Contents

Lectures

Contents

Lectures **September 20, 2021**

Supervised Learning and Empirical Risk Minimization

Table of Contents Local *Written by Brandon McKinzie*

All the basic stuff:

- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ accepts model prediction $\hat{y}$ and the true label y. We assume $\ell(\hat{y}, y) \geq 0$.
- Goal: find a model $h : \mathcal{X} \mapsto \mathcal{Y}$ that minimizes the **expected loss**:

$$L(h) \triangleq \mathbb{E}_{(x,y)\sim p}\left[\ell(h(x), y)\right] \tag{1}$$

AKA population loss, expected risk, population risk

- **Hypothesis class** $\mathcal{H}$ is a set of functions $h : \mathcal{X} \mapsto \mathcal{Y}$ that we want to consider/search over for finding the best one.
- **Excess Risk** of a specific model h wrt the hypothesis class $\mathcal{H}$ is the difference bw the population risk of h and the best possible population risk inside $\mathcal{H}$:

$$E(h) \triangleq L(h) - \inf_{g \in \mathcal{H}} L(g) \tag{2}$$

Empirical Risk Minimization [38:50].

- Although we'd like to minimize the *population* risk $L(h)$ (eq 1), we only have access to our finite training set, so we instead compute the **empirical risk**:

$$\hat{L}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x^{(i)}), y^{(i)}) \tag{3}$$

- **Empirical Risk Minimization** (ERM) is the method of finding the minimizer of $\hat{L}$:

$$\hat{h} \triangleq \arg\min_{h \in \mathcal{H}} \hat{L}(h) \tag{4}$$

- The empirical risk is an unbiased estimator of the population risk:

$$\mathbb{E}_{P(x,y)}\left[\hat{L}(h)\right] = L(h) \tag{5}$$

which is due to our assumption that each example is drawn iid from P.

The key question that we seek to answer in the first part of this course is: what guarantees do we have on the excess risk for the parameters learned by ERM? The hope with ERM is that minimizing the training error will lead to small testing error. One way to make this rigorous is by showing that the ERM minimizer's excess risk is bounded.

Asymptotic Analysis [44:00]. An asymptotic approach is one that considers $n \to \infty$ and tries to derive bounds on quantities of interest. Specifically for ERM, our goal will be to show the excess risk (eq 2) of $\hat{\theta}_{ERM}$ is small[1]. Specifically, we'll prove that the excess risk as bounded as below:

$$L(\hat{\theta}) - \arg\min_{\theta\in\Theta} L(\theta) \leq \frac{c}{n} + o\left(\frac{1}{n}\right) \tag{6}$$

where c is a problem-dependent constant that does not depend on n, and $o(\frac{1}{n})$ just means "terms that are lower-order than $\frac{1}{n}$.

Theorem 2.1

$\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}^p\}$
$\hat{\theta} \triangleq \arg\min_{\theta\in\Theta} \hat{L}(h_\theta)$
$\theta^* = \arg\min_\theta L(\theta)$.

Assume the following:

- *Consistency of $\hat{\theta}$: $\hat{\theta} \xrightarrow{p} \theta^*$ as $n \to \infty$*[2] *[57:40]*
- *The hessian*[3] *$\nabla^2 L(\theta^*)$ is full rank.*
- *"Other appropriate regularity conditions hold."*

Note that $\hat{\theta}$ is a function of n; we just are omitting that for compactness

- ► $\sqrt{n}\left(\hat{\theta} - \theta^*\right) = O_P(1)$, *where $O_P(1)$ reads "is bounded in probability to 1"*[4]. .
- ► $n\left(L(\hat{\theta}) - L(\theta^*)\right) = O_P(1)$
- ► $\sqrt{n}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{H}^{-1} Cov\left[\nabla \ell\left((x,y);\theta^*\right)\right] \boldsymbol{H}^{-1}\right)$ *where $\boldsymbol{H} = \nabla^2 L(\theta^*)$. The notation $\mathbf{x} \xrightarrow{d} p(\mathbf{x})$ means "the distribution of the random variable $\mathbf{x}$ converges to $p(\mathbf{x})$."*
- ► $n\left(L(\hat{\theta}) - L(\theta^*)\right) \xrightarrow{d} \frac{1}{2}||S||_2^2$ *where $S \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{H}^{-1} Cov\left[\nabla \ell\left((x,y);\theta^*\right)\right] \boldsymbol{H}^{-1}\right)$.*
- ► $\lim_{n\to\infty} \mathbb{E}\left[n\left(L(\hat{\theta}) - L(\theta^*)\right)\right] = \frac{1}{2} tr\left(\boldsymbol{H}^{-1} Cov\left[\nabla \ell\left((x,y);\theta^*\right)\right]\right)$

Thoughts/observations.

