

# McDiarmid's Inequality

Ashish Rastogi

# Motivation

- Generalization bounds:
  - capacity measures [covering numbers, Rademacher complexity, VC theory]
  - stability-based bounds
- Applications:
  - chromatic number

# McDiarmid's Inequality

- **Theorem:** Let  $X_1, \dots, X_m$  be independent random variables all taking values in the set  $\mathcal{X}$ . Further, let  $f : \mathcal{X}^m \mapsto \mathbb{R}$  be a function of  $X_1, \dots, X_m$  that satisfies  $\forall i, \forall x_1, \dots, x_m, x'_i \in \mathcal{X}$ ,

$$|f(x_1, \dots, x_i, \dots, x_m) - f(x_1, \dots, x'_i, \dots, x_m)| \leq c_i.$$

Then for all  $\epsilon > 0$ ,

$$\Pr [f - \mathbb{E}[f] \geq \epsilon] \leq \exp \left( \frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2} \right).$$

- **Corollary:** For  $X_i \in [a_i, b_i]$ ,  $f = \frac{1}{m} \sum_{i=1}^m X_i$ ,  $c_i = \frac{b_i - a_i}{m}$ .

$$\Pr [f - \mathbb{E}[f] \geq \epsilon] \leq \exp \left( \frac{-2\epsilon^2 m^2}{\sum_{i=1}^m (b_i - a_i)^2} \right).$$

Hoeffding's Inequality

# Proof Elements

- **Markov's Inequality:** For a non-negative random variable  $X$ ,

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$$

- **Proof:**

$$\begin{aligned}\mathbb{E}[X] &= \sum_x x \Pr[X = x] \\ &\geq \sum_{x \geq t} x \Pr[X = x] \\ &\geq t \sum_{x \geq t} \Pr[X = x] \\ &= t \Pr[X \geq t].\end{aligned}$$

# Law of Iterated Expectation

- For random variables  $X, Y, Z$ :

$$\mathbb{E}[\mathbb{E}[X|Y, Z]|Z] = \mathbb{E}[X|Z]$$

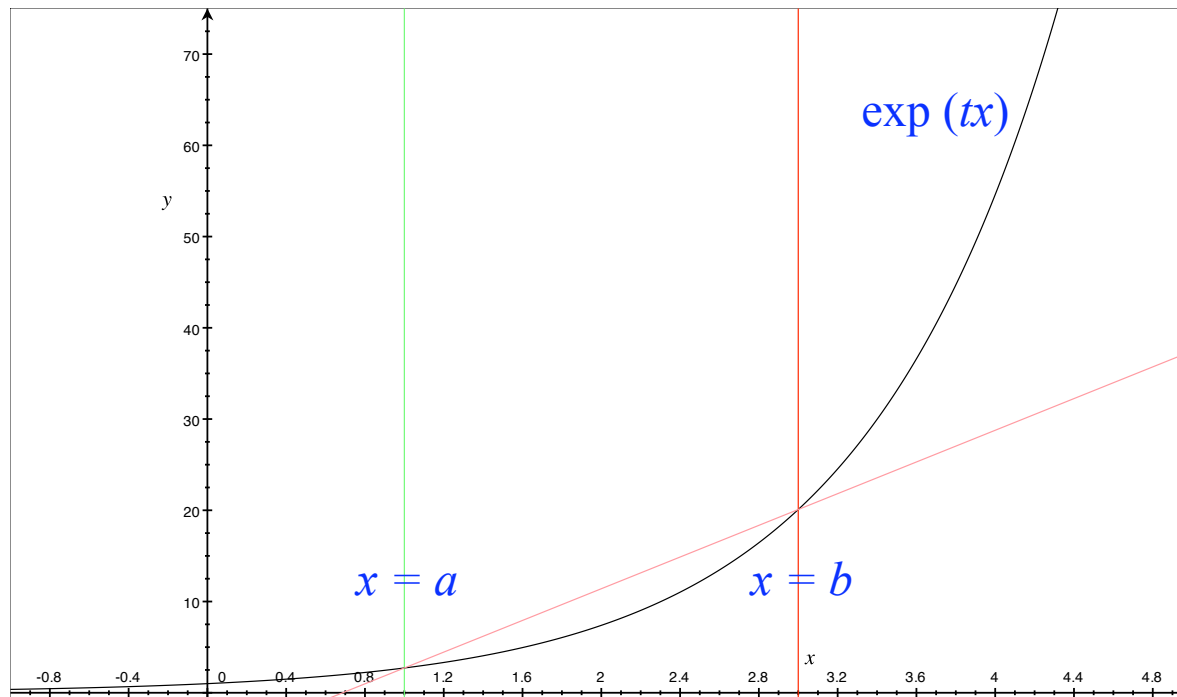
- **Proof:** follows from definitions.
- **Idea:** taking expectation conditioning over  $Y$  and then taking expectation over values of  $Y$  is the same as taking the expectation all at once.

