Learnability, Stability and Strong Convexity

Nati Srebro

Shai Shalev-Shwartz
HUJI

Ohad Shamir
Weizmann

Karthik Sridharan
Cornell

Ambuj Tewari
Michigan

Toyota Technological Institute—Chicago (2008-2011)

Outline

- Theme: Role of Stability in Learning
- Story: Necessary and sufficient condition for learnability
- Characterizing (statistical) learnability
  - Stability as the master property
- Convex Problems
  - Strong convexity as the master property
- Stability in online learning
  - From Stability to Online Mirror Descent

The General Learning Setting

Vapnik95

aka Stochastic Optimization

$$\min_{w \in \mathcal{W}} F(w) = E_{z \sim \mathcal{D}}[f(w, z)]$$

given an iid sample $z_1, z_2, \ldots, z_m \sim \mathcal{D}$

- Known objective function $f: \mathcal{W} \times \Omega \to \mathbb{R}$,
  unknown distribution $\mathcal{D}$ over $Z \in \Omega$

- Problem specified by $\mathcal{W}, \Omega, f$ is ***learnable*** if there exists a learning rule $\tilde{w}(z_1, \ldots, z_m)$ s.t. for every $\epsilon > 0$ and large enough sample size $m(\epsilon)$, for any distribution $\mathcal{D}$:

$$\mathbb{E}_{z_1, \ldots, z_m \sim \mathcal{D}}[F(\tilde{w})] \leq \underbrace{\inf_{w \in \mathcal{W}} F(w)}_{F(w^*)} + \epsilon$$

General Learning: Examples

Minimize $F(w)=E_z[f(w;z)]$ based on sample $z_1,z_2,\ldots,z_n$

- Supervised learning:

  $z = (x,y)$

  w specifies a perdictor $h_w: \mathcal{X} \to \mathcal{Y}$

  $f(\ w\ ;\ (x,y)\ ) = loss(h_w(x),y)$

  e.g. linear prediction: $f(w\ ;(x,y)\ ) = loss(\langle w,x\rangle,y)$

- Unsupervised learning, e.g. k-means clustering:

  $\theta = x \in \mathbb{R}^d$

  $w = (\mu[1],\ldots,\mu[k]) \in \mathbb{R}^{d\times k}$ specifies k cluster centers

  $f(\ (\mu[1],\ldots,\mu[k])\ ;\ x\ ) = \min_j |\mu[j]-x|^2$

- Density estimation:

  w specifies probability density $p_w(x)$

  $f(\ w\ ;\ x\ ) = -\log p_w(x)$

- Optimization in uncertain environment, e.g.:

  z = traffic delays on each road segment

  w = route chosen (indicator over road segments in route)

  $f(\ w\ ;\ z\ ) = \langle w,z\rangle$ = total delay along route

{ h_w | $w \in W$ } has finite fat-shattering dimension

Uniform convergence: $\sup_{\mathbf{w}\in\mathbf{W}} \left| F(\mathbf{w}) - \hat{F}(\mathbf{w}) \right| \xrightarrow{n\to\infty} 0$

Learnable using ERM:

$\hat{\mathbf{w}} = \arg\min \hat{F}(\mathbf{w})$

$F(\hat{\mathbf{w}}) \xrightarrow{n\to\infty} F(\mathbf{w}^\star)$

$$\hat{F}(w) = \frac{1}{m}\sum_i f(w, z_i)$$

$$\hat{w} = \arg\min_w \hat{F}(w)$$

Supervised Classification

$f(w;(x,y)) = \text{loss}(h_w(x),y)$:

$\{ h_w \mid w \in W \}$ has finite fat-shattering dimension

Uniform convergence: $\sup_{\mathbf{w}\in\mathbf{W}} \left| F(\mathbf{w}) - \hat{F}(\mathbf{w}) \right| \stackrel{n\to\infty}{\longrightarrow} 0$

Learnable using ERM:

