Optimization for Machine Learning

Proximal operator and proximal gradient methods

Lecturers: Francis Bach & Robert M. Gower

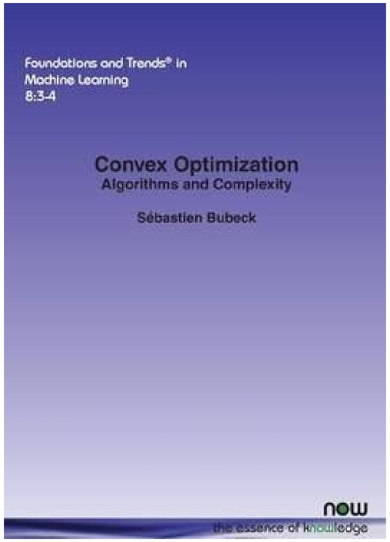
Tutorials: Hadrien Hendrikx, Rui Yuan, Nidham Gazagnadou

AIMS | African Institute for Mathematical Sciences NEXT EINSTEIN INITIATIVE

African Master's in Machine Intelligence (AMMI), Kigali

References classes today

Sébastien Bubeck (2015)
Convex Optimization: Algorithms and Complexity



Foundations and Trends® in Machine Learning
8:3-4

Convex Optimization
Algorithms and Complexity

Sébastien Bubeck

now
the essence of knowledge

Chapter 1 and Section 5.1

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

Optimization Sum of Terms

A Datum Function

$$f_i(w) := \ell\left(h_w(x^i), y^i\right) + \lambda R(w)$$

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right) + \lambda R(w) &= \frac{1}{n}\sum_{i=1}^{n} \left(\ell\left(h_w(x^i), y^i\right) + \lambda R(w)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} f_i(w) \end{aligned}$$

Finite Sum Training Problem

$$\min_{w \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

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

Theorem

Let *f* be μ-strongly convex and *L*-smooth.

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

Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t), \quad \text{for } t = 1, \ldots, T$$

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

Convergence GD I

Not true for many problems

Theorem

Let *f* be convex and *L*-smooth.

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

Where

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

Is *f* always differentiable?

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

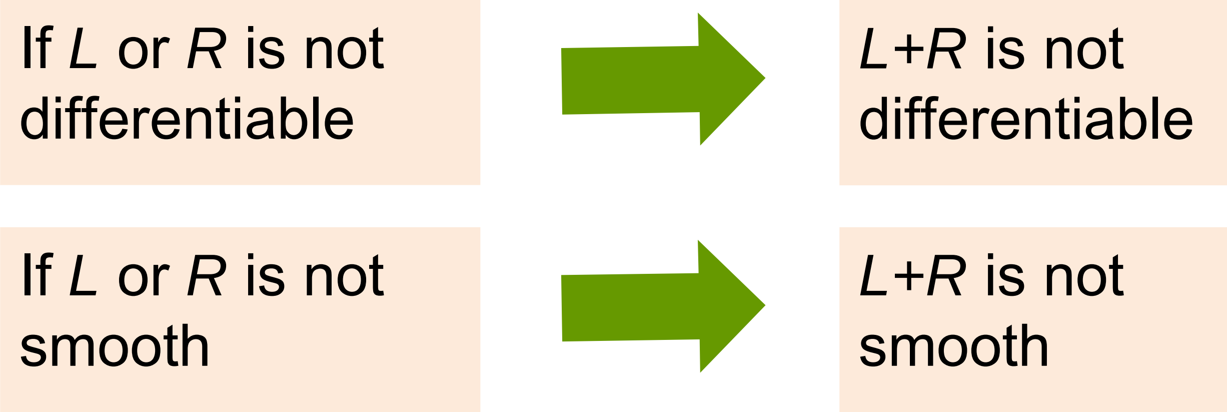
Change notation: Keep loss and regularizor separate

