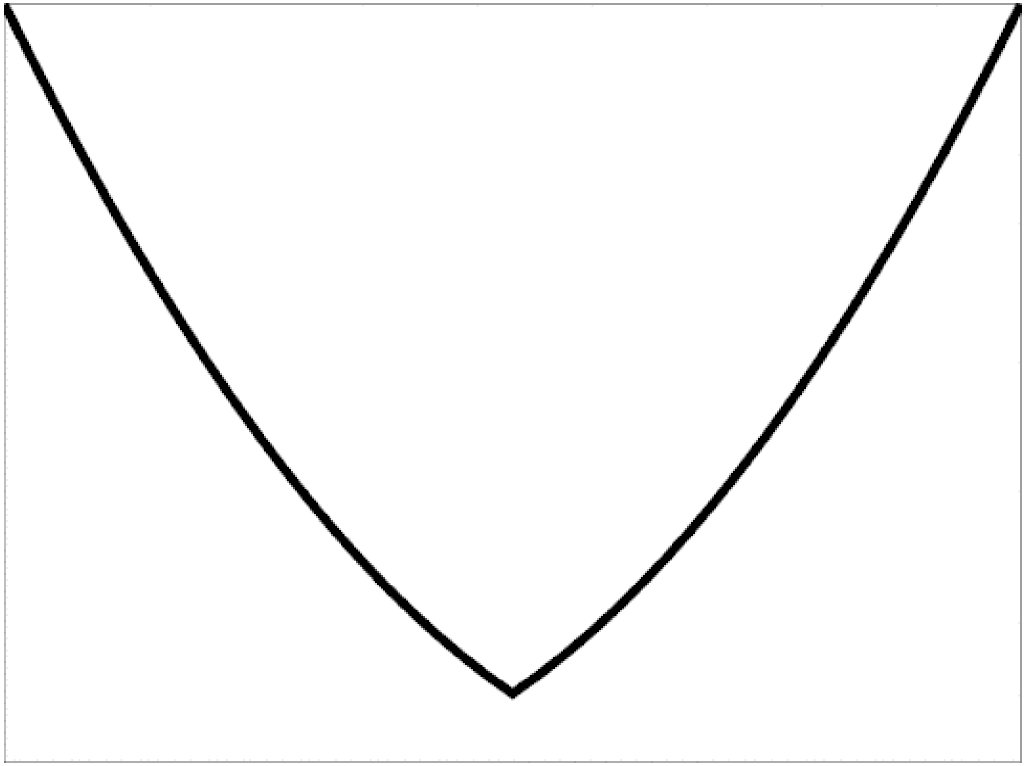
If F or R is not differentiable → F+R is not differentiable

If F or R is not smooth → F+R is not smooth

(In most cases)

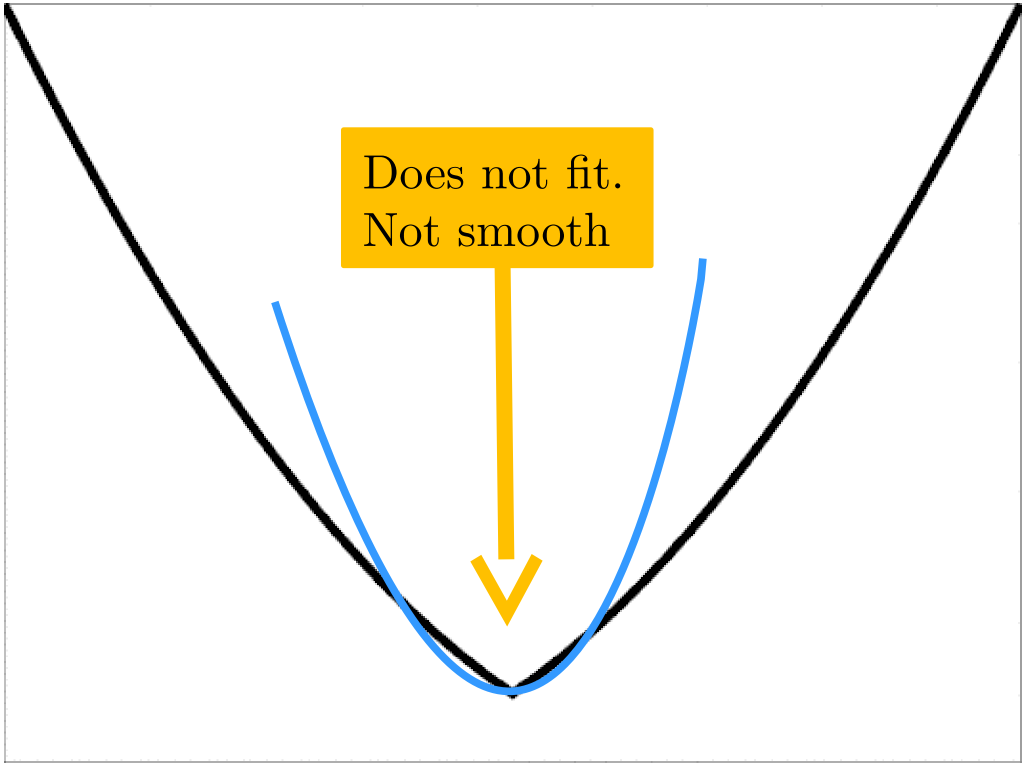
Non-smooth Example

$$F(w) + R(w) = \frac{1}{2}||w||_2^2 + ||w||_1$$



Non-smooth Example

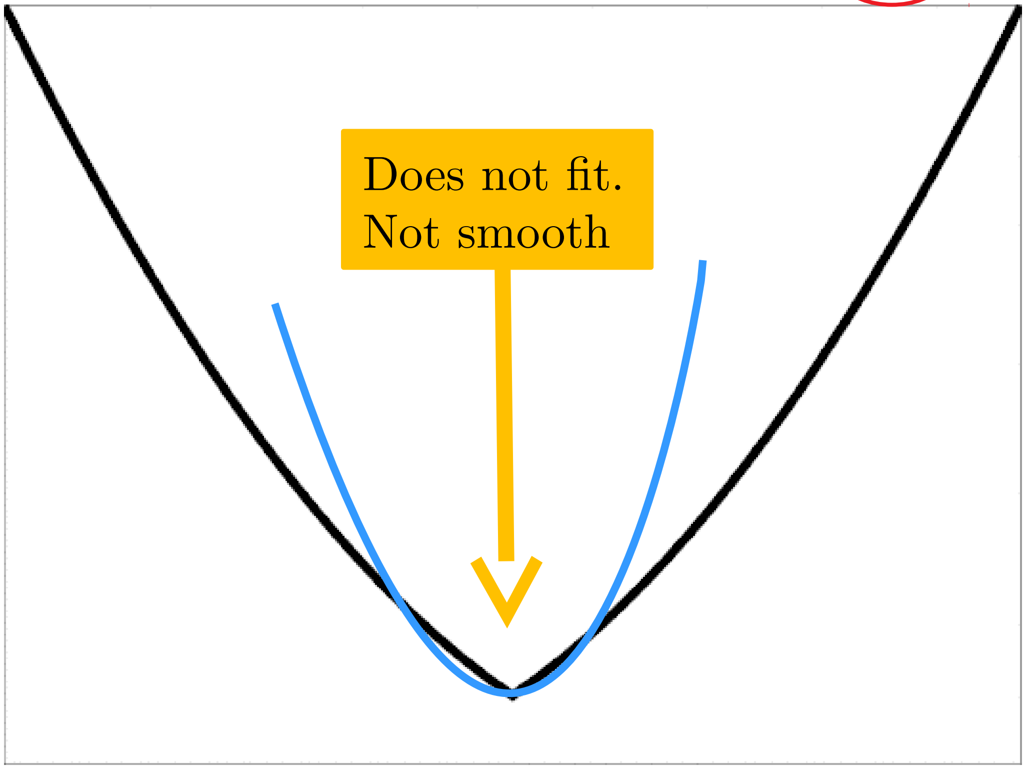
$$F(w) + R(w) = \frac{1}{2}||w||_2^2 + ||w||_1$$



Non-smooth Example

The problem

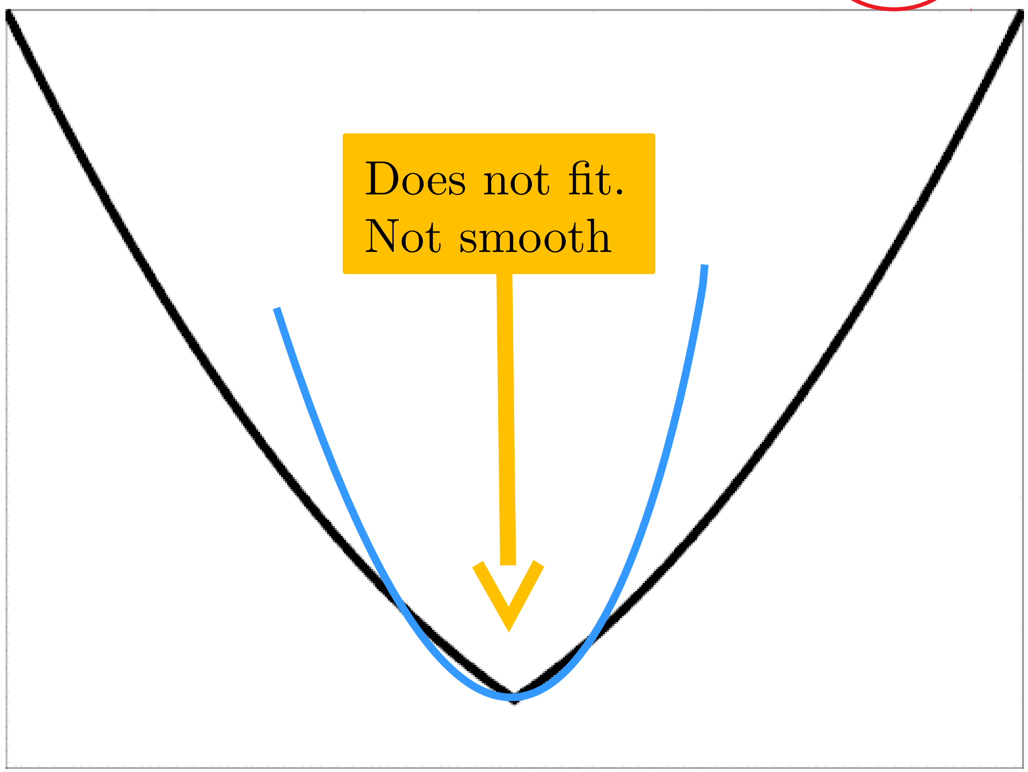
$$F(w) + R(w) = \frac{1}{2}||w||_2^2 + ||w||_1$$



Non-smooth Example

The problem

$$F(w) + R(w) = \frac{1}{2}||w||_2^2 + ||w||_1$$



Need more tools

Assumptions for this class

The Training problem

$$\min_{w} F(w) + \lambda R(w)$$

$F(w)$ is differentiable, L–smooth and convex

$R(w)$ is convex and "easy to optimize"

Assumptions for this class

The Training problem

$$\min_{w} F(w) + \lambda R(w)$$

$F(w)$ is differentiable, L–smooth and convex

$R(w)$ is convex and "easy to optimize"

What does this mean?

Examples

Lasso

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}\|Xw - y\|^2 + \lambda ||w||_1$$

Low Rank Matrix Recovery

$$\min_{W \in \mathbf{R}^{d \times d}} \frac{1}{2}||AW - Y||_F^2 + \lambda ||W||_*$$

SVM with soft margin

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y^i \langle w, a^i \rangle\} + \lambda ||w||_2^2$$

Not smooth

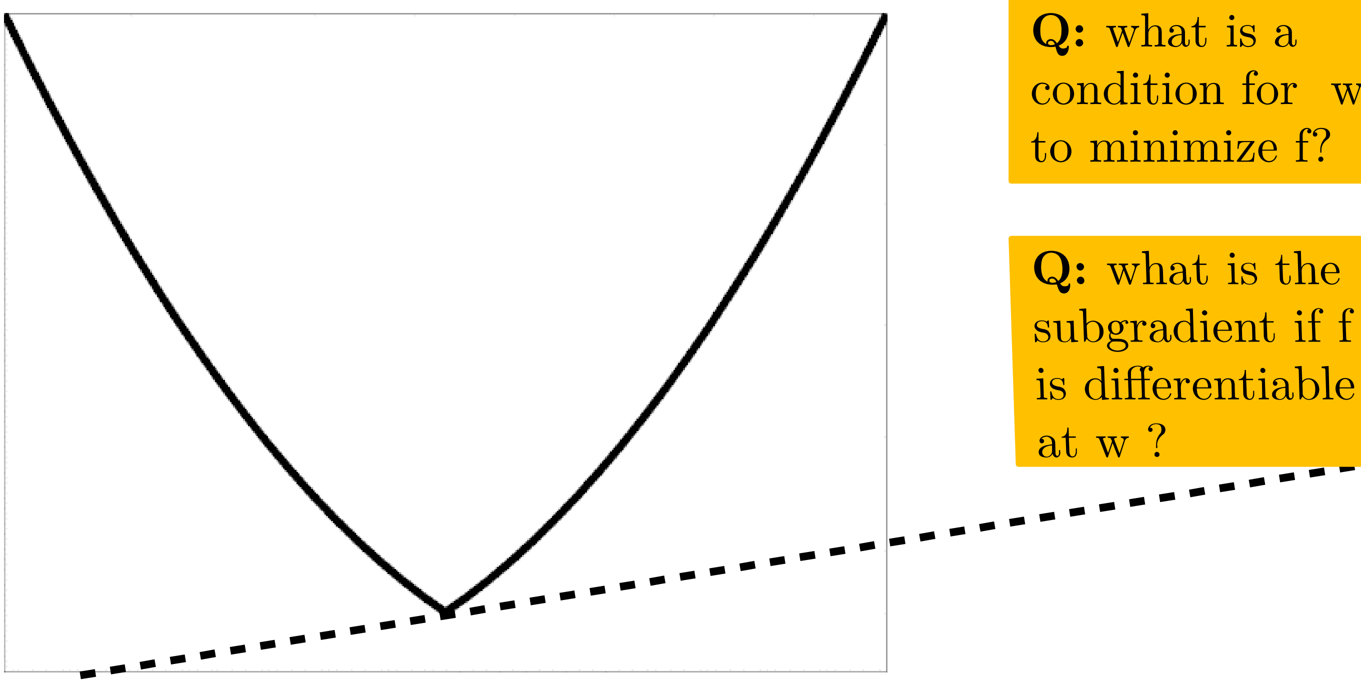
Not smooth

$$\|W\|_* = \mathrm{trace}(\sqrt{W^\top W}) = \sum_{i=1}^{d} \sigma_i(W)$$

Convexity without smoothness: Subgradient

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex

$$\partial f(w) := \{g \in \mathbb{R}^n \ : \ f(y) \geq f(w) + \langle g, y - w \rangle, \forall y \in \mathrm{dom}(f)\}$$