# Proof Elements

- **Hoeffding's Lemma:** Let  $X$  be a random variable with  $\mathbb{E}[X] = 0$  and  $a \leq X \leq b$ . Then for  $t > 0$ ,

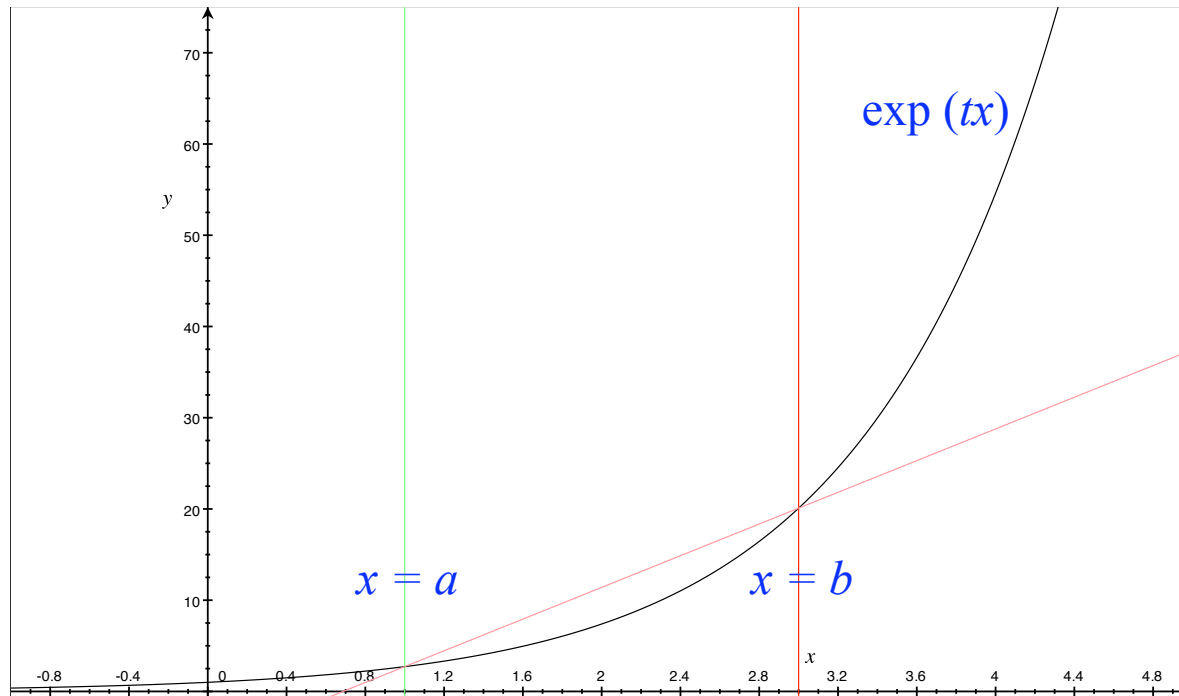
$$\mathbb{E}[e^{tX}] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right).$$

- **Proof:** Convexity and Taylor's Theorem (do on the board).



# Hoeffding's Lemma

- Convexity implies:  $e^{tx} \leq \frac{b-x}{b-a} e^{ta} + \frac{x-a}{b-a} e^{tb}$
- Expectation on both sides:  $\mathbb{E}[e^{tx}] \leq \frac{b}{b-a} e^{ta} - \frac{a}{b-a} e^{tb}$
- Set  $e^{\phi(t)} := \frac{b}{b-a} e^{ta} - \frac{a}{b-a} e^{tb}$
- Observe  $\phi(0) = 0, \phi'(0) = 0, \phi''(t) \leq \frac{(b-a)^2}{4}$ .



