Proximal operators and proximal gradient methods

Pierre Ablin

ENS | PSL

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.

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

Where

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

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

Convergence GD I

Theorem

Let f be convex and L-smooth.

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

Where

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

Is f always differentiable?

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

Convergence GD I

Theorem

Let f be convex and L-smooth.

Not true for many problems

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

Where

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

Is f always differentiable?

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

Change notation: Keep loss and regularizer separate

Loss function

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

The Training problem

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

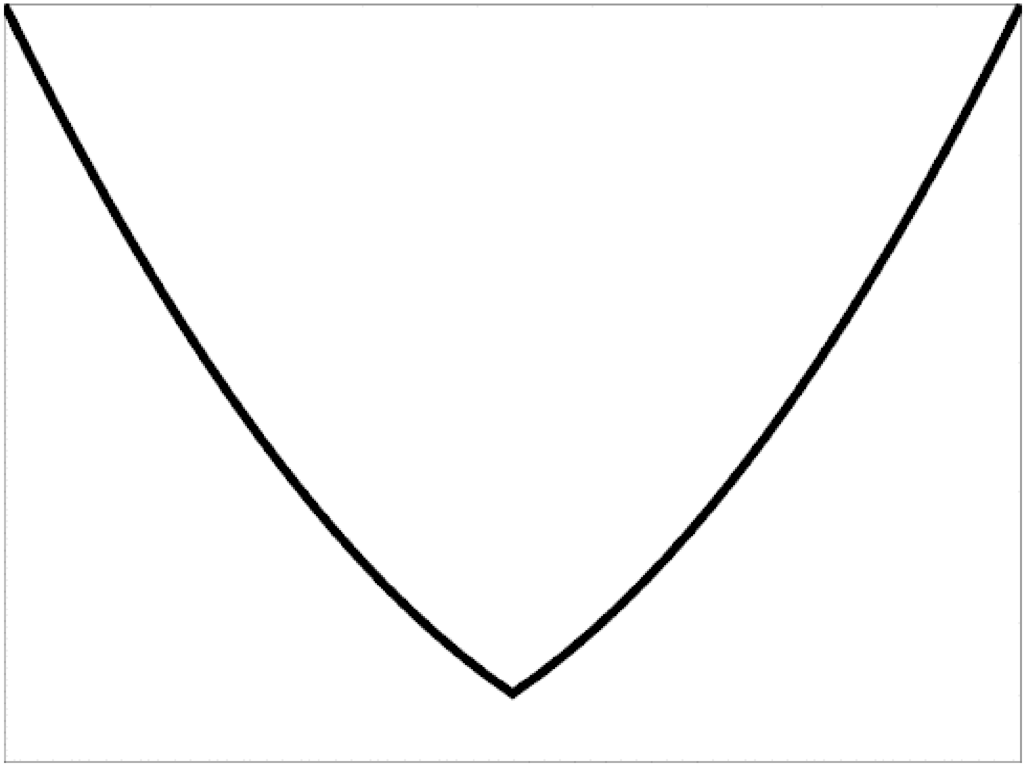
If L or R is not differentiable → $L+R$ is not differentiable

If L or R is not smooth → $L+R$ is not smooth

(In most cases)

