Generalization bounds for uniformly stable algorithms

Vitaly Feldman

Google Brain

with Jan Vondrak

Stanford University

Uniform stability

- Domain Z (e.g. $X \times Y$)
- Dataset $S = (z_1, \ldots, z_n) \in Z^n$
- Learning algorithm $A: Z^n \to W$
- Loss function $\ell: W \times Z \to \mathbb{R}^+$

Uniform stability **[Bousquet,Elisseeff 02]**:

A has uniform stability γ w.r.t. ℓ if

For all neighboring $S, S', z \in Z$

$$|\ell(A(S), z) - \ell(A(S'), z)| \leq \gamma$$

Generalization error

- Probability distribution P over Z, $S \sim P^n$
- Population loss: $\mathbf{E}_P[\ell(w)] = \mathbf{E}_{z\sim P}[\ell(w,z)]$
- Empirical loss:

$$\boldsymbol{\varepsilon}_S[\ell(w)] = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$$

- Generalization error/gap:

$$\Delta_S\big(\ell(A)\big) = \mathbf{E}_P\big[\ell\big(A(S)\big)\big] - \boldsymbol{\varepsilon}_S\big[\ell\big(A(S)\big)\big]$$

Stochastic convex optimization

- $W = \mathbb{B}_2^d(1) \doteq \{w \mid \|w\|_2 \leq 1\}$
- For all $z \in Z$, $\ell(w, z)$ is convex 1-Lipschitz in w
- Minimize $\mathbf{E}_P[\ell(w)] \doteq \mathbf{E}_{z\sim P}[\ell(w, z)]$ over W

$$L^* = \min_{w\in W} \mathbf{E}_P[\ell(w)]$$

For all P, A being SGD with rate $\eta = 1/\sqrt{n}$:

$$\Pr_{S\sim P^n}\left[\mathbf{E}_P[\ell(A(S))] \geq L^* + O\left(\frac{\sqrt{\log(1/\delta)}}{\sqrt{n}}\right)\right] \leq \delta$$

Uniform convergence error: $\gtrsim \sqrt{\frac{d}{n}}$

ERM might have generalization error: $\gtrsim \frac{d}{n}$

[Shalev-Shwartz,Shamir,Srebro,Sridharan '09; Feldman 16]

Stable optimization

Strongly convex ERM [BE 02, SSSS 09]

$$A_\lambda(S) = \text{argmin}_{w \in W} \left\{ \boldsymbol{\mathcal{E}}_S[\ell(w)] + \frac{\lambda}{2} \|w\|_2^2 \right\}$$

A_λ is $\frac{1}{\lambda n}$ uniformly stable and minimizes $\boldsymbol{\mathcal{E}}_S[\ell(w)]$ within $\frac{\lambda}{2}$

Gradient descent on smooth losses [Hardt,Recht,Singer 16]

$A_T(S)$: T steps of GD on $\boldsymbol{\mathcal{E}}_S[\ell(w)]$ with $\eta = \frac{1}{\sqrt{T}}$

A_T is $\frac{\sqrt{T}}{n}$ uniformly stable and minimizes $\boldsymbol{\mathcal{E}}_S[\ell(w)]$ within $\frac{2}{\sqrt{T}}$

Generalization bounds

For A, ℓ w/ range $[0,1]$ and uniform stability $\gamma \in \left[\frac{1}{n}, 1\right]$ $\left(\frac{1}{\sqrt{n}}\right)$

$$\left|\mathop{\mathbf{E}}_{S\sim P^n}[\Delta_S(\ell(A))]\right| \leq \gamma$$

[Rogers,Wagner 78]

$$\mathop{\mathbf{Pr}}_{S\sim P^n}\left[\Delta_S(\ell(A)) \geq \gamma\sqrt{\boldsymbol{n}\log(1/\delta)}\right] \leq \delta$$

[Bousquet,Elisseeff 02]