$f(w) + \langle g, y - w \rangle$

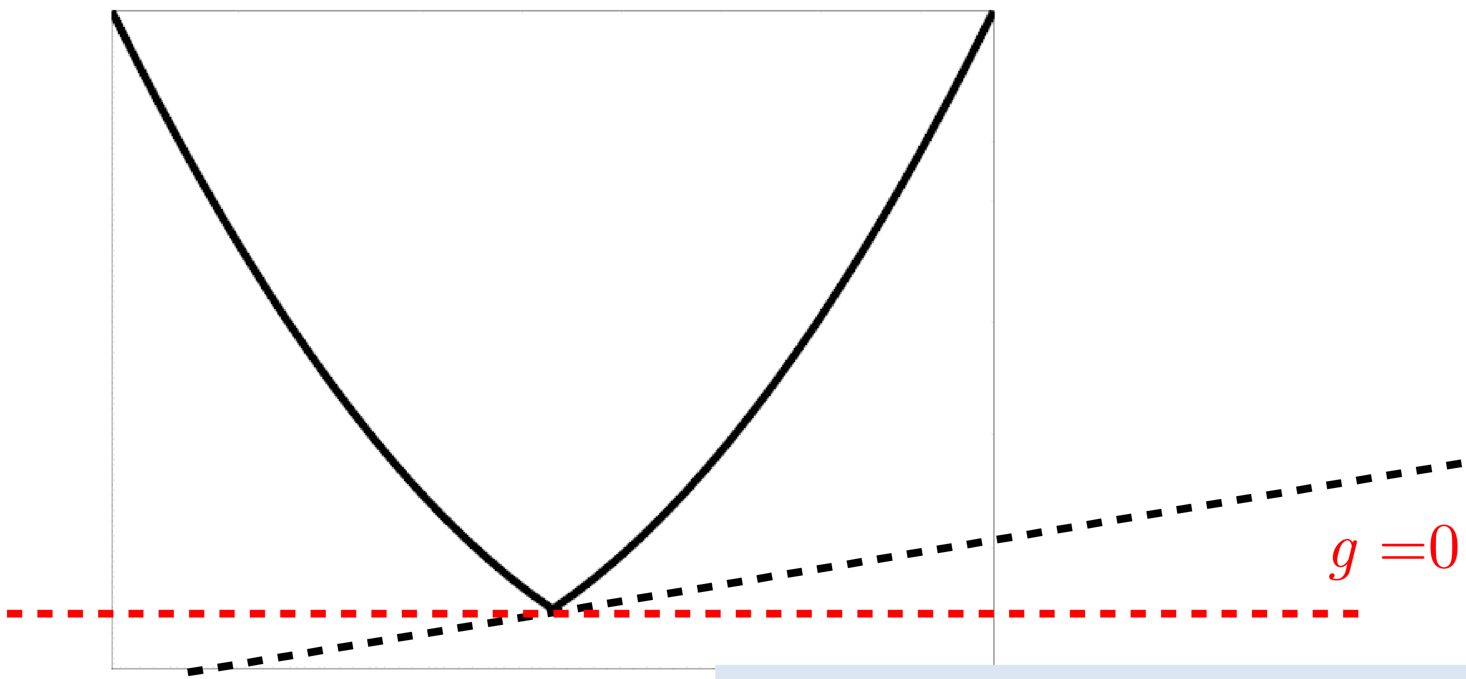
Q: what is a condition for w to minimize f?

Q: what is the subgradient if f is differentiable at w ?

Convexity without smoothness: Subgradient

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex

$$\partial f(w) := \{g \in \mathbb{R}^n \;:\; f(y) \geq f(w) + \langle g, y - w\rangle, \forall y \in \mathrm{dom}(f)\}$$



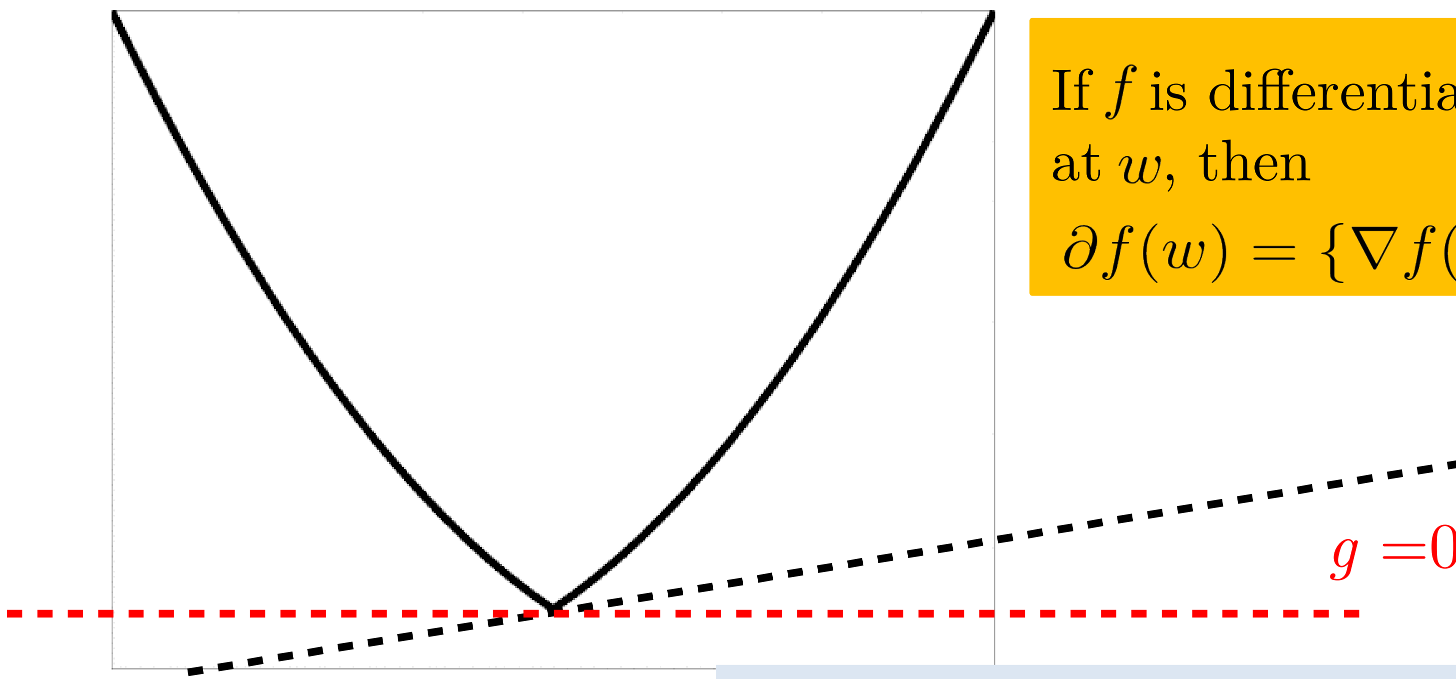
$$w^* = \arg\min_{w} f(w) \Leftrightarrow 0 \in \partial f(w^*)$$

Convexity without smoothness: Subgradient

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex

$$\partial f(w) := \{g \in \mathbb{R}^n \, : \, f(y) \geq f(w) + \langle g, y - w \rangle, \forall y \in \mathrm{dom}(f)\}$$

If f is differentiable at w, then $\partial f(w) = \{\nabla f(w)\}$

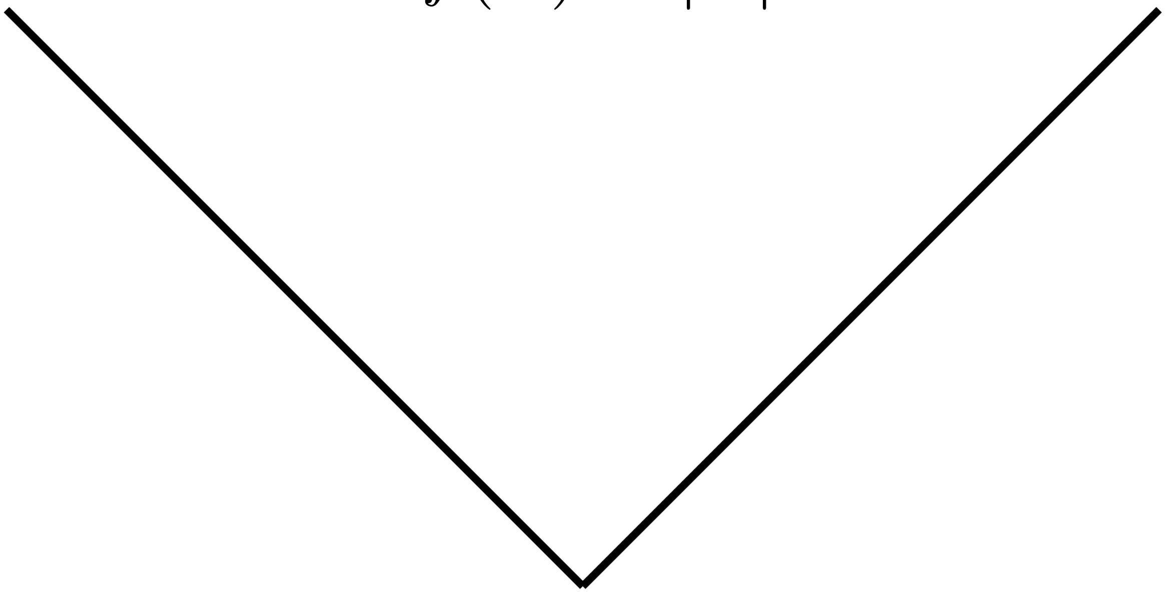


$f(w) + \langle g, y - w \rangle$

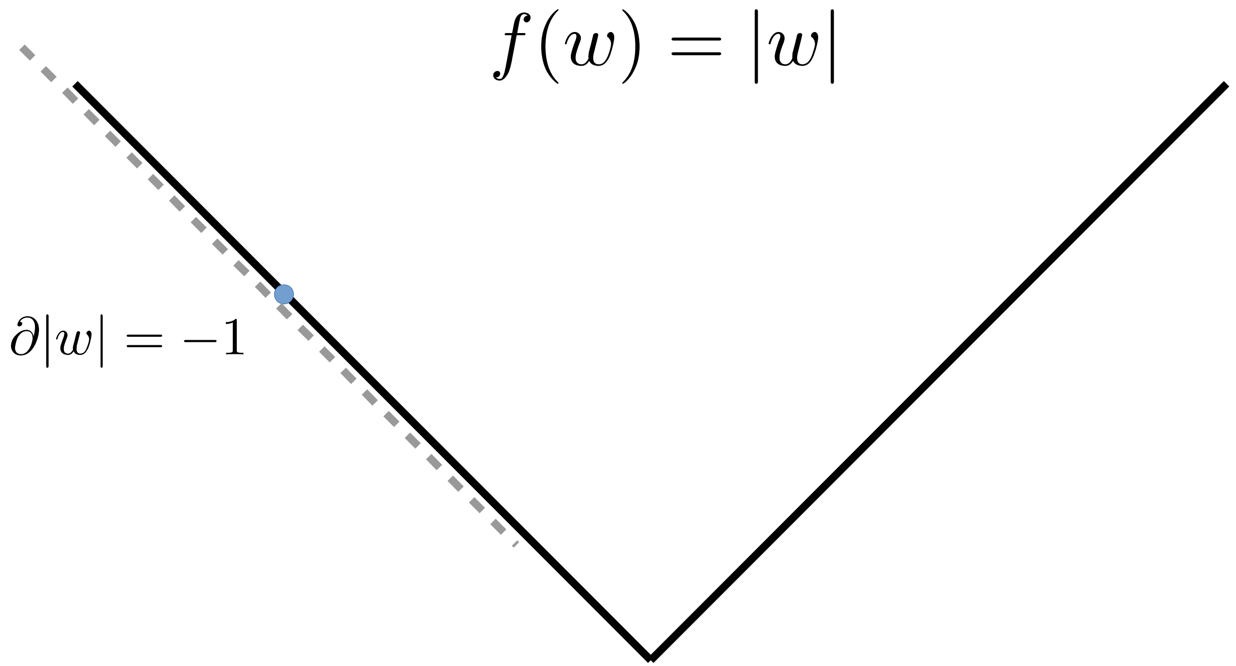
$$w^* = \arg\min_w f(w) \Leftrightarrow 0 \in \partial f(w^*)$$