$\hat{\mathbf{w}} = \arg\min \hat{F}(\mathbf{w})$

$F(\hat{\mathbf{w}}) \stackrel{n\to\infty}{\longrightarrow} F(\mathbf{w}^\star)$

Learnable (using some rule): $F(\tilde{\mathbf{w}}) \stackrel{n\to\infty}{\longrightarrow} F(\mathbf{w}^\star)$

[Alon Ben-David Cesa-Bianchi Haussler 93]

Beyond Supervised Learning

- Supervised learning:

$$f\big(w,(x,y)\big) = loss(h_w(x),y)$$

  - Combinatorial necessary and sufficient condition of learning
  - Uniform convergence necessary and sufficient for learning
  - ERM universal (if learnable, can do it with ERM)

- General learning / stochastic optimization:

$$f(w,z)$$

????

Online Learning (Optimization)

Adversary: $f(\cdot;z_1)$ $f(\cdot;z_2)$ $f(\cdot;z_3)$

Learner: w_1 w_2 w_3 ….

$$w_1 \nearrow f(\cdot;z_1) \searrow w_2 \nearrow f(\cdot;z_2) \searrow w_3 \nearrow f(\cdot;z_3) \searrow \ldots$$

- Known function $f(\cdot,\cdot)$
- Unknown sequence $z_1, z_2, \ldots$
- Online learning rule: $w_i(z_1, \ldots, z_{i-1})$
- Goal: $\sum_i f(w_i, z_i)$

Differences vs stochastic setting:

- Any sequence—not necessarily iid
- No distinction between “train” and “test”

Online and Stochastic Regret

- Online Regret: for any sequence,

$$\frac{1}{m}\sum_{i=1}^{m} f(w_i(z_1,\ldots,z_{i-1}),z_i) \le \underbrace{\inf_{w\in\mathcal{W}} \frac{1}{m}\sum_{i=1}^{m} f(w,z_i)}_{\hat{F}(\hat{w})} + Reg(m)$$

- Statistical Regret: for any distribution $\mathcal{D}$,

$$\mathbb{E}_{z_1,\ldots,z_m\sim\mathcal{D}}\left[F_{\mathcal{D}}\left(\tilde{w}(z_1,\ldots,z_m)\right)\right] \le \underbrace{\inf_{w\in\mathcal{W}} F_{\mathcal{D}}(w)}_{F(w^*)} + \epsilon(m)$$

- Online-To-Batch:

$$\tilde{w}(z_1,\ldots,z_m) = w_i \text{ with prob } 1/m$$

$$\mathbb{E}[F(\tilde{w})] \le F(w^*) + Reg(m)$$

Supervised Classification

f(w;(x,y)) = loss(h_w(x),y):

{ h_w | w $\in$ W } has finite fat-shattering dimension

Uniform convergence: $\sup_{\mathbf{w}\in\mathbf{W}} \left| F(\mathbf{w}) - \hat{F}(\mathbf{w}) \right| \xrightarrow{n\to\infty} 0$

Learnable using ERM: $F(\hat{\mathbf{w}}) \xrightarrow{n\to\infty} F(\mathbf{w}^\star)$

$\hat{\mathbf{w}} = \arg\min \hat{F}(\mathbf{w})$

Online Learnable

Learnable (using some rule): $F(\tilde{\mathbf{w}}) \xrightarrow{n\to\infty} F(\mathbf{w}^\star)$

[Alon Ben-David Cesa-Bianchi Haussler 93]

Convex Lipschitz Problems

- $\mathcal{W}$ convex bounded subset of Hilbert space (or $\mathbb{R}^d$)
$$\forall_{w\in\mathcal{W}} \|w\|_2 \leq \boldsymbol{B}$$
- For each z, $f(w,z)$ convex Lipschitz w.r.t w
$$|f(w,z) - f(w',z)| \leq \boldsymbol{L} \cdot \|w - w'\|_2$$
- E.g., $f\big(w,(x,y)\big) = loss(\langle w,x\rangle; y)$, $|loss'| \leq 1$
$$\|x\|_2 \leq L$$
- Online Gradient Descent: $Reg(m) \leq \sqrt{\frac{B^2L^2}{m}}$
- Stochastic Setting:
  - For generalized linear (including supervised): matches ERM rate
  - For general Convex Lipschitz Problems?
    - Learnable via online-to-batch (SGD)
    - Using ERM?