Loss function

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

The Training problem

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

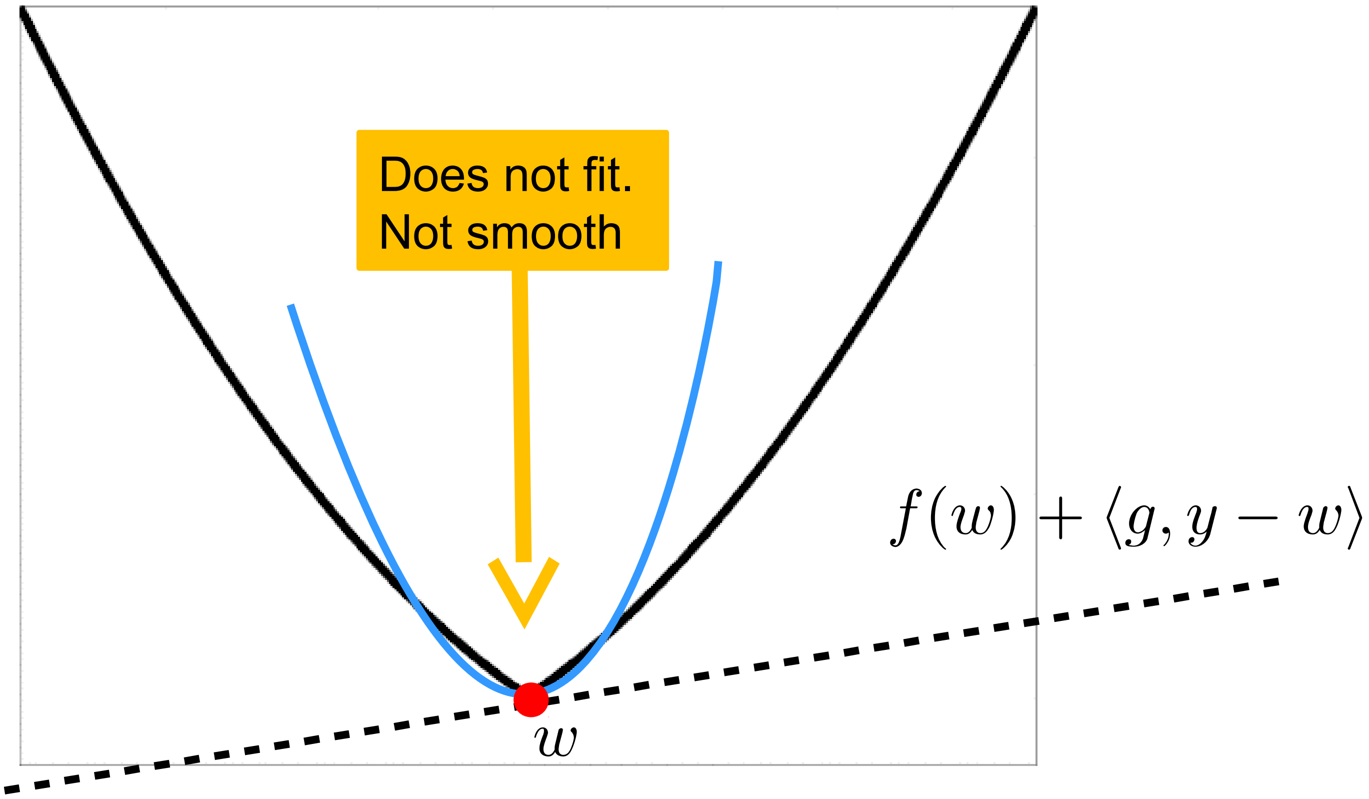


If *L* or *R* is not differentiable → *L*+*R* is not differentiable

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

Non-smooth Example

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



Does not fit.
Not smooth

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"

What does this mean?

Examples all together!!

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

Assume this is easy to solve

Examples

Lasso

$$\min_{w \in \mathbf{R}^d} \frac{1}{2n} \sum_{i=1}^{n} (y^i - \langle w, a^i \rangle)^2 + \lambda ||w||_1$$

