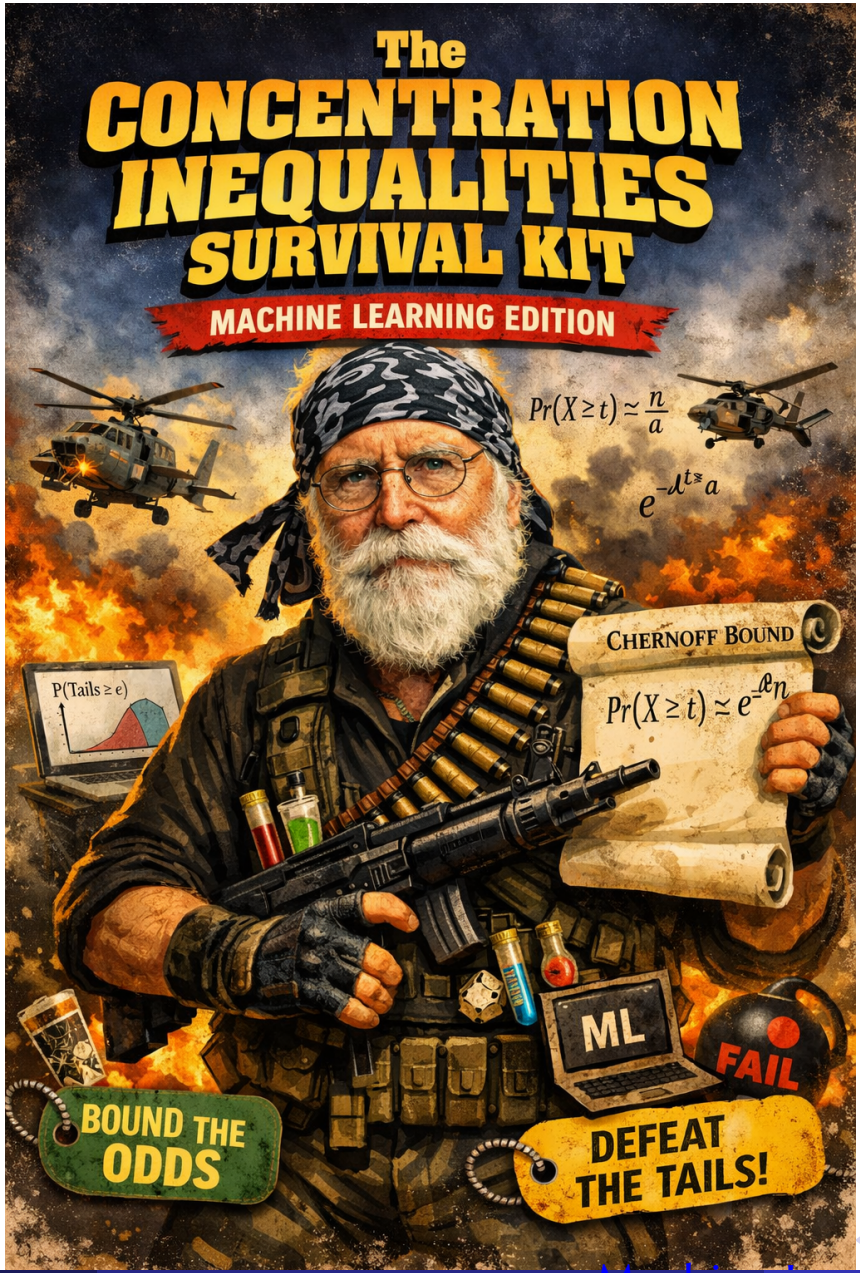
The Concentration Inequalities Survival Kit

Machine Learning Edition

Thiago Rodrigo Ramos

26/02/2026

Introduction



Supervised learning

- **Data:** unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and an i.i.d. sample

$$S = \{(X_i, Y_i)\}_{i=1}^{n} \sim \mathcal{D}^n.$$

- **Hypothesis class:** $\mathcal{H}$ of predictors $h : \mathcal{X} \to \mathcal{Y}$.
- **Loss function:** $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.
- **True risk (population):**

$$R(h) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}\big[\ell(h(X), Y)\big].$$

