Class intro

Objective. Develop graduate students into researchers who can
• understand and criticize papers in ML theory
• conjecture and prove new theorems that with high impact

Course structure

• First half: classical techniques for non-asymptotic risk bounds
  • Core reference: Martin Wainwright: High-dimensional statistics
    (available for free online via ETH)
• Second half: projects that review and extend current papers

Logistics

• Class website <sml.inf.ethz.ch/gml20/syllabus.html>
• Lecture slides will be uploaded after lectures
• TAs: Alexandru Tifrea, Amir Joudaki (Mondays 4-5 pm in G 69.3)
• Other platforms (everyone on waitlist / enrolled will be added):
  Piazza (announcements, questions), Gradescope (assignments)
• Install eduApp now! https://eduapp-app1.ethz.ch/
Evaluation & enrollment

Evaluation

• 3 homeworks (10%), midterm (50%), project (40%)
• HWs:
  • select questions graded by TAs, always self-graded
  • Homeworks out Fridays, assignments always due Fridays of the week
• Project (in groups of two):
  • Discussion & extension of one theoretical paper
  • 15-20 min Presentation in last four weeks
  • report (due June 14th)

Enrollment

• Criteria: Courses taken, degree program, research interest
• Current waitlist: ~90. Admitted: 37. Limit for admissions: ~50
• Final deadline to de-register: March 13th else no-show
• Others welcome to audit as long as there is space

Who is from which department?

Guarantees for Machine learning

1. Task: Make prediction or decisions on some data points \( z \in \mathcal{D}' \)
   • prediction, estimation
   • compression
   • anomaly detection . . .
2. Method: Train model \( f \) on data set \( \mathcal{D} \) (this class: similar to \( \mathcal{D}' \))
3. Evaluation: How well it works on \( \mathcal{D}' \)

Can you characterize when an algorithm does “well”?

Statistical translation

1. Task: assuming statistical truth \( \mathbb{P} \), interested in some “true” function \( f^* \) of \( \mathbb{P} \)
2. Method: learn estimate \( \hat{f} \) given samples \( z_i \sim \mathbb{P} \) in \( \mathcal{D} = \{z_i\}_{i=1}^n \)
3. How small is \( R_n(\hat{f}, \mathcal{D}') \)? How close is it to truth \( R_n(f^*, \mathcal{D}') \)?

Interesting generalized settings we will not cover:
• adversarial, distribution shift (course: Reliable and Interpretable AI)
• online setting where data set \( \mathcal{D} \) is not available at once
Machine Learning via empirical risk minimization

Define $f^*$ in terms of minimizer or some $\ell$:

1. Task: Given population risk $R(f) := R(f ; \mathbb{P}) = \mathbb{E}_z \ell(z ; f)$ truth is
   
   $$f^* := \min_f R(f)$$

2. Method: Given empirical risk $R_n(f) := R(f ; D') = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i ; f)$
   estimate is
   
   $$\hat{f}_n := \arg \min_{f \in \mathcal{F}} R_n(f)$$

3. Evaluation: Excess risk $E_R(n) := R(\hat{f}_n) - R(f^*) > 0$

Supervised example

- for $z = (x, y) \sim \mathbb{P}_{f^*}$: $y = f^*(x) + \epsilon$ for random $\epsilon$ and random $x$
- losses: $\ell(x, y; f) = (y - f(x))^2$, 0-1 loss: $\ell(x, y; f) = 1_{f(x) = y}$

Unsupervised example

- for $z \sim \mathbb{P}_{f^*}$: $z \sim \mathcal{N}(\mu, \Sigma)$ and $f^* = \mu$
- losses: $\ell(z; f) = \log \mathbb{P}_f(z)$ (log probability)

Factors that influence the excess risk

For the excess risk

$$E_R(n) := R(\hat{f}_n) - R(f^*) > 0$$

Questions to be discussed with neighbor

1. How is excess risk related to train and test error?
2. What are components of excess risk?
3. What does it depend on and what are tradeoffs?
4. Why do we need uniform laws to bound it?
Factors that influence excess risk

