“Optimistic” Rates

Nati Srebro

Based on work with Karthik Sridharan and Ambuj Tewari

Examples based on work with Andy Cotter, Elad Hazan, Tomer Koren, Percy Liang, Shai Shalev-Shwartz, Ohad Shamir, Karthik Sridharan

Outline

- What?
- When?

(How?)

- Why?

Estimating the Bias of a Coin

$$|p - \hat{p}| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

$$|p - \hat{p}| \leq \sqrt{\frac{2p \log(2/\delta)}{n}} + \frac{2 \log(2/\delta)}{3n}$$

Optimistic VC bound
(aka L^*-bound, multiplicative bound)

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{L}(h) \qquad L^* = \inf_{h \in \mathcal{H}} L(h)$$

- For a hypothesis class with VC-dim *D*, w.p. 1-δ over *n* samples:

$$L(\hat{h}) \leq L^* + \sqrt{8 \frac{D \log \frac{2en}{D} + \log \frac{2}{\delta}}{n}}$$

Optimistic VC bound
(aka L^*-bound, multiplicative bound)

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \hat{L}(h) \qquad L^* = \inf_{h\in\mathcal{H}} L(h)$$

- For a hypothesis class with VC-dim *D*, w.p. 1-δ over *n* samples:

$$L(\hat{h}) \le L^* + 2\sqrt{L^* \frac{D\log\frac{2en}{D} + \log\frac{2}{\delta}}{n}} + 4\frac{D\log\frac{2en}{D} + \log\frac{2}{\delta}}{n}$$

$$= \inf_{\alpha} \ (1+\alpha)L^* + (1+\tfrac{1}{\alpha})4\frac{D\log\frac{2en}{D} + \log\frac{2}{\delta}}{n}$$

- Sample complexity to get $L(h) \le L^* + \epsilon$:

$$n(\epsilon) = O\left(\frac{D}{\epsilon}\cdot\frac{L^*+\epsilon}{\epsilon}\log(1/\epsilon)\right) = \tilde{O}\left(\frac{D}{\epsilon}\cdot\frac{L^*+\epsilon}{\epsilon}\right)$$

- Extends to bounded real-valued loss, *D*=VC subgraph dim

From Parametric to Scale Sensitive Classes

$$L(h) = \mathbf{E}_{x,y}\left[\phi(h(x), y)\right]$$

- Instead of VC-dim or VC-subgraph-dim (≈ #params), rely on metric scale to control complexity, e.g.:

$$\mathcal{H} = \{h_{\mathbf{w}} : \mathbf{w} \to \langle \mathbf{w}, \mathbf{x} \rangle \mid \|\mathbf{w}\|_2 \le B\}$$

$$\mathcal{R}_n(\mathcal{H}) = \sqrt{\frac{\overset{R}{\overset{\shortparallel}{B^2 \sup \|\mathbf{x}\|^2}}}{n}}$$

- Learning depends on:
  - Metric complexity measures: fat shattering dimension, covering numbers, Rademacher Complexity
  - Scale sensitivity of loss ϕ (bound on derivatives or "margin")
- For $\mathcal{H}$ with Rademacher Complexity $\mathcal{R}_n$, and $|\phi'| \le G$:

$$L(\hat{h}) \le L^* + 2G\mathcal{R}_n + \sqrt{\frac{\log(2/\delta)}{2n}}$$

$$\mathcal{R}_n \le \sqrt{\frac{R}{n}}$$

$$\le L^* + O\left(\sqrt{\frac{G^2 R + \log(2/\delta)}{n}}\right)$$

Non-Parametric Optimistic Rate for Smooth Loss

- **Theorem:** for any $\mathcal{H}$ with (worst case) Rademacher Complexity $\mathcal{R}_n(\mathcal{H})$, and any smooth loss with $|\phi''| \leq H$, $|\phi| \leq b$, w.p. $1-\delta$ over *n* samples: **[S Sridharan Tewari 2010]**

$$L(\hat{h}) \leq \inf_{\alpha} \ (1+\alpha)L^* + (1+\tfrac{1}{\alpha})K\left(H\mathcal{R}_n^2 \log^3(n) + \frac{b\log(1/\delta)}{n}\right)$$

$$= L^* + \tilde{O}\left(\sqrt{L^* H}\mathcal{R}_n + H\mathcal{R}_n^2\right)$$

$\mathcal{R}_n \leq \sqrt{\frac{R}{n}}$

$$= L^* + \tilde{O}\left(\sqrt{\frac{L^* H R}{n}} + \frac{HR}{n}\right)$$

- Sample complexity

$$n(\epsilon) = O\left(\frac{R}{\epsilon} \cdot \frac{L^* + \epsilon}{\epsilon} \log^3(R/\epsilon)\right) = \tilde{O}\left(\frac{R}{\epsilon} \cdot \frac{L^* + \epsilon}{\epsilon}\right)$$

