The Duality of Strong Convexity and Strong Smoothness Applications to Machine Learning

Shai Shalev-Shwartz

Toyota Technological Institute at Chicago

Talk at AFOSR workshop, January, 2009

January 2009

Outline

Lemma

f *is strongly convex w.r.t.* $\|\cdot\| \iff f^\star$ *is strongly smooth w.r.t.* $\|\cdot\|_\star$

Applications:

- Rademacher Bounds ($\Rightarrow$ Generalization Bounds)

Outline

Lemma

f *is strongly convex w.r.t.* $\|\cdot\|$ $\iff$ $f^\star$ *is strongly smooth w.r.t.* $\|\cdot\|_\star$

Applications:

- Rademacher Bounds ($\Rightarrow$ Generalization Bounds)
- Low regret online algorithms ($\Rightarrow$ runtime of SGD/SMD)

Outline

Lemma

f is strongly convex w.r.t. $\|\cdot\|$ $\iff$ $f^\star$ is strongly smooth w.r.t. $\|\cdot\|_\star$

Applications:

- Rademacher Bounds ($\Rightarrow$ Generalization Bounds)
- Low regret online algorithms ($\Rightarrow$ runtime of SGD/SMD)
- Boosting

Outline

Lemma

f is strongly convex w.r.t. $\|\cdot\|$ $\iff$ $f^\star$ is strongly smooth w.r.t. $\|\cdot\|_\star$

Applications:

- Rademacher Bounds ($\Rightarrow$ Generalization Bounds)
- Low regret online algorithms ($\Rightarrow$ runtime of SGD/SMD)
- Boosting
- Sparsity and ℓ_1 norm

Outline

Lemma

f is strongly convex w.r.t. $\|\cdot\|$ $\iff$ $f^\star$ is strongly smooth w.r.t. $\|\cdot\|_\star$

Applications:

- Rademacher Bounds ($\Rightarrow$ Generalization Bounds)
- Low regret online algorithms ($\Rightarrow$ runtime of SGD/SMD)
- Boosting
- Sparsity and ℓ_1 norm
- Concentration inequalities

Outline

Lemma

f is strongly convex w.r.t. $\|\cdot\|$ $\iff$ $f^\star$ is strongly smooth w.r.t. $\|\cdot\|_\star$

Applications:

- Rademacher Bounds ($\Rightarrow$ Generalization Bounds)
- Low regret online algorithms ($\Rightarrow$ runtime of SGD/SMD)
- Boosting
- Sparsity and ℓ_1 norm
- Concentration inequalities
- Matrix regularization (Multi-task, group Lasso, dynamic bounds)

Motivating Problem – Generalization Bounds

- Linear predictor is a mapping $\mathbf{x} \mapsto \phi(\langle \mathbf{w}, \mathbf{x} \rangle)$
  - E.g. $\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$ or $\mathbf{x} \mapsto \mathrm{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle)$
- Loss of $\mathbf{w}$ on $(\mathbf{x}, y)$ is assessed by $\ell(\langle \mathbf{w}, \mathbf{x} \rangle, y)$
- Goal: minimize expected loss $L(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)}[\ell(\langle \mathbf{w}, \mathbf{x} \rangle, y)]$
- Instead, minimize empirical loss $\hat{L}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\langle \mathbf{w}, \mathbf{x}_i \rangle, y_i)$
- Bartlett and Mendelson [2002]:
  If ℓ Lipschitz and bounded, w.p. at least $1 - \delta$

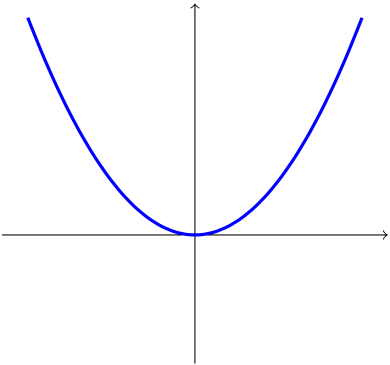
$$\forall \mathbf{w} \in S, \quad L(\mathbf{w}) \le \hat{L}(\mathbf{w}) + \frac{2}{n} \mathcal{R}_n(S) + \sqrt{\frac{\log(1/\delta)}{2\,n}}$$

where

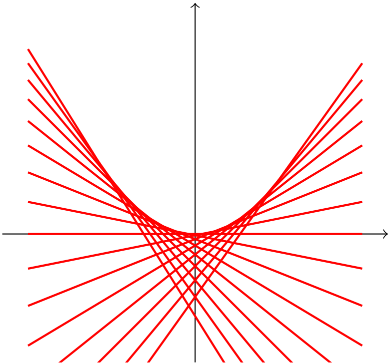
$$\mathcal{R}_n(S) \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{\epsilon} \stackrel{\text{iid}}{\sim} \{\pm 1\}^n} \left[\sup_{\mathbf{u} \in S} \sum_{i=1}^{n} \epsilon_i \langle \mathbf{u}, \mathbf{x}_i \rangle \right]$$