- **Empirical risk (training):**

$$\widehat{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i).$$

Example: binary classification

- $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0, 1\}$ and $h : \mathbb{R}^d \to \{0, 1\}$.
- **0–1 loss:** $\ell(h(x), y) = \mathbf{1}\{h(x) \neq y\}$.
- **True risk (test error):**

$$R(h) = \mathbb{P}_{(X,Y)\sim\mathcal{D}}\big(h(X) \neq Y\big).$$

- **Empirical risk (training error):**

$$\widehat{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{h(X_i) \neq Y_i\}.$$

Example: regression

- $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$ and $h : \mathbb{R}^d \to \mathbb{R}$.
- **Squared loss:** $\ell(h(x), y) = (h(x) - y)^2$.
- **True risk (population MSE):**

$$R(h) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[(h(X) - Y)^2\right].$$

- **Empirical risk (training MSE):**

$$\widehat{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n}(h(X_i) - Y_i)^2.$$

Empirical Risk Minimization (ERM)

- We observe a training sample $S = \{(X_i, Y_i)\}_{i=1}^n \sim \mathcal{D}^n$.
- For a hypothesis $h \in \mathcal{H}$, define the empirical risk

$$\widehat{R}_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(h(X_i), Y_i\big).$$

- **ERM principle:** choose a predictor that minimizes the training loss:

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \widehat{R}_S(h).$$

- **Main question:** when does minimizing $\widehat{R}_S(h)$ also lead to small *true risk*

$$R(h) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}\big[\ell(h(X), Y)\big] \ ?$$

From ERM to uniform convergence

- Define the ERM solution: $\hat{h}_S \in \arg\min_{h \in \mathcal{H}} \widehat{R}_S(h)$.
- Define the best-in-class predictor: $h^\star \in \arg\min_{h \in \mathcal{H}} R(h)$.
- We want to study the excess risk of ERM:

$$\mathbb{P}\left[R(\hat{h}_S) - \inf_{h \in \mathcal{H}} R(h) > \varepsilon \right].$$

- Using a standard inequality:

$$\mathbb{P}\left[R(\hat{h}_S) - \inf_{h \in \mathcal{H}} R(h) > \varepsilon \right] \leq \mathbb{P}\left[\sup_{h \in \mathcal{H}} \left| R(h) - \widehat{R}_S(h) \right| > \frac{\varepsilon}{2} \right].$$

- **Our goal:** to control the highlighted probability.

A simplification step

- For each $h \in \mathcal{H}$ define $Z_i(h) := \ell\big(h(X_i), Y_i\big)$, where (X_i, Y_i) are i.i.d.
- Then

$$\widehat{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n} Z_i(h) \quad \text{and} \quad R(h) = \mathbb{E}\big[Z_1(h)\big].$$

- **From now on:** we ignore the ML interpretation and focus on concentration of empirical processes:

$$\sup_{h\in\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \right|.$$

Baby steps: single hypothesis

- To begin from the basics, assume $\mathcal{H}$ contains only **one** hypothesis h (so there is no $\sup_{h \in \mathcal{H}}$ yet).
- Then we only need to control

$$\left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \right|.$$

- In this setting, classical inequalities give high-probability bounds. For example,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \right| \geq t \right) \leq$$

Baby steps: single hypothesis

- To begin from the basics, assume $\mathcal{H}$ contains only **one** hypothesis h (so there is no $\sup_{h \in \mathcal{H}}$ yet).
- Then we only need to control

$$\left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \right|.$$

- In this setting, classical inequalities give high-probability bounds. For example,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \right| \geq t \right) \leq \frac{\mathrm{Var}(Z_1(h))}{n\, t^2},$$

by **Chebyshev**, if $\mathrm{Var}(Z_1(h)) < \infty$.