Center of Mass with Missing Data

$$f\big(w,(I,x_I)\big) = \sum_{i\in I}(w[i]-x[i])^2$$

$w \in \mathbb{R}^d, \|w\| \leq 1$

$I \subseteq [d], x[i], i \in I, \|x\| \leq 1$

Consider $P(i \in I) = 1/2$ independently for all i, $x = 0$

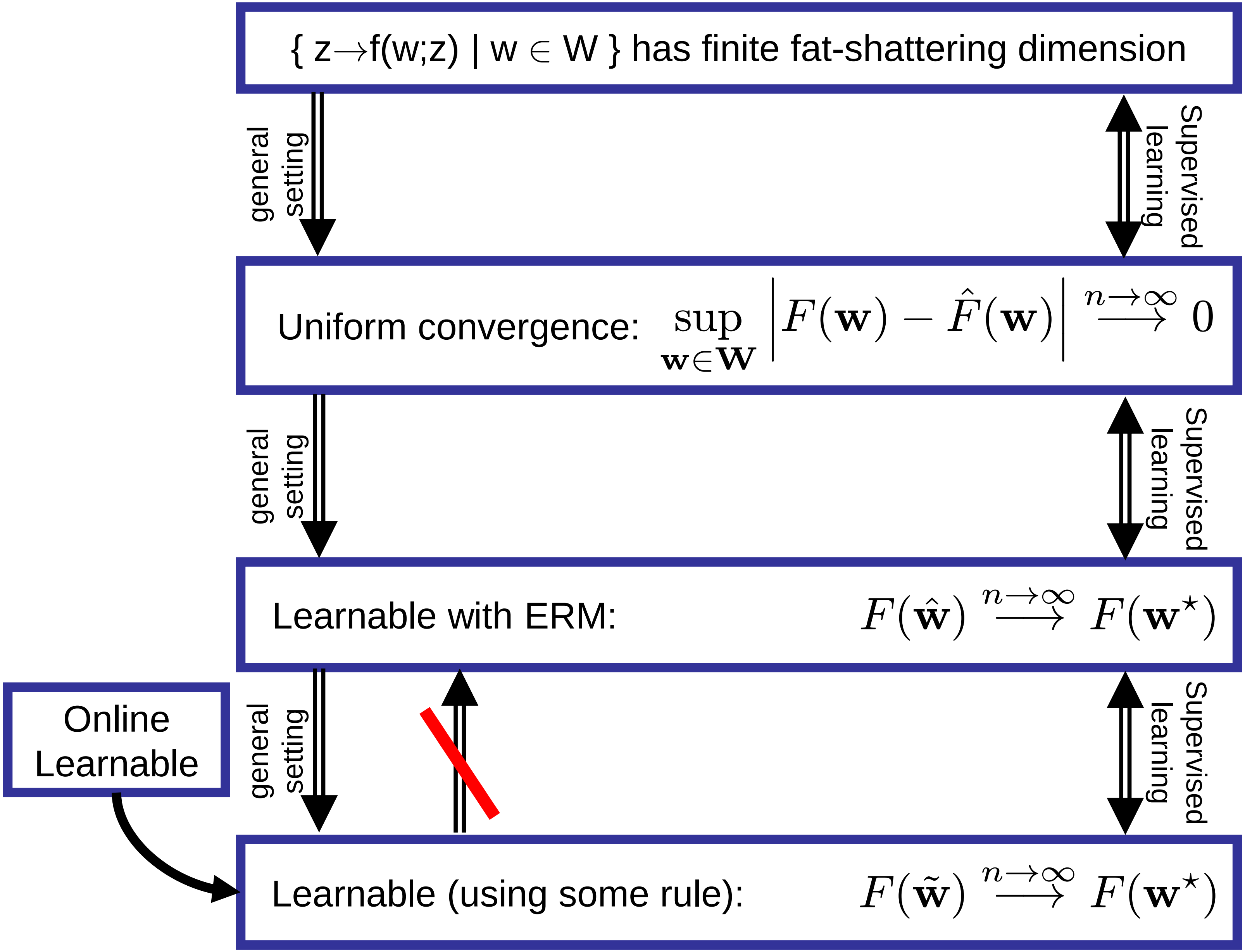
If d>>2^m (think of d=∞) then with high probability there exists a coordinate j that is never seen in the sample, i.e. $j \notin I$ for all i=1..m