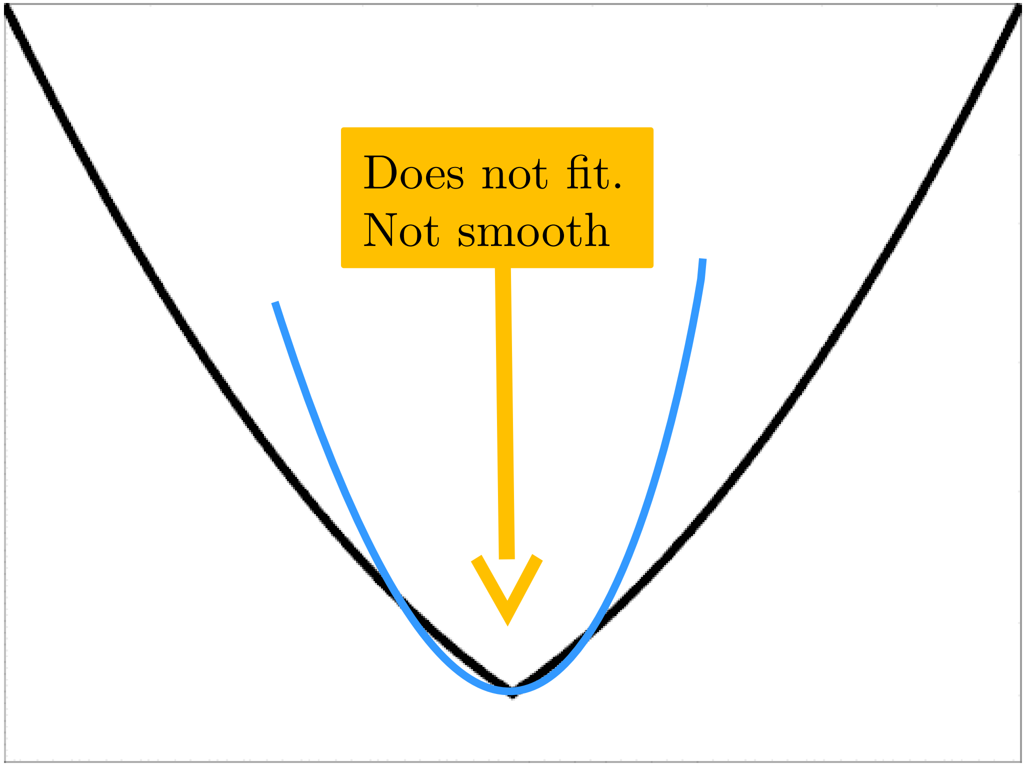
Non-smooth Example

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



Non-smooth Example

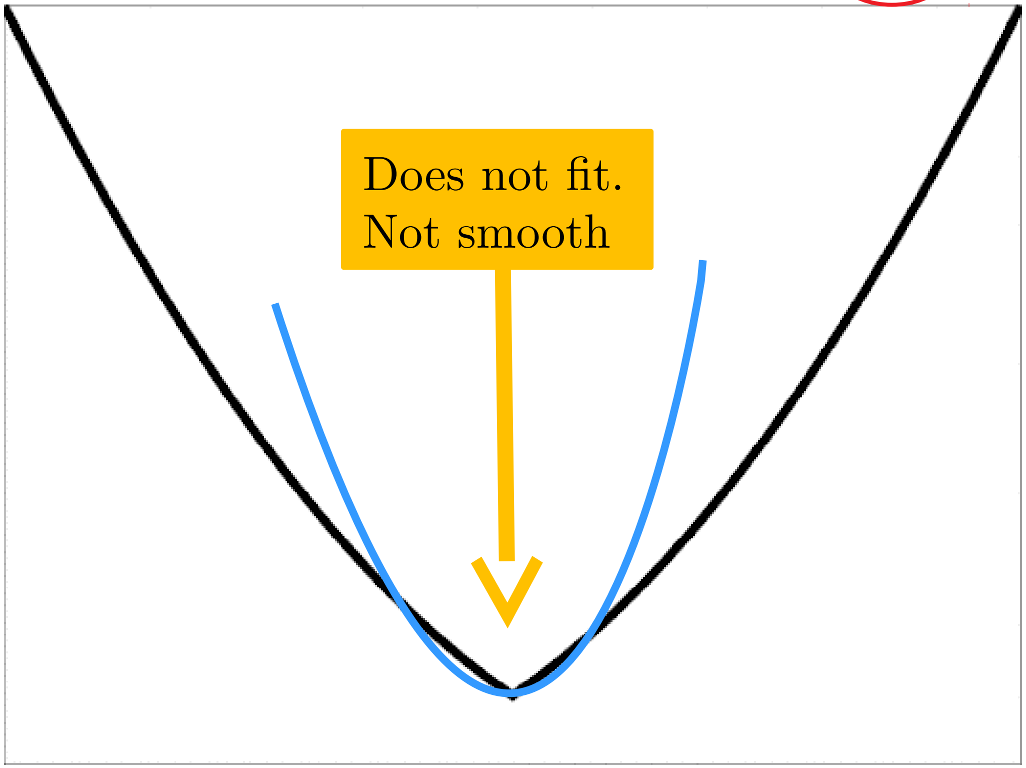
$$L(w) + R(w) = \frac{1}{2}||w||_2^2 + ||w||_1$$



The problem

Non-smooth Example

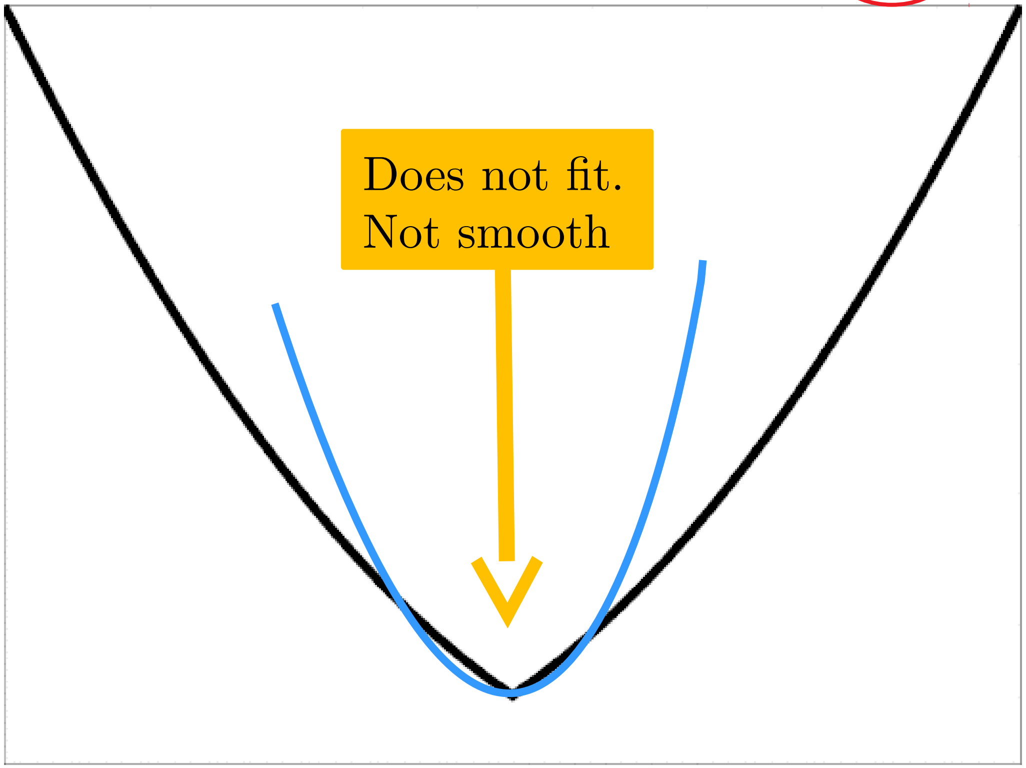
$$L(w) + R(w) = \frac{1}{2}||w||_2^2 + ||w||_1$$



Non-smooth Example

The problem

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



Need more tools

Assumptions for this class

The Training problem

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

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

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

Assumptions for this class

The Training problem

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

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

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

What does this mean?

Assumptions for this class

The Training problem

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

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

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

What does this mean?

$$\operatorname{prox}_{\gamma R}(y) := \arg\min_w \frac{1}{2}||w - y||_2^2 + \gamma R(w)$$