Two equivalent representations of a convex function

Set of Points

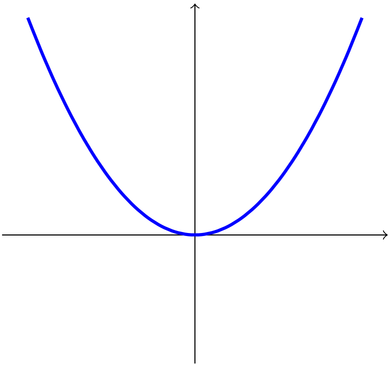


Set of Tangents

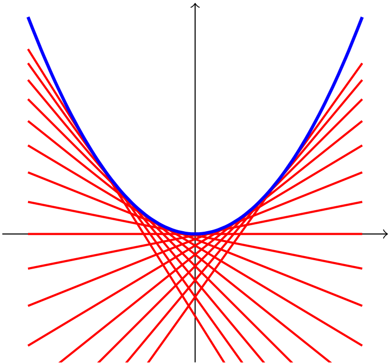


Two equivalent representations of a convex function

Set of Points

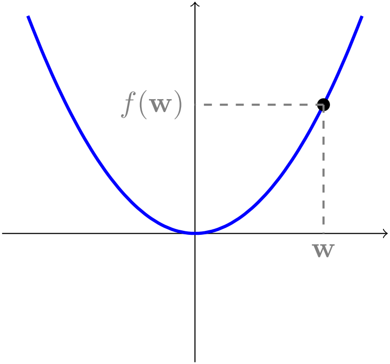


Set of Tangents

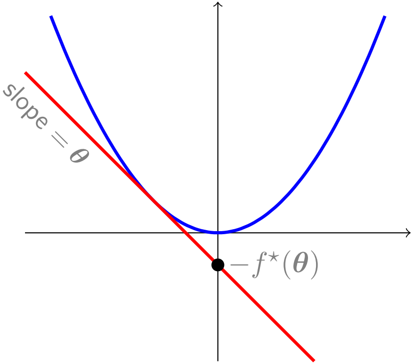


Two equivalent representations of a convex function

Point $(\mathbf{w}, f(\mathbf{w}))$



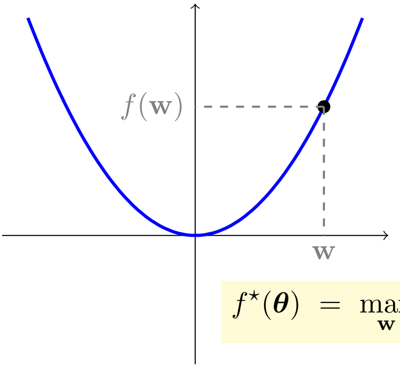
Tangent $(\boldsymbol{\theta}, -f^\star(\boldsymbol{\theta}))$



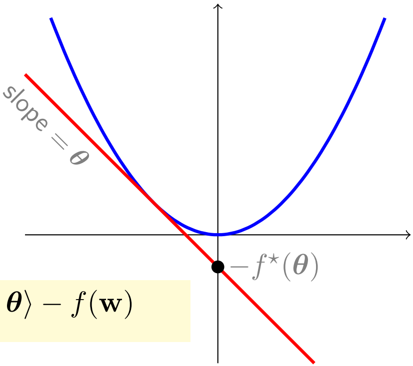
Background – Fenchel Conjugate

Two equivalent representations of a convex function

Point $(\mathbf{w}, f(\mathbf{w}))$



Tangent $(\boldsymbol{\theta}, -f^\star(\boldsymbol{\theta}))$



$$f^\star(\boldsymbol{\theta}) \;=\; \max_{\mathbf{w}} \; \langle \mathbf{w}, \boldsymbol{\theta} \rangle - f(\mathbf{w})$$

Background – Fenchel Conjugate

- The definition immediately implies Fenchel-Young inequality:

$$\forall \mathbf{u}, \qquad \begin{aligned} f^\star(\boldsymbol{\theta}) &= \max_{\mathbf{w}} \langle \mathbf{w}, \boldsymbol{\theta} \rangle - f(\mathbf{w}) \\ &\geq \langle \mathbf{u}, \boldsymbol{\theta} \rangle - f(\mathbf{u}) \end{aligned}$$

Background – Fenchel Conjugate

- The definition immediately implies Fenchel-Young inequality:

$$\forall \mathbf{u}, \qquad f^\star(\boldsymbol{\theta}) = \max_{\mathbf{w}} \langle \mathbf{w}, \boldsymbol{\theta} \rangle - f(\mathbf{w}) \geq \langle \mathbf{u}, \boldsymbol{\theta} \rangle - f(\mathbf{u})$$

- If f is closed and convex then $f^{\star\star} = f$

Background – Fenchel Conjugate

- The definition immediately implies Fenchel-Young inequality:

$$\begin{aligned} \forall \mathbf{u}, \qquad f^\star(\boldsymbol{\theta}) &= \max_{\mathbf{w}} \langle \mathbf{w}, \boldsymbol{\theta} \rangle - f(\mathbf{w}) \\ &\geq \langle \mathbf{u}, \boldsymbol{\theta} \rangle - f(\mathbf{u}) \end{aligned}$$

- If f is closed and convex then $f^{\star\star} = f$
- By the way, this implies Jensen's inequality:

$$\begin{aligned} f(\mathbb{E}[\mathbf{w}]) &= \max_{\boldsymbol{\theta}} \langle \boldsymbol{\theta}, \mathbb{E}[\mathbf{w}] \rangle - f^\star(\boldsymbol{\theta}) \\ &= \max_{\boldsymbol{\theta}} \mathbb{E}\left[\langle \boldsymbol{\theta}, \mathbf{w} \rangle - f^\star(\boldsymbol{\theta})\right] \\ &\leq \mathbb{E}[\ \max_{\boldsymbol{\theta}} \langle \boldsymbol{\theta}, \mathbf{w} \rangle - f^\star(\boldsymbol{\theta})\ ] = \mathbb{E}[f(\mathbf{w})] \end{aligned}$$