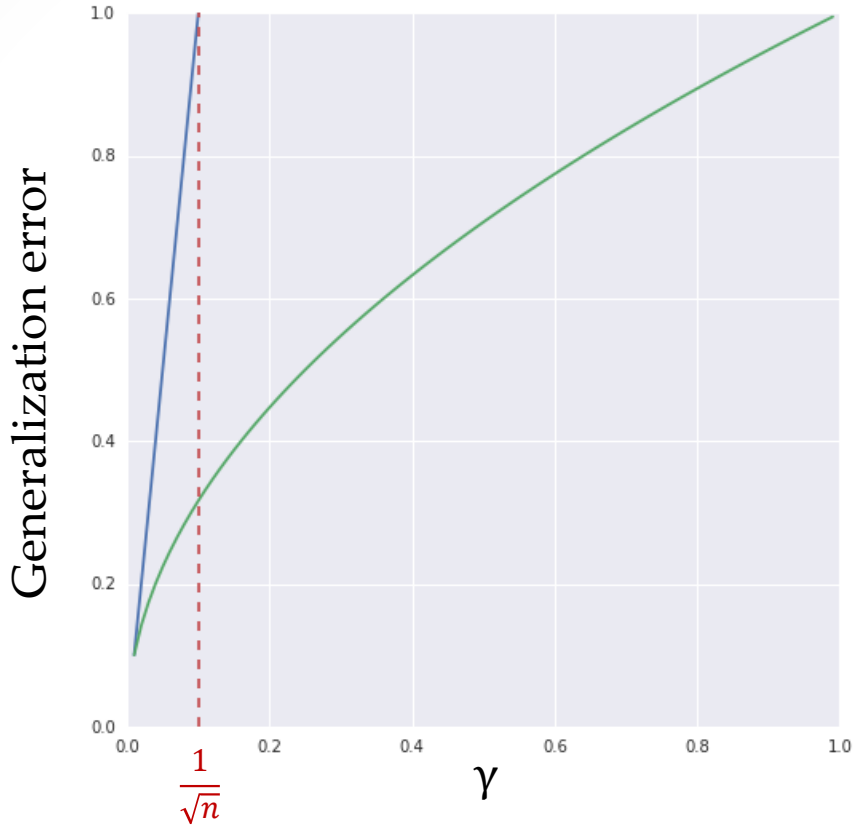
Vacuous when $\gamma \geq 1/\sqrt{n}$

NEW!

$$\mathop{\mathbf{Pr}}_{S\sim P^n}\left[\Delta_S(\ell(A)) \geq \sqrt{\gamma\log(1/\delta)}\right] \leq \delta$$

Comparison

$n = 100$



[BE 02]

This work

Second moment

NEW!

$$\mathop{\mathbf{E}}_{S\sim P^n}[\Delta_S(\ell(A))^2] \le \gamma^2 + \frac{1}{n}$$

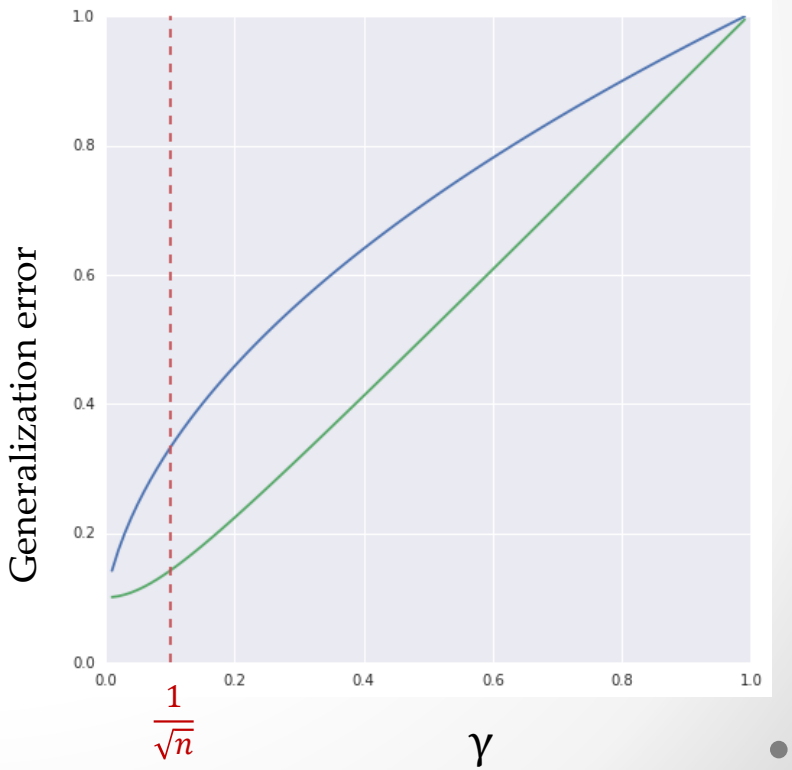
TIGHT!