- In English, the first property just reads "the difference between the ERM minimizer and the true minimizer doesn't blow up as $n \to \infty$; it is bounded/converges in some sense." However, it's weird that you need to tack on the $\sqrt{n}$. Isn't the following equivalent: $(\hat{\theta} - \theta^*) = O_P(\frac{1}{\sqrt{n}})$.
- I have no idea how/why you can say that $\Delta L \approx \frac{1}{n}$ but $\Delta\theta \approx \frac{1}{\sqrt{n}}$, i.e. that $\Delta\theta / \Delta L \approx \sqrt{n}$

[1] $\hat{\theta}_{ERM} \triangleq \arg\min_{\theta\in\Theta} \hat{L}(\theta)$. Also note that $\theta \in \Theta$ is used interchangeably with $h \in \mathcal{H}$.

[2] The notation $a \xrightarrow{p} b$ means "converges in probability," which is what we say when a and b are random variables. Formally:

$$\lim_{n\to\infty} \Pr\left[||\hat{\theta} - \theta^*|| \geq \epsilon\right] = 0 \quad \forall \epsilon > 0 \tag{7}$$

[3] The Hessian matrix the symmetric matrix defined by

$$\left(\nabla^2 L(\theta^*)\right)_{i,j} \triangleq \frac{\partial^2 L}{\partial\theta_i^* \partial\theta_j^*} \tag{8}$$

It is also the Jacobian matrix of the gradient of the function f: $\boldsymbol{H}(f(\boldsymbol{x})) = \boldsymbol{J}(\nabla f(\boldsymbol{x}))$.

[4] A sequence of random variables $\{\boldsymbol{x}_n\}$ is **bounded in probability** if for any $\epsilon > 0$, there exist M and N such that $\Pr\left[||\boldsymbol{x}_n|| > M\right] < \epsilon$ for all $n > N$.

Review: Taylor Series

A **power series** is an infinite series of the form $\sum_{k=0}^{\infty} c_k(x-a)^k$ where the coefficients c_k and the **center** of the series a are constants. Recall that if a function f is differentiable at a point a, it can be approximated near a by its tangent line:

$$f(x) \approx f(a) + f'(a)(x-a) \tag{9}$$

Let's denote this by $p_1(x)$ – this is a first-order approximation of f at a, since $p_1(a) = f(a)$ and $p_1'(a) = f'(a)$. We can find successively higher-order approximations by adding the next term in the power series, which here would be $c_2(x-a)^2$, and solving for c based on the constraints that all derivatives (from order 0 to n – where n is 2 here) must be the same as the original function's. For the 2nd order approximation, we end up finding that $c_2 = \frac{1}{2}f''(a)$.
The general form of a Taylor expansion for some function $f(x)$ about x=a is thus

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \tag{10}$$

So, how good are these approximations? **Taylor's Theorem** states:

$$f(x) = p_n(x) + R_n(x) \tag{11}$$

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1} \tag{12}$$

for some point c between x and a.

Review: Central Limit Theorem (CLT) (Theorem 2.2)

Let $X_1, \cdots, X_n$ be iid random variables, and let $\hat{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Assume the covariance matrix Σ **is finite** (?). Then, as $n \to \infty$, we have

$$\hat{X} \xrightarrow{p} \mathbb{E}[X] \tag{13}$$

$$\sqrt{n}\left(\hat{X} - \mathbb{E}[X]\right) \xrightarrow{d} \mathcal{N}(0, \Sigma) \tag{14}$$

$$\sqrt{n}\left(\hat{X} - \mathbb{E}[X]\right) = O_P(1) \tag{15}$$

Lemma 2.3

- *If $Z \sim \mathcal{N}(0, \Sigma)$ and A is a* ***deterministic matrix****, then $AZ \sim \mathcal{N}\left(0, A\Sigma A^T\right)$.*
- *If $Z \sim \mathcal{N}\left(0, \Sigma^{-1}\right)$ and $Z \in \mathbb{R}^p$, then $Z^T\Sigma Z \sim \chi^2(p)$, where $\chi^2(p)$ is the chi-squared distribution with p degrees of freedom.*