1. Train vs. test error, see risk decomposition next slide
2. Consists of:
   - approximation error (larger $F$, smaller $d$ better)
   - optimization error (Lipschitz, (strong) convexity better)
   - statistical error (larger $n$ better) ← this course
3. Depends on:
   - $\mathcal{F}$ (model class) & dimension $d$, $R$ (loss function), $n$ (sample size)
   - Optimization algorithm and stopping time
4. Why uniform bound, see risk decomposition next slide

---

Risk decomposition and generalization gap

Given $\hat{f}_n := \arg \min_{f \in \mathcal{F}} R_n(f)$ and $f^* \in \mathcal{F}$, does $\mathcal{E}_R(n) \to 0$?

First step: Decompose the risk

$$R(\hat{f}_n) - R(f^*) = R(\hat{f}_n) - R_n(\hat{f}_n) + R_n(\hat{f}_n) - R_n(f^*) + R_n(f^*) - R(f^*) \leq 0 \text{ by optimality}$$

$$\leq R(\hat{f}_n) - R_n(\hat{f}_n) + \underbrace{R_n(f^*) - R(f^*)}_{T_2} \leq \underbrace{R(\hat{f}_n) - R_n(\hat{f}_n)}_{T_1} + \underbrace{R_n(f^*) - R(f^*)}_{T_2}$$

(1)

- Can write $T_2 = \frac{1}{n} \sum_{i=1}^{n} t_i - \mathbb{E}t$ with $t_i := \ell(z_i; f^*)$
- $t_i$ i.i.d. R.V. for fixed $f \to T_2$ converges by concentration of averages
- $T_1$, often called generalization gap, is “similar” but more intricate

Test error $R_n'(\hat{f}_n)$ vs. train error $R(\hat{f}_n)$:

- $R_n'(\hat{f}_n) - R_n(\hat{f}_n) \leq R_n'(\hat{f}_n) - R(\hat{f}_n) + R(\hat{f}_n) - R_n(\hat{f}_n) \sim T_2 + T_1$
- Test minus train error is statistically similar to excess risk
- Since $T_2$ always concentrates, it’s a proxy for generalization gap
Different kinds of concentration

But how fast does it vanish? Is it optimal? For \( x \in \mathbb{R}^d, n \) sample size

- Asymptotics: e.g. for “large enough” \( n \), arbitrarily close to rate \( d/n \)
- Non-asymptotics: e.g. for any \( n, d \), w.p. \( \geq 1 - \delta \), \( \mathcal{E}_R(n) \leq O(d/n) \)
- Upper bounds: what an estimator achieves
- Lower bounds: what the best estimator can achieve for its hardest instance (minimax optimality)

In this course, we study the latter three.

Before we start with non-asymptotic bounds → a short primer on asymptotics

Asymptotic normality

Classical asymptotic theory: fixed \( d, n \to \infty \), for example

- (strong) Central Limit Theorem for empirical means: For i.i.d. \( X_i \) with \( \text{Var}(X_i) = \sigma^2 < \infty \), we have

\[
\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}X \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)
\]

- Asymptotic normality of Maximum Likelihood Estimator \( \hat{\theta}_{\text{MLE}} \) under regularity conditions:

\[
\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})
\]

with \( I(\theta) = \mathbb{E}\left( \frac{\partial \log p_\theta(X)}{\partial \theta} \right)^2 \) the Fisher information and \( p_\theta(X) \) the likelihood of \( X \).

But doesn’t hold for large \( d \) (on the order of \( n \))!
High-dimensional asymptotics

Instead of fixed $d$, we have both $d, n \to \infty$, for example

- High-dimensional unregularized linear regression, for $\frac{d}{n} \to \gamma > 1$:

\[ R(\theta_{\text{int}}) \xrightarrow{a.s.} r^2 \left( 1 - \frac{1}{\gamma} \right) + \sigma^2 \frac{1}{\gamma - 1} \]

- Debiased Lasso, for $\frac{\log p}{n} \to 0$

\[ \sqrt{n}(\hat{\theta}_{\text{Lasso}} - \theta^*) \xrightarrow{P} \mathcal{N}(0, \Sigma^{-1}) \]

Clicker-Q: What can we deduce about finite $d, n$?
With asymptotics, can we obtain a guarantee like

“with some constant probability we have $R(\hat{f}_n) - R(f^*) \leq O(\frac{d}{n})$”?