Assume this is easy to solve

Examples

Lasso

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

Low Rank Matrix Recovery

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

Not smooth, but prox is easy

SVM with soft margin

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

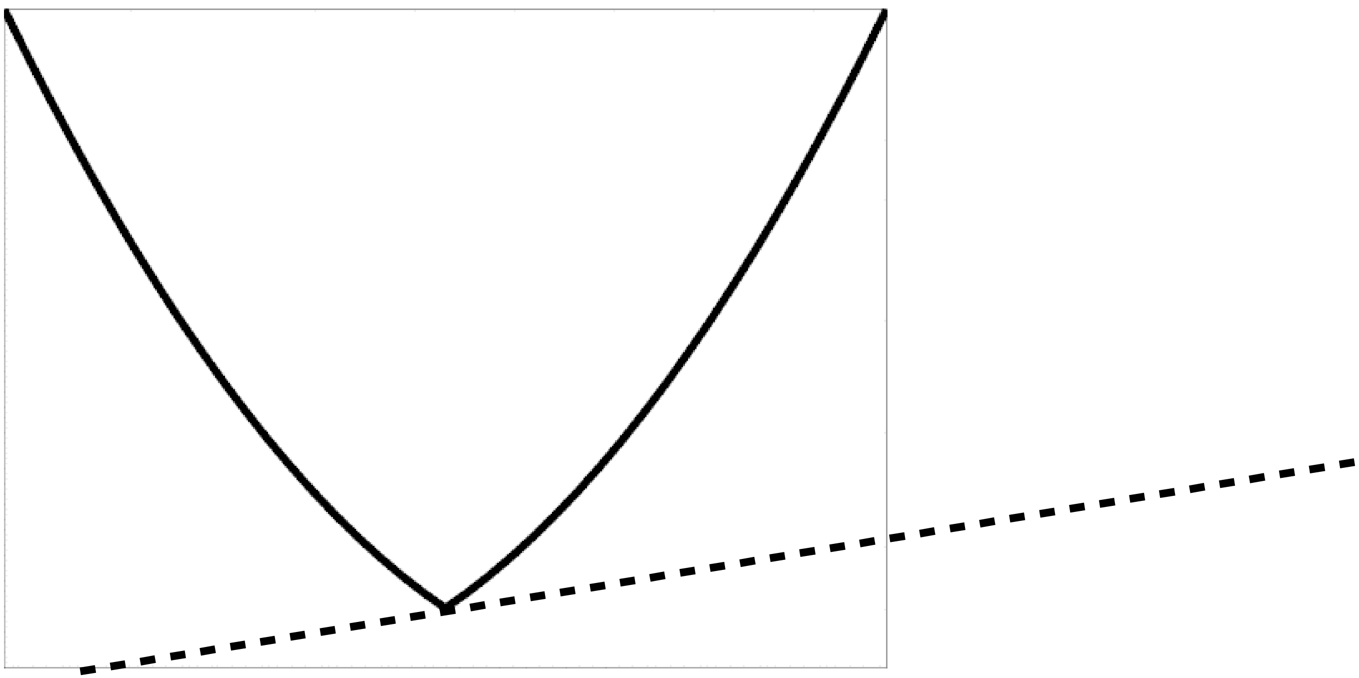
Not smooth

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

Convexity without smoothness: Subgradient

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

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

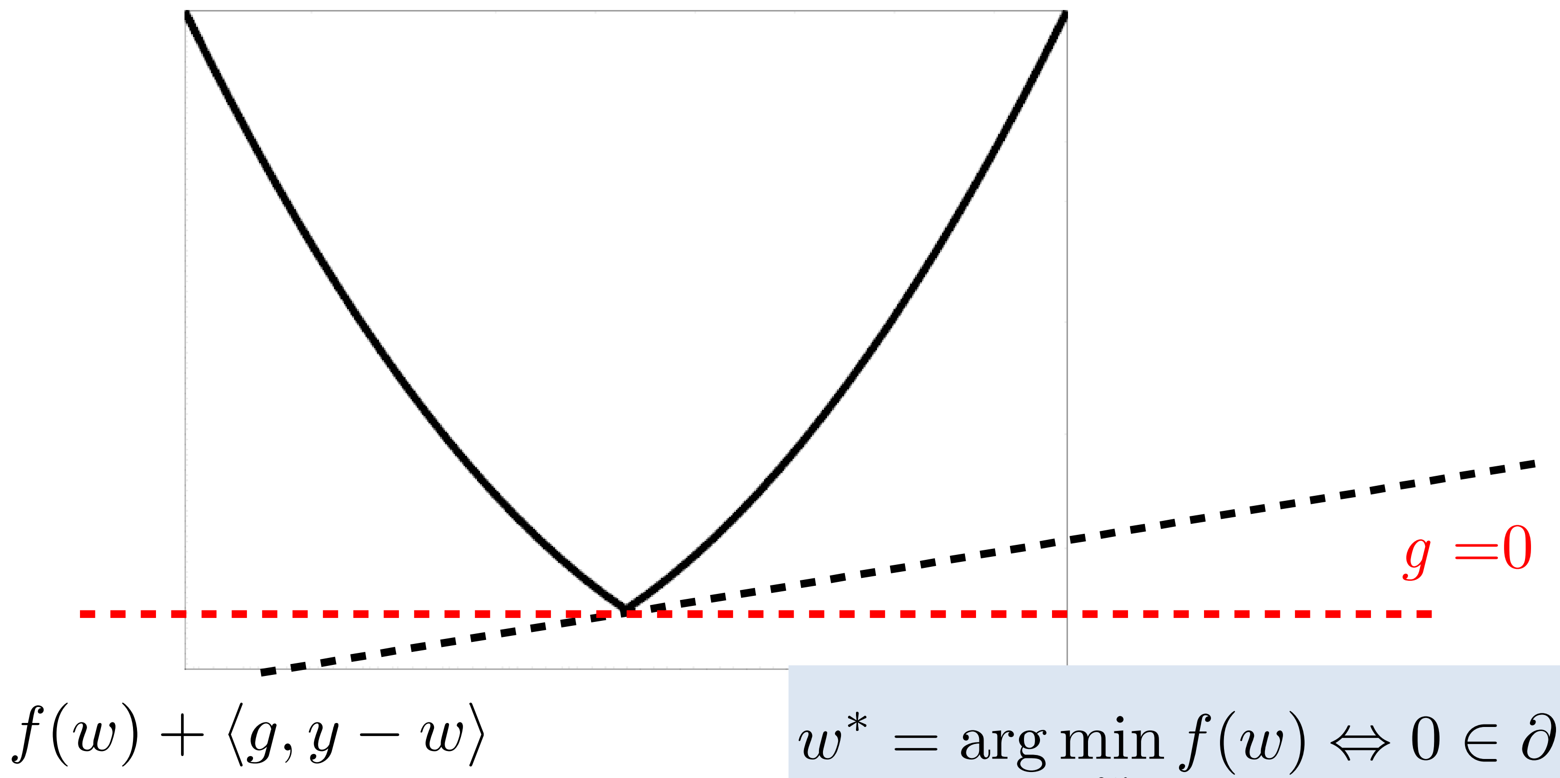


$f(w) + \langle g, y - w \rangle$

Convexity without smoothness: Subgradient

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

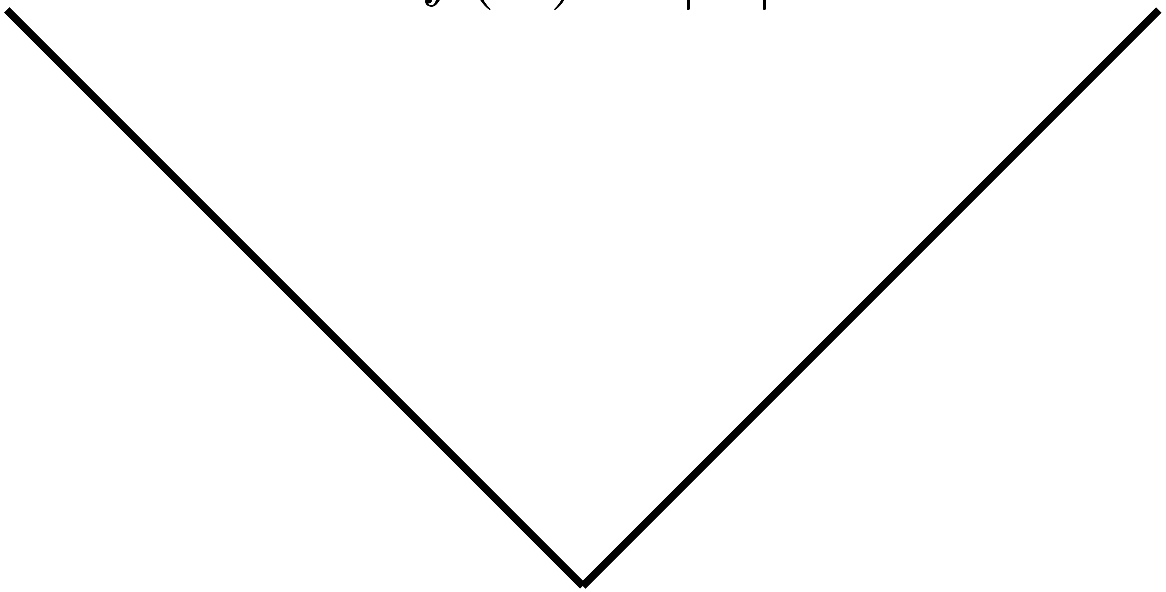
$$\partial f(w) := \{g \in \mathbb{R}^n \ : \ f(y) \geq f(w) + \langle g, y - w \rangle, \forall y \in \mathrm{dom}(f)\}$$