Proof of Theorem 2.1 [1:12:00].

$$0 = \nabla\hat{L}(\hat{\theta}) = \nabla\hat{L}(\theta^*) + \nabla^2\hat{L}(\theta^*)(\hat{\theta} - \theta^*) + O(||\hat{\theta} - \theta^*||_2^2) \tag{16}$$

$$\hat{\theta} - \theta^* \approx -\left(\nabla^2\hat{L}(\theta^*)\right)^{-1}\nabla\hat{L}(\theta^*) \tag{17}$$

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \approx -\left(\nabla^2\hat{L}(\theta^*)\right)^{-1}\sqrt{n}\nabla\hat{L}(\theta^*) \tag{18}$$

Summary. First, the main notation that was new to me:

- **Converges in Probability**: (wiki) A sequence $\{X_n\}$ of random variables **converges in probability** towards the random variable X if for all $\epsilon > 0$:

$$\lim_{n \to \infty} \Pr\left[||X_n - X|| > \epsilon\right] = 0 \tag{19}$$

- **Bounded in Probability** (stackexchange) (wiki) The notation

$$X_n = O_p(a_n) \tag{20}$$

means that the set of values $\{X_n / a_n\}$ is **stochastically bounded**. That is, for any $\epsilon > 0$, there exists a finite $M > 0$ and a finite $N > 0$ such that

$$\Pr\left[|X_n / a_n| > M\right] < \epsilon \quad (\forall n > N) \tag{21}$$

The main equations to remember:

$$\textbf{[Population Risk]} \qquad L(\theta) \triangleq \mathbb{E}_{(x,y)\sim p}\left[\ell\left(x, y; \theta\right)\right] \tag{22}$$

$$\textbf{[Excess Risk]} \qquad E(\theta) \triangleq L(\theta) - \inf_{\theta' \in \Theta} L(\theta') \tag{23}$$

$$\textbf{[Empirical Risk]} \qquad \hat{L}(\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell\left(x^{(i)}, y^{(i)}; \theta\right) \tag{24}$$

$$\textbf{[ERM]} \qquad \hat{\theta} \triangleq \arg\min_{\theta \in \Theta} \hat{L}(\theta) \tag{25}$$

$$\textbf{[Bound on Excess Risk]} \qquad L(\hat{\theta}) - \arg\min_{\theta \in \Theta} L(\theta) \leq \frac{c}{n} + o\left(\frac{1}{n}\right) \tag{26}$$

Lectures **October 18, 2021**

Covering Number, Dudley Theorem, Generalization Bounds for Deep Nets

Table of Contents Local *Written by Brandon McKinzie*

Proof of Dudley's Theorem [28:40]

wow this is extremely long, it lasts until **[1:00:00]**
then interpretation stuff until **[1:15:00]**

Covering Number Bounds for Linear Models [1:15:00].

Theorem 3 of Zhang'02

Suppose $\{x_1, \ldots, x_n\}$ *are* n *data points. Let* p,q *satisfy* $\frac{1}{p} + \frac{1}{q} = 1$, *with* $2 \leq p \leq \infty$. *Assume that* $||x_i||_p \leq C$ $(\forall i)$. *Let* $\mathcal{F}_q \triangleq \{x \mapsto w^x : ||w||_q \in B\}$. *Let* $\rho := L_2(p_n)$. *Then*

$$\log N(\epsilon, \mathcal{F}_q, \rho) \leq \left[\frac{B^2C^2}{\epsilon^2}\right] \lg(2d+1) \tag{27}$$

When $p = q = 2$, *can strengthen slightly to*

$$\log N(\epsilon, \mathcal{F}_2, \rho) \leq \left[\frac{B^2C^2}{\epsilon^2}\right] \lg(2\min\{n, d\}+1) \tag{28}$$

Lectures **November 15, 2021**

Implicit Regularization of Noise in SGD

Table of Contents Local *Written by Brandon McKinzie*

Noisy SGD. Loss function $g(\theta)$. SGD does

$$\theta_{t+1} = \theta_t - \eta \left(\nabla g(\theta_t) + \xi_t \right) \tag{29}$$

where $\mathbb{E}\left[\xi_t\right] = 0$, and the distribution of ξ_t might depend on θ_t.