Background – Fenchel Conjugate

Examples:

$f(\mathbf{w})$	$f^\star(\boldsymbol{\theta})$
$\frac{1}{2}\|\mathbf{w}\|^2$	$\frac{1}{2}\|\boldsymbol{\theta}\|_\star^2$
$\|\mathbf{w}\|$	Indicator of unit $\|\cdot\|_\star$ ball
$\sum_i w_i \log(w_i)$	$\log\left(\sum_i e^{\theta_i}\right)$
Indicator of prob. simplex	$\max_i \theta_i$
$c\, g(\mathbf{w})$ for $c > 0$	$c\, g^\star(\boldsymbol{\theta}/c)$
$\inf_{\mathbf{x}} g_1(\mathbf{w}) + g_2(\mathbf{w} - \mathbf{x})$	$g_1^\star(\boldsymbol{\theta}) + g_2^\star(\boldsymbol{\theta})$

Examples:

	$f(\mathbf{w})$	$f^\star(\boldsymbol{\theta})$
	$\frac{1}{2}\|\mathbf{w}\|^2$	$\frac{1}{2}\|\boldsymbol{\theta}\|_\star^2$
	$\|\mathbf{w}\|$	Indicator of unit $\|\cdot\|_\star$ ball
$\Rightarrow$	$\sum_i w_i \log(w_i)$	$\log\left(\sum_i e^{\theta_i}\right)$
$\Rightarrow$	Indicator of prob. simplex	$\max_i \theta_i$
	$c\, g(\mathbf{w})$ for $c > 0$	$c\, g^\star(\boldsymbol{\theta}/c)$
	$\inf_{\mathbf{x}} g_1(\mathbf{w}) + g_2(\mathbf{w} - \mathbf{x})$	$g_1^\star(\boldsymbol{\theta}) + g_2^\star(\boldsymbol{\theta})$

(used for boosting)

Examples:

$f(\mathbf{w})$	$f^\star(\boldsymbol{\theta})$
$\frac{1}{2}\|\mathbf{w}\|^2$	$\frac{1}{2}\|\boldsymbol{\theta}\|_\star^2$
$\|\mathbf{w}\|$	Indicator of unit $\|\cdot\|_\star$ ball
$\sum_i w_i \log(w_i)$	$\log\left(\sum_i e^{\theta_i}\right)$
Indicator of prob. simplex	$\max_i \theta_i$
$c\,g(\mathbf{w})$ for $c > 0$	$c\,g^\star(\boldsymbol{\theta}/c)$
$\Rightarrow$ $\inf_{\mathbf{x}} g_1(\mathbf{w}) + g_2(\mathbf{w}-\mathbf{x})$	$g_1^\star(\boldsymbol{\theta}) + g_2^\star(\boldsymbol{\theta})$

(infimal convolution theorem)

f is strongly convex $\iff$ $f^\star$ is strongly smooth

The following properties are equivalent:

- $f(\mathbf{w})$ is σ-strongly convex w.r.t. $\|\cdot\|$
- $f^\star(\mathbf{w})$ is $\frac{1}{\sigma}$-strongly smooth w.r.t. $\|\cdot\|_\star$

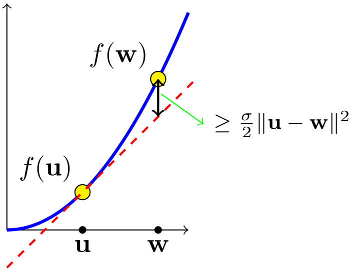
f is strongly convex $\iff$ $f^\star$ is strongly smooth

The following properties are equivalent:

- $f(\mathbf{w})$ is σ-strongly convex w.r.t. $\|\cdot\|$, that is

$$\forall \mathbf{w}, \mathbf{u}, \boldsymbol{\theta} \in \partial f(\mathbf{u}), \quad f(\mathbf{w}) - f(\mathbf{u}) - \langle \boldsymbol{\theta}, \mathbf{w} - \mathbf{u} \rangle \geq \frac{\sigma}{2} \|\mathbf{u} - \mathbf{w}\|^2 \ .$$

- $f^\star(\mathbf{w})$ is $\frac{1}{\sigma}$-strongly smooth w.r.t. $\|\cdot\|_\star$



f is strongly convex $\iff$ $f^\star$ is strongly smooth

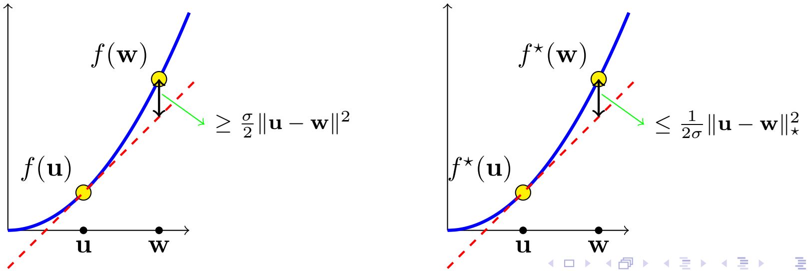
The following properties are equivalent:

- $f(\mathbf{w})$ is σ-strongly convex w.r.t. $\|\cdot\|$, that is

$$\forall \mathbf{w}, \mathbf{u}, \boldsymbol{\theta} \in \partial f(\mathbf{u}), \quad f(\mathbf{w}) - f(\mathbf{u}) - \langle \boldsymbol{\theta}, \mathbf{w} - \mathbf{u} \rangle \geq \frac{\sigma}{2} \|\mathbf{u} - \mathbf{w}\|^2 .$$

- $f^\star(\mathbf{w})$ is $\frac{1}{\sigma}$-strongly smooth w.r.t. $\|\cdot\|_\star$, that is

$$\forall \mathbf{w}, \mathbf{u}, \quad f^\star(\mathbf{w}) - f^\star(\mathbf{u}) - \langle \nabla f^\star(\mathbf{u}), \mathbf{w} - \mathbf{u} \rangle \leq \frac{1}{2\,\sigma} \|\mathbf{u} - \mathbf{w}\|_\star^2 .$$



f is strongly convex $\iff$ $f^\star$ is strongly smooth

Examples:

$f(\mathbf{w})$	$f^\star(\boldsymbol{\theta})$	w.r.t. norm	σ
$\frac{1}{2}\|\mathbf{w}\|_2^2$	$\frac{1}{2}\|\boldsymbol{\theta}\|_2^2$	$\|\cdot\|_2$	1

f is strongly convex $\iff$ $f^\star$ is strongly smooth

Examples:

$f(\mathbf{w})$	$f^\star(\boldsymbol{\theta})$	w.r.t. norm	σ
$\frac{1}{2}\|\mathbf{w}\|_2^2$	$\frac{1}{2}\|\boldsymbol{\theta}\|_2^2$	$\|\cdot\|_2$	1
$\frac{1}{2}\|\mathbf{w}\|_q^2$	$\frac{1}{2}\|\boldsymbol{\theta}\|_p^2$	$\|\cdot\|_q$	$(q-1)$

(where $q \in (1,2]$ and $\frac{1}{q} + \frac{1}{p} = 1$)

f is strongly convex $\iff$ $f^\star$ is strongly smooth

Examples:

$f(\mathbf{w})$	$f^\star(\boldsymbol{\theta})$	w.r.t. norm	σ
$\frac{1}{2}\|\mathbf{w}\|_2^2$	$\frac{1}{2}\|\boldsymbol{\theta}\|_2^2$	$\|\cdot\|_2$	1
$\frac{1}{2}\|\mathbf{w}\|_q^2$	$\frac{1}{2}\|\boldsymbol{\theta}\|_p^2$	$\|\cdot\|_q$	$(q-1)$
$\sum_i w_i \log(w_i)$	$\log\left(\sum_i e^{\theta_i}\right)$	$\|\cdot\|_1$	1

Importance

Theorem (1)

Let

- f *be* σ *strongly convex w.r.t.* $\|\cdot\|$
- *Assume* $f^\star(\mathbf{0}) = 0$ *(for simplicity)*
- $\mathbf{v}_1, \ldots, \mathbf{v}_n$ *be arbitrary sequence of vectors*
- *Denote* $\mathbf{w}_t = \nabla f^\star(\sum_{j<t} \mathbf{v}_j)$

Then, for any $\mathbf{u}$ *we have*

$$\sum_t \langle \mathbf{u}, \mathbf{v}_t \rangle - f(\mathbf{u}) \le f^\star(\sum_t \mathbf{v}_t) \le \sum_t \left(\langle \mathbf{w}_t, \mathbf{v}_t \rangle + \tfrac{1}{2\,\sigma} \|\mathbf{v}_t\|_\star^2 \right) \ .$$

Importance

Theorem (1)

Let

- f *be* σ *strongly convex w.r.t.* $\|\cdot\|$
- *Assume* $f^\star(\mathbf{0}) = 0$ *(for simplicity)*
- $\mathbf{v}_1, \ldots, \mathbf{v}_n$ *be arbitrary sequence of vectors*
- *Denote* $\mathbf{w}_t = \nabla f^\star(\sum_{j<t} \mathbf{v}_j)$

Then, for any $\mathbf{u}$ *we have*

$$\sum_t \langle \mathbf{u}, \mathbf{v}_t \rangle - f(\mathbf{u}) \le f^\star\Big(\sum_t \mathbf{v}_t\Big) \le \sum_t \left(\langle \mathbf{w}_t, \mathbf{v}_t \rangle + \tfrac{1}{2\,\sigma} \|\mathbf{v}_t\|_\star^2 \right) \ .$$

Proof.