Not smooth, but prox is easy

SVM with soft margin

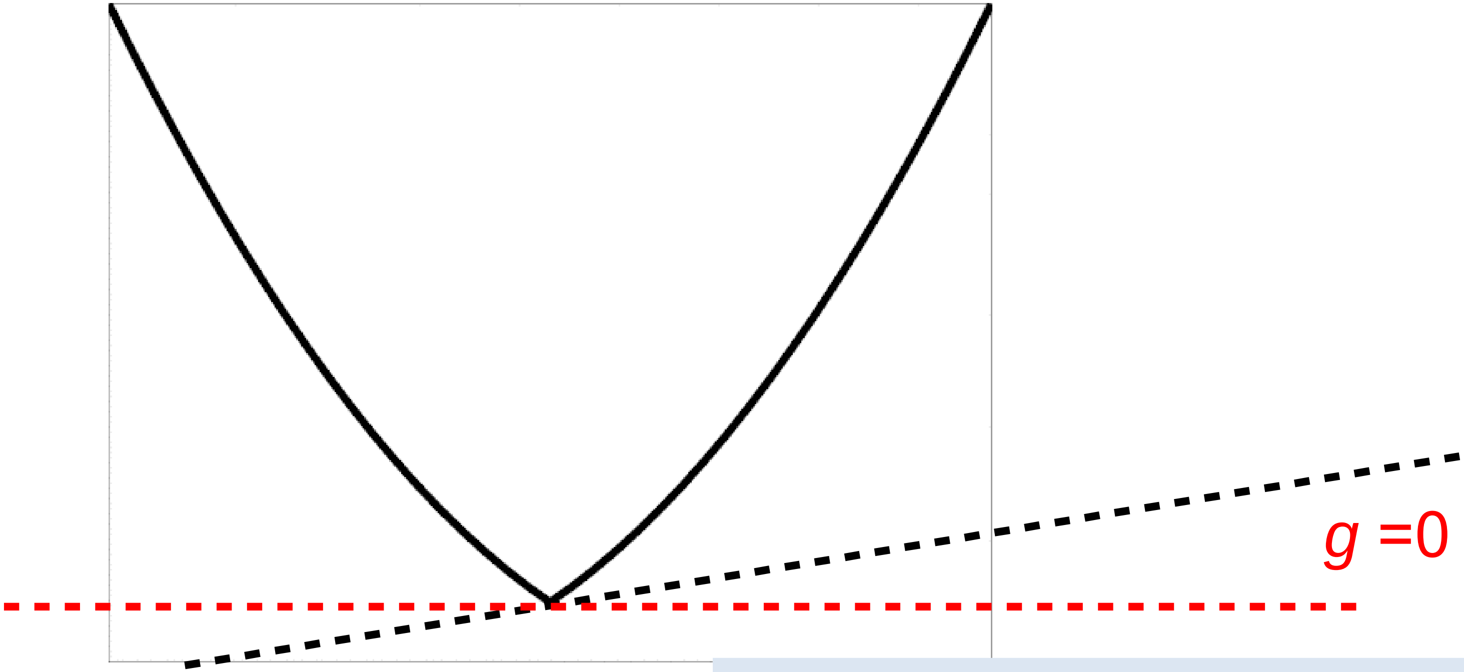
$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y^i \langle w, a^i \rangle\} + \lambda ||w||_2^2$$

Not smooth

Convexity: Subgradient

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

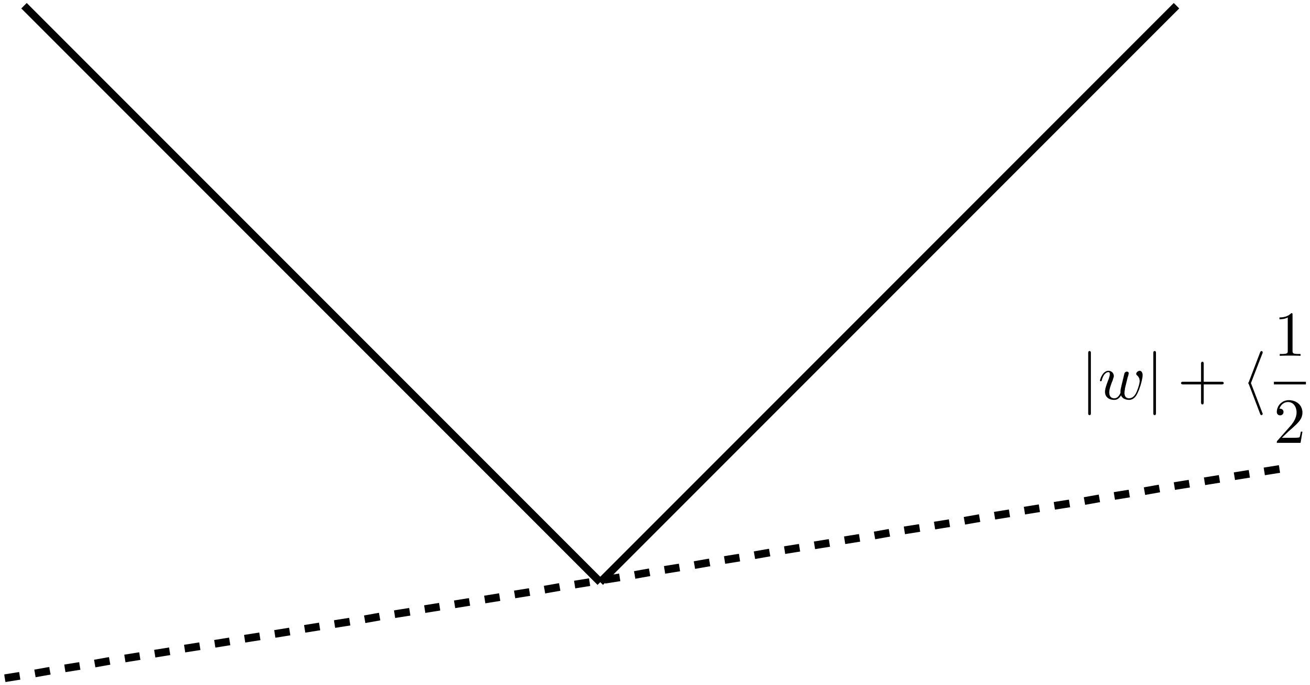
$$\partial f(w) := \{g \in \mathbb{R}^n \ : \ f(y) \geq f(w) + \langle g, y - w \rangle, \forall y \in \mathrm{dom}(f)\}$$