Non-asymptotics key ingredients

No! At most something like

- w/ some prob., for any $\epsilon$: $R(\hat{f}_n) - R(f^*) \leq \frac{d}{n} + \epsilon$ for large enough $n$

Wanted:

- w/ some prob., for all $d, n$ we have: $R(\hat{f}_n) - R(f^*) \leq O(\frac{d}{n})$

But asymptotics can be useful sometimes depending on the setting!

Today

1. Tight concentration bounds via Chernoff technique
2. Uniform bounds using Rademacher/Gaussian complexities
Tight tail and concentration bounds

Refresher (AML)

- Key idea: Markov $\Pr(X \geq t) \leq \frac{EX}{t}$ for $X \geq 0$; used on $e^{\lambda(X-EX)}$
- Chernoff $\Pr(X - EX \geq t) \leq \inf_{\lambda} \frac{\mathbb{E}[e^{\lambda(X-EX)}]}{e^{\lambda t}}$ for $\lambda$ well-defined $\mathbb{E}e^{\lambda X}$
- Bounds for random variables $X$
  - $\sigma$-subgaussian: $\mathbb{E}e^{\lambda X} \leq e^{\lambda^2 \sigma^2/2}$ $\Rightarrow$ $\Pr(X - EX \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}$
  - Hoeffding for $n$ i.i.d. $\sigma$-subgaussian: $\Pr(\frac{1}{n} \sum_{i=1}^{n} X_i - EX \geq t) \leq e^{-\frac{nt^2}{2\sigma^2}}$

Bounded variable $X \in [a, c]$

- is subgaussian with $\delta = \frac{c-a}{2}$
- if also $\mathbb{E}(X - \mu)^k \leq \nu^2$, get tighter:

**Theorem (Bernstein’s inequality)**

*If moments of $X$ are bounded as $|\mathbb{E}(X - \mu)^k| \leq \frac{1}{2} k! \nu^2 b^{k-2}$ for $k = 2, 3, \ldots$ then*

$$
\Pr(|X - EX| \geq t) \leq 2e^{-c \frac{t^2}{2(\nu^2 + bt)}}
$$

Bernstein’s inequality (ctd)

- For bounded $X \in [a, c]$, it holds that $|\mathbb{E}(X - \mu)^k| \leq \nu^2 (c - a)^{k-2}$
- For i.i.d. sums this *could* be tighter than subgaussian bound, e.g.
  - if $t \ll \frac{\nu^2}{c-a}$ and $\nu^2 \ll (c - a)^2$
- Writing $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

<table>
<thead>
<tr>
<th>Subgaussian concentration</th>
<th>Bernstein concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Pr(</td>
<td>\bar{X} - EX</td>
</tr>
</tbody>
</table>

*Proof sketch:*

- Taylor’s theorem on MGF and sum of geometric series with $|\lambda| < \frac{1}{b}$
  $$
  \mathbb{E}e^{\lambda X} \leq 1 + \frac{\lambda^2 \sigma^2/2}{1 - b|\lambda|} \leq e^{\frac{\lambda^2 \sigma^2}{1 - b|\lambda|}}
  $$
- Set $\lambda = \frac{t}{bt + \sigma^2}$

**Even tighter:** Bennett’s inequality (MW Exercise 2.7.)
Functional concentration bounds
Back to eq. 1, how about term $T_1$?

$$R_n(\hat{f}_n) = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i; \hat{f}_n) \neq \text{not an empirical mean of i.i.d. R.V.}$$

$=: \tilde{t}_i \ \text{non- i.i.d.}$

$\rightarrow$ for uniform laws people use the hammer inequality:

$$R(\hat{f}_n) - R_n(\hat{f}_n) \leq \sup_{f \in \mathcal{F}} R(f) - R_n(f) =: Z_n$$