$$\hat{F}(\mathbf{e}_j) = 0 \qquad\qquad F(\mathbf{e}_j) = 1/2$$

$$\sup_{\mathbf{w}\in\mathbf{W}} \left|F(\mathbf{w}) - \hat{F}(\mathbf{w})\right| \geq 1/2$$

No uniform convergence!

e_j is an empirical minimizer with $F(e_j)$ = ½, far from $F(w^*)=F(0)=0$



{ z→f(w;z) | w ∈ W } has finite fat-shattering dimension
general setting
Supervised learning
Uniform convergence: $\sup_{\mathbf{w}\in\mathbf{W}} \left| F(\mathbf{w}) - \hat{F}(\mathbf{w}) \right| \xrightarrow{n\to\infty} 0$
general setting
Supervised learning
Learnable with ERM: $F(\hat{\mathbf{w}}) \xrightarrow{n\to\infty} F(\mathbf{w}^\star)$
general setting
Supervised learning
Online Learnable
Learnable (using some rule): $F(\tilde{\mathbf{w}}) \xrightarrow{n\to\infty} F(\mathbf{w}^\star)$

Stochastic Convex Optimization

- Empirical minimization might not be consistent
- Learnable using specific procedural rule (online-to-batch conversion of online gradient descent)
- ??????????

Strongly Convex Objectives

$f(w,z)$ is λ-strongly convex in w iff:

$$f\left(\frac{w+w'}{2},z\right) \le \frac{f(w,z)+f(w',z)}{2} - \frac{\lambda}{8}\|w-w'\|_2^2$$

Equivalent to $\nabla_w^2 f(w,z) \succcurlyeq \lambda$

If $f(w,z)$ is λ-convex and L-Lipschitz w.r.t. w

- Online Gradient Descent [Hazan Kalai Kale Agarwal 2006]

$$Reg \le O\left(\frac{L^2 \log(m)}{\lambda m}\right)$$

- Empirical Risk Minimization:

Stochastic Setting?

$$\mathbb{E}[F(\hat{w})] \le F(w^*) + O\left(\frac{L^2}{\lambda m}\right)$$

ERM?

Strong Convexity and Stability

- Definition: rule $\tilde{w}(z_1, \ldots z_m)$ is $\beta(m)$-*stable* if:
  $$|f(\tilde{w}(z_1, \ldots, z_{m-1}), z_m) - f(\tilde{w}(z_1, \ldots, z_m), z_m)| \le \beta(m)$$

- Symmetric $\tilde{w}$ is β−stable $\Rightarrow \mathbb{E}[F(\tilde{w}_{m-1})] \le \mathbb{E}\left[\hat{F}(\tilde{w}_m)\right] + \beta$

  For ERM: $\mathbb{E}\left[\hat{F}(\hat{w})\right] \le \mathbb{E}\left[\hat{F}(w^*)\right] = F(w^*)$

- f is λ-strongly convex and L-Lipschitz $\Rightarrow$
  $$|f(\hat{w}(z_1, \ldots, z_{m-1}), z_m) - f(\hat{w}(z_1, \ldots, z_m), z_m)| \le \beta = \frac{4L^2}{\lambda m}$$

