
Costis Daskalakis
Aleksander Mądry
what, when, how do deep NNs learn?
e.g. Classification

• Basic learning task: design function $h: \mathcal{X} \to \mathcal{C}$, mapping objects from some set $\mathcal{X}$ to their class label in $\mathcal{C}$

• e.g. $\mathcal{X}$: images of cats and dogs, $\mathcal{C} = \{0, 1\}$

• How to do this?
  1. identify “expressive enough” family of functions $\mathcal{H}$
  2. use examples to choose some “good” $h \in \mathcal{H}$
e.g. Classification

• Basic learning task: design function \( h: \mathcal{X} \rightarrow \mathcal{C} \), mapping objects from some set \( \mathcal{X} \) to their class label in \( \mathcal{C} \)

• e.g. \( \mathcal{X} \): images of cats and dogs, \( \mathcal{C} = \{0,1\} \)

• How to do this?
  1. identify “expressive enough” family of functions \( \mathcal{H} \)
     • e.g. \( \mathcal{H} \) all convolutional nets of certain width and depth
  2. use examples to choose some “good” \( h \in \mathcal{H} \)
     • each example is a pair \((x, y)\) of an image and its label
     • output empirical risk minimizer:
       \[
       \hat{h} \in \arg\max_{h \in \mathcal{H}} \sum_{\text{examples } (x_i,y_i)} 1_{h(x_i)=y_i}
       \]
e.g. Classification

- identify “expressive enough” family of functions $\mathcal{H}$
  - e.g. $\mathcal{H}$ all convolutional nets of certain width and depth
- use examples to choose some “good” $h \in \mathcal{H}$
  - output empirical risk minimizer $\hat{h} \in \operatorname{argmax}_{h \in \mathcal{H}} \sum_{(x_i, y_i) \in \mathcal{E}} 1_{h(x_i) = y_i}$
- hope: $\mathbb{E}_{(X,Y) \sim F} \left[ 1_{\hat{h}(X) = Y} \right] \geq \max_{h \in \mathcal{H}} \mathbb{E}_{(X,Y) \sim F} \left[ 1_{h(X) = Y} \right] - \epsilon$ (*)
  - $F$: true distribution of (image, class label) pairs to be encountered in the future
  - presumably training set of examples were drawn from $F$
- Two questions:
  1. How close is $\max_{h \in \mathcal{H}} \mathbb{E}_{(X,Y) \sim F} \left[ 1_{h(X) = Y} \right]$ to $\max_{h: \text{unrestricted}} \mathbb{E}_{(X,Y) \sim F} \left[ 1_{h(X) = Y} \right]$?
  2. How fast does $\epsilon$ in (*) decay in the number of examples $N$?

- Rich $\mathcal{H}$ ⇒ 1 good, 2 bad
- Poor $\mathcal{H}$ ⇒ 1 bad, 2 maybe good
- For 1, use a rich enough family $\mathcal{H}$
- For 2, bound the “dimensionality” of $\mathcal{H}$, get “generalization bounds”

These imply small $\epsilon$ in (*)
Generalization Bounds

• Skip what a “generalization bound” is for a moment
  • central topic in ML theory
  • **Today:** Vapnik–Chervonenkis (VC) theory of generalization

• Consider a class of Boolean functions \( \mathcal{H} = \{h: \mathcal{X} \to \{0,1\}\} \)

• **Def:** VC dimension of \( \mathcal{H} = \max \) #points \( \mathcal{H} \) can **shatter**
  • points \( x_1, \ldots, x_k \in \mathcal{X} \) are shattered by \( \mathcal{H} \) iff \( \forall \) 0/1 patterns \( \sigma \in \{0,1\}^k \) \( \exists \) a function \( h \in \mathcal{H} \) whose values on the points \( x_1, \ldots, x_k \) equal \( \sigma \), i.e. \( h(x_i) = \sigma_i, \forall i \)
  • e.g. say \( \mathcal{H} = \{\text{halfplanes in } \mathbb{R}^2\} \)
  • \( \text{VC}(\mathcal{H}) = 3 \)
Generalization Bounds

• Skip what a "generalization bound" is for a moment
  • central topic in ML theory
  • Here: Vapnik–Chervonenkis (VC) theory of generalization