Chebyshev: $\mathop{\mathbf{Pr}}_{S\sim P^n}\left[\Delta_S(\ell(A)) \ge \frac{\gamma+1/\sqrt{n}}{\sqrt{\delta}}\right] \le \delta$

Previously

[Devroye,Wagner 79; BE 02]

$$\mathop{\mathbf{E}}_{S\sim P^n}[\Delta_S(\ell(A))^2] \le \gamma + \frac{1}{n}$$



Implications

$$\Pr_{S\sim P^n}\left[\mathbf{E}_P[\ell(A(S))] \geq L^* + O\left(\frac{1}{\delta^{1/4}\sqrt{n}}\right)\right] \leq \delta$$

$$\Pr_{S\sim P^n}\left[\mathbf{E}_P[\ell(A(S))] \geq L^* + O\left(\frac{\sqrt{\log(1/\delta)}}{n^{1/3}}\right)\right] \leq \delta$$

Differentially-private prediction **[Dwork,Feldman 18]**

$Z = X \times Y$. A randomized algorithm $A(S, x)$ has ϵ-DP prediction if for all $S, S', x \in X$

$$D_\infty\big(A(S,x)||A(S',x)\big) \leq \epsilon$$

For any loss $L: Y \times Y \to [0,1]$,

$\mathbf{E}_A[L(A(S,x),y)]$ has uniform stability $e^\epsilon - 1 \approx \epsilon$

Stronger generalization bounds for DP prediction algorithms

Data-dependent functions

Consider M: $Z^n \times Z \to \mathbb{R}$. E.g. $M(S,z) \equiv \ell(A(S),z)$

M has uniform stability γ if

For all neighboring $S, S', z \in Z$

$$|M(S,z) - M(S',z)| \leq \gamma$$

$$\|M(S,\cdot) - M(S',\cdot)\|_\infty \leq \gamma$$

Generalization error/gap:

$$\Delta_S(M) = \mathbf{E}_P[M(S)] - \boldsymbol{\mathcal{E}}_S[M(S)]$$

Generalization in expectation

Goal: $\left| \mathop{\mathbf{E}}_{S\sim P^n}\left[\mathbf{E}_P[M(S)] - \boldsymbol{\varepsilon}_S[M(S)]\right]\right| \le \gamma$

$$\mathop{\mathbf{E}}_{S\sim P^n}\left[\boldsymbol{\varepsilon}_S[M(S)]\right] = \mathop{\mathbf{E}}_{S\sim P^n, i\sim[n]}[M(S, z_i)]$$

For all i and $z \in Z$, $M(S, z_i) \le M\left(S^{i\leftarrow z}, z_i\right) + \gamma$

$$M(S, z_i) \le \mathop{\mathbf{E}}_{z\sim P}\left[M\left(S^{i\leftarrow z}, z_i\right)\right] + \gamma$$

$$\mathop{\mathbf{E}}_{S\sim P^n, i\sim[n]}[M(S, z_i)] \le \mathop{\mathbf{E}}_{S\sim P^n, z\sim P, i\sim[n]}\left[M\left(S^{i\leftarrow z}, z_i\right)\right] + \gamma$$

$$= \mathop{\mathbf{E}}_{S\sim P^n, z\sim P}[M(S, z)] + \gamma$$

$$= \mathop{\mathbf{E}}_{S\sim P^n}\left[\mathbf{E}_P[M(S)]\right] + \gamma$$

$\left(S^{i\leftarrow z}, z_i\right) \equiv P^{n+1} \equiv (S, z)$

Concentration via McDiarmid

For all neighboring S, S'

$$|\Delta_S(M) - \Delta_{S'}(M)| \leq 2\gamma + \frac{1}{n}$$

1. $\left| \mathop{\mathbf{E}}_{z\sim P}[M(S,z)] - \mathop{\mathbf{E}}_{z\sim P}[M(S',z)] \right| \leq \gamma$
2. $|\boldsymbol{\mathcal{E}}_S[M(S)] - \boldsymbol{\mathcal{E}}_{S'}[M(S',z)]|$
   $\leq |\boldsymbol{\mathcal{E}}_S[M(S)] - \boldsymbol{\mathcal{E}}_S[M(S',z)]| + |\boldsymbol{\mathcal{E}}_S[M(S')] - \boldsymbol{\mathcal{E}}_{S'}[M(S',z)]|$
   $\leq \gamma + \frac{1}{n}$