Proof Ideas

- Smooth functions are self bounding: for any *H*-smooth non-negative *f*:

$$|f'(t)| \le \sqrt{4Hf(t)}$$

- 2nd order version of Lipschitz composition Lemma, restricted to predictors with low loss:

$$\mathcal{R}_n(\mathcal{L}(r)) \le 21\sqrt{6Hr}\log^{3/2}(64n)\mathcal{R}_n(\mathcal{H})$$

$$\mathcal{L}(r) = \left\{ (x,y) \mapsto \phi(h(x),y) \;\middle|\; h \in \mathcal{H}, \hat{L}(h) \le r \right\}$$

Rademacher → fat shattering → L_∞ covering → (compose with loss and use smoothness) → L_2 covering → Rademacher

- Local Rademacher analysis

Non-Parametric Optimistic Rate for Smooth Loss

- **Theorem:** for any $\mathcal{H}$ with (worst case) Rademacher Complexity $\mathcal{R}_n(\mathcal{H})$, and any **smooth loss** with $|\phi''| \le H$, $|\phi| \le b$, w.p. $1-\delta$ over n samples: **[S Sridharan Tewari 2010]**

$$L(\hat{h}) \le \inf_{\alpha} \; (1+\alpha)L^* + (1+\tfrac{1}{\alpha})K\left(H\mathcal{R}_n^2 \log^3(n) + \frac{b\log(1/\delta)}{n}\right)$$

$$= L^* + \tilde{O}\left(\sqrt{L^* H}\mathcal{R}_n + H\mathcal{R}_n^2\right)$$

$$\mathcal{R}_n \le \sqrt{\frac{R}{n}}$$

$$= L^* + \tilde{O}\left(\sqrt{\frac{L^* H R}{n}} + \frac{HR}{n}\right)$$

- Sample complexity

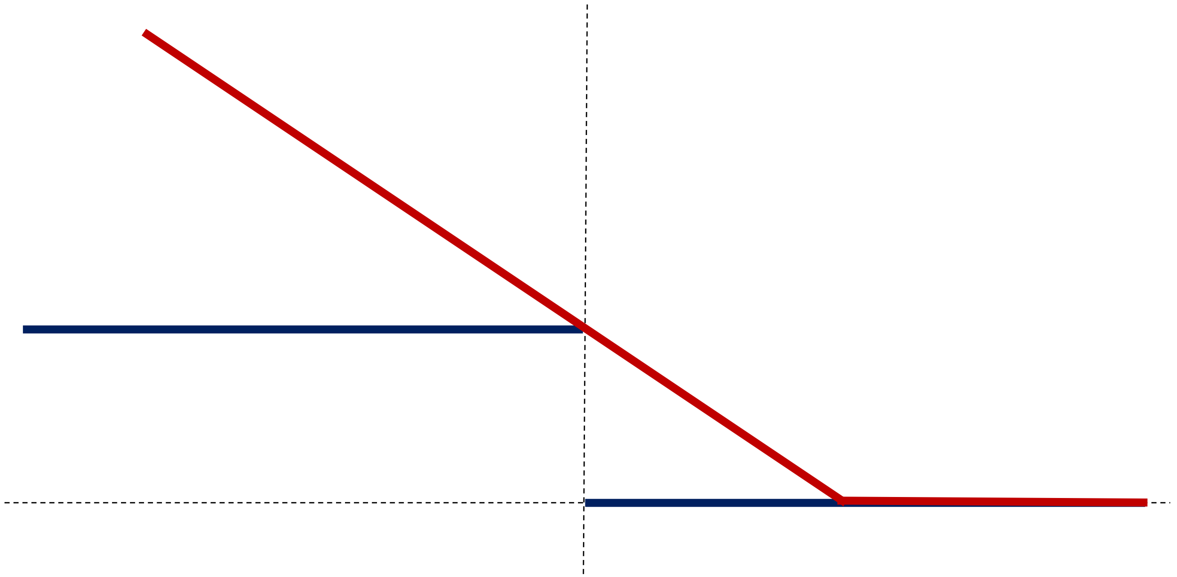
$$n(\epsilon) = O\left(\frac{R}{\epsilon} \cdot \frac{L^* + \epsilon}{\epsilon} \log^3(R/\epsilon)\right) = \tilde{O}\left(\frac{R}{\epsilon} \cdot \frac{L^* + \epsilon}{\epsilon}\right)$$