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

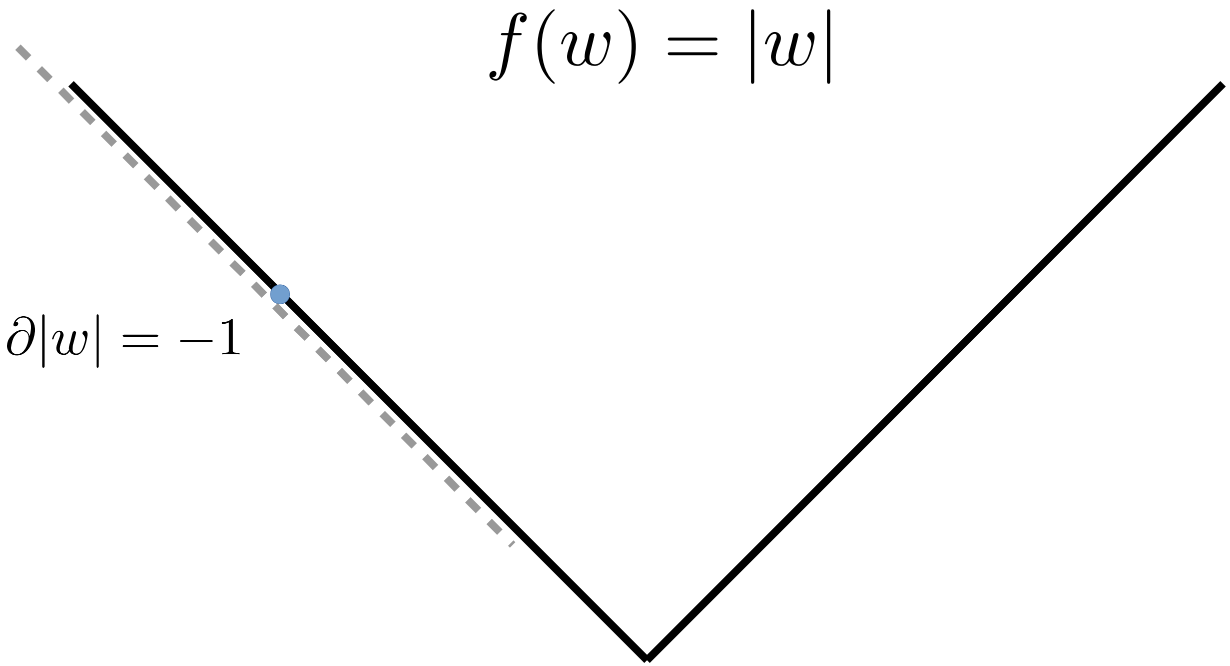
Examples: L1 norm

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



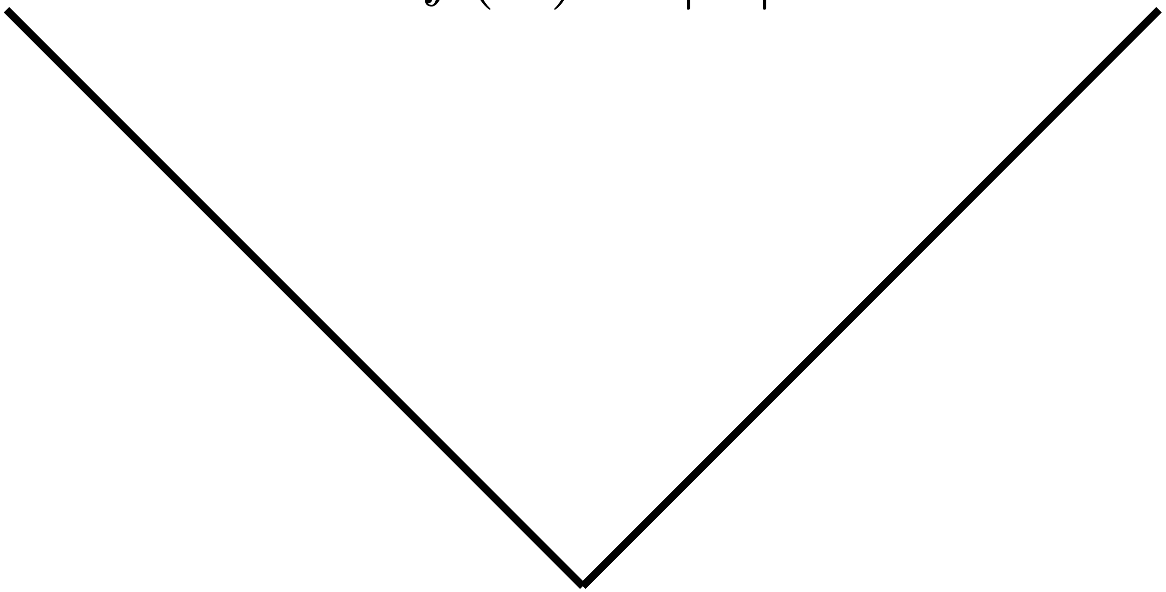
Q: what is the sub-gradient?