Chernoff bound

- A slightly less well-known (but extremely useful) tool is the **Chernoff bound**.
- For any $\lambda > 0$ and any random variable W,

$$\mathbb{P}(W \geq t) = \mathbb{P}\big(e^{\lambda W} \geq e^{\lambda t}\big) \leq e^{-\lambda t}\,\mathbb{E}\big[e^{\lambda W}\big] \qquad \text{(Markov on } e^{\lambda W}\text{)}.$$

- We apply it with

$$W = \sum_{i=1}^{n} \big(Z_i(h) - \mathbb{E}[Z_1(h)]\big),$$

which yields the explicit bound

$$\mathbb{P}\left(\sum_{i=1}^{n} \big(Z_i(h) - \mathbb{E}[Z_1(h)]\big) \geq t\right) \leq e^{-\lambda t}\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} \big(Z_i(h) - \mathbb{E}[Z_1(h)]\big)\right)\right].$$

Chernoff bound

- Using independence,

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(Z_i(h)-\mathbb{E}[Z_1(h)])\right)\right]=\prod_{i=1}^{n}\mathbb{E}[\exp(\lambda(Z_i(h)-\mathbb{E}[Z_1(h)]))]\,.$$

- Since the $Z_i(h)$'s are i.i.d.,

$$\prod_{i=1}^{n}\mathbb{E}[\exp(\lambda(Z_i(h)-\mathbb{E}[Z_1(h)]))]=\left(\mathbb{E}\left[e^{\lambda(Z_1(h)-\mathbb{E}[Z_1(h)])}\right]\right)^n.$$

- Therefore,

$$\mathbb{P}\left(\sum_{i=1}^{n}(Z_i(h)-\mathbb{E}[Z_1(h)])\geq t\right)\leq\exp(-\lambda t)\left(\mathbb{E}\left[e^{\lambda(Z_1(h)-\mathbb{E}[Z_1(h)])}\right]\right)^n.$$

- **Key point:** everything reduces to bounding the...

Chernoff bound

- Using independence,

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(Z_i(h)-\mathbb{E}[Z_1(h)])\right)\right]=\prod_{i=1}^{n}\mathbb{E}[\exp(\lambda(Z_i(h)-\mathbb{E}[Z_1(h)]))]\,.$$

- Since the $Z_i(h)$'s are i.i.d.,

$$\prod_{i=1}^{n}\mathbb{E}[\exp(\lambda(Z_i(h)-\mathbb{E}[Z_1(h)]))]=\left(\mathbb{E}\left[e^{\lambda(Z_1(h)-\mathbb{E}[Z_1(h)])}\right]\right)^n.$$

- Therefore,

$$\mathbb{P}\left(\sum_{i=1}^{n}(Z_i(h)-\mathbb{E}[Z_1(h)])\geq t\right)\leq\exp(-\lambda t)\left(\mathbb{E}\left[e^{\lambda(Z_1(h)-\mathbb{E}[Z_1(h)])}\right]\right)^n.$$

- **Key point:** everything reduces to bounding the...MGF $\mathbb{E}\left[e^{\lambda(Z_1(h)-\mathbb{E}[Z_1(h)])}\right]$.

Chernoff bound: example

- Assume that for some $\sigma^2 > 0$,

$$\mathbb{E}\big[\exp\big(\lambda\big(Z_1(h) - \mathbb{E}[Z_1(h)]\big)\big)\big] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \qquad \text{for all } \lambda \in \mathbb{R}.$$

- Plugging this into the previous bound gives, for any $\lambda > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n}(Z_i(h) - \mathbb{E}[Z_1(h)]) \geq t\right) \leq \exp\left(-\lambda t + \frac{n\lambda^2\sigma^2}{2}\right).$$

- Optimize over λ (take $\lambda^\star = \frac{t}{n\sigma^2}$) to obtain

$$\mathbb{P}\left(\sum_{i=1}^{n}(Z_i(h) - \mathbb{E}[Z_1(h)]) \geq t\right) \leq \exp\left(-\frac{t^2}{2n\sigma^2}\right).$$

- In terms of the empirical mean,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

Chernoff bound: takeaway

- If this MGF satisfies, for some $\sigma^2 > 0$ and all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\big[\exp\big(\lambda\big(Z_1 - \mathbb{E}[Z_1]\big)\big)\big] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right),$$

  then for any $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_1] \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

- Inverting the bound: for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_1] \leq \sigma\sqrt{\frac{2\log(1/\delta)}{n}}.$$

- **Interpretation:** with high probability $(1-\delta)$, the gap is at most $\sigma\sqrt{2\log(1/\delta)/n}$, shrinking as n grows.