• Consider a class of Boolean functions $\mathcal{H} = \{ h: \mathcal{X} \rightarrow \{0,1\} \}$

• Def: VC dimension of $\mathcal{H}$ = max #points $\mathcal{H}$ can shatter
  • points $x_1, \ldots, x_k \in \mathcal{X}$ are shattered by $\mathcal{H}$ iff $\forall$ 0/1 patterns $\sigma \in \{0,1\}^k \exists$ a function $h \in \mathcal{H}$ whose values on the points $x_1, \ldots, x_k$ equal $\sigma$, i.e. $h(x_i) = \sigma_i$, $\forall i$
  • e.g. say $\mathcal{H} = \{\text{halfplanes in } \mathbb{R}^2\}$
  • $\text{VC}(\mathcal{H}) = 3$ (proved lower bound in previous slide, upper bound is left an exercise)

• VC Theorem: Suppose $\mathcal{H}$ is a class of Boolean functions w/ VC-dimension $d$. Then given:
  \[
  N \approx \frac{(d \cdot \ln(1/\epsilon) + \ln(1/\delta))}{\epsilon^2}
  \]
  samples $(X_1, Y_1), \ldots, (X_N, Y_N) \sim F$ we have that, w/ prob $\geq 1 - \delta$,
  \[
  \forall h \in \mathcal{H}: \left| \mathbb{E}_{(X,Y) \sim F}[1_{h(X)=Y}] - \frac{1}{N} \sum_i 1_{h(x_i)=y_i} \right| \leq \epsilon
  \]
Generalization Bounds

• How to prove?
  • Many ways, central topic in ML theory
  • **Here**: Vapnik–Chervonenkis (VC) theory of generalization
  • Similar generalization theorems exist for real-valued functions via Rademacher complexity, pseudo-dimension, ...
  • they also exist for different access to examples
  • It is a well-developed theory

• Disconnect with practical performance of Deep NNs:
  • VC/Rademacher complexity/Pseudo-dimension of Deep NNs too large compared to sample size: is there overfitting?
  • Finding ERM is sort of hopeless; maybe SGD finds local optimum:
    • maybe a good thing?
    • Is there an optimality vs overfitting tradeoff?
    • Is stochasticity in GD also a good thing?
  • Role of optimization method, max pooling, dropout?
  • Training set: attacks because training set non-representative or because of overfitting?
Generative Adversarial Networks

- Algorithms mapping white noise to high-dimensional objects with structure:
  \[ z \sim N(0, I_{100 \times 100}) \]
  face GAN

- If you want, what human imagination does (presumably)
- Trained using samples (e.g. faces) from true high-dimensional distribution with structure (e.g. natural face images)
- **Statistical Question**: after GAN has been trained, did it really learn the underlying structured high-dimensional distribution?
- Or did it “memorize” the training set?
A Hypothesis Testing Problem

- Sample access to $F$: distribution of true faces
- Sample + white-box access to $Q$: GAN, and its output
- **Goal**: distinguish $d(F, Q) \leq \varepsilon_1$ vs $d(F, Q) \geq \varepsilon_2$
- Really well-studied problem in Statistics, Information Theory, TCS
- Trouble is:
  - what is the right distance $d$ to use?
  - $F, Q$: high-dimensional (e.g. face image distributions)
    - Statistical tests commonly require exponentially many samples in the dimension, unless one has deeper understanding of structure in both $F$ and $Q$
    - e.g. even if $Q$ is trivial (product measure), and $d$ is total variation distance, answering above question requires exponentially many samples in the dimension.
- What is the right statistical lens via which to approach this question?
Game Theory
GAN Training