- Conclusion:
  $$\mathbb{E}[F(\hat{w})] \le \beta(m)$$

Empirical Minimization Consistent, but is there Uniform Convergence?

$$f\big(w,(I,x_I)\big) = \sum_{i\in I}(w[i]-x[i])^2 + \lambda\|w\|_2^2$$

$w \in \mathbb{R}^d, \|w\| \leq 1$

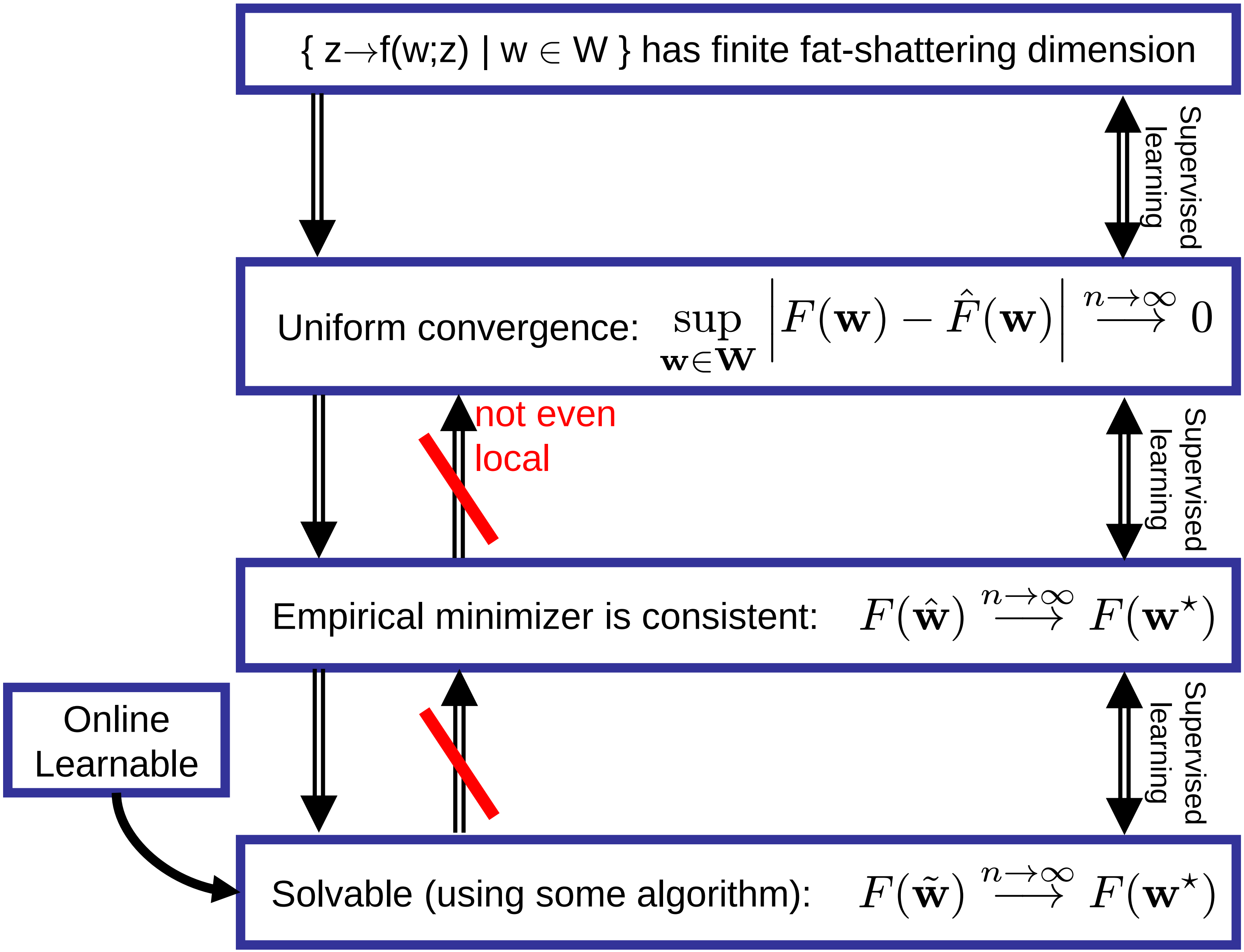
$I \subseteq [d], x[i], i \in I, \|x\| \leq 1$