Subgaussian distributions

- The MGF condition we used before is a very convenient way to express **Gaussian-like tails**.
- A random variable Z is called **subgaussian** if there exists $\sigma^2 > 0$ such that, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda\left(Z - \mathbb{E}[Z]\right)\right)\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

- This assumption is useful because it **immediately implies** concentration for empirical means:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z] \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

- Many common variables are subgaussian (often after centering), e.g.:
  - Gaussian $\mathcal{N}(\mu, \sigma^2)$,
  - bounded variables $Z \in [a, b]$,
  - sums/averages of independent subgaussians.

Hoeffding's inequality

- Previously, we saw that an MGF bound implies strong concentration.
- A particularly important case is when the variables are **bounded**, since then we get an MGF bound *for free*.
- Let $Z_1, \ldots, Z_n$ be independent with $Z_i \in [a, b]$ almost surely. Then, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_1] \geq \varepsilon\right) \leq \exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right).$$

- Equivalently,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_1]\right| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right).$$

Back to many hypotheses

- Now we return to the case $1 < |\mathcal{H}| < \infty$ and aim to control

$$\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \right|.$$

- Assume **bounded losses**; then by Hoeffding, for a fixed h,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \right| \geq \varepsilon \right) \leq 2 \exp\left(\frac{-2n\varepsilon^2}{(b-a)^2} \right).$$

- Using the union bound,

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \right| \geq \varepsilon \right) \leq 2|\mathcal{H}| \exp\left(\frac{-2n\varepsilon^2}{(b-a)^2} \right).$$

Back to many hypotheses

- From the union bound + Hoeffding, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)]\right| \geq \varepsilon\right) \leq 2|\mathcal{H}| \exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right).$$

- Invert the bound: set the right-hand side equal to δ and solve for ε.
- For any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)]\right| \leq (b-a)\sqrt{\frac{\log\left(2|\mathcal{H}|/\delta\right)}{2n}}.$$

Back to many hypotheses

- From the union bound + Hoeffding, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)]\right| \geq \varepsilon\right) \leq 2|\mathcal{H}|\exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right).$$

- Invert the bound: set the right-hand side equal to δ and solve for ε.
- For any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)]\right| \leq (b-a)\sqrt{\frac{\log\left(2|\mathcal{H}|/\delta\right)}{2n}}.$$

- **But:** this still does not solve our original problem, since in most ML settings $\mathcal{H}$ is **infinite** (e.g., linear predictors, neural networks), so we need more refined notions of complexity!

Next step: McDiarmid's inequality

- To handle **possibly infinite**, we will apply concentration to suitable functions of the sample; a key tool is **McDiarmid's inequality**.
- **Bounded differences property:** let $X_1, \ldots, X_n$ be independent and $f : \mathcal{X}^n \to \mathbb{R}$. We say f has bounded differences with constants $c_1, \ldots, c_n$ if for every i and for any two samples $x, x' \in \mathcal{X}^n$ that differ only in coordinate i,

$$\left| f(x_1, \ldots, x_n) - f(x'_1, \ldots, x'_n) \right| \leq c_i.$$

Next step: McDiarmid's inequality

- To handle **possibly infinite**, we will apply concentration to suitable functions of the sample; a key tool is **McDiarmid's inequality**.
- **Bounded differences property:** let $X_1, \ldots, X_n$ be independent and $f : \mathcal{X}^n \to \mathbb{R}$. We say f has bounded differences with constants $c_1, \ldots, c_n$ if for every i and for any two samples $x, x' \in \mathcal{X}^n$ that differ only in coordinate i,

$$\left| f(x_1, \ldots, x_n) - f(x'_1, \ldots, x'_n) \right| \leq c_i.$$

- **McDiarmid:** if f has bounded differences and $X_1, \ldots, X_n$ are independent, then for any $t > 0$,

$$\mathbb{P}\Big(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \geq t \Big) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2} \right).$$