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

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

Examples: L1 norm



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

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

$$\partial ||w||_1 = (\partial |w_1|, \ldots, \partial |w_d|)$$

Optimality conditions

The Training problem

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

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

$$\Downarrow$$

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

Working example: Lasso

Lasso

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

$$A = [a^1, \ldots, a^n]^\top \; \Rightarrow \sum_{i=1}^{n} (y^i - \langle w, a^i \rangle)^2 = ||Aw - y||_2^2$$

$$-\nabla L(w^*) \in \partial R(w^*) \quad \Longrightarrow \quad -\frac{1}{n} A^\top (Aw^* - y) \in \partial ||w^*||_1$$

Difficult inclusion, do iteratively.

Proximal method I

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 II

Minimizing the right-hand side of

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

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

What is this prox operator?

$$\operatorname{prox}_{\frac{\lambda}{\mathcal{L}}R}(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + \frac{\lambda}{\mathcal{L}}R(w)$$

Proximal Operator I

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 II: Optimality conditions

The Training problem

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

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

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

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

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

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

Optimal is a fixed point.

Proximal Operator III: 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 is closed and conve

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 IV: 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

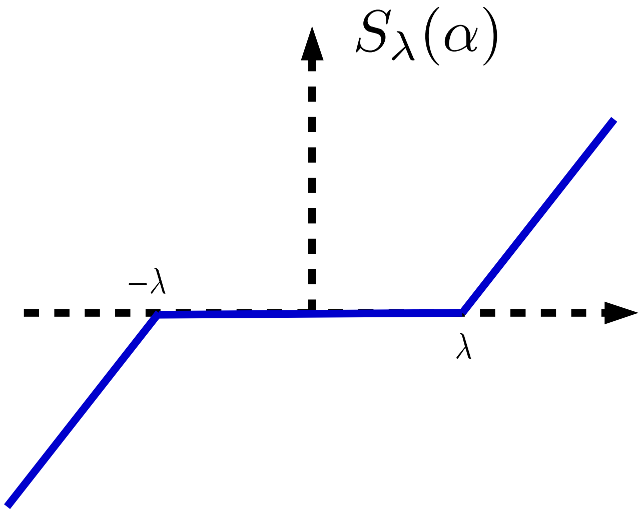
$$\alpha^* \in v - \lambda\partial|\alpha^*| \qquad (I)$$

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 V: Singular value thresholding

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

Similarily, the prox of the nuclear norm for matrices:

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

where $A = U\text{diag}(\sigma(A))V^\top$ is a SVD decomposition.

Proximal method V

Minimizing the right-hand side of

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

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

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

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 and prox friendly

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 $w^{T+1}$
```

Iterative Soft Thresholding Algorithm (ISTA)

Lasso

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

$$A=[a^1,\ldots,a^n]^\top \Rightarrow \sum_{i=1}^n (y^i-\langle w,a^i\rangle)^2 = ||Aw-y||_2^2$$

ISTA:

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

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

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

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

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

Can we do better?

Then

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

where

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

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

The FISTA Method

Solving the *training problem*:

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

The FISTA Algorithm

```
Set $w^1 = 0 = z^1, \beta^1 = 1$
for $t = 1, 2, 3, \ldots, T$
    $w^{t+1} = \text{prox}_{\lambda R/\mathcal{L}} \left( z^t - \frac{1}{\mathcal{L}} \nabla L(z^t) \right)$
    $\beta^{t+1} = \frac{1 + \sqrt{1 + 4(\beta^t)^2}}{2}$
    $z^{t+1} = w^{t+1} + \frac{\beta^t - 1}{\beta^{t+1}} (w^{t+1} - w^t)$
Output $w^{T+1}$
```

Weird, but it works

Convergence of FISTA

Theorem (Beck Teboulle 2009)

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

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

$$R(w) \text{ is convex and prox friendly}$$

Is this as good as it gets?

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.

Lab Session 01.10

Bring your laptop!

Introduction to Stochastic Gradient Descent

Optimization Sum of Terms

A Datum Function

$$f_i(w) := \ell\left(h_w(x^i), y^i\right) + \lambda R(w)$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right) + \lambda R(w) &= \frac{1}{n}\sum_{i=1}^{n} \left(\ell\left(h_w(x^i), y^i\right) + \lambda R(w)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} f_i(w)\end{aligned}$$

Finite Sum Training Problem

$$\min_{w\in\mathbf{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(w) =: f(w)$$

The Training Problem

Solving the *training problem*:

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

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 $w^0 = 0$, 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 $w^{T+1}$
```