# McDiarmid's Inequality

- **Theorem:** Let  $X_1, \dots, X_m$  be independent random variables all taking values in the set  $\mathcal{X}$ . Further, let  $f : \mathcal{X}^m \mapsto \mathbb{R}$  be a function of  $X_1, \dots, X_m$  that satisfies  $\forall i, \forall x_1, \dots, x_m, x'_i \in \mathcal{X}$ ,

$$|f(x_1, \dots, x_i, \dots, x_m) - f(x_1, \dots, x'_i, \dots, x_m)| \leq c_i.$$

Then for all  $\epsilon > 0$ ,

$$\Pr [f - \mathbb{E}[f] \geq \epsilon] \leq \exp \left( \frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2} \right).$$

- **Proof:** Let  $\mathbf{X}_1^i$  be the sequence of random variables  $X_1, \dots, X_i$ . Define random variables  $Z_i = \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_1^i]$ . Observe that  $Z_0 = \mathbb{E}[f]$ ,  $Z_m = f(\mathbf{X})$ .



# Proof continued

- Consider the random variable  $Z_i - Z_{i-1} \mid \mathbf{X}_1^{i-1}$
- **Observation 1:**  $\mathbb{E}[Z_i - Z_{i-1} \mid \mathbf{X}_1^{i-1}] = 0$ .
- **Observation 2:**
  - Let  $U_i = \sup_u \{\mathbb{E}[f \mid \mathbf{X}_1^{i-1}, u] - \mathbb{E}[f \mid \mathbf{X}_1^{i-1}]\}$ .
  - Let  $L_i = \inf_l \{\mathbb{E}[f \mid \mathbf{X}_1^{i-1}, l] - \mathbb{E}[f \mid \mathbf{X}_1^{i-1}]\}$ .
  - Note that  $L_i \leq (Z_i - Z_{i-1}) \mid \mathbf{X}_1^{i-1} \leq U_i$ .
  - Finally,  $U_i - L_i \leq c_i$ .
  - Thus,  $\mathbb{E}[e^{t(Z_i - Z_{i-1})} \mid \mathbf{X}_1^{i-1}] \leq e^{\frac{t^2 c_i^2}{8}}$ .

# Proof continued

$$\begin{aligned}\Pr [f - \mathbb{E}[f] \geq \epsilon] &= \Pr \left[ e^{t(f - \mathbb{E}[f])} \geq e^{t\epsilon} \right] \\ \text{Markov's Inequality} &\leq e^{-t\epsilon} \mathbb{E} \left[ e^{t(f - \mathbb{E}[f])} \right] \\ \text{Telescoping} &= e^{-t\epsilon} \mathbb{E} \left[ e^{t \sum_{i=1}^m (Z_i - Z_{i-1})} \right] \\ \text{Iterative Expectation} &= e^{-t\epsilon} \mathbb{E} \left[ \mathbb{E} \left[ e^{t \sum_{i=1}^m (Z_i - Z_{i-1})} \mid \mathbf{X}_1^{m-1} \right] \right] \\ &= e^{-t\epsilon} \mathbb{E} \left[ e^{t \sum_{i=1}^{m-1} (Z_i - Z_{i-1})} \mathbb{E} \left[ e^{t(Z_m - Z_{m-1})} \mid \mathbf{X}_1^{m-1} \right] \right] \\ &\leq e^{-t\epsilon} e^{\frac{t^2 c_m^2}{8}} \mathbb{E} \left[ e^{t \sum_{i=1}^{m-1} (Z_i - Z_{i-1})} \right]\end{aligned}$$