Consider $P(i \in I) = 1/2$ independently for all i, $x = 0$

For j that is never seen in the sample:

$$\hat{F}\big(te_j\big) = \lambda t^2 \qquad F\big(te_j\big) = \frac{1}{2}t + \lambda t^2$$

No uniform convergence: $\sup_{\mathbf{w}\in\mathbf{W}} \left|F(\mathbf{w}) - \hat{F}(\mathbf{w})\right| \geq 1/2$


{ z→f(w;z) | w ∈ W } has finite fat-shattering dimension
Supervised learning
Uniform convergence: $\sup_{\mathbf{w}\in\mathbf{W}} \left| F(\mathbf{w}) - \hat{F}(\mathbf{w}) \right| \xrightarrow{n\to\infty} 0$
not even local
Supervised learning
Empirical minimizer is consistent: $F(\hat{\mathbf{w}}) \xrightarrow{n\to\infty} F(\mathbf{w}^\star)$
Supervised learning
Online Learnable
Solvable (using some algorithm): $F(\tilde{\mathbf{w}}) \xrightarrow{n\to\infty} F(\mathbf{w}^\star)$

Back to Weak Convexity

$f(w,z)$ L-Lipschitz (and convex), $\|w\|_2 \le B$

- Use Regularized ERM:

$$\hat{w}_\lambda = \arg\min_{w \in \mathcal{W}} \hat{F}(w) + \frac{\lambda}{2}\|w\|_2^2$$

- Setting $\lambda = \sqrt{\frac{L^2}{B^2 m}}$:

$$\mathbb{E}[F(\hat{w}_\lambda)] \le F(w^*) + O\left(\sqrt{\frac{L^2 B^2}{m}}\right)$$

- Key: strongly convex regularizer ensures ***stability***

The Role of Regularization

- Structure Risk Minimization view:
  - Adding regularization term effectively constrains domain to lower complexity domain $\mathcal{W}_r = \{w \mid \|w\| \leq r\}$
  - Learning guarantees (e.g. for SVMs, LASSO) are actually for empirical minimization inside $\mathcal{W}_r$, and are based on uniform convergence in $\mathcal{W}_r$.

- In our case:
  - No uniform convergence in $\mathcal{W}_r$, for any r>0
  - No uniform convergence even of regularized loss
  - Cannot solve stochastic optimization problem by restricting to $\mathcal{W}_r$, for any r.
  - What regularization buys is **stability**

Stability Characterizes Learnability

Theorem: Learnable with (symmetric) ERM $\hat{w}$ iff $\forall\mathcal{D}$

$$\mathbb{E}[|f(\hat{w}(z_1,\ldots,z_{m-1}),z_m) - f(\hat{w}(z_1,\ldots,z_m),z_m)|] \leq \beta(m)$$

For some $\beta(m) \to 0$

Theorem: Learnable iff $\exists$ symetric $\tilde{w}$ s.t. $\forall\mathcal{D}$:

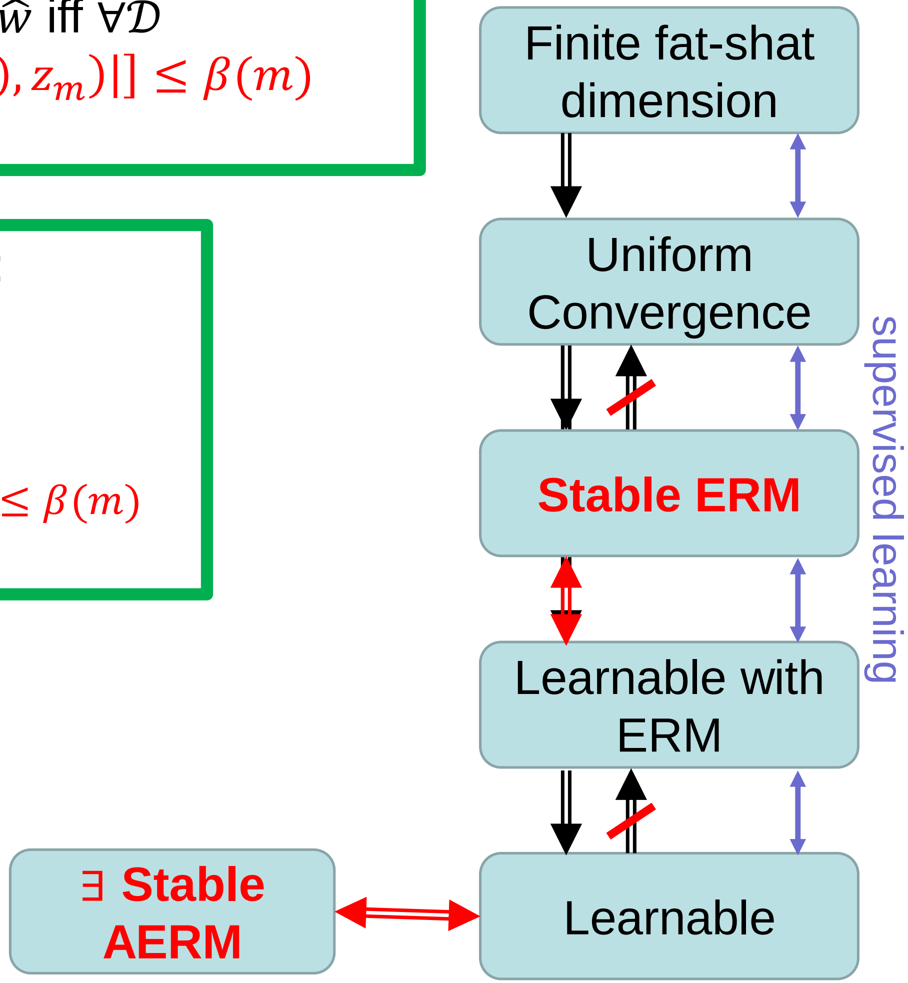
- $\tilde{w}$ is an "**almost ERM**":

$$\mathbb{E}\left[\hat{F}(\tilde{w}) - \hat{F}(\hat{w})\right] \leq \epsilon(m)$$

- $\tilde{w}$ is **stable**:

$$|\mathbb{E}[f(\hat{w}(z_1,\ldots,z_{m-1}),z_m) - f(\hat{w}(z_1,\ldots,z_m),z_m)]| \leq \beta(m)$$

For some $\epsilon(m) \to 0$, $\beta(m) \to 0$



Strong Convexity and Stability

- For any norm $\|w\|$:
  - $\Psi(w) \geq 0$ is strongly convex w.r.t. $\|w\|$, i.e.
  $$\Psi\left(\frac{w+w'}{2}\right) \geq \frac{\Psi(w)+\Psi(w')}{2} + \frac{1}{4}\|w\|^2$$
  - $f(w,z)$ is L-Lipschitz w.r.t. $\|w\|$:
  $$|f(w,z) - f(w',z)| \leq L \cdot \|w - w'\|$$

  ➔ $\hat{w}_\lambda = \arg\min_w \hat{F}(w) + \frac{\lambda}{2}\Psi(w)$ is $\frac{L^2\lambda}{m}$-stable

- With $\lambda = \sqrt{L^2/(\Psi(w^*)m)}$:
$$F(\hat{w}_\lambda) \leq F(w^*) + \sqrt{\frac{L^2\Psi(w^*)}{m}}$$