Examples: L1 norm



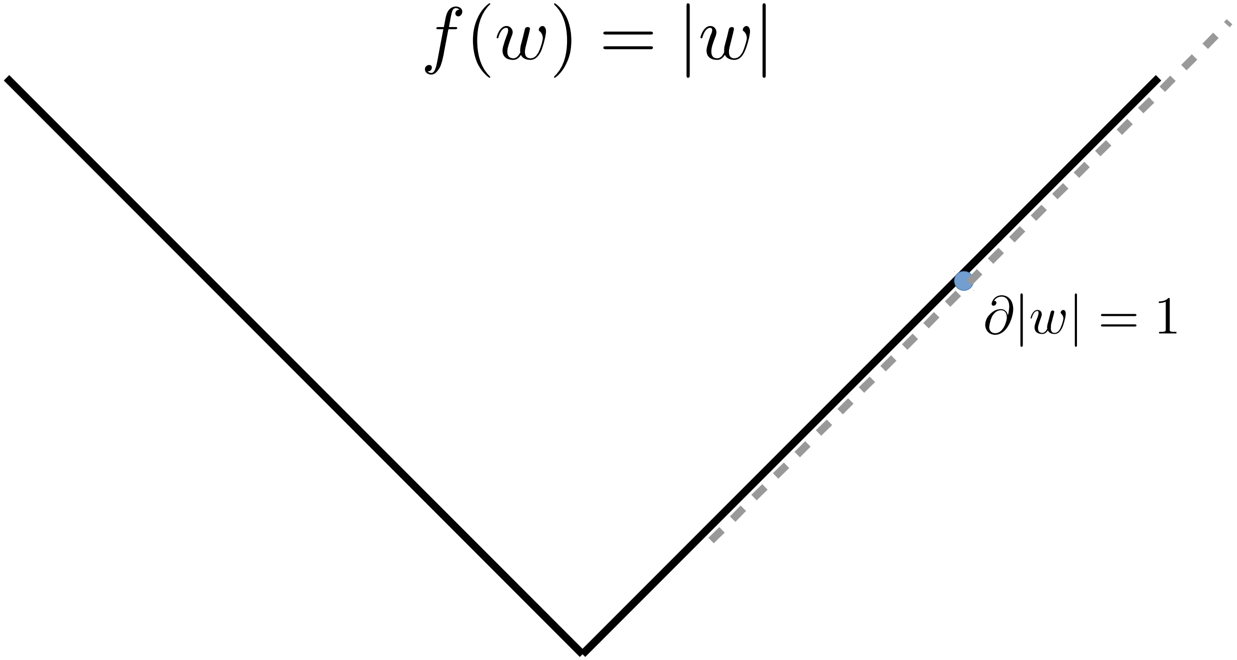
Examples: L1 norm

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



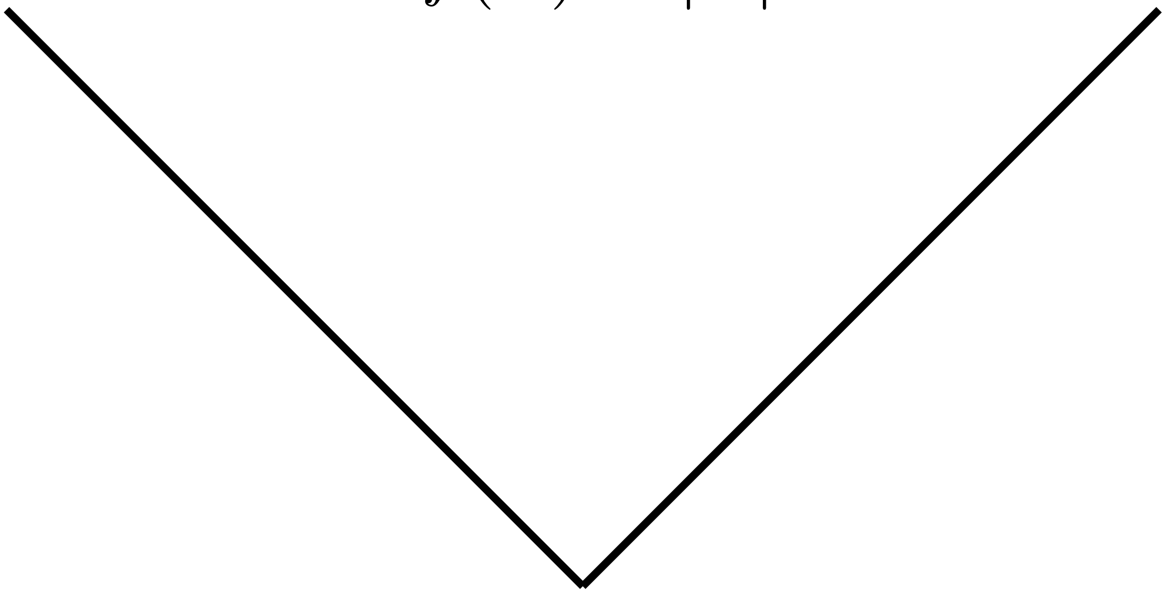
Examples: L1 norm

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

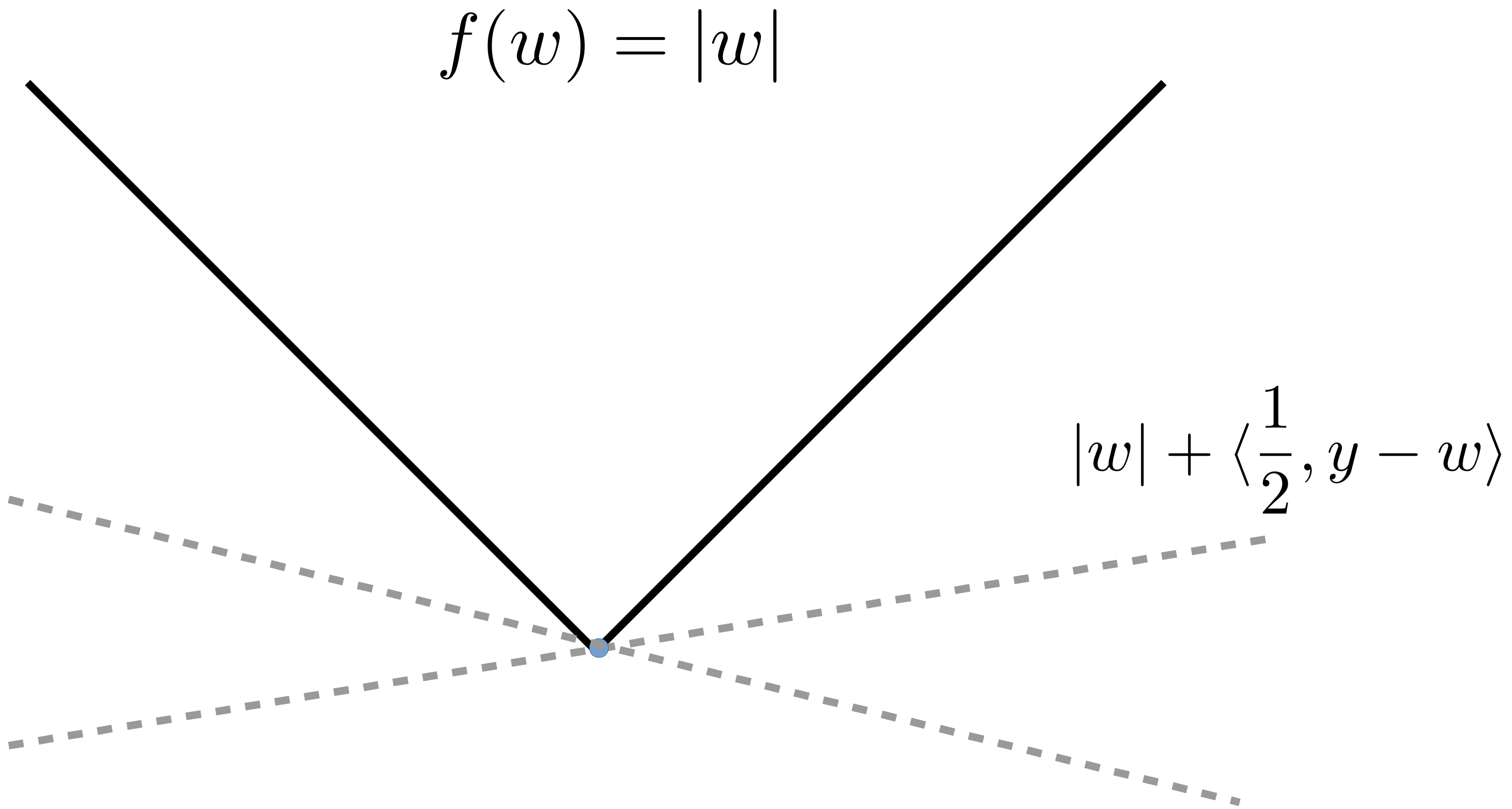


Examples: L1 norm

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

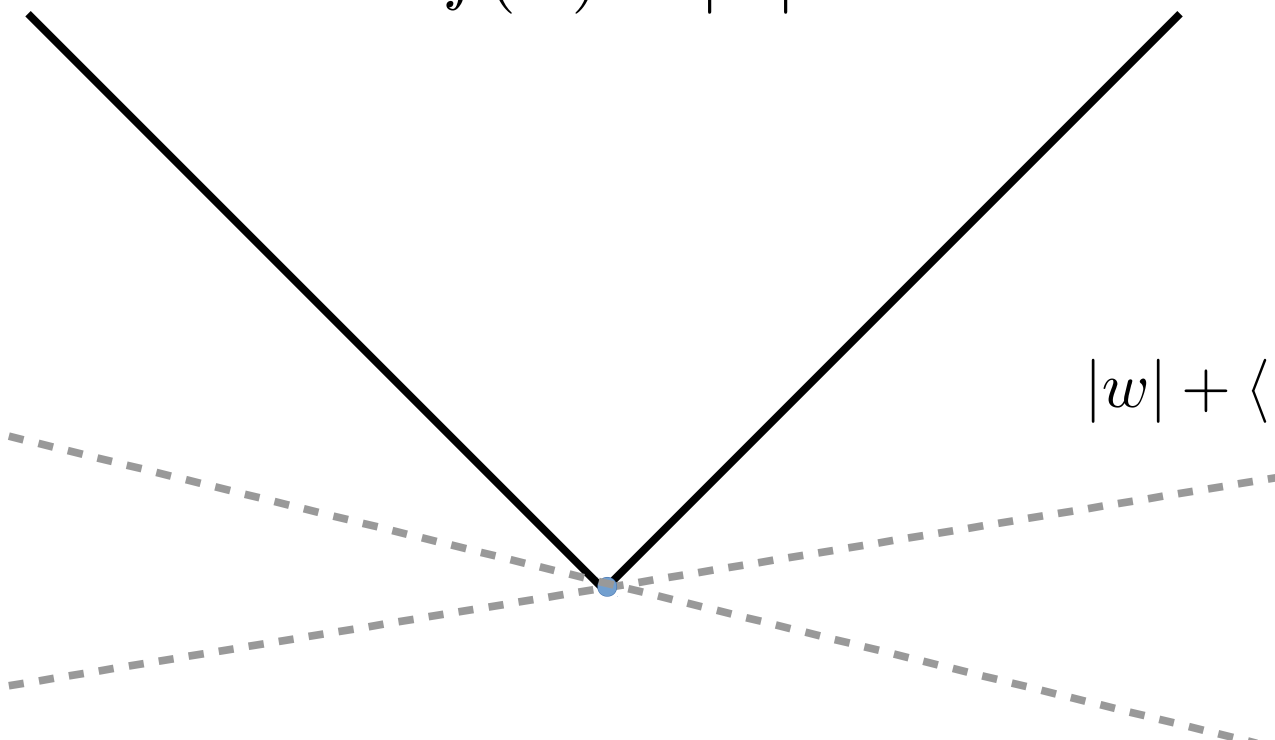


Examples: L1 norm



Examples: L1 norm

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



$$|w| + \langle \frac{1}{2}, y - w \rangle$$

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

Optimality conditions

The Training problem

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

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

$R(w)$ is convex

Optimality conditions

The Training problem

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

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

$R(w)$ is convex

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

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

Working example: Lasso

Lasso

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

$$-\nabla L(w^*) \in \partial R(w^*) \quad \Rightarrow \quad -\frac{1}{n} X^\top (Xw^* - y) \in \lambda \partial ||w^*||_1$$

$$\forall i, \frac{1}{n} \left[X^\top (Xw - y) \right]_i \in \begin{cases} \{-\lambda\} & \text{if } w_i < 0 \\ [-\lambda, \lambda] & \text{if } w_i = 0 \\ \{\lambda\} & \text{if } w_i > 0 \end{cases}$$

Working example: Lasso

Lasso

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

$$-\nabla L(w^*) \in \partial R(w^*) \Rightarrow -\frac{1}{n} X^\top (Xw^* - y) \in \lambda \partial ||w^*||_1$$

Difficult inclusion → Solve iteratively

Proximal method I: iteratively minimizes an upper bound

Using $\mathcal{L}$–smoothness of L :

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

The w that minimizes the upper bound gives gradient descent

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

Proximal method I: iteratively minimizes an upper bound

Using $\mathcal{L}$–smoothness of L :

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

The w that minimizes the upper bound gives gradient descent

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

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

Proximal method I: iteratively minimizes an upper bound

Using $\mathcal{L}$–smoothness of L :

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

The w that minimizes the upper bound gives gradient descent