How to pplying McDiarmid

- Consider the function of the sample

$$\Phi(S) := \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \right|.$$

- If losses are bounded, $Z_i(h) \in [a, b]$, then changing a single example in S changes $\Phi(S)$ by at most

$$c_i = \frac{b-a}{n} \qquad (i = 1, \ldots, n).$$

- Therefore Φ satisfies the bounded differences property:

$$\mathbb{P}\big(\Phi(S) - \mathbb{E}[\Phi(S)] \geq t\big) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right) = \exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

- So, with probability at least $1 - \delta$,

$$\Phi(S) \leq \mathbb{E}[\Phi(S)] + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}}.$$

How to pplying McDiarmid

- McDiarmid gives concentration *around the mean*:

$$\Phi(S) \leq \mathbb{E}[\Phi(S)] + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}} \qquad \text{w.p. } \geq 1-\delta.$$

- This does **not** fully solve the problem yet, because...

How to pplying McDiarmid

- McDiarmid gives concentration *around the mean*:

$$\Phi(S) \leq \mathbb{E}[\Phi(S)] + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}} \qquad \text{w.p. } \geq 1-\delta.$$

- This does **not** fully solve the problem yet, because...we introduced a new quantity:

$$\mathbb{E}[\Phi(S)] = \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)]\right|\right].$$

- **Question:** How do we bound this quantity? We will see one approach for classification problems.

Back to the finite hypothesis again

- Assume $\mathcal{H}$ is finite and $Z_i(h) \in [a, b]$ for all $h \in \mathcal{H}$.
- Define

$$W(h) := \frac{1}{n} \sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)].$$

- Boundedness implies an MGF bound (Hoeffding):

$$\mathbb{E}[\exp(tW(h))] \leq \exp\left(\frac{t^2\sigma^2}{2}\right) \qquad \text{for all } t > 0,\ h \in \mathcal{H},$$

  where we can take $\sigma^2 = \dfrac{(b-a)^2}{4n}$.
- Now we prove an inequality for $\mathbb{E}\big[\max_{h\in\mathcal{H}} W(h)\big]$.

Back to the finite hypothesis again

- For any $t > 0$, by convexity of $\exp$ and Jensen's inequality,

$$\exp\left(t\,\mathbb{E}\left[\max_{h\in\mathcal{H}} W(h)\right]\right) \leq \mathbb{E}\left[\exp\left(t \max_{h\in\mathcal{H}} W(h)\right)\right].$$

- Moreover,

$$\mathbb{E}\left[\exp\left(t \max_{h\in\mathcal{H}} W(h)\right)\right] = \mathbb{E}\left[\max_{h\in\mathcal{H}} e^{tW(h)}\right] \leq \mathbb{E}\left[\sum_{h\in\mathcal{H}} e^{tW(h)}\right] \leq |\mathcal{H}|\, e^{t^2\sigma^2/2}.$$

- Taking $\log$ gives

$$\mathbb{E}\left[\max_{h\in\mathcal{H}} W(h)\right] \leq \frac{\log|\mathcal{H}|}{t} + \frac{t\sigma^2}{2}.$$

- Choosing $t = \sqrt{2\log|\mathcal{H}|}/\sigma$ yields

$$\mathbb{E}\left[\max_{h\in\mathcal{H}} W(h)\right] \leq \sigma\sqrt{2\log|\mathcal{H}|}.$$

Reducing to the finite case: symmetrization

- Goal: control probabilities when $\mathcal{H}$ may be infinite.

Reducing to the finite case: symmetrization

- Goal: control probabilities when $\mathcal{H}$ may be infinite.
- Introduce an **independent ghost sample** $S' = \{(X_i', Y_i')\}_{i=1}^n$ and the corresponding $Z_i'(h) := \ell(h(X_i'), Y_i')$.
- Since $\mathbb{E}[Z_i'(h)] = \mathbb{E}[Z_i(h)]$, we can rewrite the expectation as

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\left(Z_i(h) - \mathbb{E}[Z_i'(h)]\right)\right|\right].$$

Reducing to the finite case: symmetrization

- Goal: control probabilities when $\mathcal{H}$ may be infinite.
- Introduce an **independent ghost sample** $S' = \{(X_i', Y_i')\}_{i=1}^n$ and the corresponding $Z_i'(h) := \ell(h(X_i'), Y_i')$.
- Since $\mathbb{E}[Z_i'(h)] = \mathbb{E}[Z_i(h)]$, we can rewrite the expectation as

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\left(Z_i(h) - \mathbb{E}[Z_i'(h)]\right)\right|\right].$$

- Applying Jensen to pull the expectation inside and using independence of S' gives the **symmetrization bound**

$$\mathbb{E}[\Phi(S)] \leq \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\left(Z_i(h) - Z_i'(h)\right)\right|\right].$$

- This transforms the problem into controlling the behavior of $\mathcal{H}$ on **finite samples** (no population expectation).