We write $Z_n = g(z_1, \ldots, z_n)$, a function of R.V. $z_i \in \mathcal{Z}$

and define $z \backslash^k$ with $z_j \backslash^k = \begin{cases} z_j & \text{if } j \neq k \\ z' & \text{if } j = k \end{cases}$

**Theorem (McDiarmid / bounded differences)**

Given $z, z' \in \mathcal{Z}^n$. If $\forall k, z, z' |g(z) - g(z \backslash^k)| \leq \sigma_k$ holds, we have

$$\mathbb{P}(g(z) - \mathbb{E}g(z) \geq t) \leq e^{-\frac{2t^2}{\sum \sigma_k^2}}$$

More on (uniform) functional bounds

**Theorem (Lipschitz $g$)**

$g$ is $L$-Lipschitz (wrt Euclidean norm) for i.i.d. $x_i$ then if

(i) $x_i \sim \mathcal{N}(0, \sigma^2)$ or

(ii) $x_i \in [a, b]$ (so that $\sigma = b - a$) and $g$ is also separately convex

$$\mathbb{P}(g(x) - \mathbb{E}g(x) \geq t) \leq e^{-\frac{ct^2}{L^2\sigma^2}}$$

see MW Thm. 2.26 for (i), and 3.4. for (ii)

$n$ enters in (i) via $t \sim \frac{1}{n}$, in (ii) via $\sigma \sim \frac{1}{n}$

Specifically for the sup of empiric. proc. can give tighter bounds:

- bounded (Functional Hoeffding MW 3.26)
- bounded + variance (Functional Bernstein / Talagrand MW 3.27)
Uniform tail bounds and Rademacher complexity

- Define $\mathcal{H} = \{ h : h(\cdot) = \ell(\cdot; f) \; \forall f \in \mathcal{F} \}$
- $\epsilon_i$ are i.i.d. Rademacher R.V.: $\epsilon_i = 1$ w/ prob. 1/2 and else $-1$

**Theorem (Uniform law)**

For $b$-uniformly bounded $\mathcal{H}$ with $R_n(\mathcal{H}) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \ell(z_i; f)$

$$
\mathbb{P} \left( \sup_{f \in \mathcal{F}} R(f) - R_n(f) \geq 2R_n(\mathcal{H}) + t \right) \leq e^{-\frac{nt^2}{2b^2}}
$$

Consequently, if $R_n(\mathcal{H}) = o(1)$, then $\sup_{f \in \mathcal{F}} R(f) - R_n(f) \xrightarrow{a.s.} 0$.

**Proof steps using** $Z_n := \sup_{f \in \mathcal{F}} R_n(f) - R(f) = \sup_{h \in \mathcal{H}} \bar{h} - \mathbb{E} h$

1. McDiarmid w/ $Z_n = g(z)$, $z \in \mathcal{Z}^n \Rightarrow \mathbb{P}(Z_n - \mathbb{E} Z_n \geq \delta) \leq e^{-\frac{n\delta^2}{2b^2}}$
   - For $b$-uniformly bounded $\mathcal{H}$: $|g(z) - g(z^{\backslash k})| \leq \frac{2b}{n} =: \sigma \; \forall k, z, z'$
2. Symmetrization: $\mathbb{E} \sup_{h \in \mathcal{H}} \bar{h} - \mathbb{E} h \leq \mathbb{E} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h(z_i)$

**References**

Tail bounds and uniform bounds using Rademacher complexity:

- MW Chapters 2, 4

Concentration bounds on suprema of empirical processes:

- MW Chapter 3
- Ledoux, Talagrand: Probability for Banach spaces for functional
  Bernstein

Asymptotics:

- Keener: Theoretical Statistics
- For debiased Lasso: On asymptotically optimal confidence regions
  and tests for high-dimensional models, Sara van de Geer, Peter
  Bühlmann, Ya’acov Ritov, and Ruben Dezeure, Annals of Statistics,
  2014.
- Ridge regression: High-dimensional asymptotics of prediction: Ridge
  regression and classification, Edgar Dobriban and Stefan Wager,