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

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

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

Proximal method I: iteratively minimizes an upper bound

Using $\mathcal{L}$–smoothness of L :

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

The w that minimizes the upper bound gives gradient descent

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

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

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

Can we minimize the right-hand side?

Proximal method I: iteratively minimizes an upper bound

Minimizing the right-hand side of

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

$$\begin{aligned}
&\arg\min_w L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} ||w - y||^2 + \lambda R(w) \\
&\quad = \arg\min_w \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} ||w - y||^2 + \lambda R(w) \\
&\quad = \arg\min_w \frac{1}{2} ||w - (y - \frac{1}{\mathcal{L}} \nabla L(y))||^2 + \frac{\lambda}{\mathcal{L}} R(w) \\
&\quad =: \operatorname{prox}_{\frac{\lambda}{\mathcal{L}} R}(y - \frac{1}{\mathcal{L}} \nabla L(y)))
\end{aligned}$$

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

Proximal method I: iteratively minimizes an upper bound

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

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

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

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

This suggests an iterative method

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

Proximal method I: iteratively minimizes an upper bound

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

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

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

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

This suggests an iterative method

What is this prox operator?

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

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 II: 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 $w_v = \mathrm{prox}_f(v)$.

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

Proximal Operator II: Inclusion definition

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

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

Let $w_v = \operatorname{prox}_f(v)$. 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 II: 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 III: fixed point

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