The first inequality is Fenchel-Young and the second inequality follows from the $\frac{1}{\sigma}$ smoothness of $f^\star$ by induction. □

Back to Rademacher Complexities

- Theorem 1:

$$\sum_t \langle \mathbf{u}, \mathbf{v}_t \rangle - f(\mathbf{u}) \;\leq\; \sum_t \left(\langle \mathbf{w}_t, \mathbf{v}_t \rangle + \tfrac{1}{2\,\sigma} \|\mathbf{v}_t\|_\star^2 \right) \;.$$

Based on Kakade, Sridharan, Tewari [2008]

Back to Rademacher Complexities

- Theorem 1:

$$\sum_t \langle \mathbf{u}, \mathbf{v}_t \rangle - f(\mathbf{u}) \;\leq\; \sum_t \left(\langle \mathbf{w}_t, \mathbf{v}_t \rangle + \tfrac{1}{2\,\sigma} \|\mathbf{v}_t\|_\star^2 \right) \;.$$

- Therefore, for all S:

$$\sup_{\mathbf{u} \in S} \sum_t \langle \mathbf{u}, \mathbf{v}_t \rangle \;\leq\; \tfrac{1}{2\,\sigma} \sum_t \|\mathbf{v}_t\|_\star^2 \; + \sup_{\mathbf{u} \in S} f(\mathbf{u}) + \sum_t \langle \mathbf{w}_t, \mathbf{v}_t \rangle$$

Based on Kakade, Sridharan, Tewari [2008]

Back to Rademacher Complexities

- Theorem 1:

$$\sum_t \langle \mathbf{u}, \mathbf{v}_t \rangle - f(\mathbf{u}) \;\leq\; \sum_t \left(\langle \mathbf{w}_t, \mathbf{v}_t \rangle + \tfrac{1}{2\,\sigma} \|\mathbf{v}_t\|_\star^2 \right) \;.$$

- Therefore, for all S:

$$\sup_{\mathbf{u}\in S} \sum_t \langle \mathbf{u}, \mathbf{v}_t \rangle \;\leq\; \tfrac{1}{2\,\sigma} \sum_t \|\mathbf{v}_t\|_\star^2 \;+ \sup_{\mathbf{u}\in S} f(\mathbf{u}) + \sum_t \langle \mathbf{w}_t, \mathbf{v}_t \rangle$$

- Applying with $\mathbf{v}_t = \epsilon_t \mathbf{x}_t$ and taking expectation we obtain:

$$\mathcal{R}_n(S) \;\leq\; \tfrac{1}{2\,\sigma} \sum_t \mathbb{E}[\epsilon_t^2] \,\|\mathbf{x}_t\|_\star^2 \;+ \sup_{\mathbf{u}\in S} f(\mathbf{u}) + \underbrace{\mathbb{E}\left[\sum_t \langle \mathbf{w}_t, \epsilon_t\, \mathbf{x}_t \rangle \right]}_{=0}$$

Based on Kakade, Sridharan, Tewari [2008]

Rademacher Bounds – Examples

S	$f(\mathbf{w})$	X	$R_n(S)$
$\{\mathbf{w} : \|\mathbf{w}\|_2 \le W\}$	$\frac{\sigma}{2}\|\mathbf{w}\|_2^2$	$\frac{\sum_i \|\mathbf{x}_i\|_2^2}{n}$	$X\,W\,\sqrt{n}$

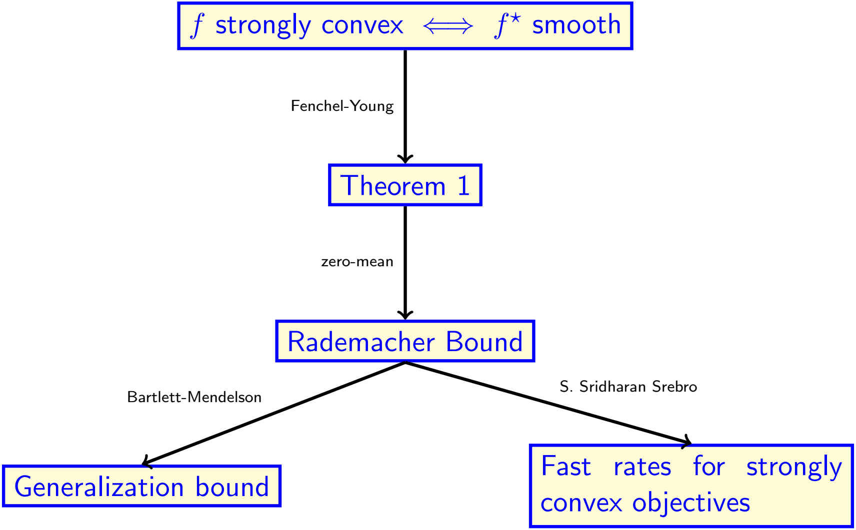
Rademacher Bounds – Examples

S	$f(\mathbf{w})$	X	$R_n(S)$
$\{\mathbf{w} : \|\mathbf{w}\|_2 \leq W\}$	$\frac{\sigma}{2}\|\mathbf{w}\|_2^2$	$\frac{\sum_i \|\mathbf{x}_i\|_2^2}{n}$	$X\,W\,\sqrt{n}$
$\{\mathbf{w} : \|\mathbf{w}\|_q \leq W\}$	$\frac{\sigma}{2}\|\mathbf{w}\|_q^2$	$\frac{\sum_i \|\mathbf{x}_i\|_p^2}{n}$	$X\,W\,\sqrt{(p-1)\,n}$