Reducing to the finite case: growth function

- From symmetrization, the key quantity depends on how $\mathcal{H}$ behaves on a **finite sample**.
- In **binary classification**, we can measure this via the **growth function** (also called the shattering coefficient).
- For a fixed set of inputs $z_1, \ldots, z_n \in \mathcal{Z}$, define the set of labelings induced by $\mathcal{H}$:

$$\mathcal{H}(z_1^n) := \Big\{ \big(z_1(h), \ldots, z_n(h)\big) \ : \ h \in \mathcal{H} \Big\} \subseteq \{0,1\}^n.$$

- The growth function is the maximum number of distinct labelings over all samples of size n:

$$\Pi_{\mathcal{H}}(n) := \sup_{z_1,\ldots,z_n \in \mathcal{Z}} |\mathcal{H}(z_1^n)| < 2^n.$$

- **Key point:** even if $\mathcal{H}$ is infinite, $\Pi_{\mathcal{H}}(n)$ is a **finite** number that captures the complexity of $\mathcal{H}$ *on finite samples.*

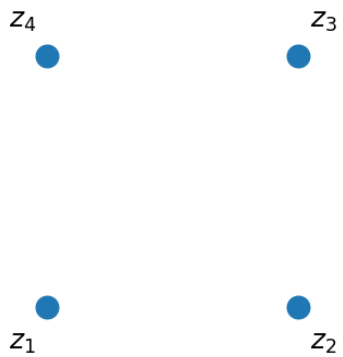
Reducing to the finite case: growth function

- For a fixed set of inputs $z_1, \ldots, z_n \in \mathcal{Z}$, define the set of labelings induced by $\mathcal{H}$:

$$\mathcal{H}(z_1^n) := \left\{ \big(z_1(h), \ldots, z_n(h)\big) \ : \ h \in \mathcal{H} \right\} \subseteq \{0,1\}^n.$$

- The growth function is the maximum number of distinct labelings over all samples of size n:

$$\Pi_{\mathcal{H}}(n) := \sup_{z_1,\ldots,z_n \in \mathcal{Z}} |\mathcal{H}(z_1^n)| < 2^n.$$



From maximal inequality to a growth-function bound

- In binary classification with 0–1 loss, we have $Z_i(h) \in [0, 1]$.
- Applying the maximal inequality with $|\mathcal{H}(z_1^n)|$ in place of $|\mathcal{H}|$ and then taking the worst case over samples yields a bound in terms of the growth function $\Pi_{\mathcal{H}}(n)$.
- Concretely,

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)]\right|\right] \leq \sqrt{\frac{\log \Pi_{\mathcal{H}}(n)}{n}}.$$

Application: uniform deviation of the CDF

- Let $Z_i(h) := \mathbf{1}\{X_i \leq h\} \in \{0, 1\}$, then

$$\sup_{h \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \right| = \sup_{h \in \mathbb{R}} \left| F_n(h) - F(h) \right|,$$

  is the uniform deviation between the empirical CDF F_n and the true CDF F.

Application: uniform deviation of the CDF

- Let $Z_i(h) := \mathbf{1}\{X_i \leq h\} \in \{0, 1\}$, then

$$\sup_{h \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \right| = \sup_{h \in \mathbb{R}} \left| F_n(h) - F(h) \right|,$$

  is the uniform deviation between the empirical CDF F_n and the true CDF F.