- Think $F$: true high-dimensional distribution (e.g. faces) in $\mathbb{R}^n$
- $Q$: output of a Deep NN $G$, of certain architecture, with parameters $\theta$
  - i.e. $G_\theta(z)$, where $z \sim N(0, I)$
- Suppose interested in Wasserstein distance:
  $$W(F, Q) = \sup_{D: \mathbb{R}^n \rightarrow \mathbb{R}, 1-\text{Lipschitz}} (\mathbb{E}_{X \sim F}[D(X)] - \mathbb{E}_{X \sim Q}[D(X)])$$
- In a perfect world, $G_\theta$ should minimize:
  $$\inf_{\theta} \sup_{D: \mathbb{R}^n \rightarrow \mathbb{R}, 1-\text{Lipschitz}} (\mathbb{E}_{X \sim F}[D(X)] - \mathbb{E}_{Z \sim N(0, I)}[D(G_\theta(Z))])$$
- In practice, hard to compute sup over all Lipschitz functions, so only take sup over all Deep NNs $D$, of certain architecture, w/ parameters $w$:
  $$\inf_{\theta} \sup_{w} (\mathbb{E}_{X \sim F}[D_w(X)] - \mathbb{E}_{Z \sim N(0, I)}[D_w(G_\theta(Z))])$$
- In other words, set up a game between a Generator deep NN, and a Discriminator deep NN
GAN Training

• A **game** between a *Generator* deep NN, w/ parameters $\theta$, and a *Discriminator* deep NN, w/ parameters $w$:

$$\inf_{\theta} \sup_w \mathbb{E}_{X \sim F} [D_w(X)] - \mathbb{E}_{Z \sim N(0,I)} [D_w(G_{\theta}(Z))]$$

• **Training**: generator and discriminator run some variant of gradient descent each to update their parameters $\theta, w$; expectations are approximated by finite sample averages.
GAN Training

• A game between a Generator deep NN, w/ parameters $\theta$ and a Discriminator deep NN, w/ parameters $w$:
  \[
  \inf_{\theta} \sup_{w} \left( \mathbb{E}_{X \sim F} [D_{w}(X)] - \mathbb{E}_{Z \sim N(0,I)} [D_{w}(G_{\theta}(Z))] \right)
  \]

• Training: generator and discriminator run some variant of gradient descent each to update their parameters $\theta, w$; expectations are approximated by finite sample averages (even ignore errors coming from this – assume access to true expectations)

• Will gradient descent converge?

• If yes, to what?
The Min-Max Theorem

- **[von Neumann 1928]:** If \( X \subset \mathbb{R}^n, Y \subset \mathbb{R}^m \) are compact and convex, and \( f: X \times Y \rightarrow \mathbb{R} \) is convex-concave (i.e. \( f(x, y) \) is convex in \( x \) for all \( y \) and is concave in \( y \) for all \( x \)), then

\[
\min_{x \in X} \max_{y \in Y} f(x, y) = \max_{y \in Y} \min_{x \in X} f(x, y)
\]

- Min-max optimal \((x, y)\) is essentially unique (unique if \( f \) is strictly convex-concave, o.w. a convex set of solutions)

- von Neumann: "As far as I can see, there could be no theory of games ... without that theorem ... I thought there was nothing worth publishing until the Minimax Theorem was proved"

- Equivalent to strong LP duality

- **[Blackwell,...]:** A host of uncoupled update-rules (dynamics) applied by the min and the max players “converge” to min-max equilibrium

- *no-regret learning dynamics:* e.g. Multiplicative-weights-update, follow-the-regularized-leader, follow-the-perturbed-leader, etc.

- Follow-the-regularized-leader with \( \ell_2 \)-regularization \( \equiv \) gradient descent
Challenges

• “Convergence” of online learning to min-max solutions for convex-concave functions $f(x, y)$ only happens in an average sense
  • E.g. gradient descent for $f(x, y) = x \cdot y$

![Diagram showing spiral convergence]

• Objective function in Wasserstein GAN training isn’t even convex-concave

• Questions:
  • Stability: how to converge to local saddles?
  • Generalization: Effects of approximation of expectation with sample averages?
Game Playing
Deep Mind

• Stated Mission: Solve intelligence, use it to make the world a better place.
• ...
• We’ll take a look at the guts of AlphaGo, and AlphaGo Zero
• Connection to Reinforcement Learning, Policy and Value Iteration, and the Min-Max Theorem
Outlook

• Really small sample size: health data
• Robust Statistics
• Causality + Counterfactuals
• Privacy concerns
• Fairness
• Ethical Considerations
• Philosophical ramifications of unreasonable practical success of Deep Learning
6.883 Statement of Purpose:
- to entice the practically-minded into theory as a means to understand and improve practice
- to entice the theoretically-minded into the deep questions motivated by practical experience