Examples: L1 norm

$$f(w) = |w|$$

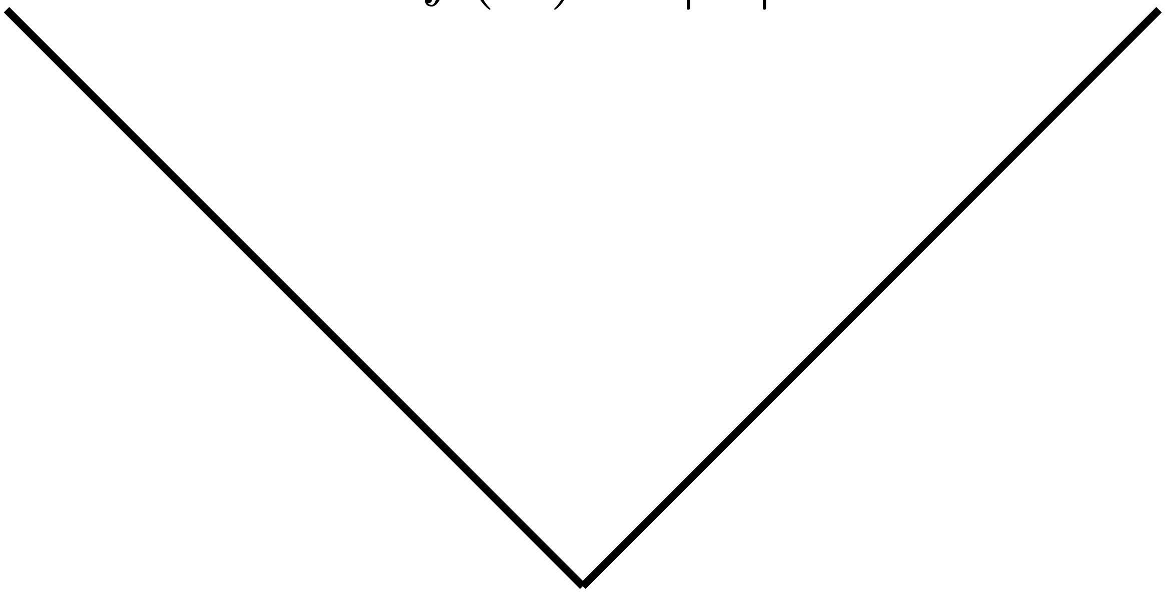


Examples: L1 norm

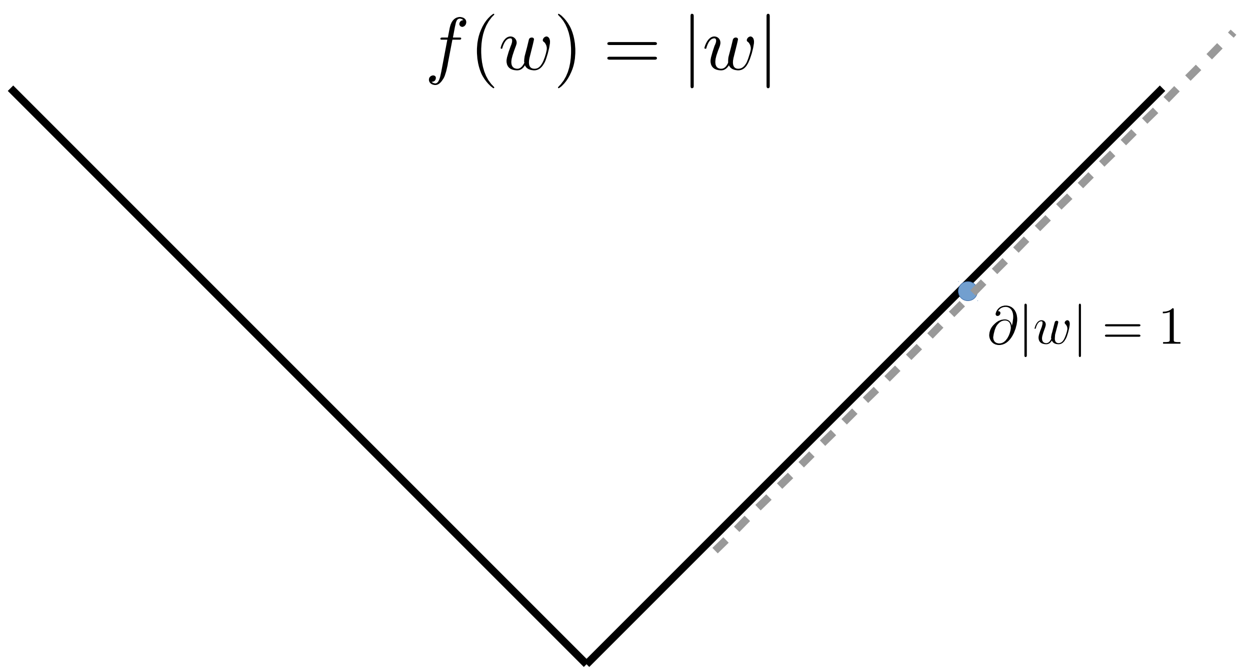


Examples: L1 norm

$$f(w) = |w|$$

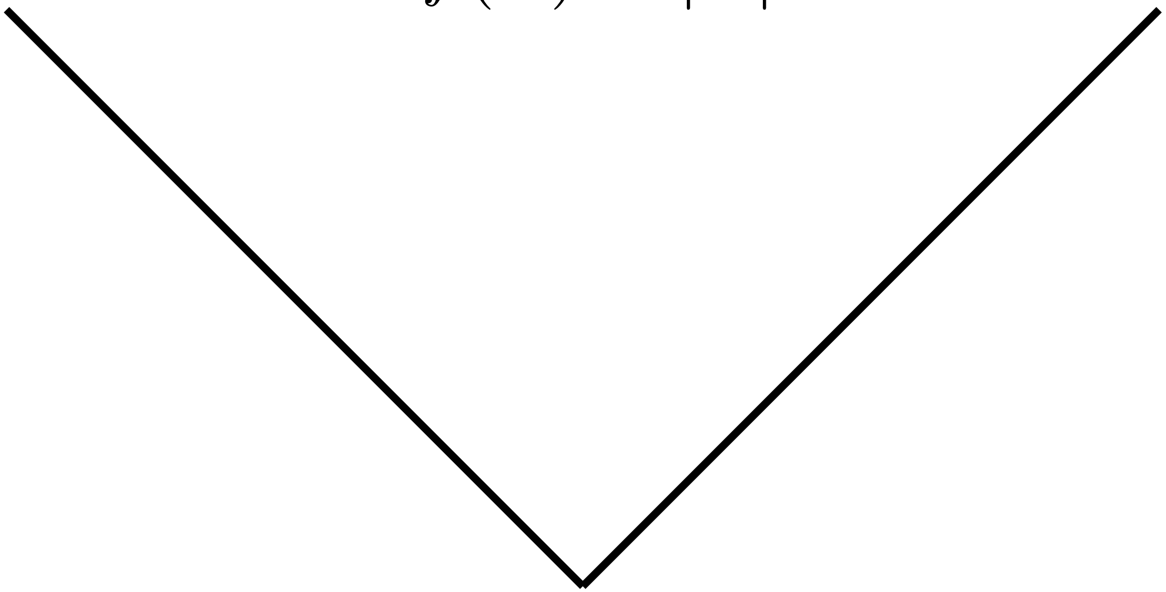


Examples: L1 norm

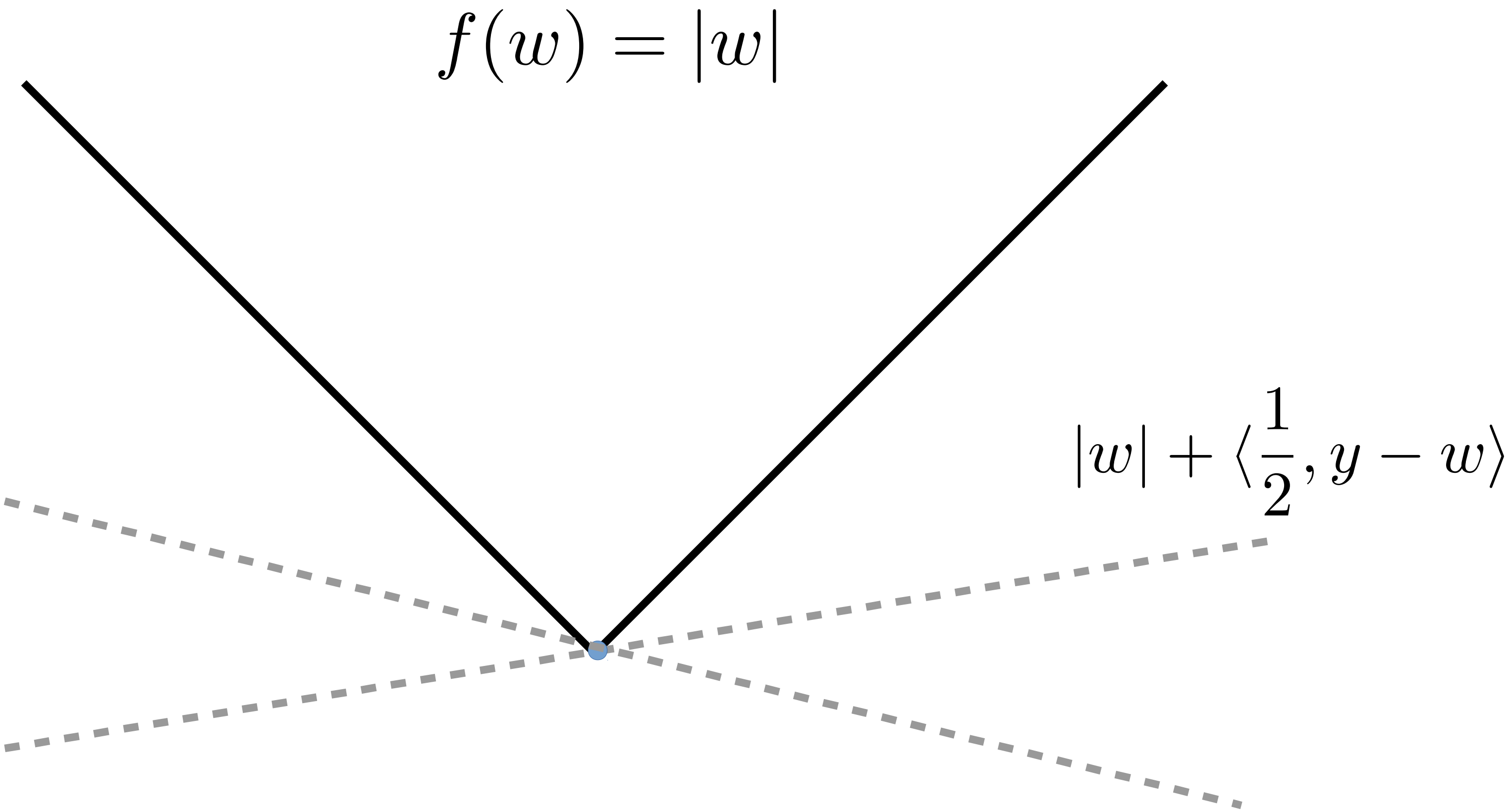


Examples: L1 norm

$$f(w) = |w|$$

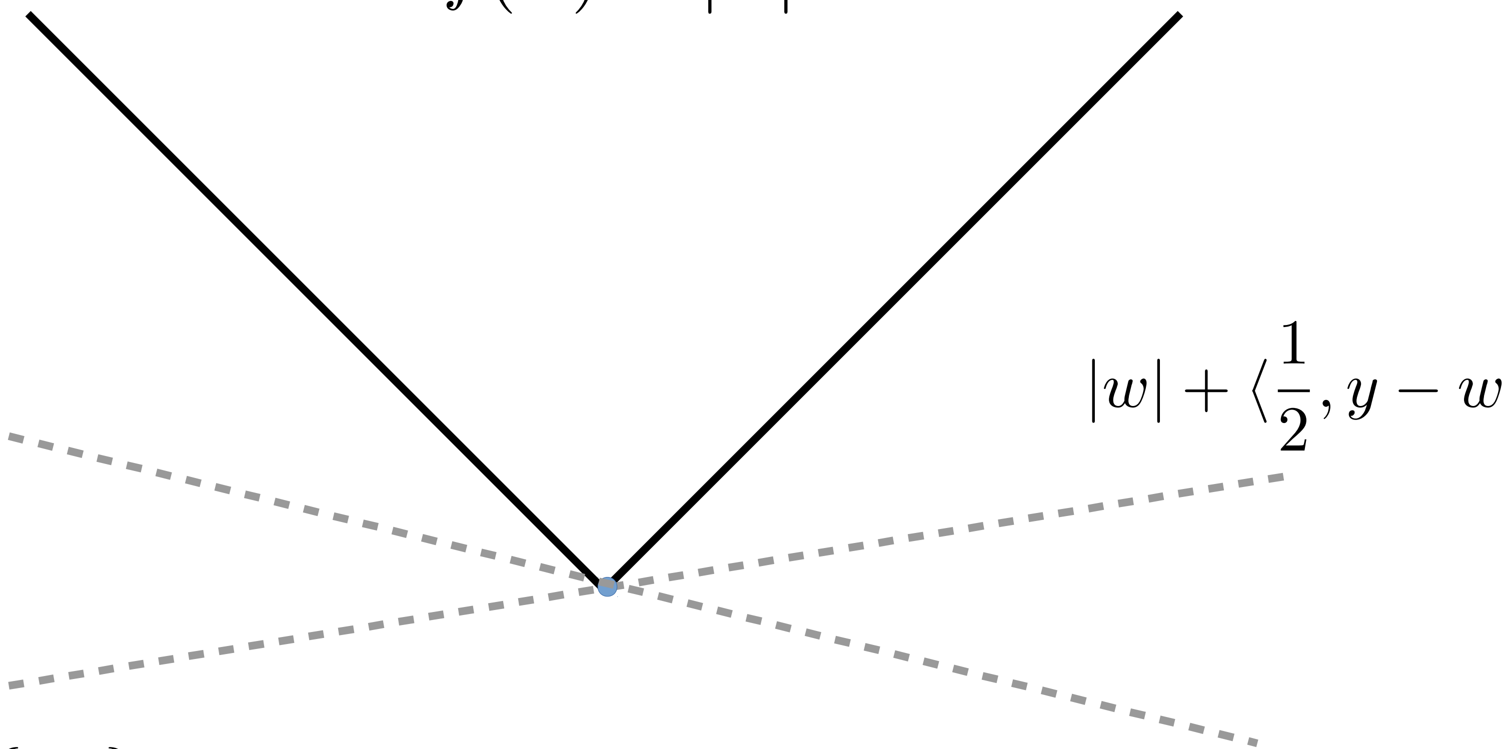


Examples: L1 norm



Examples: L1 norm

$$f(w) = |w|$$



$$\partial |w| = \begin{cases} \{-1\} & \text{if } w < 0 \\ [-1,\ 1] & \text{if } w = 0 \\ \{1\} & \text{if } w > 0 \end{cases}$$