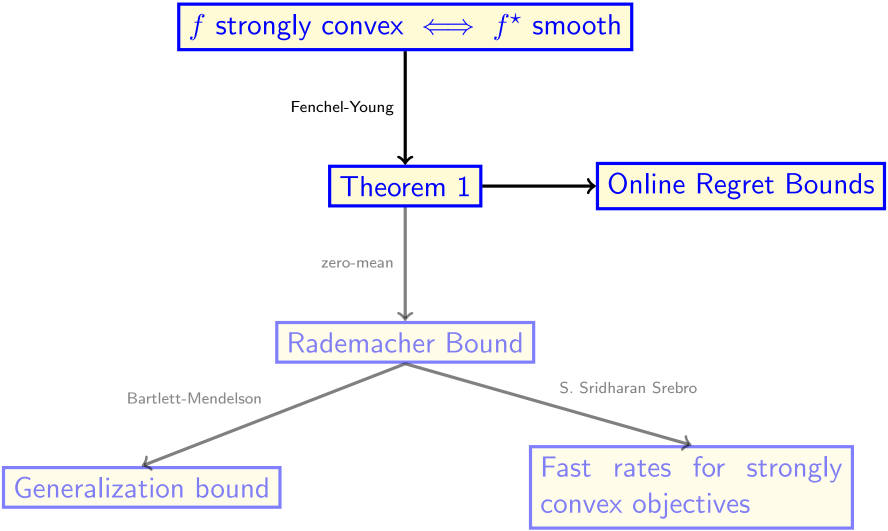
Rademacher Bounds – Examples

S	$f(\mathbf{w})$	X	$R_n(S)$
$\{\mathbf{w} : \|\mathbf{w}\|_2 \leq W\}$	$\frac{\sigma}{2}\|\mathbf{w}\|_2^2$	$\frac{\sum_i \|\mathbf{x}_i\|_2^2}{n}$	$X\,W\,\sqrt{n}$
$\{\mathbf{w} : \|\mathbf{w}\|_q \leq W\}$	$\frac{\sigma}{2}\|\mathbf{w}\|_q^2$	$\frac{\sum_i \|\mathbf{x}_i\|_p^2}{n}$	$X\,W\,\sqrt{(p-1)\,n}$
Prob. simplex	$\sigma \sum_i w_i \log(d\,w_i)$	$\frac{\sum_i \|\mathbf{x}_i\|_\infty^2}{n}$	$X\,\sqrt{\log(d)\,n}$

Intermediate Summary



Coming Next ...



f strongly convex $\iff$ $f^\star$ smooth
Fenchel-Young
Theorem 1
Online Regret Bounds
zero-mean
Rademacher Bound
Bartlett-Mendelson
S. Sridharan Srebro
Generalization bound
Fast rates for strongly convex objectives

Online Learning – Brief Background

- Studied in game theory, information theory, and machine learning
- Examples:
  - Repeated 2-players games (Hannan [57], Blackwell [56])
  - Predicting with side information (Rosenblatt's Perceptron [58], Weighted Majority of Littlestone and Warmuth [88,94])
  - Predicting of individual sequences (Cover [78], Feder, Merhav and Gutman [92])
- Online convex optimization – a general abstract prediction model (Gordon [99], Zinkevich [03])
- Using our lemma, we can easily derived optimal low regret algorithms

Online Learning

Prediction Game – Online Optimization

For $t = 1, \ldots, n$

- Learner chooses a decision $\mathbf{w}_t \in S$
- Environment chooses a loss function $\ell_t : S \to \mathbb{R}$
- Learner pays loss $\ell_t(\mathbf{w}_t)$

- Regret of learner for not always following the best decision in S

$$\sum_{t=1}^{n} \ell_t(\mathbf{w}_t) - \min_{\mathbf{u} \in S} \sum_{t=1}^{n} \ell_t(\mathbf{u})$$

- Goal: Conditions on S and loss functions that guarantee low regret learning strategy

Online Learning

- Assume f σ-strongly convex on S w.r.t. $\|\cdot\|$
- Recall Theorem 1: For $\mathbf{w}_t = \nabla f^\star(\sum_{j<t} \mathbf{v}_j)$ we have

$$\sum_t \langle \mathbf{u} - \mathbf{w}_t, \mathbf{v}_t \rangle \leq f(\mathbf{u}) + \tfrac{1}{2\sigma} \sum_t \|\mathbf{v}_t\|_\star^2$$

- Assume ℓ_t convex and apply with $\mathbf{v}_t \in \partial \ell_t(\mathbf{w}_t)$, thus

$$\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u}) \leq \langle \mathbf{u} - \mathbf{w}_t, \mathbf{v}_t \rangle$$

- Assume ℓ_t Lipschitz w.r.t. dual norm, thus $\|\mathbf{v}_t\|_\star \leq V$
- We obtain the regret bound (S. and Singer [06]):

$$\sum_{t=1}^n \ell_t(\mathbf{w}_t) - \min_{\mathbf{u} \in S} \sum_{t=1}^n \ell_t(\mathbf{u}) \ \leq \ \max_{\mathbf{u} \in S} f(\mathbf{u}) + \frac{n\,V^2}{2\sigma} \ .$$

Online Learning – Example I

Predicting the next bit of a sequence

For $t = 1, \ldots, n$

- Learner predict $\hat{y}_t \in \{0, 1\}$
- Environment responds with $y_t \in \{0, 1\}$
- Learner pays 1 if $\hat{y}_t \neq y_t$

Modeling:

- $S = [0, 1]$, $f(w) = \frac{\sigma}{2} w^2$, $\sigma = \sqrt{n}$
- Predict $\hat{y}_t = 1$ with probability $w_t \in S$
- Then, probability of $\hat{y}_t \neq y_t$ is $\ell_t(w_t) = |y_t - w_t|$, which is convex
- The expected regret is thus bounded by $\sqrt{n}$