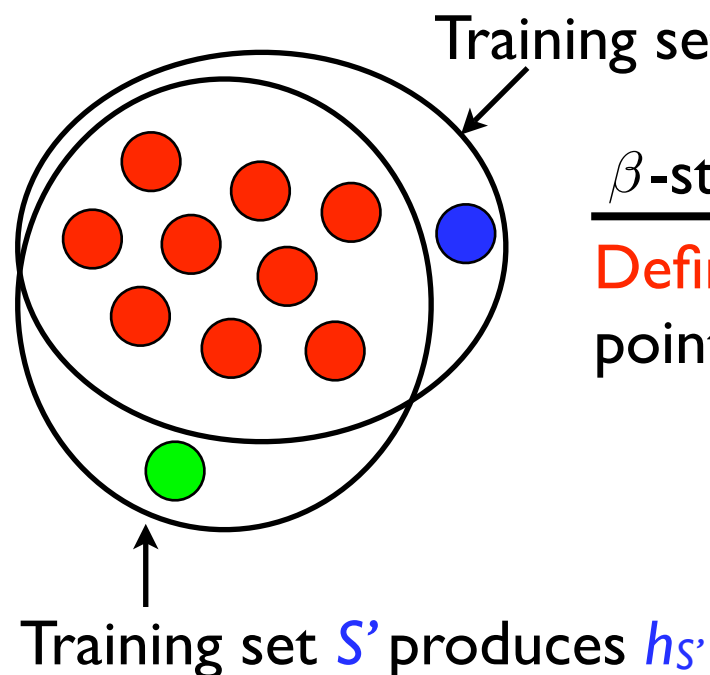
Thus, 
$$\Pr[f - \mathbb{E}[f] \geq \epsilon] \leq \exp \left( -t\epsilon + \frac{t^2}{8} \sum_{i=1}^m c_i^2 \right)$$

# Proof continued

- Choose  $t$  that minimizes  $-t\epsilon + \frac{t^2}{8} \sum_{i=1}^m c_i^2$ .
- This leads to  $t = \frac{4\epsilon}{\sum_{i=1}^m c_i^2}$ .
- And therefore,  $-t\epsilon + \frac{t^2}{8} \sum_{i=1}^m c_i^2 = \frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}$ .
- Thus,  $\Pr[f - \mathbb{E}[f] \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right)$ .

# Stability of an Algorithm

- **Idea:** small change in training set ( $\Rightarrow$ ) small change in hypothesis.
- “Sufficient” stability leads to generalization (McDiarmid’s ineq.)



## $\beta$ -stability

**Definition:** When  $S$  and  $S'$  differ in exactly one point, then for all  $\forall x \in \mathcal{X}$ ,

$$|c(h_S, x) - c(h_{S'}, x)| \leq \beta.$$

- **Advantage:** algorithm specific, analysis independent of any capacity term.

# Ingredients of a Generalization Bound

- **Errors:**

- test error:  $R(h, S) = \mathbb{E}_{x \sim D}[c(h_S, x)]$

- training error:  $\hat{R}(h, S) = \frac{1}{m} \sum_{i=1}^m c(h_S, x_i)$

- Shape of the generalization bound:

$$R(h, S) \leq \hat{R}(h, S) + \text{stability-dependent terms.}$$

- **Key step:** for a hypothesis  $h$ , deriving a bound on

$$\Pr_{S \sim X} \left[ |R(h, S) - \hat{R}(h, S)| \geq \epsilon \right].$$

# From Stability to Generalization

- Apply **McDiarmid's inequality** to the random variable:

$$f(S) = R(h, S) - \widehat{R}(h, S)$$

- Need to bound:
  - for  $S$  and  $S'$  differing in one point,  $|f(S) - f(S')|$ .
  - the expectation,  $\mathbb{E}_{S \sim D^m}[f(S)]$ .
- Let  $A$  be a  $\beta$ -stable learning algorithm with respect to a cost-function  $c$  and the cost-function  $c$  is bounded, i.e.  $\forall x \in \mathcal{X}, \forall h \in \mathcal{H}, c(h, x) \leq M$  for some  $M > 0$ . Then,
  - $|f(S) - f(S')| \leq 2\beta + \frac{M}{m}$
  - $\mathbb{E}[f(S)] \leq \beta$

# Generalization Bound

- Applying **McDiarmid's Inequality** leads to, for all  $\epsilon > 0$ ,

$$\Pr[R(h, S) - \hat{R}(h, S) - \beta \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{m(2\beta + \frac{M}{m})^2}\right)$$

- Or,

$$\Pr[R(h, S) - \hat{R}(h, S) \geq \beta + \epsilon] \leq \exp\left(\frac{-2\epsilon^2 m}{(2\beta m + M)^2}\right)$$

- Note that for effective bound, need  $\beta = o(1/\sqrt{m})$ .
- With confidence  $1 - \delta$ ,

$$R(h, S) \leq \hat{R}(h, S) + \beta + (2\beta m + M) \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