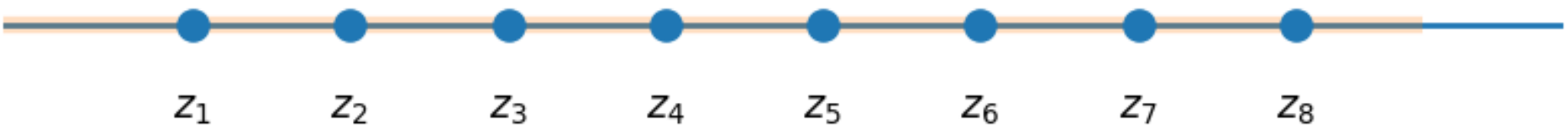
- We have shown that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sup_{h \in \mathbb{R}} \left| F_n(h) - F(h) \right| \leq \sqrt{\frac{\log \Pi_{\mathcal{H}}(n)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Application: uniform deviation of the CDF

- We have shown that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sup_{h \in \mathbb{R}} \left| F_n(h) - F(h) \right| \leq \sqrt{\frac{\log \Pi_{\mathcal{H}}(n)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$



Application: uniform deviation of the CDF

- Let $Z_i(h) := \mathbf{1}\{X_i \leq h\} \in \{0,1\}$, then

$$\sup_{h\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} Z_i(h) - \mathbb{E}[Z_1(h)] \right| = \sup_{h\in\mathbb{R}} \left| F_n(h) - F(h) \right|,$$

  a uniform deviation between the empirical CDF F_n and the true CDF F.

- We have shown that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{h\in\mathbb{R}} \left| F_n(h) - F(h) \right| \;\leq\; \sqrt{\frac{\log \Pi_{\mathcal{H}}(n)}{n}} \;+\; \sqrt{\frac{\log(1/\delta)}{2n}}.$$

- The class of half-intervals satisfies $\Pi_{\mathcal{H}}(n) \leq n+1$, and therefore

$$\sup_{h\in\mathbb{R}} \left| F_n(h) - F(h) \right| \;\leq\; \sqrt{\frac{\log(n+1)}{n}} \;+\; \sqrt{\frac{\log(1/\delta)}{2n}} \qquad \text{w.p. } \geq 1-\delta.$$

- **Interpretation:** a *Glivenko–Cantelli-type* guarantee with an explicit high-probability rate.

Main techniques recap

- **Chernoff bound:** turns tail probabilities into controlling a **MGF**; in many cases the MGF is bounded by a Gaussian-type condition (subgaussian assumption).
- **Hoeffding's inequality:** a sharp concentration result for **bounded** random variables.
- **McDiarmid's inequality:** controls deviations of functions with **bounded differences**; useful for **infinite suprema** by applying it to $\Phi(S) = \sup_{h \in \mathcal{H}}(\cdot)$, at the cost of introducing $\mathbb{E}[\Phi(S)]$.
- **Symmetrization:** a key tool to bound $\mathbb{E}[\Phi(S)]$ by rewriting it in terms of a **ghost sample** and reducing population expectations to behavior on **finite samples**.
- **Binary classification:** on a fixed sample, only finitely many labelings are possible; studying these via the **growth function** (and VC dimension bounds) yields explicit generalization guarantees.

What we did not cover

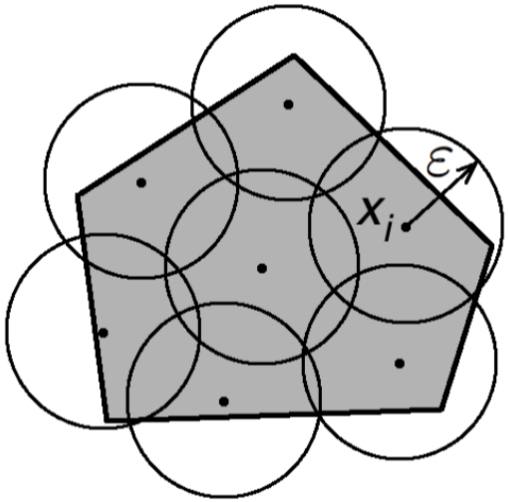
- **Bernstein/Bennett inequalities:** variance-sensitive bounds that can be much tighter than Hoeffding when $\mathrm{Var}(Z)$ is small.