Online Learning – Example II

Predicting with expert advice

For $t = 1, \ldots, n$

- Learner receives a vector $\mathbf{x}_t \in [0,1]^d$ of experts advice
- Learner need to predict $\hat{y}_t \in \{0,1\}$
- Environment responds with $y_t \in \{0,1\}$
- Learner pays 1 if $\hat{y}_t \neq y_t$

Modeling:

- S is d dimensional probability simplex, $f(\mathbf{w}) = \sigma \sum_i w_i \log(w_i)$, $\sigma = \sqrt{n/\log(d)}$
- Predict $\hat{y}_t = 1$ with probability $\langle \mathbf{w}_t, \mathbf{x}_t \rangle$
- Then, probability of $\hat{y}_t \neq y_t$ is $\ell_t(w_t) = |y_t - \langle w_t, \mathbf{x}_t \rangle|$, which is convex
- The expected regret is thus bounded by $\sqrt{\log(d)\, n}$

Optimization from a Machine Learning Perspective

- Assume we'd like to solve regularized loss minimization:

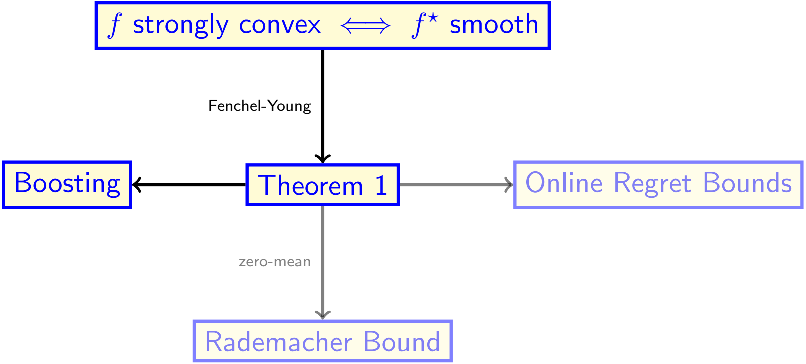
$$\min_{\mathbf{w}} \; f(\mathbf{w}) + \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{z}_i)$$

- Stochastic Mirror Descent
  - At each step, sample i uniformly at random and feed an online learner the loss $\ell_t(\mathbf{w}) = \ell(\mathbf{w}, \mathbf{z}_i)$
  - Return averaged $\mathbf{w}_t$ of the online learner
- Number of iterations required to achieve accuracy ϵ is order of sample complexity !
- Optimality (S. and Srebro [08])

Intermediate Summary

$$
\begin{tikzcd}
\boxed{f \text{ strongly convex} \iff f^\star \text{ smooth}} \arrow[d, "\text{Fenchel-Young}"'] & \\
\boxed{\text{Theorem 1}} \arrow[r] \arrow[d, "\text{zero-mean}"'] & \boxed{\text{Online Regret Bounds}} \\
\boxed{\text{Rademacher Bound}} &
\end{tikzcd}
$$

Coming Next ...


f strongly convex $\iff$ $f^\star$ smooth
Fenchel-Young
Boosting
Theorem 1
Online Regret Bounds
zero-mean
Rademacher Bound

Boosting

Input:

- Training set of examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
- d weak hypotheses $h_1, \ldots, h_d$

Output:

- Strong hypothesis: $H_{\mathbf{w}}(\cdot) = \sum_{i=1}^{d} w_i h_i(\cdot)$

Weak Learnability Assumption

- For any probability $\mathbf{p} \in \mathbb{S}^m$ over examples
- Exists h_j with edge at least γ,

$$\sum_i p_i y_i h_j(\mathbf{x}_i) = \Pr[h_j = y] - \Pr[h_j \neq y] \;\geq\; \gamma$$

Deriving Boosting Algorithm from Theorem 1

Goal: Find $\mathbf{w}$ s.t. $\min_i y_i H_{\mathbf{w}}(\mathbf{x}_i) > 0$.

- Equivalently: Find $\mathbf{w}, \boldsymbol{\mu}$ s.t. $\mu_i = y_i H_{\mathbf{w}}(\mathbf{x}_i)$ and $\min_i \mu_i > 0$
- Define: $L(\boldsymbol{\mu}) = \log\left(\frac{1}{m}\sum_i \exp(-\mu_i)\right)$ (1-smooth w.r.t. $\|\cdot\|_\infty$)
- Observe: $L(\boldsymbol{\mu}) \le -\log(m) \;\Rightarrow\; \min_i \mu_i > 0$
- Recall from Theorem 1: $L(\boldsymbol{\mu}_n) \le \sum_t \left(\langle \nabla L(\boldsymbol{\mu}_t), \mathbf{v}_t\rangle + \frac{1}{2}\|\mathbf{v}_t\|_\infty^2\right)$
- Observe: $\mathbf{p} \stackrel{\text{def}}{=} \nabla L(\boldsymbol{\mu}_t) \in \mathbb{S}^m$.
- Weak learnability $\Rightarrow$ exists r_t s.t. $\sum_i p_i y_i h_{r_t}(\mathbf{x}_i) \ge \gamma$
- Apply Theorem 1 with $v_{t,i} = -\gamma y_i h_{r_t}(\mathbf{x}_i)$ gives $L(\boldsymbol{\mu}_n) \le -\frac{n\gamma^2}{2}$
- Therefore, $n \ge \frac{2\log(m)}{\gamma^2} \;\Rightarrow\; \min_i \mu_{n,i} > 0$

Boosting – Brief history

Is weak learnability equivalent to strong learnability ?