Parametric vs Non-Parametric

	Parametric $\mathbf{dim}(\mathcal{H}) \leq \mathbf{D}, \|h\| \leq \mathbf{1}$	**Scale-Sensitive** $\mathcal{R}_n(\mathcal{H}) \leq \sqrt{R/n}$
Lipschitz: $\|\phi'\| \leq G$ (e.g. hinge, ℓ_1)	$\frac{G\ D}{n} + \sqrt{L^* \frac{GD}{n}}$	$\sqrt{\frac{G^2 R}{n}}$
Smooth: $\|\phi''\| \leq H$ (e.g. logistic, Huber, smoothed hinge)	$\frac{H\ D}{n} + \sqrt{L^* \frac{HD}{n}}$	$\frac{H\ R}{n} + \sqrt{L^* \frac{HR}{n}}$
Smooth & **strongly convex**: $\lambda \leq \phi'' \leq H$ (e.g. square loss)	$\frac{H}{\lambda} \cdot \frac{H\ D}{n}$	$\frac{H\ R}{n} + \sqrt{L^* \frac{HR}{n}}$

Min-max tight up to poly-log factors

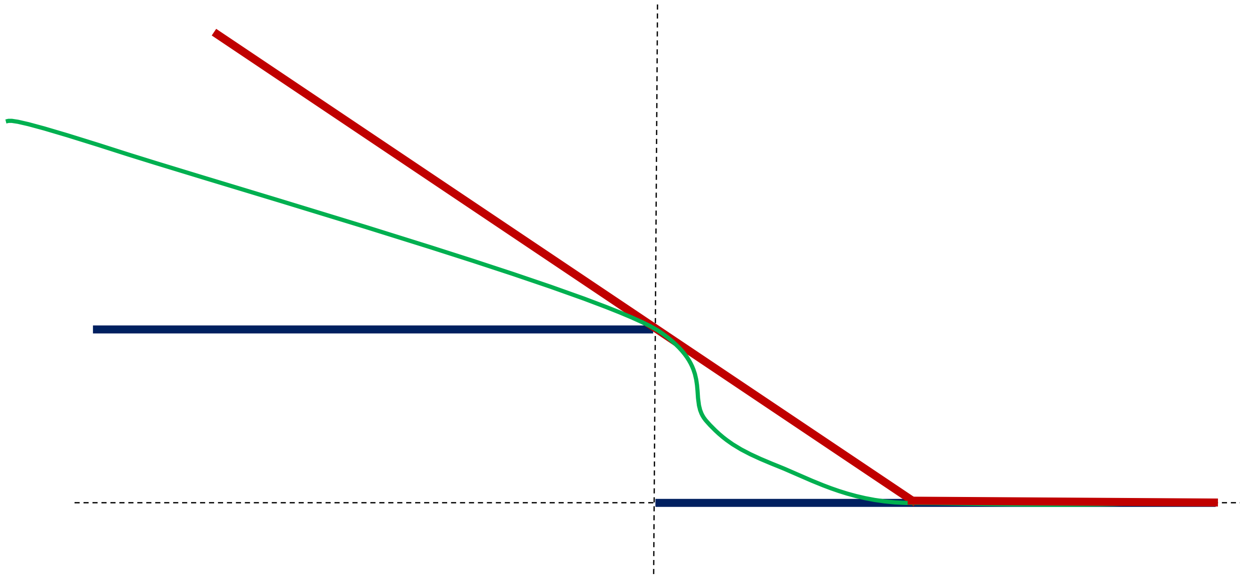
Optimistic SVM-Type Bounds



$\phi_{01} \quad \leq \quad \phi_{\text{hinge}}$

- Optimize
- Generalize

Optimistic SVM-Type Bounds



$$\phi_{01} \leq \phi_{\text{smooth}} \leq \phi_{\text{hinge}}$$

- Generalize

- Optimize

$$L_{01}(\hat{h}_{\text{hinge}}) \leq L^*_{\text{hinge}} + \tilde{O}\left(\sqrt{L^*_{\text{hinge}}\frac{R}{n}} + \frac{R}{n}\right)$$