Optimality conditions

The Training problem

$$w^* = \arg\min_{w \in \mathbf{R}^d} F(w) + \lambda R(w)$$

$F(w)$ is differentiable, L–smooth and convex

$R(w)$ is convex

Optimality conditions

The Training problem

$$w^* = \arg\min_{w \in \mathbf{R}^d} F(w) + \lambda R(w)$$

$F(w)$ is differentiable, L–smooth and convex

$R(w)$ is convex

$$0 \in \partial \left(F(w^*) + \lambda R(w^*)\right) = \nabla F(w^*) + \lambda \partial R(w^*)$$

$$\Downarrow$$

$$-\nabla F(w^*) \in \lambda \partial R(w^*)$$

Working example: Lasso

Lasso

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda||w||_1$$

Working example: Lasso

Lasso

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda ||w||_1$$

$$-\nabla F(w^*) \in \partial R(w^*) \quad \Rightarrow \quad -X^\top (Xw^* - y) \in \lambda \partial ||w^*||_1$$

$$\forall i, \left[X^\top (Xw - y)\right]_i \in \begin{cases} \{\lambda\} & \text{if } w_i < 0 \\ [-\lambda, \lambda] & \text{if } w_i = 0 \\ \{-\lambda\} & \text{if } w_i > 0 \end{cases}$$

Working example: Lasso

Lasso

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda ||w||_1$$

$$-\nabla F(w^*) \in \partial R(w^*) \quad \Rightarrow \quad -X^\top (Xw^* - y) \in \lambda \partial ||w^*||_1$$

$$\forall i, \left[X^\top (Xw - y)\right]_i \in \begin{cases} \{\lambda\} & \text{if } w_i < 0 \\ [-\lambda, \lambda] & \text{if } w_i = 0 \\ \{-\lambda\} & \text{if } w_i > 0 \end{cases}$$

Q: Show that 0 is solution if and only if $\lambda \geq \max_i |[X^\top y]_i|$

Working example: Lasso

Lasso

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda ||w||_1$$

$$-\nabla F(w^*) \in \partial R(w^*)$$

$$-X^\top (Xw^* - y) \in \lambda \partial ||w^*||_1$$

Difficult inclusion

Solve iteratively

Solving the problem by iterative minimization

Proximal method I: iteratively minimizes an upper bound

Using L–smoothness of F :

$$F(w) \leq F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2} ||w - y||^2, \quad \forall w, y \in \mathbb{R}^d$$

The w that minimizes the upper bound gives ...

Proximal method I: iteratively minimizes an upper bound

Using L–smoothness of F :

$$F(w) \leq F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2, \quad \forall w, y \in \mathbb{R}^d$$

The w that minimizes the upper bound gives gradient descent

$$w = y - \frac{1}{L}\nabla F(y)$$

Proximal method I: iteratively minimizes an upper bound

Using L–smoothness of F :

$$F(w) \leq F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2, \quad \forall w, y \in \mathbb{R}^d$$

The w that minimizes the upper bound gives gradient descent

$$w = y - \frac{1}{L}\nabla F(y)$$

But what about *R(w)*? Adding on $+\lambda R(w)$ to upper bound:

Proximal method I: iteratively minimizes an upper bound

Using L–smoothness of F :

$$F(w) \leq F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2, \quad \forall w, y \in \mathbb{R}^d$$

The w that minimizes the upper bound gives gradient descent

$$w = y - \frac{1}{L}\nabla F(y)$$

But what about $R(w)$? Adding on $+\lambda R(w)$ to upper bound:

$$\boxed{F(w) + \lambda R(w) \leq F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2 + \lambda R(w)}$$

Proximal method I: iteratively minimizes an upper bound

Using L–smoothness of F :

$$F(w) \leq F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2, \quad \forall w, y \in \mathbb{R}^d$$

The w that minimizes the upper bound gives gradient descent

$$w = y - \frac{1}{L}\nabla F(y)$$

But what about *R(w)*? Adding on $+\lambda R(w)$ to upper bound:

$$\boxed{F(w) + \lambda R(w) \leq F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2 + \lambda R(w)}$$

Can we minimize the right-hand side?

Proximal method I: iteratively minimizes an upper bound

Minimizing the right-hand side of

$$F(w) + \lambda R(w) \leq F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2 + \lambda R(w)$$

Proximal method I: iteratively minimizes an upper bound

Minimizing the right-hand side of

$$F(w) + \lambda R(w) \leq F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2 + \lambda R(w)$$

Factorization ! Let $w' = y - \dfrac{1}{L}\nabla F(y)$

Proximal method I: iteratively minimizes an upper bound

Minimizing the right-hand side of

$$F(w) + \lambda R(w) \leq F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2 + \lambda R(w)$$

Factorization ! Let $w' = y - \frac{1}{L}\nabla F(y)$

$$F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2 = \frac{L}{2}||w - w'||^2 + \text{cst}$$

$$F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2 + \lambda R(w) = \frac{L}{2}||w - w'||^2 + \lambda R(w) + \text{cst}$$

Optimality:

$$w \in \arg\min_{w} \frac{1}{2}\|w - w'\|^2 + \frac{\lambda}{L} R(w)$$

Proximal method I: iteratively minimizes an upper bound

Minimizing the right-hand side of

$$F(w) + \lambda R(w) \leq F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2 + \lambda R(w)$$

Factorization ! Let $w' = y - \frac{1}{L}\nabla F(y)$

$$F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2 = \frac{L}{2}||w - w'||^2 + \mathrm{cst}$$

$$F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2 + \lambda R(w) = \frac{L}{2}||w - w'||^2 + \lambda R(w) + \mathrm{cst}$$

Optimality:

$$w = \mathrm{prox}_{\frac{\lambda}{L} R}(w')$$

Proximal operator

Proximal Operator: Inclusion definition

Let $f(x)$ be a convex function. The proximal operator is

$$\operatorname{prox}_f(v) := \arg\min_w \frac{1}{2}||w-v||_2^2 + f(w)$$

EXE: Is this Proximal operator well defined? Is it even a function?

Proximal Operator: Inclusion definition

Let $f(x)$ be a convex function. The proximal operator is

$$\mathrm{prox}_f(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + f(w)$$

Let $w_v = \mathrm{prox}_f(v)$. Using optimality conditions

$$0 \in \partial\left(\tfrac{1}{2}||w_v - v||_2^2 + f(w)\right) = w_v - v + \partial f(w_v)$$

EXE: Is this Proximal operator well defined? Is it even a function?

Proximal Operator: Inclusion definition

Let $f(x)$ be a convex function. The proximal operator is

$$\mathrm{prox}_f(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + f(w)$$

Let $w_v = \mathrm{prox}_f(v)$. Using optimality conditions

$$0 \in \partial\left(\tfrac{1}{2}||w_v - v||_2^2 + f(w)\right) = w_v - v + \partial f(w_v)$$

Rearranging

$$\mathrm{prox}_f(v) = w_v \in v - \partial f(w_v)$$

EXE: Is this Proximal operator well defined? Is it even a function?

Proximal Operator: fixed point

Let $f(x)$ be a convex function. The proximal operator is

$$\mathrm{prox}_f(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + f(w)$$

EXE: **Show that** $w^* \in \arg\min f(w)$ **if and only if** $\mathrm{prox}_f(w^*) = w^*$

Gradient Descent using proximal map

$$\mathrm{prox}_f(y) := \arg\min_w \frac{1}{2}||w-y||_2^2 + f(w)$$

EXE : Let

$$R(w) \;=\; f(y) + \langle \nabla f(y), w-y\rangle$$

Show that

$$\mathrm{prox}_{\gamma R}(y) = y - \gamma \nabla f(y)$$

A gradient step is also a proximal step

Proximal Operator: Properties

$$\text{prox}_f(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + f(w)$$

Exe:

1) If $f(w) = \sum_{i=1}^{d} f_i(w_i)$

2) If $f(w) = I_C(w) := \begin{cases} 0 & \text{if } w \in C \\ \infty & \text{if } w \notin C \end{cases}$ where C closed and convex

3) If $f(w) = \langle b, w \rangle + c$

4) If $f(w) = \frac{\lambda}{2} w^\top A w + \langle b, w \rangle$ where $A \succeq 0,\ A = A^\top, \lambda \geq 0$

Proximal Operator: Properties

$$\mathrm{prox}_f(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + f(w)$$

Exe:

1) If $f(w) = \sum_{i=1}^{d} f_i(w_i)$ then $\mathrm{prox}_f(v) = (\mathrm{prox}_{f_1}(v_1), \ldots, \mathrm{prox}_{f_d}(v_d))$

2) If $f(w) = I_C(w) := \begin{cases} 0 & \text{if } w \in C \\ \infty & \text{if } w \notin C \end{cases}$ where C closed and convex

then $\mathrm{prox}_f(v) = \mathrm{proj}_C(v)$

3) If $f(w) = \langle b, w \rangle + c$ then $\mathrm{prox}_f(v) = v - b$

4) If $f(w) = \frac{\lambda}{2} w^\top A w + \langle b, w \rangle$ where $A \succeq 0,\ A = A^\top, \lambda \geq 0$ then

$$\mathrm{prox}_f(v) = (I + \lambda A)^{-1}(v - b)$$

Proximal Operator: Soft thresholding

$$\text{prox}_{\lambda||w||_1}(v) := \arg\min_{w} \frac{1}{2}||w - v||_2^2 + \lambda||w||_1$$

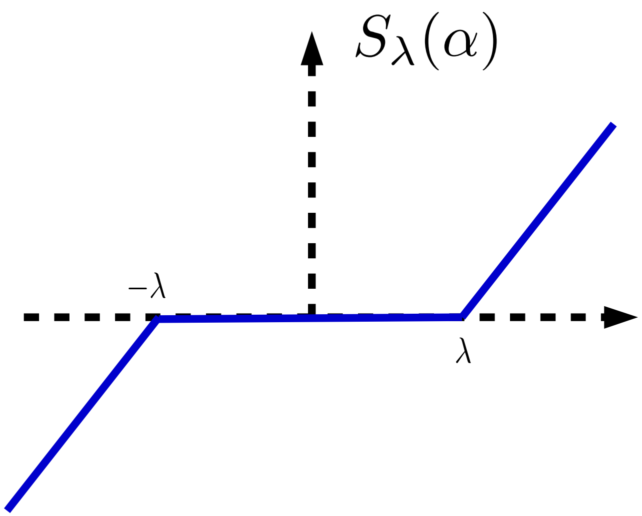
Exe:

1) Let $\alpha \in \mathbf{R}$. If $\alpha^* = \arg\min_{\alpha} \frac{1}{2}(\alpha - v)^2 + \lambda|\alpha|$ then

$$\alpha^* \in v - \lambda\partial|\alpha^*| \qquad (I)$$

2) If $\lambda < v$ show (I) gives $\alpha^* = v - \lambda$
3) If $v < -\lambda$ show (I) gives $\alpha^* = v + \lambda$
4) Show that

$$\text{prox}_{\lambda|\alpha|}(v) = \begin{cases} v - \lambda & \text{if } \lambda < v \\ 0 & \text{if } -\lambda \leq v \leq \lambda \\ v + \lambda & \text{if } v < -\lambda. \end{cases}$$



Proximal Operator: Soft thresholding

$$\mathrm{prox}_{\lambda||w||_1}(v) := \arg\min_{w} \frac{1}{2}||w-v||_2^2 + \lambda||w||_1$$

Exe:

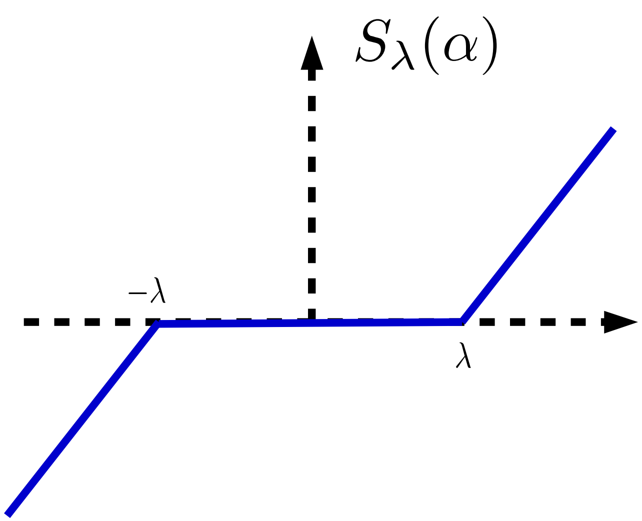
1) Let $\alpha \in \mathbf{R}$. If $\alpha^* = \arg\min_{\alpha} \frac{1}{2}(\alpha - v)^2 + \lambda|\alpha|$ then

$$\alpha^* \in v - \lambda\partial|\alpha^*| \qquad (I)$$

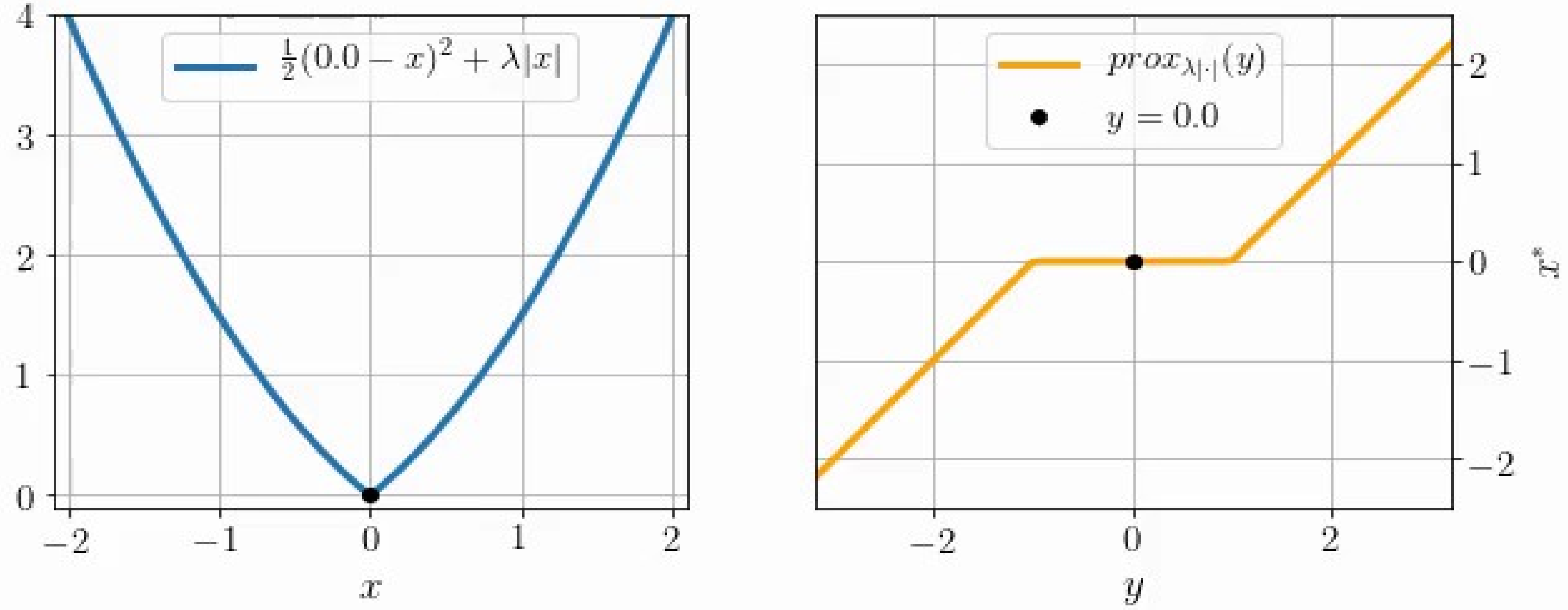
2) If $\lambda < v$ show (I) gives $\alpha^* = v - \lambda$
3) If $v < -\lambda$ show (I) gives $\alpha^* = v + \lambda$
4) Show that

Induces **sparsity**

$$\mathrm{prox}_{\lambda|\alpha|}(v) = \begin{cases} v - \lambda & \text{if } \lambda < v \\ 0 & \text{if } -\lambda \leq v \leq \lambda \\ v + \lambda & \text{if } v < -\lambda. \end{cases}$$



$$prox_{\lambda|\cdot|}(y) = \arg\min_x \tfrac{1}{2}(y-x)^2 + \lambda|x|$$



Proximal Operator: Singular value thresholding

$$\mathrm{ST}_\lambda(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + \lambda||w||_1$$

Similarly, the prox operator of the nuclear norm for matrices:

$$U\mathrm{ST}_\lambda(\Sigma)V^\top := \arg\min_{W \in \mathbf{R}^{d\times d}} \frac{1}{2}||W - A||_F^2 + \lambda||W||_*$$

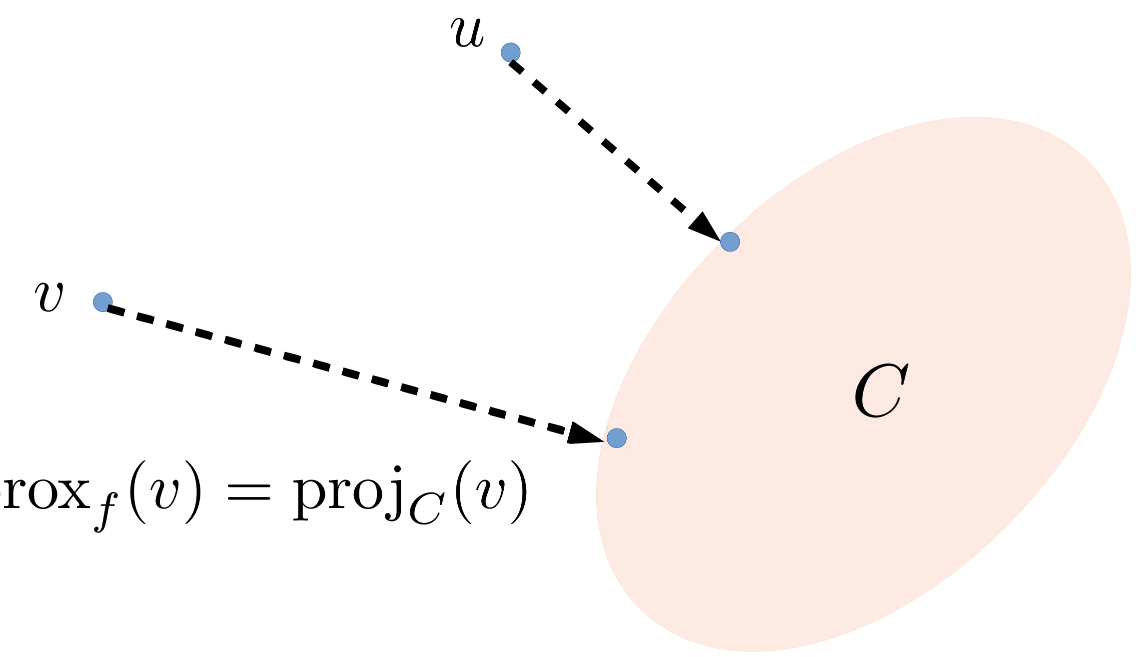
where $A = U\Sigma V^\top$ is a SVD decomposition,

and $\|W\|_* = \mathrm{trace}(\sqrt{W^\top W}) = \sum_i \sigma_i(W)$ is the nuclear norm

EXE: This is a HARD exercise ! Use lemma:
For W, W' orthogonal, D, D' diagonal with >0 entries, $\langle WDW', D'\rangle \leq \langle D, D'\rangle$

Proximal Operator: Non-expansiveness

$f(w) = I_C(w)$ $\qquad$ $||\text{proj}_C(v) - \text{proj}_C(u)||_2 \ \leq \ ||u - v||_2$



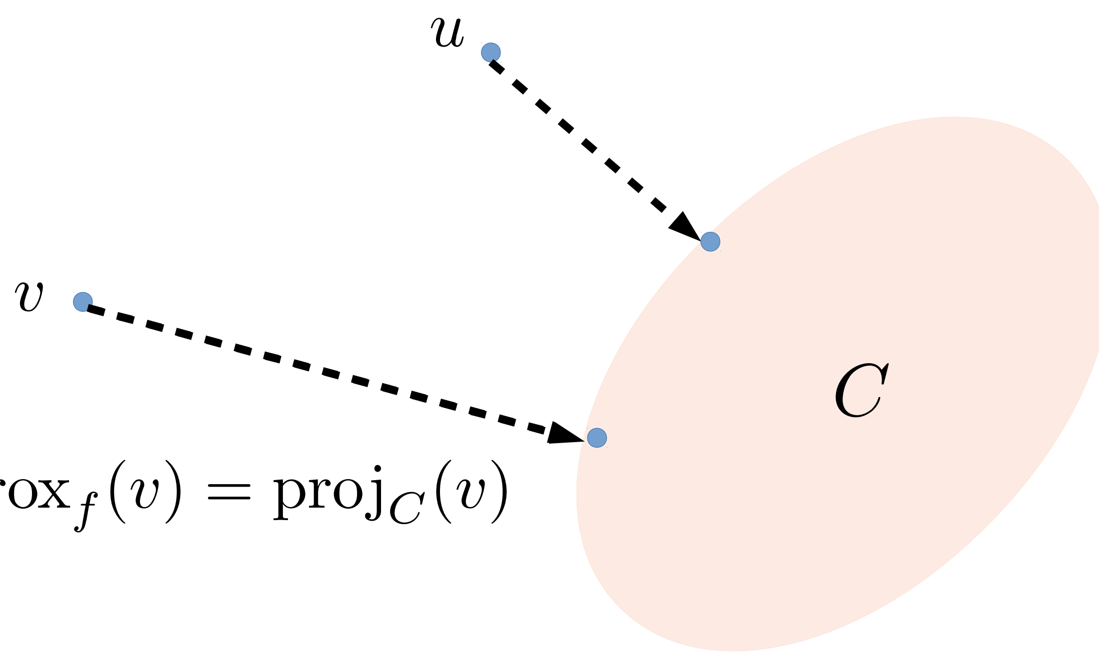
Proximal Operators are nonexpansive

$$||\text{prox}_f(u) - \text{prox}_f(v)||_2 \ \leq \ ||u - v||_2$$

Proximal Operator: Non-expansiveness

$f(w) = I_C(w)$

$$||\text{proj}_C(v) - \text{proj}_C(u)||_2 \;\leq\; ||u - v||_2$$



$$\text{prox}_f(v) = \text{proj}_C(v)$$

This will be used to show that proximal steps do not hurt the convergence of gradient descent

Proximal Operators are nonexpansive

$$||\text{prox}_f(v) - \text{prox}_f(u)||_2 \;\leq\; ||u - v||_2$$

Proximal Operator: Non-expansiveness

Proximal Operators are nonexpansive

$$||\mathrm{prox}_f(v) - \mathrm{prox}_f(u)||_2 \;\leq\; ||u - v||_2$$

Proof: Let $p_v = \mathrm{prox}_f(v) \quad \text{and} \quad p_u = \mathrm{prox}_f(u)$

Using subgradient characterization

Proximal Operator: Non-expansiveness

Proximal Operators are nonexpansive

$$||\text{prox}_f(v) - \text{prox}_f(u)||_2 \leq ||u - v||_2$$

Proof: Let $p_v = \text{prox}_f(v)$ and $p_u = \text{prox}_f(u)$

Using subgradient characterization

$$\text{prox}_f(v) = p_v \in v - \partial f(p_v) \Rightarrow v - p_v \in \partial f(p_v)$$

Proximal Operator: Non-expansiveness

Proximal Operators are nonexpansive

$$||\text{prox}_f(v) - \text{prox}_f(u)||_2 \;\leq\; ||u - v||_2$$

Proof: Let $p_v = \text{prox}_f(v)$ and $p_u = \text{prox}_f(u)$

Using subgradient characterization

$$\text{prox}_f(v) = p_v \in v - \partial f(p_v) \;\Rightarrow\; v - p_v \in \partial f(p_v)$$
$$\text{prox}_f(u) = p_u \in u - \partial f(p_u) \;\Rightarrow\; u - p_u \in \partial f(p_u)$$

Proximal Operator: Non-expansiveness

Proximal Operators are nonexpansive

$$||\text{prox}_f(v) - \text{prox}_f(u)||_2 \leq ||u - v||_2$$

Proof: Let $p_v = \text{prox}_f(v)$ and $p_u = \text{prox}_f(u)$

Using subgradient characterization

$$\text{prox}_f(v) = p_v \in v - \partial f(p_v) \Rightarrow v - p_v \in \partial f(p_v)$$
$$\text{prox}_f(u) = p_u \in u - \partial f(p_u) \Rightarrow u - p_u \in \partial f(p_u)$$

Using convexity and subgradient

Proximal Operator: Non-expansiveness

Proximal Operators are nonexpansive

$$||\mathrm{prox}_f(v) - \mathrm{prox}_f(u)||_2 \le ||u - v||_2$$

Proof: Let $p_v = \mathrm{prox}_f(v)$ and $p_u = \mathrm{prox}_f(u)$

Using subgradient characterization

$$\mathrm{prox}_f(v) = p_v \in v - \partial f(p_v) \Rightarrow v - p_v \in \partial f(p_v)$$
$$\mathrm{prox}_f(u) = p_u \in u - \partial f(p_u) \Rightarrow u - p_u \in \partial f(p_u)$$