Boosting – Brief history

Is weak learnability equivalent to strong learnability ?

Yes!

Boosting – Brief history

Is weak learnability equivalent to strong learnability ?

Yes!

You can use AdaBoost

Boosting – Brief history

Is weak learnability equivalent to strong learnability ?

Yes!

You can use AdaBoost

Boosting is related to margin

Boosting – Brief history

Is weak learnability equivalent to strong learnability ?

Yes!

You can use AdaBoost

Boosting is related to margin

Of course, it is a corollary of the minimax theorem

Weak Learnability = Separability with ℓ_1 margin

$$A = \begin{pmatrix} y_1\, h_1(\mathbf{x}_1) & \dots & y_1\, h_d(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ y_m\, h_1(\mathbf{x}_m) & \dots & y_m\, h_d(\mathbf{x}_m) \end{pmatrix}$$

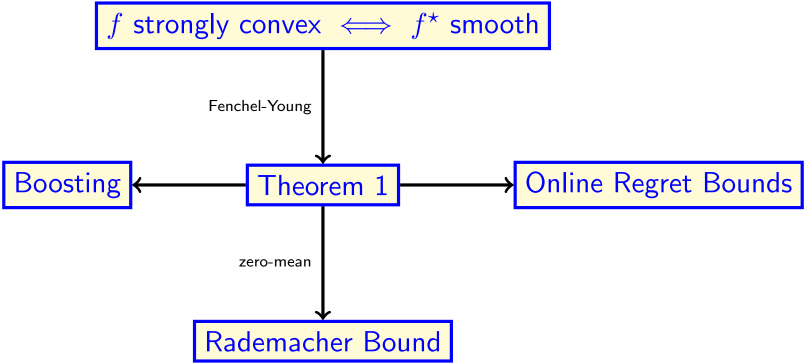
Minimax theorem

$$\max_{\mathbf{w} \in \mathbb{S}^d} \underbrace{\min_i \, (A\,\mathbf{w})_i}_{\text{margin}} \;=\; \gamma \;=\; \underbrace{\min_{\mathbf{p} \in \mathbb{S}^m} \; \max_j \, (\mathbf{p}^T A)_j}_{\gamma \text{ Weak learnability}}$$

Reinterpreting Boosting Result

	Assumption	#iterations	runtime
AdaBoost	ℓ_1 margin γ	$\frac{\log(m)}{\gamma^2}$	$\frac{m\ \log(m)\,d}{\gamma^2}$
Perceptron	ℓ_2 margin γ	$\frac{d}{\gamma^2}$	$\frac{d^2}{\gamma^2}$
Winnow	ℓ_1 margin γ	$\frac{\log(d)}{\gamma^2}$	$\frac{d\ \log(d)}{\gamma^2}$

Summary


f strongly convex $\iff$ $f^\star$ smooth
Fenchel-Young
Boosting
Theorem 1
Online Regret Bounds
zero-mean
Rademacher Bound

More Applications

Sparsification

- Theorem: For smooth loss functions, any low ℓ_1 linear predictor can be converted into sparse linear predictor
- Proof idea: definition of smoothness + probabilistic construction
- Theorem: Also true for non-smooth but Lipschitz loss functions
- Proof idea: infimal-convolution + our main lemma $\Rightarrow$ it's possible to approximate any Lipschitz function by a smooth function

More Applications

Sparsification

- Theorem: For smooth loss functions, any low ℓ_1 linear predictor can be converted into sparse linear predictor
- Proof idea: definition of smoothness + probabilistic construction
- Theorem: Also true for non-smooth but Lipschitz loss functions
- Proof idea: infimal-convolution + our main lemma $\Rightarrow$ it's possible to approximate any Lipschitz function by a smooth function

Concentration Inequalities

- Pinelis-like concentration results for martingales in Banach spaces

More Applications – Matrix Regularization

- Lemma: The matrix function $F(A) = f(\sigma(A))$, where f is strongly convex w.r.t. $\|\mathbf{w}\|$, is strongly convex w.r.t. $\|\sigma(A)\|$
- Corollaries:
    - Generalization bounds for multi-task learning
    - Regret bounds for multi-task learning

More Applications – Matrix Regularization

- Lemma: The matrix function $F(A) = f(\sigma(A))$, where f is strongly convex w.r.t. $\|\mathbf{w}\|$, is strongly convex w.r.t. $\|\sigma(A)\|$
- Corollaries:
  - Generalization bounds for multi-task learning
  - Regret bounds for multi-task learning

- Lemma: The matrix function $F(A) = \|\ (\|A_{1,\cdot}\|_2, \ldots, \|A_{m,\cdot}\|_2)\ \|_q^2$ is strongly convex w.r.t. the matrix norm $\|\ (\|A_{1,\cdot}\|_2, \ldots, \|A_{m,\cdot}\|_2)\ \|_q$
- Corollaries:
  - Generalization bounds for group Lasso, kernel learning, multi-task learning
  - Regret bounds for the above and also shifting regret bounds

Summary

Lemma

f is strongly convex w.r.t. $\|\cdot\|$ $\iff$ $f^\star$ is strongly smooth w.r.t. $\|\cdot\|_\star$

- Isolating a single useful property of regularization functions
- Deriving many known result easily based on this property
- Good theory should also predict new results – we derived new algorithms and bounds from the generalized theory