Optimistic Learning Guarantees

$$L(\hat{h}) \leq (1+\alpha)L^* + (1+\tfrac{1}{\alpha})\tilde{O}\left(\frac{R}{n}\right)$$

$$n(\epsilon) = \tilde{O}\left(\frac{R}{\epsilon} \cdot \frac{L^* + \epsilon}{\epsilon}\right)$$

✓Parametric classes
✓Scale-sensitive classes with smooth loss
✓SVM-type bounds
✓Margin Bounds
✓Online Learning/Optimization with smooth loss
✓Stability-based guarantees with smooth loss

× Non-param (scale sensitive) classes with non-smooth loss
× Online Learning/Optimization with non-smooth loss

Why Optimistic Guarantees?

$$L(\hat{h}) \leq (1+\alpha)L^* + (1+\tfrac{1}{\alpha})\tilde{O}\left(\frac{R}{n}\right)$$

$$n(\epsilon) = \tilde{O}\left(\frac{R}{\epsilon} \cdot \frac{L^* + \epsilon}{\epsilon}\right)$$

- Optimistic regime typically relevant regime:
  - Approximation error L^* $\approx$ Estimation error ϵ
  - If $\epsilon \ll L^*$, better to spend energy on lowering approx. error (use more complex class)

- Important in understanding statistical learning

Training Kernel SVMs

Kernel evaluations to get excess error ϵ: $(R = \|w^*\|^2)$

- Using SGD:

$$T(\epsilon) = O\left(n(\epsilon)^2\right) = \tilde{O}\left(\frac{R^2}{\epsilon^2}\left(\frac{L^*+\epsilon}{\epsilon}\right)^2\right)$$

- Using the Stochastic Batch Perceptron **[Cotter et al 2012]**:

$$T(\epsilon) = \tilde{O}\left(\frac{R^2}{\epsilon}\left(\frac{L^*+\epsilon}{\epsilon}\right)^3\right)$$

(is this the best possible?)

Training Linear SVMs

Runtime (# feature evaluations): $\quad (R = \|w^*\|^2)$

- Using SGD:

$$T(\epsilon) = \tilde{O}\left(dn(\epsilon)\right) = \tilde{O}\left(d\frac{R}{\epsilon}\left(\frac{L^*+\epsilon}{\epsilon}\right)\right)$$

- Using SIMBA **[Hazan et al 2011]**:

$$T(\epsilon) = \tilde{O}\left((d+n(\epsilon)) \cdot R\left(\frac{L^*+\epsilon}{\epsilon}\right)^2\right)$$
$$= \tilde{O}\left(dR\left(\frac{L^*+\epsilon}{\epsilon}\right)^2 + R\frac{R}{\epsilon}\left(\frac{L^*+\epsilon}{\epsilon}\right)^3\right)$$

(is this the best possible?)

Mini-Batch SGD

- Stochastic optimization of smooth $L(w)$ using *n* training-points, doing *T=n/b* iterations of SGD with mini-batches of size *b*
- Pessimistic Analysis (ignoring L^*):

$$L(\bar{w}_T) \leq O\left(\sqrt{\frac{G^2 R}{n}} + \frac{HRb}{n}\right)$$

➔ Can use minibatch of size $b \propto \sqrt{n}$, with $T \propto \sqrt{n}$ iterations and get same error (up to constant factor) as sequential SGD

[Dekel et al 2010][Agarwal Duchi 2011]

- But taking into account L^*:

$$L(\bar{w}_T) \leq O\left(\sqrt{\frac{L^* HR}{n}} + \frac{HR}{n} + \frac{HRb}{n}\right)$$