Warmup: $\frac{1}{2}x^2$

Let $g(x) := \frac{1}{2}x^2$. Also, denote the scale of the noise as σ, and let $\xi_t \sim \mathcal{N}\left(0, 1\right)$.

$$x_{t+1} = x_t - \eta \left(\nabla g(x_t) + \sigma \xi_t \right) \tag{30}$$

$$x_{t+1} = (1-\eta)^{t+1} x_0 - \eta\sigma \sum_{k=0}^{t} \xi_{t-k} (1-\eta)^k \tag{31}$$

$$\implies \lim_{t\to\infty} x_t \sim \mathcal{N}\left(0, \Theta(\eta\sigma^2)\right) \tag{32}$$

In other words, x_t eventually reaches global minimum and bounces around it. Note that this “bouncing” is more affected by the noise than the learning rate. **[23:45]**

Summaries

Contents

Summaries **October 16, 2021**

Notation

Table of Contents Local *Written by Brandon McKinzie*

Small-O Notation

The statement $f(x) = o(g(x))$ means

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0 \tag{33}$$

Converges in Probability

NB: ok so it seems like the two notations $X_n \xrightarrow{p} X$ and $X_n = o_p(a_n)$ are both called "converges in probability" but mean COMPLETELY different things.

Convergence to a random variable *($X_n \xrightarrow{p} X$). A sequence $\{X_n\}$ of random variables* ***converges in probability*** *towards the random variable X if for all $\epsilon > 0$:*

$$\lim_{n \to \infty} Pr\left[||X_n - X|| > \epsilon\right] = 0 \tag{34}$$

(wiki)

Convergence to a constant *($X_n = o_p(a_n)$). The following two equivalent notations*

$$X_n = o_p(a_n) \qquad \textit{and} \qquad \frac{X_n}{a_n} = o_p(1) \tag{35}$$

both mean

$$\lim_{n \to \infty} Pr\left[\left|\frac{X_n}{a_n}\right| \geq \epsilon\right] = 0 \qquad (\forall \epsilon > 0) \tag{36}$$

i.e. "For large enough n, X_n converges to something [much] less than a_n."

Bounded in Probability $X_n = O_p(a_n)$

The notation

$$X_n = O_p(a_n) \tag{37}$$

means that the set of values $\{X_n/a_n\}$ is ***stochastically bounded****. That is, for any $\epsilon > 0$, there exists a finite $M > 0$ and a finite $N > 0$ such that*

$$Pr\left[|X_n/a_n| > M\right] < \epsilon \quad (\forall n > N) \tag{38}$$

(stackexchange) (wiki)

Comparison between O_p and o_p

Specifically, difference between $X_n = O_p(a_n)$ and $X_n = o_p(a_n)$:

$$[O_p] \qquad (\forall \epsilon > 0)(\exists \delta_\epsilon)(\exists N_\epsilon) \text{ s.t. } \Pr\left[\left|\frac{X_n}{a_n}\right| \geq \delta_\epsilon\right] \ \leq \ \epsilon \quad (\forall n > N_\epsilon) \tag{39}$$

$$[o_p] \qquad (\forall \epsilon > 0)(\forall \delta)(\exists N_{\epsilon,\delta}) \text{ s.t. } \Pr\left[\left|\frac{X_n}{a_n}\right| \geq \delta\right] \ \leq \ \epsilon \quad (\forall n > N_{\epsilon,\delta}) \tag{40}$$

Basically, the difference is all in the δ, and we can see that

$$X_n = o_p(a_n) \quad \Longrightarrow \quad X_n = O_p(a_n) \tag{41}$$

Summaries **October 16, 2021**

Asymptotic Analysis

Table of Contents Local *Written by Brandon McKinzie*

Goal: bound the excess risk $E(\hat{\theta}) \triangleq L(\hat{\theta}) - L(\theta^*)$[5] as follows:

$$E(\hat{\theta}) \leq \frac{c}{n} + o(\frac{1}{n}) \tag{42}$$

Asymptotic Bounds for Excess Risk

Assume the following:

- *Consistency of $\hat{\theta}$: $\hat{\theta} \xrightarrow{p} \theta^*$.*
- *The hessian $\boldsymbol{H} = \nabla^2 L(\theta^*)$ is full rank.*
- *"Other appropriate regularity conditions hold."*

Let $\nabla \ell^* := \nabla \ell\left((x, y); \theta^*\right)$.