Using convexity and subgradient

$$f(p_u) \ge f(p_v) + \langle \underbrace{v - p_v}_{\in \partial f(p_v)}, p_u - p_v \rangle$$

Proximal Operator: Non-expansiveness

Proximal Operators are nonexpansive

$$||\text{prox}_f(v) - \text{prox}_f(u)||_2 \leq ||u - v||_2$$

Proof: Let $p_v = \text{prox}_f(v)$ and $p_u = \text{prox}_f(u)$

Using subgradient characterization

$$\text{prox}_f(v) = p_v \in v - \partial f(p_v) \Rightarrow v - p_v \in \partial f(p_v)$$
$$\text{prox}_f(u) = p_u \in u - \partial f(p_u) \Rightarrow u - p_u \in \partial f(p_u)$$

Using convexity and subgradient

$$f(p_u) \geq f(p_v) + \langle \underbrace{v - p_v}_{\in \partial f(p_v)}, p_u - p_v \rangle$$

$$f(p_v) \geq f(p_u) + \langle u - p_u, p_v - p_u \rangle$$

Proximal Operator: Non-expansiveness

Proximal Operators are nonexpansive

$$||\text{prox}_f(v) - \text{prox}_f(u)||_2 \;\leq\; ||u - v||_2$$

Proof: Let $p_v = \text{prox}_f(v)$ and $p_u = \text{prox}_f(u)$

Using subgradient characterization

$$\text{prox}_f(v) = p_v \in v - \partial f(p_v) \;\Rightarrow\; v - p_v \in \partial f(p_v)$$

$$\text{prox}_f(u) = p_u \in u - \partial f(p_u) \;\Rightarrow\; u - p_u \in \partial f(p_u)$$

Using convexity and subgradient

$$f(p_u) \geq f(p_v) + \langle \underbrace{v - p_v}_{\in \partial f(p_v)}, p_u - p_v \rangle$$

$$+$$

$$f(p_v) \geq f(p_u) + \langle u - p_u, p_v - p_u \rangle$$

$$0 \leq \langle v - u - (p_v - p_u), p_u - p_v \rangle$$

Proximal Operator: Non-expansiveness

Proximal Operators are nonexpansive

$$||\mathrm{prox}_f(v) - \mathrm{prox}_f(u)||_2 \ \leq \ ||u - v||_2$$

Proof: Let $p_v = \mathrm{prox}_f(v)$ and $p_u = \mathrm{prox}_f(u)$

Using subgradient characterization

$$\mathrm{prox}_f(v) = p_v \in v - \partial f(p_v) \ \Rightarrow \ v - p_v \in \partial f(p_v)$$
$$\mathrm{prox}_f(u) = p_u \in u - \partial f(p_u) \ \Rightarrow \ u - p_u \in \partial f(p_u)$$

Using convexity and subgradient

$$f(p_u) \geq f(p_v) + \langle \underbrace{v - p_v}_{\in \partial f(p_v)}, p_u - p_v \rangle$$

$$+$$

$$f(p_v) \geq f(p_u) + \langle u - p_u, p_v - p_u \rangle$$

$$0 \leq \langle v - u - (p_v - p_u), p_u - p_v \rangle$$

$$\Updownarrow$$

$$\|p_u - p_v\|^2 \leq \langle v - u, p_u - p_v \rangle$$
$$\leq \|v - u\| \|p_u - p_v\|$$

Proximal Operator: Non-expansiveness

Proximal Operators are nonexpansive

$$||\text{prox}_f(v) - \text{prox}_f(u)||_2 \;\leq\; ||u - v||_2$$

Proof: Let $p_v = \text{prox}_f(v)$ and $p_u = \text{prox}_f(u)$

Using subgradient characterization

$$\text{prox}_f(v) = p_v \in v - \partial f(p_v) \;\Rightarrow\; v - p_v \in \partial f(p_v)$$
$$\text{prox}_f(u) = p_u \in u - \partial f(p_u) \;\Rightarrow\; u - p_u \in \partial f(p_u)$$

Using convexity and subgradient

$$f(p_u) \geq f(p_v) + \langle \underbrace{v - p_v}_{\in \partial f(p_v)}, p_u - p_v\rangle$$
$$+$$
$$f(p_v) \geq f(p_u) + \langle u - p_u, p_v - p_u\rangle$$

$$0 \leq \langle v - u - (p_v - p_u), p_u - p_v\rangle$$
$$\Updownarrow$$
$$\|p_u - p_v\|^2 \leq \langle v - u, p_u - p_v\rangle$$
$$\leq \|v - u\|\|p_u - p_v\|$$

Now divide both sides by $\|p_u - p_v\|$ ■

Proximal method : iteratively minimizes an upper bound

Set $y = w^t$ and minimize the right-hand side in w

$$F(w) + \lambda R(w) \le F(y) + \langle \nabla F(y), w - y \rangle + \frac{L}{2}||w - y||^2 + \lambda R(w)$$

$$\arg\min_w F(w^t) + \langle \nabla F(w^t), w - w^t \rangle + \frac{L}{2}||w - w^t||^2 + \lambda R(w)$$

$$=: \operatorname{prox}_{\frac{\lambda}{L}R}(w^t - \frac{1}{L}\nabla F(w^t)))$$

This suggests an iterative method

$$w^{t+1} = \operatorname{prox}_{\frac{\lambda}{L}R}(w^t - \frac{1}{L}\nabla F(w^t)))$$

Proximal method: A fixed point viewpoint

The Training problem

$$w^* \in \arg\min_w F(w) + \lambda R(w)$$

$$-\nabla F(w^*) \in \lambda \partial R(w^*)$$

Proximal method: A fixed point viewpoint

The Training problem

$$w^* \in \arg\min_{w} F(w) + \lambda R(w)$$

$$-\nabla F(w^*) \in \lambda \partial R(w^*) \iff w^* + \gamma \nabla F(w^*) \in w^* - (\lambda\gamma)\partial R(w^*)$$

Proximal method: A fixed point viewpoint

The Training problem

$$w^* \in \arg\min_w F(w) + \lambda R(w)$$

$$-\nabla F(w^*) \in \lambda \partial R(w^*) \iff w^* + \gamma \nabla F(w^*) \in w^* - (\lambda\gamma)\partial R(w^*)$$

$$\iff w^* \in (w^* - \gamma \nabla F(w^*)) - (\lambda\gamma)\partial R(w^*)$$

Proximal method: A fixed point viewpoint

The Training problem

$$w^* \in \arg\min_w F(w) + \lambda R(w)$$

$$-\nabla F(w^*) \in \lambda \partial R(w^*) \iff w^* + \gamma \nabla F(w^*) \in w^* - (\lambda\gamma)\partial R(w^*)$$

$$\iff w^* \in (w^* - \gamma \nabla F(w^*)) - (\lambda\gamma)\partial R(w^*)$$

$$\iff w^* = \mathrm{prox}_{\lambda\gamma R}(w^* - \gamma \nabla F(w^*))$$

$$\mathrm{prox}_f(v) = w_v \in v - \partial f(w_v)$$

Proximal method: A fixed point viewpoint

The Training problem

$$w^* \in \arg\min_w F(w) + \lambda R(w)$$

$$-\nabla F(w^*) \in \lambda \partial R(w^*) \iff w^* + \gamma \nabla F(w^*) \in w^* - (\lambda\gamma)\partial R(w^*)$$

$$\iff w^* \in (w^* - \gamma \nabla F(w^*)) - (\lambda\gamma)\partial R(w^*)$$

$\mathrm{prox}_f(v) = w_v \in v - \partial f(w_v)$

$$\iff w^* = \mathrm{prox}_{\lambda\gamma R}\left(w^* - \gamma \nabla F(w^*)\right)$$

Optimal is a fixed point

$$\Rightarrow w^{k+1} = \mathrm{prox}_{\lambda\gamma R}\left(w^k - \gamma \nabla F(w^k)\right)$$

Upper bound viewpoint

$$\Rightarrow w^{t+1} = \mathrm{prox}_{\frac{\lambda}{L} R}(w^t - \frac{1}{L} \nabla F(w^t)))$$

The Proximal Gradient Method

Solving the *training problem*:

$$\min_{w} F(w) + \lambda R(w)$$

$F(w)$ is differentiable, L–smooth and convex

$R(w)$ is convex and $\mathrm{prox}_{\mathrm{R}}$ is available

Proximal Gradient Descent

Set $w^1 = 0$.
for $t = 1, 2, 3, \ldots, T$

$$w^{t+1} = \mathrm{prox}_{\lambda R/L}\left(w^t - \frac{1}{L}\nabla F(w^t)\right)$$

Output w^{T+1}

Example of prox gradient: Iterative Soft Thresholding Algorithm (ISTA)

Lasso

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda||w||_1$$

ISTA:

$$w^{t+1} = \operatorname{prox}_{\lambda||\cdot||_1/L}\left(w^t - \frac{1}{L}X^\top(Xw^t - y)\right)$$

$L = \sigma_{\max}(X)^2$

$$= \mathrm{ST}_{\frac{\lambda}{L}}\left(w^t - \frac{1}{\sigma_{\max}(X)^2}X^\top(Xw^t - y)\right)$$

Amir Beck and Marc Teboulle (2009), SIAM J. IMAGING SCIENCES,
A Fast Iterative Shrinkage-Thresholding Algorithm
for Linear Inverse Problems.

Example of prox gradient: Iterative Soft Thresholding Algorithm (ISTA)

Lasso

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda ||w||_1$$

ISTA:

$$w^{t+1} = \mathrm{prox}_{\lambda||\cdot||_1/L}\left(w^t - \frac{1}{L}X^\top(Xw^t - y)\right)$$

$L = \sigma_{\max}(X)^2$

$$= \mathrm{ST}_{\frac{\lambda}{L}}\left(w^t - \frac{1}{\sigma_{\max}(X)^2}X^\top(Xw^t - y)\right)$$

Soft-thresholding: induces Sparsity

Amir Beck and Marc Teboulle (2009), SIAM J. IMAGING SCIENCES,
A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems.

Convergence of Prox-GD for convex

Theorem

Let $f(w) = F(w) + \lambda R(w)$ where

$F(w)$ is differentiable, L–smooth and μ–strongly convex

$R(w)$ is convex

Then

$$\|w^t - w^*\| \;\leq\; \left(1 - \frac{\mu}{L}\right)^t \|w^0 - w^*\|$$

where

$$w^{t+1} \;=\; \text{prox}_{\lambda R/L}\left(w^t - \frac{1}{L}\nabla F(w^t)\right)$$

Amir Beck and Marc Teboulle (2009), SIAM J. IMAGING SCIENCES,
A Fast Iterative Shrinkage-Thresholding Algorithm
for Linear Inverse Problems.

Proof sketch

$$\|w^{t+1} - w^*\|_2 \quad = \quad \|\mathrm{prox}_{\frac{\lambda}{L}R}(w^t - \frac{1}{L}\nabla F(w^t))) - w^*\|_2$$

Proof sketch

Fixed point viewpoint

$$w^* = \mathrm{prox}_{\lambda\gamma R}\left(w^* - \gamma \nabla L(w^*)\right)$$

$$\begin{aligned}
\|w^{t+1} - w^*\|_2 &= \|\mathrm{prox}_{\frac{\lambda}{L}R}(w^t - \tfrac{1}{L}\nabla F(w^t)) - w^*\|_2 \\
&= \|\mathrm{prox}_{\frac{\lambda}{L}R}(w^t - \tfrac{1}{L}\nabla F(w^t)) - \mathrm{prox}_{\frac{\lambda}{L}R}\left(w^* - \tfrac{1}{L}\nabla F(w^*)\right)\|_2
\end{aligned}$$

Proof sketch

Fixed point viewpoint
$$w^* = \text{prox}_{\lambda\gamma R}\left(w^* - \gamma \nabla L(w^*)\right)$$

$$\begin{aligned}
\|w^{t+1} - w^*\|_2 &= \|\text{prox}_{\frac{\lambda}{L}R}(w^t - \tfrac{1}{L}\nabla F(w^t)) - w^*\|_2 \\
&= \|\text{prox}_{\frac{\lambda}{L}R}(w^t - \tfrac{1}{L}\nabla F(w^t)) - \text{prox}_{\frac{\lambda}{L}R}\left(w^* - \tfrac{1}{L}\nabla F(w^*)\right)\|_2 \\
&\leq \|(w^t - \tfrac{1}{L}\nabla F(w^t)) - \left(w^* - \tfrac{1}{L}\nabla F(w^*)\right)\|_2 \\
&= \|w^t - w^* - \tfrac{1}{L}\left(\nabla F(w^t) - \nabla F(w^*)\right)\|_2
\end{aligned}$$

Non-expansive
$$||\text{prox}_f(v) - \text{prox}_f(u)||_2 \leq ||u - v||_2$$

Proof sketch

Fixed point viewpoint

$$w^* = \mathrm{prox}_{\lambda\gamma R}(w^* - \gamma \nabla L(w^*))$$

$$\begin{aligned}
\|w^{t+1} - w^*\|_2 &= \|\mathrm{prox}_{\frac{\lambda}{L}R}(w^t - \tfrac{1}{L}\nabla F(w^t)) - w^*\|_2 \\
&= \|\mathrm{prox}_{\frac{\lambda}{L}R}(w^t - \tfrac{1}{L}\nabla F(w^t)) - \mathrm{prox}_{\frac{\lambda}{L}R}\left(w^* - \tfrac{1}{L}\nabla F(w^*)\right)\|_2 \\
&\leq \|(w^t - \tfrac{1}{L}\nabla F(w^t)) - \left(w^* - \tfrac{1}{L}\nabla F(w^*)\right)\|_2 \\
&= \|w^t - w^* - \tfrac{1}{L}\left(\nabla F(w^t) - \nabla F(w^*)\right)\|_2
\end{aligned}$$

The rest similar to standard proof of conv. Of standard GD without prox term

Non-expansive

$$\|\mathrm{prox}_f(v) - \mathrm{prox}_f(u)\|_2 \leq \|u - v\|_2$$

Convergence of Prox-GD

Theorem (Beck Teboulle 2009)

Let $f(w) = F(w) + \lambda R(w)$ where

$F(w)$ is differentiable, L–smooth and convex

$R(w)$ is convex and prox friendly

Then

$$f(w^T) - f(w^*) \leq \frac{L||w^1 - w^*||_2^2}{2T} = O\left(\frac{1}{T}\right).$$

where

$$w^{t+1} \;=\; \mathrm{prox}_{\lambda R/L}\left(w^t - \frac{1}{L}\nabla F(w^t)\right)$$

Amir Beck and Marc Teboulle (2009), SIAM J. IMAGING SCIENCES,
A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems.