What we did not cover

- **Bernstein/Bennett inequalities:** variance-sensitive bounds that can be much tighter than Hoeffding when $\mathrm{Var}(Z)$ is small.
- **Rademacher complexity:** after symmetrization, a standard way to control $\mathbb{E}[\Phi(S)]$ for *general* (non-binary) losses by measuring how well $\mathcal{H}$ correlates with random signs.

What we did not cover

- **Bernstein/Bennett inequalities:** variance-sensitive bounds that can be much tighter than Hoeffding when $\mathrm{Var}(Z)$ is small.
- **Rademacher complexity:** after symmetrization, a standard way to control $\mathbb{E}[\Phi(S)]$ for *general* (non-binary) losses by measuring how well $\mathcal{H}$ correlates with random signs.
- **ε-nets and covering numbers:** discretize an infinite hypothesis class by an ε-cover under a suitable metric, then combine union bounds with metric entropy to obtain uniform deviation bounds.



Spectral norm of a sub-gaussian matrix

- **Theorem.** Let $A \in \mathbb{R}^{m \times n}$ have independent, mean-zero, sub-gaussian entries A_{ij}. Let $K = \max_{i,j} \|A_{ij}\|_{\psi_2}$. Then for any $t > 0$,

$$\|A\| \;\leq\; CK\big(\sqrt{m} + \sqrt{n} + t\big) \qquad \text{with probability at least } 1 - 2e^{-t^2}.$$

Key identity:

$$\|A\| \;=\; \sup_{\|x\|_2=1,\ \|y\|_2=1} \langle Ax, y\rangle \;=\; \sup_{x \in S^{n-1},\ y \in S^{m-1}} \langle Ax, y\rangle.$$

it suffices to control $\langle Ax, y\rangle$ uniformly over the two spheres.

Spectral norm of a sub-gaussian matrix

- **Proof sketch (3 steps):**
  1. **Approximation (net):** take $\varepsilon = \frac{1}{4}$ and build nets $N \subset S^{n-1}$, $M \subset S^{m-1}$ with $|N| \leq 9^n$, $|M| \leq 9^m$, and use

$$\|A\| \leq 2 \max_{x \in N,\ y \in M} \langle Ax, y \rangle.$$

Spectral norm of a sub-gaussian matrix

- **Proof sketch (3 steps):**
  1. **Approximation (net):** take $\varepsilon = \frac{1}{4}$ and build nets $N \subset S^{n-1}$, $M \subset S^{m-1}$ with $|N| \leq 9^n$, $|M| \leq 9^m$, and use

$$\|A\| \;\leq\; 2 \max_{x \in N,\; y \in M} \langle Ax, y \rangle.$$

  2. **Concentration (fixed net points):** for fixed $x \in N, y \in M$, $\langle Ax, y \rangle$ is sub-gaussian and

$$\mathbb{P}\big(\langle Ax, y \rangle \geq u\big) \;\leq\; 2 \exp\Big(- c\, u^2 / K^2 \Big).$$

Spectral norm of a sub-gaussian matrix

- **Proof sketch (3 steps):**
  1. **Approximation (net):** take $\varepsilon = \frac{1}{4}$ and build nets $N \subset S^{n-1}$, $M \subset S^{m-1}$ with $|N| \leq 9^n$, $|M| \leq 9^m$, and use

$$\|A\| \;\leq\; 2 \max_{x \in N,\; y \in M} \langle Ax, y \rangle.$$

  2. **Concentration (fixed net points):** for fixed $x \in N$, $y \in M$, $\langle Ax, y \rangle$ is sub-gaussian and

$$\mathbb{P}\big(\langle Ax, y \rangle \geq u\big) \;\leq\; 2 \exp\Big(- c\, u^2 / K^2 \Big).$$

  3. **Union bound (over the net):** apply a union bound over $N \times M$ and choose $u = CK(\sqrt{n} + \sqrt{m} + t)$ to get

$$\mathbb{P}\Big(\max_{x \in N,\; y \in M} \langle Ax, y \rangle \geq u \Big) \;\leq\; 2e^{-t^2}, \quad \Rightarrow \quad \mathbb{P}\big(\|A\| \geq 2u\big) \leq 2e^{-t^2}.$$

References

- S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
- M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, 2nd edition, 2018.
- R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Obrigado!