- ► $\sqrt{n}\left(\hat{\theta} - \theta^*\right) = O_P(1)$.
- ► $\sqrt{n}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{H}^{-1} Cov\left[\nabla \ell^*\right] \boldsymbol{H}^{-1}\right)$
- ► $nE(\hat{\theta}) = O_P(1)$
- ► $nE(\hat{\theta}) \xrightarrow{d} \frac{1}{2}||S||_2^2$ *where* $S \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{H}^{-1/2} Cov\left[\nabla \ell^*\right] \boldsymbol{H}^{-1/2}\right)$.
- ► $\lim_{n\to\infty} \mathbb{E}\left[nE(\hat{\theta})\right] = \frac{1}{2} tr\left(\boldsymbol{H}^{-1} Cov\left[\nabla \ell^*\right]\right)$

[5] Red: population quantities. Blue: empirical quantities.

Summaries **October 09, 2021**

Concentration Inequalities

Table of Contents Local *Written by Brandon McKinzie*

Unless explicitly stated otherwise, for some random variable X, let

$$\mu_X \triangleq \mathbb{E}\left[X\right] \tag{43}$$

$$\sigma_X \triangleq \sqrt{\sigma_X^2} = \sqrt{\mathrm{Var}\left[X\right]} \tag{44}$$

Big-O Notation

Every occurrence of $O(x)$ is a placeholder for some function $f(x)$ such that $\forall x$ $|f(x)| \leq Cx$ for some absolute/universal constant $C > 0$.

For example:

$$|\bar{X} - \mu_{\bar{X}}| \leq O(\sigma\sqrt{\log n}) \tag{45}$$

$$\equiv \exists\ C \ \mathit{st}\ |\bar{X} - \mu_{\bar{X}}| \leq C\sigma\sqrt{\log n} \tag{46}$$

Alternative definition given by a TA: "The statement $f(n) = O(g(n))$ means there exists a constant C such that $|f(n)| \leq Cg(n)$ for all values of n."

Hoeffding's Inequality

Let $X_1, X_2, \ldots, X_n$ be i.i.d. real-valued RVs such that $a_i \leq X_i \leq b_i$ almost surely. Define $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then for any $\epsilon > 0$

$$Pr\left[|\bar{X} - \mu_{\bar{X}}| \geq \epsilon\right] \ \leq\ 2\exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \tag{47}$$

If we let $\sigma^2 \triangleq \frac{1}{n^2}\sum_{i=1}^n (b_i - a_i)^2$ and $\epsilon = \sigma\sqrt{c\log n}$ and plug these back in, we get

$$Pr\left[|\bar{X} - \mu_{\bar{X}}| \geq \epsilon\right] \ \leq\ 2n^{-2c} \tag{48}$$

Chebyshev's Inequality

For any RV X and $k \in R$ with $k \geq 1$

$$Pr\left[|X - \mu_X| \geq k\sigma_X\right] \ \leq\ \frac{1}{k^2} \tag{49}$$

Markov's Inequality

If $\varphi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a monotonically increasing function, X is any RV, $a \geq 0$, and $\varphi(a) > 0$, then

$$Pr\left[|X| \geq a\right] \ \leq\ \frac{\mathbb{E}\left[\varphi\left(|X|\right)\right]}{\varphi(a)} \tag{50}$$

Sub-Gaussian Random Variables

A RV X with finite μ_X is ***sub-Gaussian with parameter σ*** *if*

$$\mathbb{E}\left[e^{\lambda(X-\mu_X)}\right] \le e^{\frac{1}{2}\sigma^2\lambda^2} \qquad (\forall\lambda\in\mathbb{R}) \tag{51}$$

Equivalently:

$$Pr\left[|X-\mu_X|\ge t\right] \;\le\; 2\exp\left(-\frac{t^2}{2\sigma^2}\right) \qquad (\forall t\in\mathbb{R}) \tag{52}$$

Furthermore, the sum Z of independent sub-Gaussian RVs $X_1,\ldots,X_n$ with variance proxies $\sigma_1^2,\ldots,\sigma_n^2$ is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$. As a consequence, we have the tail bound

$$Pr\left[|Z-\mu_Z|\ge t\right] \le 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right) \tag{53}$$

McDiarmid's Inequality

Suppose $f:\mathbb{R}^n\to\mathbb{R}$ satisfies the ***bounded difference condition****: $\exists c_1,\ldots,c_n\in\mathbb{R}$ s.t. $\forall x_1,\ldots,x_n,x_i'\in\mathbb{R}$*

$$|f(\boldsymbol{x}) - f(\boldsymbol{x}_{\langle 1\ldots i-1\rangle} x_i', \boldsymbol{x}_{\langle i+1\ldots n\rangle})| \le c_i \tag{54}$$