Proximal Method III: A fixed point viewpoint

The Training problem

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

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

Proximal Method III: A fixed point viewpoint

The Training problem

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

$$-\nabla L(w^*) \in \lambda \partial R(w^*) \quad \Longleftrightarrow \quad w^* + \gamma \nabla L(w^*) \in w^* - (\lambda\gamma)\partial R(w^*)$$

Proximal Method III: A fixed point viewpoint

The Training problem

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

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

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

Proximal Method III: A fixed point viewpoint

The Training problem

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

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

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

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

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

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

Proximal Method III: A fixed point viewpoint

The Training problem

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

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

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

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

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

Optimal is a fixed point

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

Proximal Method III: A fixed point viewpoint

The Training problem

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

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

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

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

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

Optimal is a fixed point $\Rightarrow$

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

Upper bound viewpoint $\Rightarrow$

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

Proximal Operator: Properties

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

Exe:

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

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

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

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

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

Proximal Operator: Soft thresholding

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

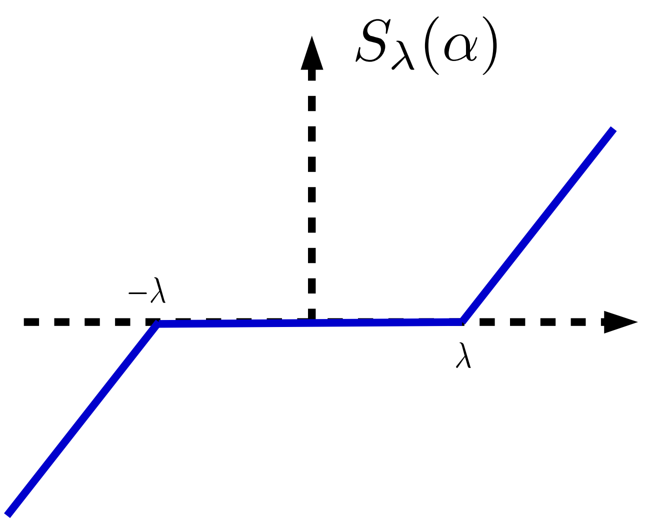
Exe:

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

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

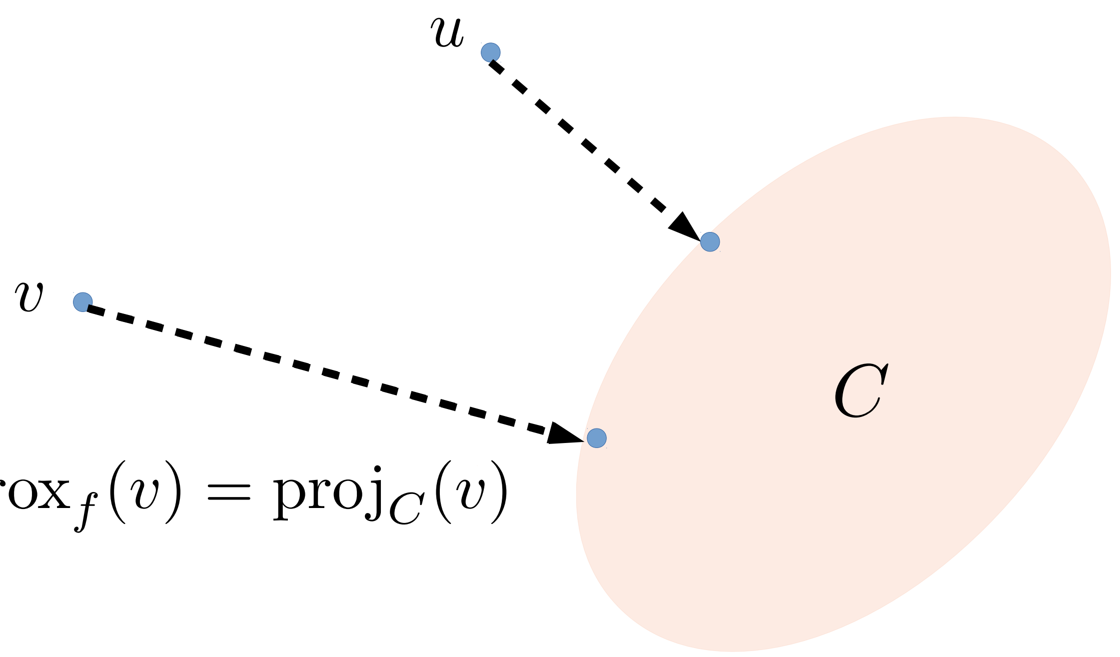
2) If $\lambda < v$ show (I) gives $\alpha^* = v - \lambda$
3) If $v < -\lambda$ show (I) gives $\alpha^* = v + \lambda$
4) Show that

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



Proximal Operator: Non-expansiveness

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



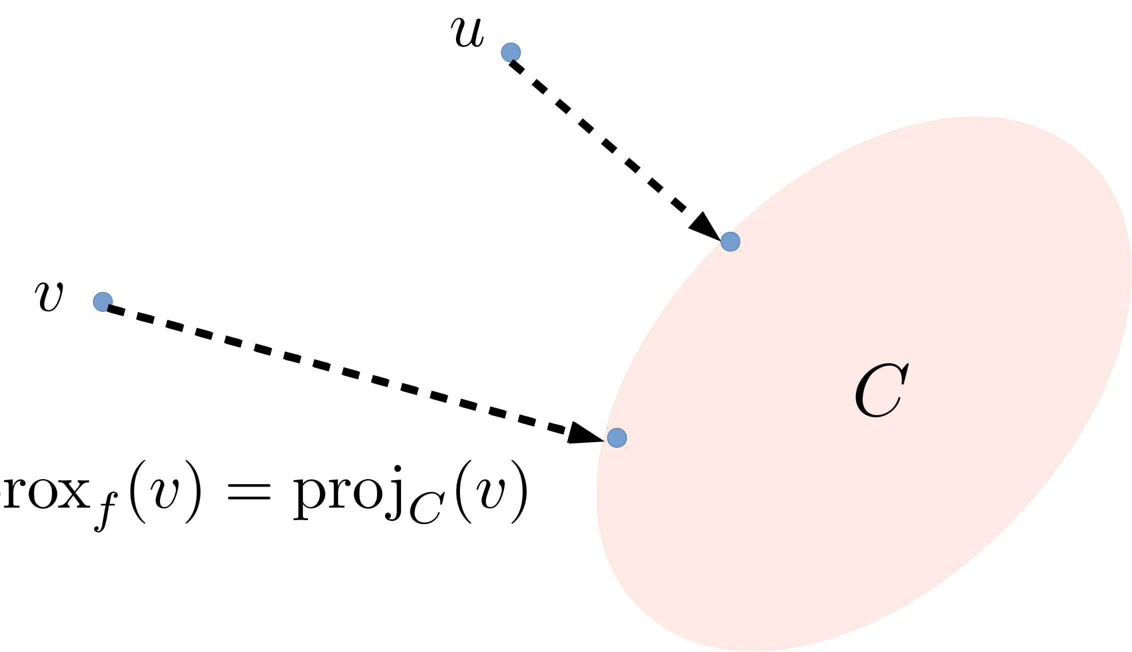
$$\mathrm{prox}_f(v) = \mathrm{proj}_C(v)$$

Proximal Operators are nonexpansive

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

Proximal Operator: Non-expansiveness

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



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

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

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

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

Using subgradient characterization

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proximal Operator: Singular value thresholding

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

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

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

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 method: iteratively minimizes an upper bound

Minimizing the right-hand side of

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

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

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

$$= \arg\min_w \frac{1}{2}||w - (y - \frac{1}{\mathcal{L}}\nabla L(y)||^2 + \frac{\lambda}{\mathcal{L}}R(w)$$

$$= \operatorname{prox}_{\frac{\lambda}{\mathcal{L}}R}\left(y - \frac{1}{\mathcal{L}}\nabla L(y)\right)$$

Make iterative method based on this upper bound minimization

The Proximal Gradient Method

Solving the *training problem*:

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

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

$R(w)$ is convex

Proximal Gradient Descent

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

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

Output w^{T+1}

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

Lasso

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

ISTA:

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

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

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

PDF Adobe

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 $f(w) = L(w) + \lambda R(w)$ where

$L(w)$ is differentiable, $\mathcal{L}$–smooth and μ–strongly convex

$R(w)$ is convex

Then

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

where

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

PDF Adobe

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

Proof sketch

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

Proof sketch

Fixed point viewpoint

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

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

Proof sketch

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

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

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

Proof sketch

Fixed point viewpoint

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

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

Non-expansive

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

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

Convergence of Prox-GD

Theorem (Beck Teboulle 2009)

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

$L(w)$ is differentiable, $\mathcal{L}$–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

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

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