McDiarmid: $\mathop{\mathbf{Pr}}_{S\sim P^n}\left[\Delta_S(M) \geq \mu + \left(2\gamma + \frac{1}{n}\right)\sqrt{\boldsymbol{n}\log(1/\delta)}\right] \leq \delta$

where $\mu = \mathop{\mathbf{E}}_{S\sim P^n}[\Delta_S(M)] \leq \gamma$

Proof technique

Based on **[Nissim,Stemmer 15; BNSSSU 16]**

Let Q be a distribution over $\mathbb{R}$:

$$\Pr_{v\sim Q}\left[v \geq 2 \mathop{\mathbf{E}}_{v_1,\ldots,v_m\sim Q}[\max\{0, v_1, \ldots, v_m\}]\right] \leq \frac{\ln 2}{m}$$

Let $S_1, \ldots, S_m \sim P^n$ for $m = \frac{\ln 2}{\delta}$

Need to bound $\mathop{\mathbf{E}}_{S_1,\ldots,S_m\sim P^n}\left[\max_{j\in[m]}\left\{\Delta_{S_j}(M)\right\}\right]$

$$\mathop{\mathbf{E}}_{S_1,\ldots,S_m\sim P^n,\ \ell=\operatorname{argmax}_j\{\Delta_{S_j}(M)\}}\left[\mathcal{E}_{S_\ell}[M(S_\ell)]\right]$$

$$\approx ?$$

$$\mathop{\mathbf{E}}_{S_1,\ldots,S_m\sim P^n,\ \ell=\operatorname{argmax}_j\{\Delta_{S_j}(M)\}}\left[\mathbf{E}_P[M(S_\ell)]\right]$$

UNSTABLE!

Stable max

Exponential mechanism **[McSherry,Talwar 07]**

EM_α: sample $j \propto e^{\alpha \Delta_{S_j}(M)}$

1. Stable: $2\alpha\left(2\gamma + \frac{1}{n}\right)$- differentially private

$$\mathop{\mathbf{E}}_{S_1,\ldots,S_m \sim P^n, \ell = \mathrm{EM}_\alpha}\left[\varepsilon_{S_\ell}[M(S_\ell)]\right] \leq$$

$$\mathop{\mathbf{E}}_{S_1,\ldots,S_m \sim P^n, \ell = \mathrm{EM}_\alpha}\left[\mathbf{E}_P[M(S_\ell)]\right] + \exp\left(2\alpha\left(2\gamma + \frac{1}{n}\right)\right) - 1 + \gamma$$

2. Approximates max

$$\mathop{\mathbf{E}}_{S_1,\ldots,S_m \sim P^n, \ell = \mathrm{EM}_\alpha}\left[\Delta_{S_\ell}(M)\right] \geq \mathop{\mathbf{E}}_{S_1,\ldots,S_m \sim P^n}\left[\max_{j \in [m]}\left\{\Delta_{S_j}(M)\right\}\right] - \frac{\ln m}{\alpha}$$

Game over

$$\mathop{\mathbf{E}}_{S_1,\ldots,S_m \sim P^n}\left[\max_{j\in[m]}\left\{\Delta_{S_j}(M)\right\}\right]$$
$$\leq \frac{\ln m}{\alpha} + \exp\left(2\alpha\left(2\gamma + \frac{1}{n}\right)\right) - 1 + \gamma$$

Pick $\alpha = \sqrt{\frac{\ln m}{\gamma + 1/n}}$ get $O\left(\sqrt{\left(\gamma + \frac{1}{n}\right)\ln\left(\frac{1}{\delta}\right)}\right)$

Let Q be a distribution over $\mathbb{R}$:

$$\Pr_{v\sim Q}\left[v \geq 2 \mathop{\mathbf{E}}_{v_1,\ldots,v_m \sim Q}[\max\{0, v_1, \ldots, v_m\}]\right] \leq \frac{\ln 2}{m}$$

Conclusions

- Better understanding of uniform stability
- New technique
- Open
  - Gap between upper and lower bounds
  - High probability generalization without strong uniformity