Then, for any independent RVs $X_1,\ldots,X_n$,

$$Pr\left[|f(X_1,\ldots,X_n)-\mu_f \ge t]\,\right| \;\le\; 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) \tag{55}$$

Bounded Difference Inequality

Let $f:\mathbb{R}^n\to\mathbb{R}$. The ***one-sided differences*** *of f are defined as*

$$D_i^+ f(\boldsymbol{x}) \triangleq \sup_z f(\boldsymbol{x}_{\langle 1\ldots i-1\rangle}, z, \boldsymbol{x}_{\langle i+1\ldots n\rangle}) - f(\boldsymbol{x}) \tag{56}$$

$$D_i^- f(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) - \inf_z f(\boldsymbol{x}_{\langle 1\ldots i-1\rangle}, z, \boldsymbol{x}_{\langle i+1\ldots n\rangle}) \tag{57}$$

For convenience, define

$$d^+ \triangleq \left\|\sum_{i=1}^n \left|D_i^+ f\right|^2\right\|_\infty = \sup_{\boldsymbol{x}} \sum_{i=1}^n \left|D_i^+ f(\boldsymbol{x})\right|^2 \tag{58}$$

$$d^- \triangleq \left\|\sum_{i=1}^n \left|D_i^- f\right|^2\right\|_\infty = \sup_{\boldsymbol{x}} \sum_{i=1}^n \left|D_i^- f(\boldsymbol{x})\right|^2 \tag{59}$$

Let $X_1,\ldots,X_n$ be independent RVs. Then $\forall t\ge 0$,

$$Pr\left[f(X_1,\ldots,X_n)-\mu_f \ge t\right] \;\le\; \exp\left(-\frac{t^2}{4d^-}\right) \tag{60}$$

$$Pr\left[f(X_1,\ldots,X_n)-\mu_f \le -t\right] \;\le\; \exp\left(-\frac{t^2}{4d^+}\right) \tag{61}$$

Lipschitz Functions of Gaussian Variables

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, *and let* $f : \mathbb{R}^n \to R$ *be* ***L-Lipschitz*** *wrt the Euclidean norm* $|| \cdot ||_2$,*i.e. that*

$$|f(\boldsymbol{x}) - f(\boldsymbol{y}) \leq L||x-y||_2 \qquad (\forall x, y \in \mathbb{R}^n) \tag{62}$$

Then the variable $f(\boldsymbol{x}) - \mu_f$ *is sub-Gaussian with parameter at most* L, *and hence*

$$Pr\left[|f(\boldsymbol{x}) - \mu_f| \geq t\right] \;\leq\; 2\exp\left(-\frac{t^2}{2L^2}\right) \tag{63}$$

Summaries **October 16, 2021**

Uniform Convergence

Table of Contents Local *Written by Brandon McKinzie*

Recall the formula for excess risk: $E(\hat{\theta}) \triangleq L(\hat{\theta}) - L(\theta^*) \geq 0$.

Uniform Convergence

A parameter set Θ *exhibits* ***uniform convergence*** *if* $\forall \theta \in \Theta$

$$Pr\left[\left|\hat{L}(\theta) - L(\theta)\right| \geq \epsilon\right] \quad \leq \quad \delta \tag{64}$$

Henceforth, I will denote

$$UC(\theta) \triangleq \left|\hat{L}(\theta) - L(\theta)\right| \tag{65}$$

Uniform Convergence Implies Generalization

Using simple telescoping sums, we can show $E(\hat{\theta}) \leq 2 \sup_\theta UC(\theta)$:

$$\begin{aligned} E(\hat{\theta}) &= \left(L(\hat{\theta}) - \hat{L}(\hat{\theta})\right) + \left(\hat{L}(\hat{\theta}) - \hat{L}(\theta^*)\right) + \left(\hat{L}(\theta^*) - L(\theta^*)\right) && (66) \\ &\leq UC(\hat{\theta}) + UC(\theta^*) && (67) \\ &\leq 2 \sup_{\theta \in \Theta} UC(\theta) && (68) \end{aligned}$$

Therefore, if we can show that our parameter family Θ exhibits UC, then we can get a bound on the excess risk.