# Determining $\beta$

- Consider regularization-based objective function:

$$F(g, S) = \|g\|_K^2 + \frac{C}{m} \sum_{i=1}^m c(g, x_i).$$

- Need two technical definitions / observations:

- $\sigma$ -admissibility:  $\forall h, h' \in \mathcal{H}, \forall x \in \mathcal{X},$

$$|c(h', x) - c(h, x)| \leq \sigma |(h' - h)(x)|.$$

- Bounded kernel:  $\forall x \in \mathcal{X}, K(x, x) \leq \kappa.$



# Determining $\beta$

- Consider regularization-based objective function:

$$F(g, S) = \|g\|_K^2 + \frac{C}{m} \sum_{i=1}^m c(g, x_i).$$

- Consider two sets,  $S$  and  $S'$  such that  $S' = S \setminus \{x_i\} \cup \{x'_i\}$  where  $x_i \in S$ .

- Let  $h = \arg \min_g F(g, S)$ ,  $h' = \arg \min_g F(g, S')$ .

- $F(g, S)$  is convex in  $g$ . Let  $\Delta h = h' - h$ .

- Thus,  $F(h, S) - F(h + t\Delta h, S) \leq 0$ , and  $F(h, S') - F(h' - t\Delta h, S') \leq 0$ .

- This leads to:

$$\|h\|_K^2 - \|h + t\Delta h\|_K^2 + \|h'\|_K^2 - \|h' - t\Delta h\|_K^2 \leq \frac{2t\sigma\kappa C \|\Delta h\|_K}{m}.$$

# Determining $\beta$

- Finally, observe that in an RHKS:

$$\|h\|_K^2 - \|h + t\Delta h\|_K^2 + \|h'\|_K^2 - \|h' - t\Delta h\|_K^2 = 2t(1 - t)\|\Delta h\|_K^2$$

- Put the pieces together to derive a bound.

# Application - Chromatic Number

- **Random Graph:** Given number of vertices  $n$  and an edge probability  $p$ , define  $G(n, p)$  as a random graph with:
  - vertices  $\{1, \dots, n\}$ .
  - edges  $E$  (random) as  $\forall i, j, (i, j) \in E$  with probability  $p$ .
- **Chromatic number:** min. number of colors to color the vertices of a graph s.t. adjacent vertices colored differently.
- **Notation:** Let  $\omega(G)$  be the chromatic number of  $G$ .
- **Vertex exposure martingale:** sequence of random variables  $Z_k, 1 \leq k \leq n$ , given the edges between the first  $k$  vertices.

$$Z_k = \mathbb{E}[\omega(G) \mid E' \subseteq E, (i, j) \in E' \Leftrightarrow (i, j) \in E \wedge i, j \leq k]$$

# Chromatic Number

- **Observation 1:**  $Z_0 = \mathbb{E}[w(G)], Z_n = w(G)$ .
- **Observation 2:**  $|Z_k - Z_{k-1}| \leq 1, 1 \leq k \leq n$ .
- Using  $Z_n - Z_0 = \sum_{k=1}^n (Z_k - Z_{k-1})$ , and setting  $\epsilon = \lambda\sqrt{n}$ , easy to show:

$$\Pr \left[ \frac{1}{\sqrt{n}} (\omega(G) - \mathbb{E}[\omega(G)]) \geq \lambda \right] \leq e^{-2\lambda^2}.$$

- **Notes:**
  - determining the chromatic number is **NP-hard**.
  - finding a  $k$ -coloring given that  $\omega(G) = k$  is also **NP-hard**.
  - there's more sophisticated analyses of  $\omega(G)$  for random  $G$ .

# Conclusion

- The condition to apply McDiarmid's inequality is relatively simple to verify.
- Provides an easy way of deriving generalization bounds.

# References

- Kazuoki Azuma. *Weighted sums of certain dependent random variables*. In *Tohoku Mathematical Journal*, volume 19, pages 357–367, 1967.
- Olivier Bousquet and Andre Elisseeff. *Stability and generalization*. *Journal of Machine Learning Research*, 2:499–526, 2002.
- Colin McDiarmid. *On the method of bounded differences*. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, Cambridge, 1989.
- N. Alon and J. H. Spencer, *The Probabilistic Method*, Wiley, New York, 1992.