The FISTA Method

Solving the *training problem*: $\min_w F(w) + \lambda R(w)$

The FISTA Algorithm

Set $w^1 = 0 = z^1, \beta^1 = 1$

for $t = 1, 2, 3, \ldots, T$

$$w^{t+1} = \text{prox}_{\lambda R/L}\left(z^t - \frac{1}{L}\nabla F(z^t)\right)$$

$$\beta^{t+1} = \frac{1 + \sqrt{1 + 4(\beta^t)^2}}{2}$$

$$z^{t+1} = w^{t+1} + \frac{\beta^t - 1}{\beta^{t+1}}(w^{t+1} - w^t)$$

Output w^{T+1}

Weird, but it works

Convergence of FISTA

Theorem (Beck Teboulle 2009)

Let $f(w) = F(w) + \lambda R(w)$ where

$F(w)$ is differentiable, L–smooth and convex

$R(w)$ is convex and prox friendly

Then

$$f(w^T) - f(w^*) \leq \frac{2L||w^1 - w^*||_2^2}{(T+1)^2} = O\left(\frac{1}{T^2}\right).$$

Where w^t are given by the FISTA algorithm

Amir Beck and Marc Teboulle (2009), SIAM J. IMAGING SCIENCES,
A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems.

More on the Lasso

L1 versus L2 regularization

Ridge regression

$$\min_{w \in \mathbf{R}^d} \tfrac{1}{2}||Xw - y||_2^2 + \frac{\lambda}{2}||w||_2$$

Lasso

$$\min_{w \in \mathbf{R}^d} \tfrac{1}{2}||Xw - y||_2^2 + \lambda||w||_1$$

L1 versus L2 regularization

Diabetes dataset

10 features (age, sex, bmi, cholesterol, ...), 442 samples. Predict disease progression.

L1 versus L2 regularization

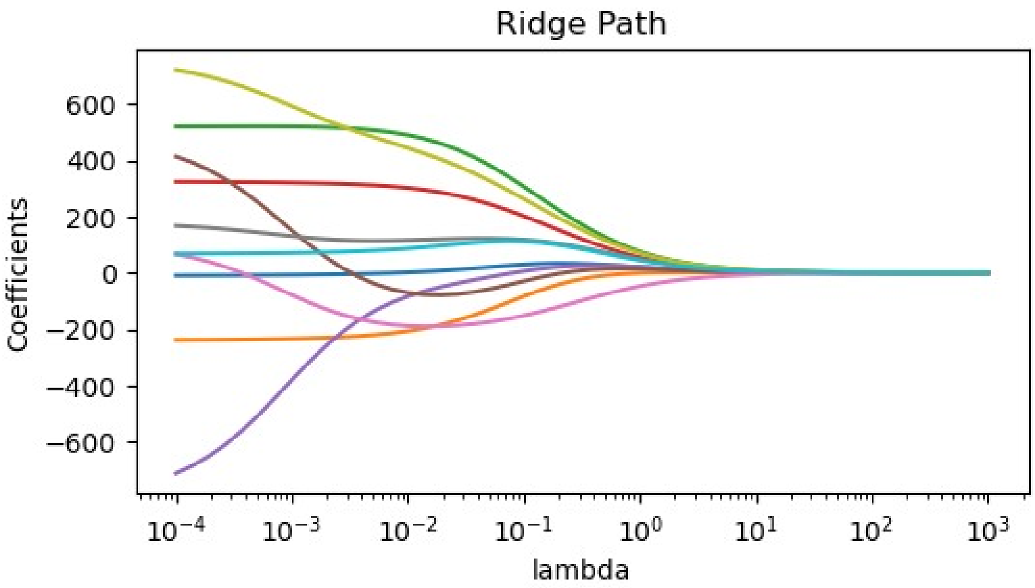
Diabetes dataset

10 features (age, sex, bmi, cholesterol, ...), 442 samples. Predict disease progression.

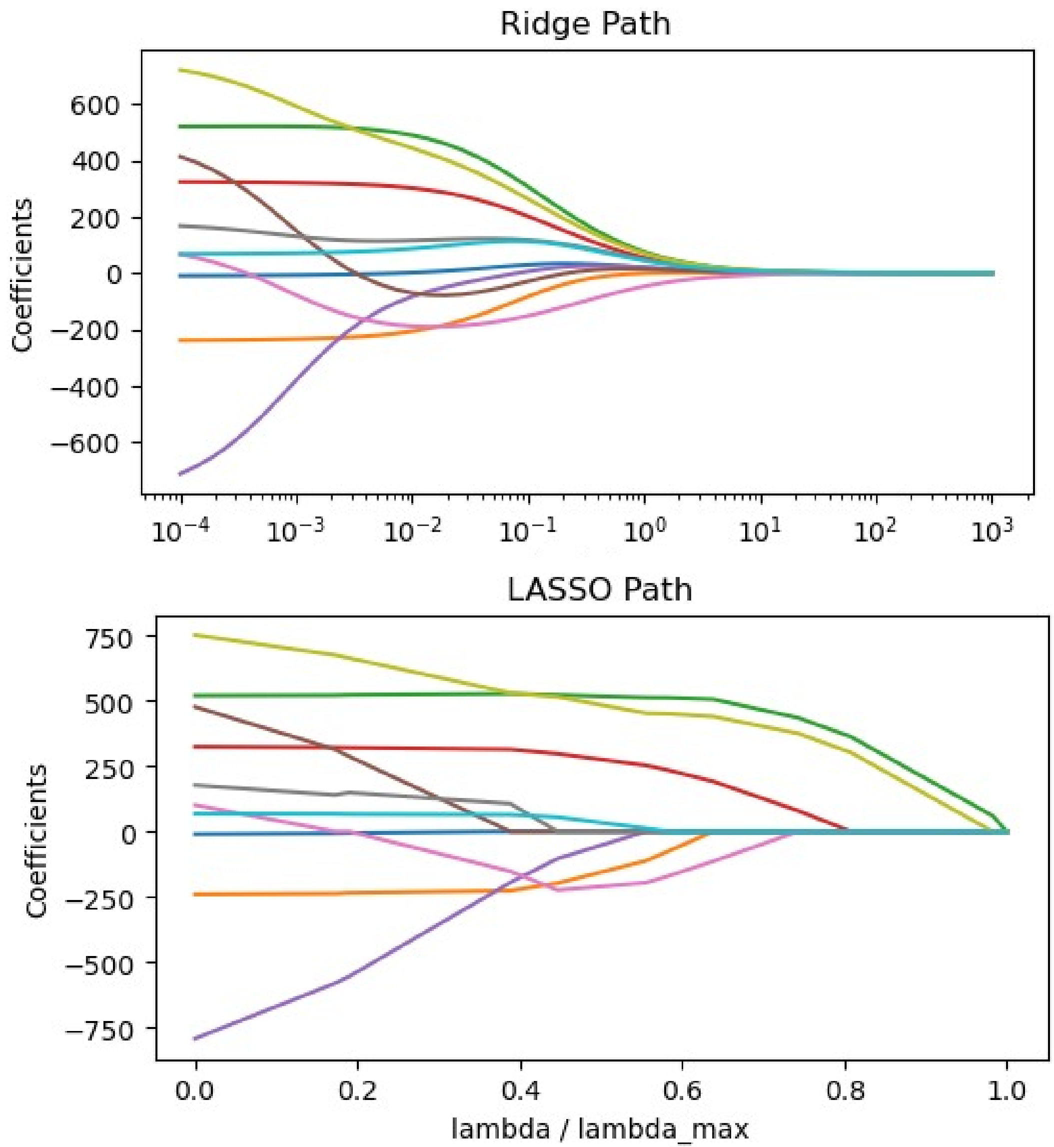
Path :