The Training Problem

Solving the *training problem*:

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Problem with Gradient Descent:
Each iteration requires computing a gradient $\nabla f_i(w)$ for each data point. One gradient for each cat on the internet!

Gradient Descent Algorithm

```
Set $w^0 = 0$, 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 $w^{T+1}$
```

Stochastic Gradient Descent

Is it possible to design a method that uses only the gradient of a **single** data function $f_i(w)$ at each iteration?

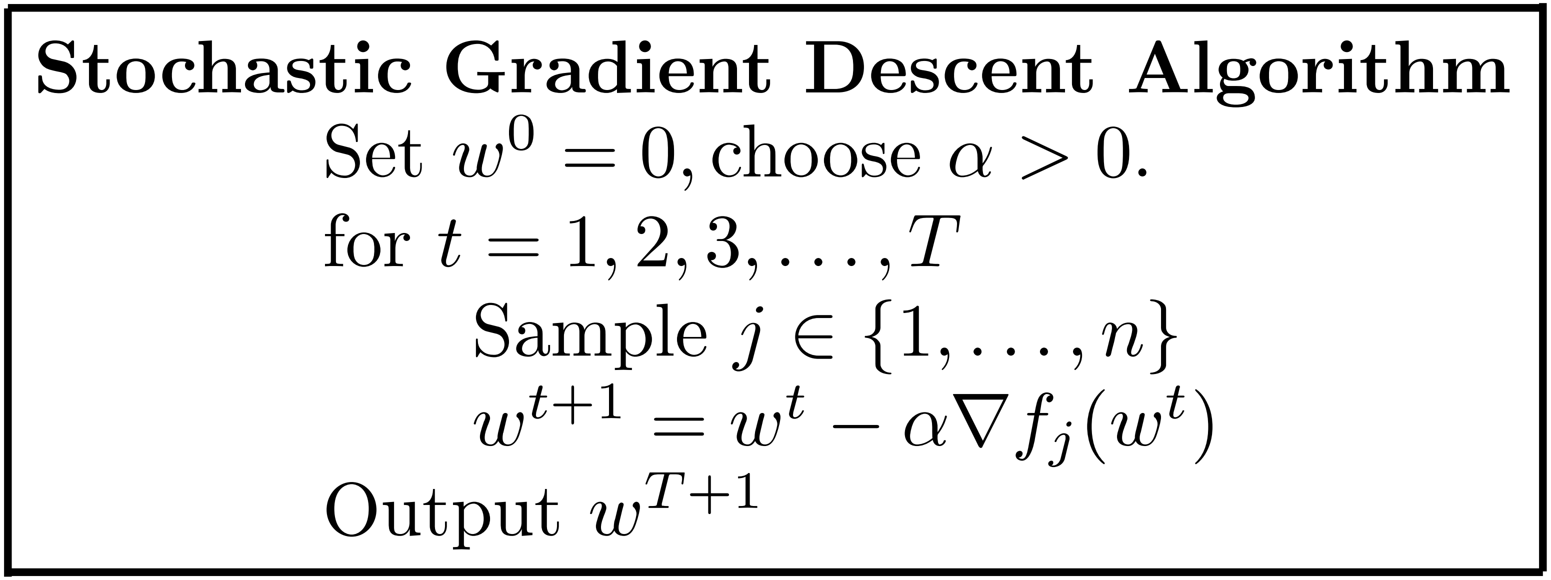
Unbiased Estimate

Let *j* be a random index sampled from {*1*, …, *n*} selected uniformly at random. Then

$$\mathbb{E}_j\left[\nabla f_j(w)\right] = \frac{1}{n}\sum \nabla f_i(w) = \nabla f(w)$$

$$\text{Use } \nabla f_j(w) \approx \nabla f(w)$$

Stochastic Gradient Descent



Stochastic Gradient Descent Algorithm

```
Set $w^0 = 0$, choose $\alpha > 0$.
for $t = 1, 2, 3, \ldots, T$
    Sample $j \in \{1, \ldots, n\}$
    $w^{t+1} = w^t - \alpha \nabla f_j(w^t)$
Output $w^{T+1}$
```