➔In Optimistic Regime: Can't use b>1, *no parallelization speedups!*

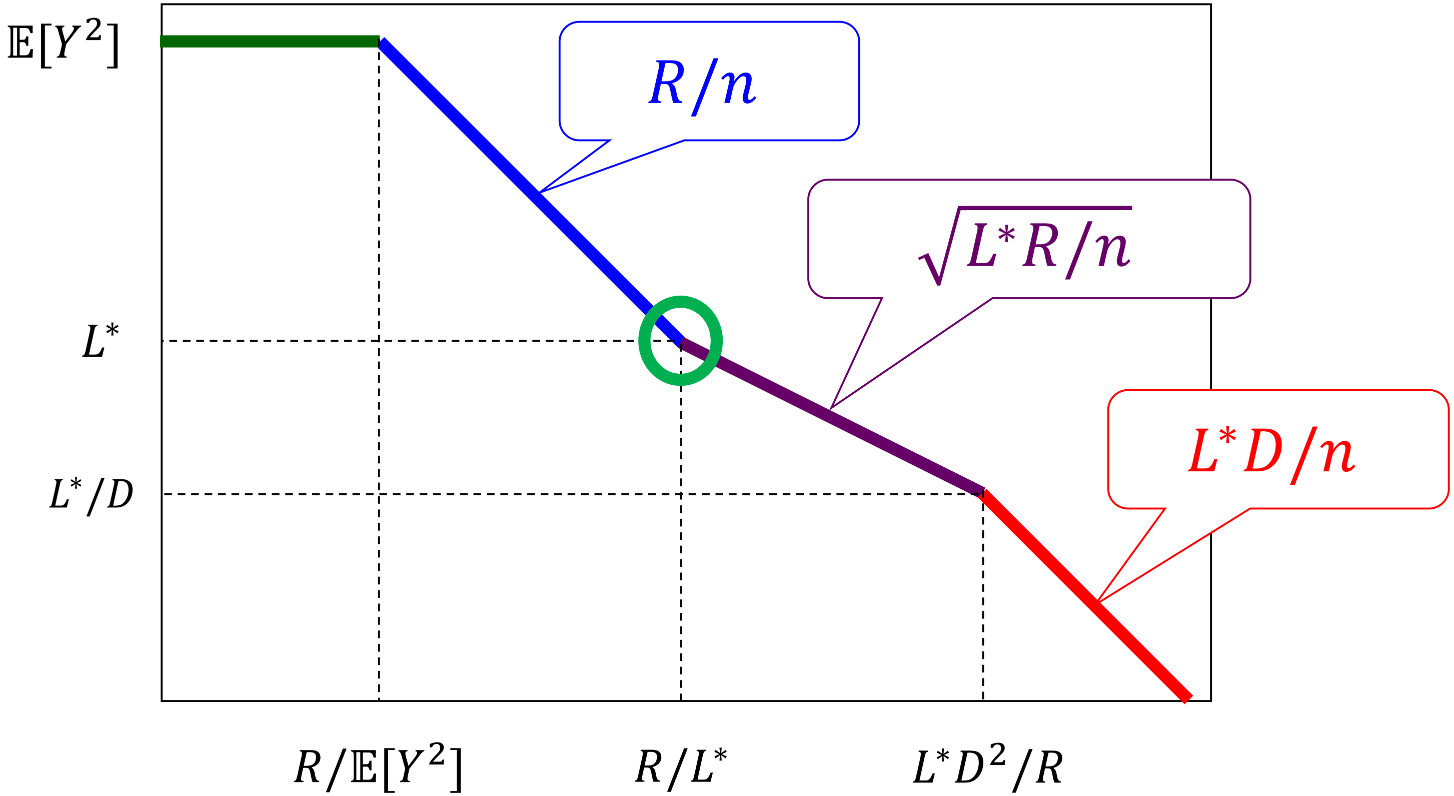
- Use acceleration to get speedup in optimistic regime ☺ **[Cotter et al 2011]**

Multiple Complexity Controls

[Liang Srebro 2010]

$$L(w) = \mathbb{E}[(\langle w, X\rangle - Y)^2], \quad Y = \langle w, X\rangle + \mathcal{N}(0, \sigma^2)$$

$$w \in \mathbb{R}^D \qquad \|w\|^2 \leq R$$



Be Optimistic

$$n(\epsilon) = \tilde{O}\left(\frac{R}{\epsilon} \cdot \frac{L^* + \epsilon}{\epsilon}\right)$$

$$L(\hat{h}) \leq (1+\alpha)L^* + (1+\tfrac{1}{\alpha})\tilde{O}\left(\frac{R}{n}\right)$$

- For scale-sensitive non-parametric classes, with smooth loss: **[Srebro Sridharan Tewari 2010]**
  - Diff vs parametric: Not possible with non-smooth loss!
- Optimistic regime typically relevant regime:
  - Approximation error $L^* \approx$ Estimation error ϵ
  - If $\epsilon \ll L^*$, better to spend energy on lowering approx. error (use more complex class)
- Important in understanding statistical learning