For both methods, plot the predicted coefficients as regularization changes

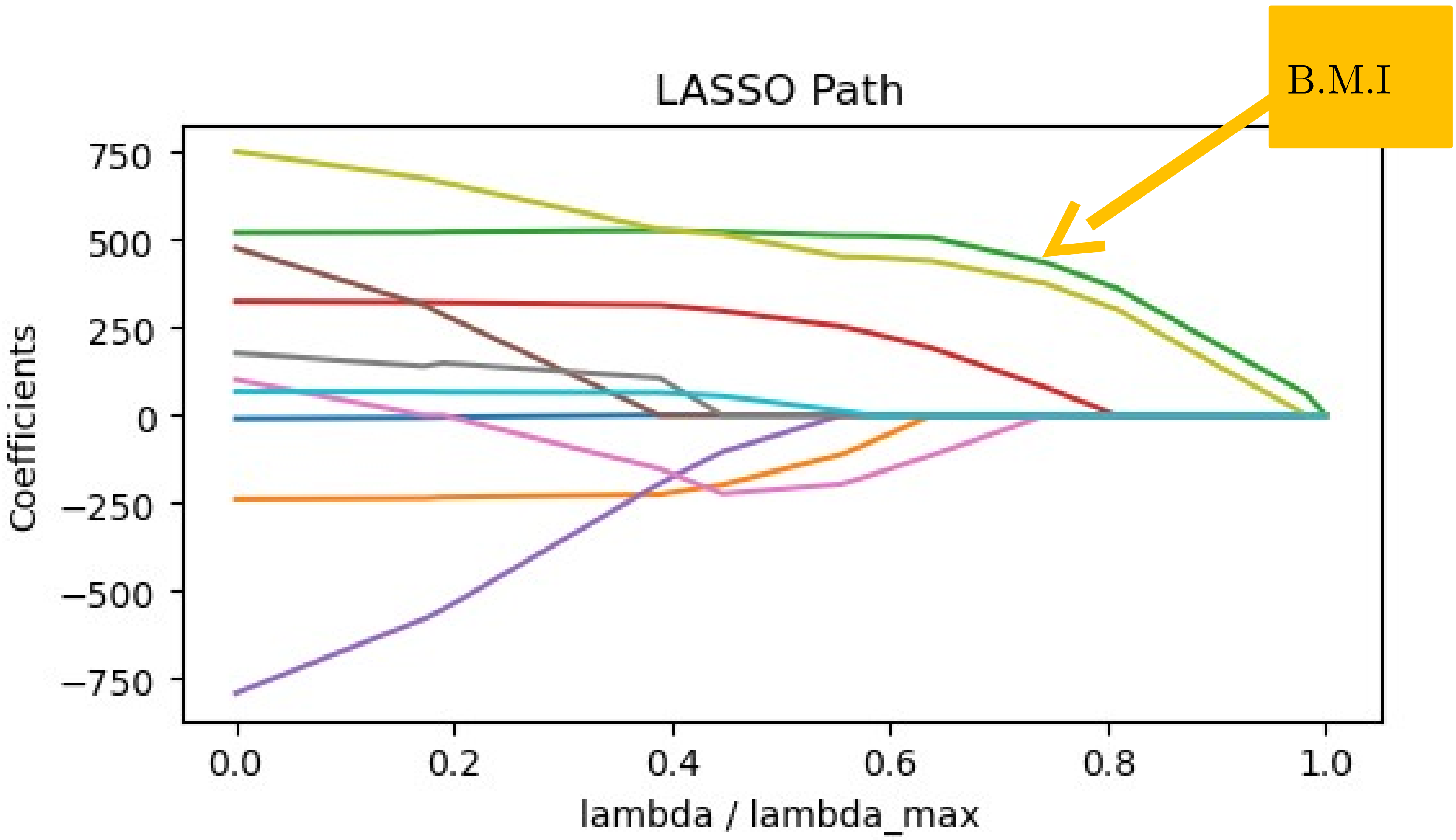
L1 versus L2 regularization



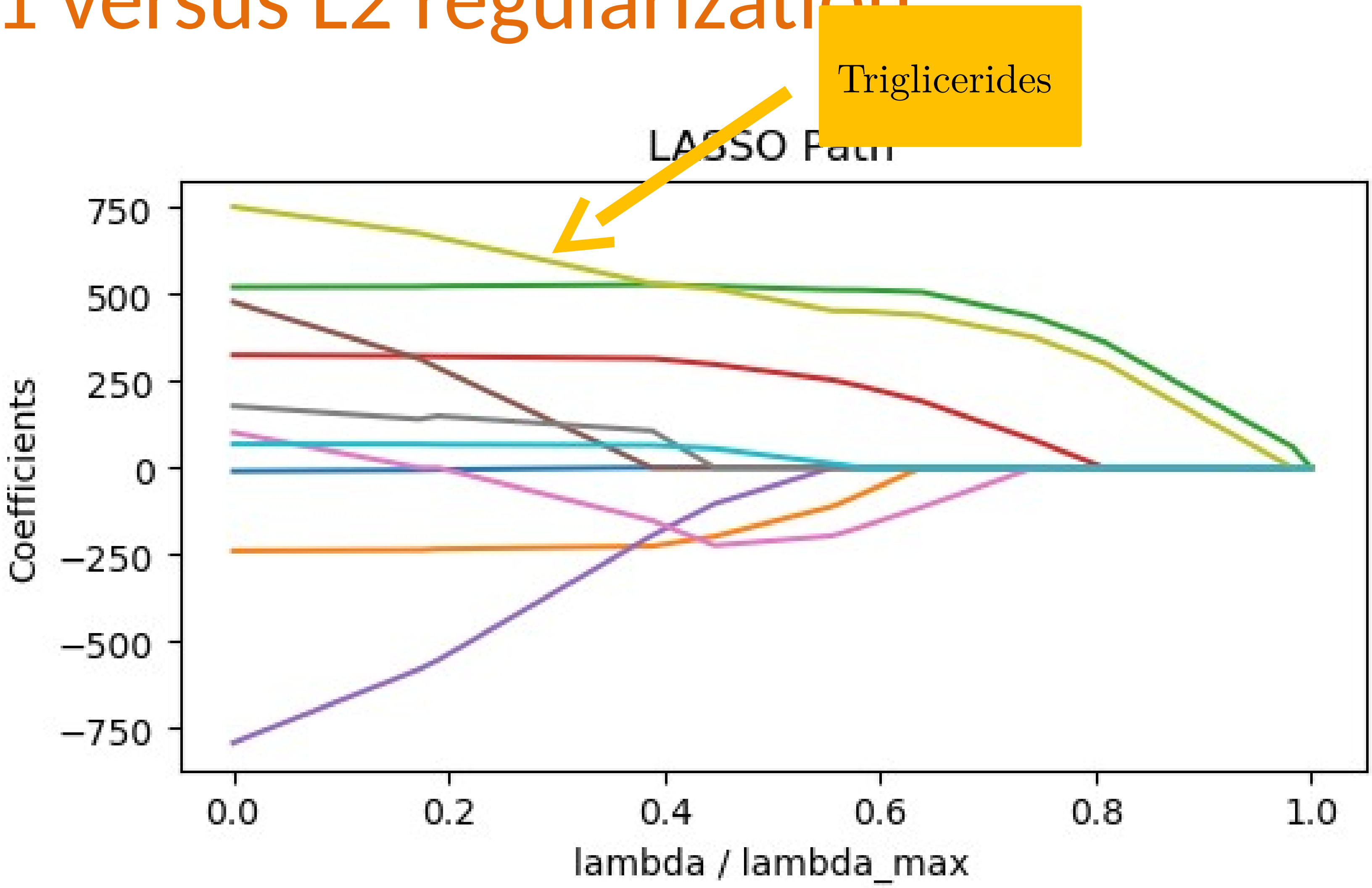
L1 versus L2 regularization



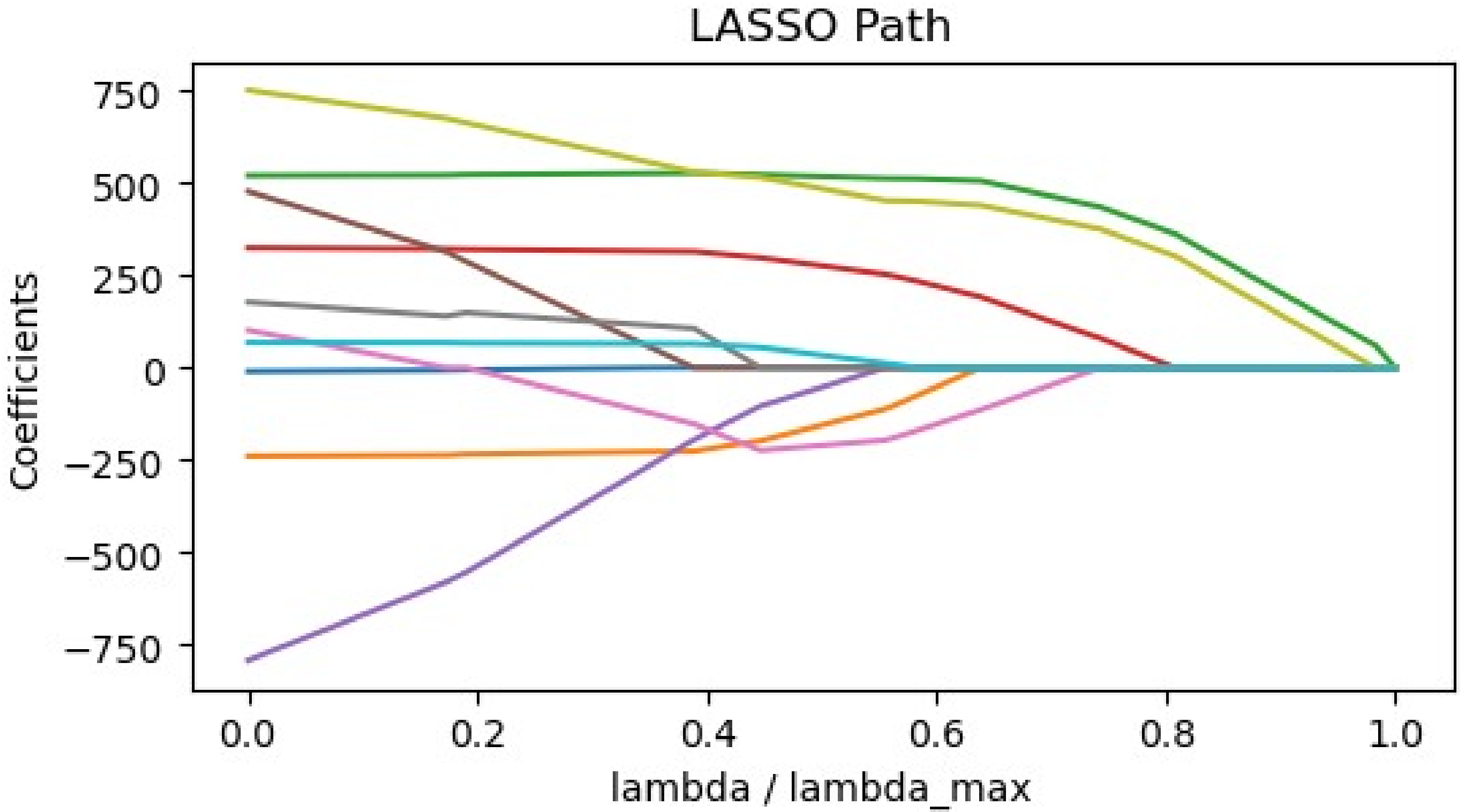
L1 versus L2 regularization



L1 versus L2 regularization



L1 versus L2 regularization



Lasso performs regularization AND feature selection !

Optimization of the Lasso

Not strongly convex when n < p !

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda||w||_1$$

Optimization of the Lasso

Not strongly convex when n < p !

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda||w||_1$$

Expe: take n = 10, p = 20, random X and y, $\lambda = 0.1\lambda_{\text{max}}$

Run ISTA and monitor $\|w^t - w^*\|^2$

We observe:

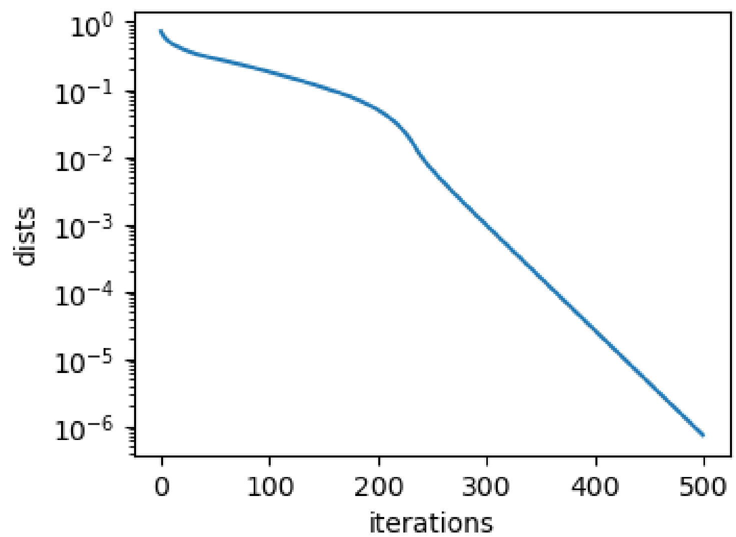
Optimization of the Lasso

Not strongly convex when n < p !

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda||w||_1$$

Expe: take n = 10, p = 20, random X and y, $\lambda = 0.1\lambda_{\max}$
Run ISTA and monitor $\|w^t - w^*\|^2$

We observe:



Optimization of the Lasso

Not strongly convex when $n < p$!

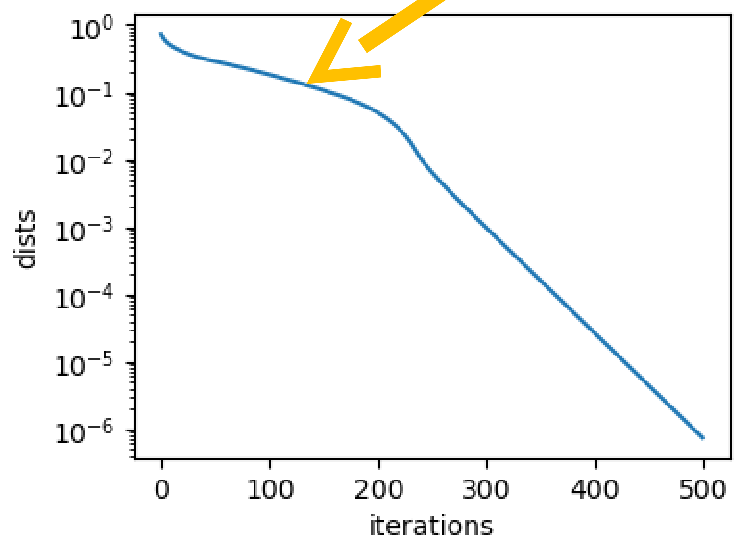
$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda ||w||_1$$

Expe: take n = 10, p = 20, random X and y, $\lambda = 0.1\lambda_{max}$

Run ISTA and monitor $\|w^t - w^*\|^2$

Slow in the beginning....

We observe:



Optimization of the Lasso

Not strongly convex when n < p !

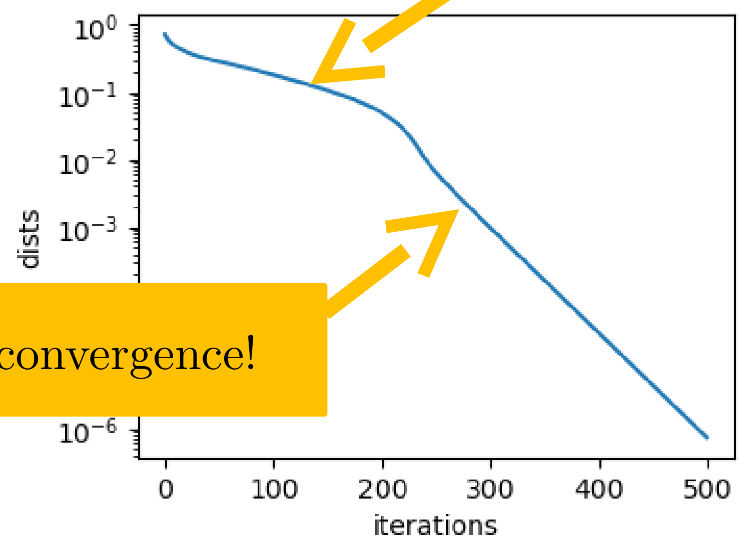
$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda ||w||_1$$

Expe: take n = 10, p = 20, random X and y, $\lambda = 0.1\lambda_{max}$

Run ISTA and monitor $\|w^t - w^*\|^2$

Slow in the beginning....

We observe:



But then, linear convergence!

Sparsity accelerates convergence!

Support identification : There exists T such that for all t > T :

$$\mathrm{supp}(w^t) = \{i| \ w_i^t \neq 0\}$$ is constant and of cardinal $\leq n$.

Sparsity accelerates convergence!

Support identification : There exists T such that for all t > T :

$\mathrm{supp}(w^t) = \{i| \ w_i^t \neq 0\}$ is constant and of cardinal $\leq n$.

For t > T, the problem $\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda||w||_1$

Becomes equivalent to

$$\min_{\tilde{w} \in \mathbf{R}^s} \frac{1}{2}||X^S \tilde{w} - y||_2^2 + \lambda||\tilde{w}||_1$$

With s the size of the support and X^S the features of X restricted to the support

Sparsity accelerates convergence!

Support identification : There exists T such that for all t > T :

$$\mathrm{supp}(w^t) = \{i|\ w_i^t \neq 0\}$$ is constant and of cardinal $\leq n$.

For t > T, the problem $\min_{w \in \mathbf{R}^d} \frac{1}{2}||Xw - y||_2^2 + \lambda||w||_1$

Becomes equivalent to

$$\min_{\tilde{w} \in \mathbf{R}^s} \frac{1}{2}||X^S\tilde{w} - y||_2^2 + \lambda||\tilde{w}||_1$$

With s the size of the support and X^S the features of X restricted to the support

Now, **strongly convex !** Fast convergence when support is identified