Problem: recall that we are assuming $0 \leq \ell \leq 1$. Therefore, we can invoke Hoeffding's inequality on $UC(\theta^*)$.

$$UC(\theta^*) \leq \widetilde{O}\left(\frac{1}{\sqrt{n}}\right) \tag{69}$$

However, we CANNOT invoke it on $UC(\hat{\theta})$, since the data-dependence of $\hat{\theta}$ means that the ℓ_i are no longer independent!

UC for Finite Hypothesis Class

Suppose $\mathcal{H}$ *is finite and* $0 \le \ell((x,y),h) \le 1$. *Then* $\forall \delta$ *s.t.* $0 < \delta < \frac{1}{2}$, *w.p. at least* $1-\delta$,

$$UC(h) \le \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2n}} \qquad (\forall h \in \mathcal{H}) \tag{70}$$

$$\textit{[corollary]} \qquad E(\hat{h}) \le \sqrt{\frac{2\left(\ln|\mathcal{H}| + \ln(2/\delta)\right)}{2n}} \tag{71}$$

where the corollary follows from 68.

Condensed Proof:

1. *For some fixed* $h \in \mathcal{H}$ *and fixed* $\epsilon > 0$:

$$Pr\left[UC(h) \ge \epsilon\right] \le 2\exp\left(-2n\epsilon^2\right) \qquad \textit{[Hoeffding]} \tag{72}$$

2. *To prove* $\forall h$, *apply the above with union-bound inequality for the event* $\{UC(h) \ge \epsilon\}$.

$$\begin{aligned} Pr\left[\exists h \text{ s.t. } UC(h) \ge \epsilon\right] &\le \sum_{h \in \mathcal{H}} Pr\left[UC(h) \ge \epsilon\right] \qquad \textit{[union-bound]} \quad (73) \\ &\le 2|\mathcal{H}|\exp\left(-2n\epsilon^2\right) \qquad \textit{[step 1]} \quad (74) \end{aligned}$$

3. *Let* $\delta = 2|\mathcal{H}|\exp\left(-2n\epsilon^2\right)$. *Solve for* ϵ *to complete the proof.*

Bounds for Infinite Hypothesis Class

Assumptions:

- *Our infinite* $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}^p, ||\theta||_2 \le B\}$ *for some fixed* $B > 0$.
- $\ell \in [0,1]$ *is* κ*-Lipschitz in* θ *w.r.t. the* ℓ_2*-norm for all* (x,y).

Then with probability at least $1 - O\left(e^{-\Omega(p)}\right)$, *we have*

$$(\forall \theta) \quad UC(\theta) \le O\left(\sqrt{\frac{p\max\left(\ln(\kappa B n),\ 1\right)}{n}}\right) \tag{75}$$

Summary of Bounds.

$$[\text{finite } \mathcal{H}]\ (\forall h \in \mathcal{H}) \text{ w.h.p.} \quad UC(h) \le \tilde{O}\left(\frac{1}{\sqrt{n}}\right) \quad \text{[Remark 3.3]} \tag{76}$$

$$[\text{finite } \mathcal{H}] \text{ w.h.p. } (\forall h \in \mathcal{H}) \quad UC(h) \le \tilde{O}\left(\sqrt{\frac{\ln|\mathcal{H}|}{n}}\right) \tag{77}$$

$$[\text{infinite } \mathcal{H}] \text{ w.h.p } (\forall h \in H) \quad UC(h) \le \tilde{O}\left(\sqrt{\frac{p}{n}}\right) \tag{78}$$

$$E(\hat{\theta}) \le 2\sup_{\theta \in \Theta} UC(\theta) \tag{79}$$

$$E(\hat{\theta}) \le \frac{c}{n} + o\left(\frac{1}{n}\right) \tag{80}$$

Summaries **October 20, 2021**

Rademacher Complexity

Table of Contents Local *Written by Brandon McKinzie*

Rademacher Complexity

Let $\mathcal{F} = \{f : Z \to \mathbb{R}\}$ and Let P be a distribution over Z. The (average) ***Rademacher complexity*** *of $\mathcal{F}$ is defined as*

$$R_n(\mathcal{F}) \triangleq \mathbb{E}_{z_1,\ldots,z_n \overset{i.i.d.}{\sim} P}\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_n \overset{i.i.d.}{\sim} \{\pm 1\}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right]\right] \tag{81}$$

Furthermore,

$$\mathbb{E}_{z_1,\ldots,z_n \overset{i.i.d.}{\sim} P}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}_{z\sim P}\left[f(z)\right]\right] \leq 2R_n(\mathcal{F}) \tag{82}$$