Convex Lipschitz Problems

- $\mathcal{W}$ convex bounded subset of normed space ($\mathbb{R}^d$ or Banach space)

- For each z, $f(w,z)$ convex Lipschitz w.r.t w
$$|f(w,z) - f(w',z)| \le \boldsymbol{L} \cdot \|w - w'\|$$

- E.g., $f\big(w,(x,y)\big) = loss(\langle w,x\rangle; y)$, $|loss'| \le 1$
$$\|x\|_* \le L$$

- To learn: need $\Psi(w)$ strongly convex w.r.t. $\|\cdot\|$

$$F(\hat{w}_\lambda) \le F(w^*) + \sqrt{\frac{L^2B^2}{m}} \qquad B^2 = \sup_{w\in\mathcal{W}} \Psi(w)$$

- Is this universal?
Can all Lipschitz problems (for all $\|\cdot\|$ and $\mathcal{W}$) be learned this way?

Stability in Online Learning

- Reminder: rule $\widetilde{w}(z_1, \dots z_m)$ is $\beta(m)$-*stable* if
$$|f(\widetilde{w}(z_1, \dots, z_{m-1}), z_m) - f(\widetilde{w}(z_1, \dots, z_m), z_m)| \le \beta(m)$$

- Follow The Leader (FTL): $\widehat{w}_m(z_1, \dots, z_{m-1}) = \arg\min_w \sum_{i=1}^{m-1} f(w, z_i)$

- Be The Leader (BTL): $w_m(z_1, \dots, z_{m-1}) = \arg\min_w \sum_{i=1}^{m} f(w, z_i)$

- If the ERM is $\beta(m)$-stable: $Reg_{FTL}(m) \le \underbrace{Reg_{BTL}(m)}_{\le 0} + \frac{1}{m}\sum_i \beta(i) \le \frac{1}{m}\sum_i \beta(i)$

- Follow The Regularized Leader (FTRL):
$$w_m(z_1, \dots, z_{m-1}) = \arg\min_w \sum_{i=1}^{m-1} f(w, z_i) + \lambda\Psi(w)$$

- If f is L-Lipschitz and Ψ strongly conv. w.r.t. $\|\cdot\|$: $Reg_{FTRL}(m) \le \sqrt{\frac{L^2 \sup \Psi(w)}{m}}$

Strong Convexity is Necessary and Sufficient

- Theorem: If a Convex Lipschitz problem (for some $\|\cdot\|$ and some convex $\mathcal{W}$) can be online learned with regret $\sqrt{\frac{L^2B^2}{m}}$, then there exists $\Psi(w)$ strongly convex w.r.t. $\|\cdot\|$ s.t. $\sup_{w\in\mathcal{W}} \Psi(w) \leq cB^2$

- More generally: For any problem, Follow The Regularized Leader with some Ψ achieves the optimal online regret (up to a constant factor), and this can be established via stability

From FTRL to Mirror Descent

- Linearized problem: $\tilde{f}_i(w) \stackrel{\text{def}}{=} f(w_i, z_i) + \langle \nabla f(w_i, z_i), w - w_i \rangle$
- Main observation: for convex f, $(Regret\ on\ f) \leq (Regret\ on\ \tilde{f})$
- Follow the Linearized Regularized Leader (aka Mirror Descent):

$$w_m = \arg\min_w \sum_{i=1}^{m-1} \langle \nabla f(w_i, z_i), w \rangle + \lambda \Psi(w)$$

$$= \nabla \Psi^{-1}\left(\nabla \Psi(w_{m-1}) - \frac{1}{\lambda} \nabla f(w_{m-1}, z_{m-1}) \right)$$

$$Reg_{MD}(m) \leq \sqrt{\frac{L^2 \sup \Psi(w)}{m}}$$

- Conclusion: Any Convex Lipschitz problem (for any $\mathcal{W}$ and $\|\cdot\|$) that is online learnable, is (optimally) learnable with this approach

Strong Convexity as the Master Property