Empirical Rademacher Complexity

Given a dataset $S = \{z_1, \ldots, z_n\}$, the ***empirical Rademacher complexity*** *is just the inner part of 81:*

$$R_S(\mathcal{F}) \triangleq \mathbb{E}_{\sigma_1,\ldots,\sigma_n \overset{i.i.d.}{\sim} \{\pm 1\}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right] \tag{83}$$

Suppose $\forall f \in \mathcal{F}$, $0 \leq f(z) \leq 1$. Then w.p. at least $1 - \delta$,

$$\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}\left[f(z)\right]\right] \leq 2R_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \tag{84}$$

$$L(h) - L(h) \leq 2R_s(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \tag{85}$$

ϵ-cover

The set $\mathcal{C}$ is an ϵ-cover of $\mathcal{Q}$ w.r.t. metric ρ if $\forall v \in \mathcal{Q}$, $\exists v' \in \mathcal{C}$ such that $\rho(v, v') \leq \epsilon$.

Covering Number

The ***covering number*** *$N(\epsilon, \mathcal{Q}, \rho)$, is the minimum size of an ϵ-cover.*

The standard metric we'll use is $\rho(v, v') = \frac{1}{\sqrt{n}}||v - v'||_2$, with the leading coefficient inserted for convenience.

Theorem 4.24

Let $\mathcal{F}$ be a family of functions $Z \mapsto [-1, 1]$. Then

$$R_S(\mathcal{F}) \leq \inf_{\epsilon > 0}\left(\epsilon + \sqrt{\frac{2\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}}\right) \tag{86}$$

Proof of stuff that doesn't seem to be in scribe notes yet (associated w/material around 4.6) starts around [23:30].

Dudley's Theorem

If $\mathcal{F}$ is a function class from Z to $\mathbb{R}$, then

$$R_S(\mathcal{F}) \leq 12 \int_0^\infty \sqrt{\frac{2 \log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} \mathrm{d}\epsilon \tag{87}$$

2.5.1 Rademacher Complexity Bounds for Concrete Models and Losses

---

Talagrand's Lemma

Let $\phi : \mathbb{R} \to \mathbb{R}$ be a κ-Lipschitz function. Then

$$R_S(\phi \circ \mathcal{H}) \leq \kappa R_S(\mathcal{H}) \tag{88}$$

Linear Models with Bounded ℓ_2 Norm

Let $\mathcal{H} = \{x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^d, ||w||_2 \leq B\}$ for some constant $B > 0$. Moreover, ***assume*** *$\mathbb{E}_{x \sim P}\left[||x||_2^2\right] \leq C^2$, where P is some distribution and $C > 0$ is a constant. Then*

$$R_s(\mathcal{H}) \leq \frac{B}{n} \sqrt{\sum_{i=1}^{n} ||x^{(i)}||_2^2} \tag{89}$$

$$R_n(\mathcal{H}) \leq \frac{BC}{\sqrt{n}} \tag{90}$$

Linear Models with Bounded ℓ_1 Norm

Let $\mathcal{H} = \{x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^d, ||w||_1 \leq B\}$ for some constant $B > 0$. Moreover, ***assume*** *$||x^{(i)}||_\infty \leq C$, for some constant $C > 0$ and all points in $S = \{x^{(i)}\}_{i=1}^n \subset \mathbb{R}^d$. Then*

$$R_s(\mathcal{H}) \leq BC \sqrt{\frac{2 \log(2d)}{n}} \tag{91}$$

$$R_n(\mathcal{H}) \leq BC \sqrt{\frac{2 \log(2d)}{n}} \tag{92}$$

Massart's Lemma

Suppose $\mathcal{Q} \subset \mathbb{R}^n$ is finite and contained in the ℓ_2-norm ball of radius $M\sqrt{n}$ for some constant $M > 0$. Then

$$\mathbb{E}_{\sigma \sim Unif\{\pm 1\}^n}\left[\sup_{q \in \mathcal{Q}} \frac{1}{n} \langle \sigma, q \rangle\right] \leq M \sqrt{\frac{2 \log |\mathcal{Q}|}{n}} \tag{93}$$

Two-Layer Neural Networks

Weak Bound. *For some constants* $B_w > 0$ *and* $B_u > 0$*, let*

$w \in \mathbb{R}^m$

$U \in \mathbb{R}^{m \times d}$

$$\mathcal{H} = \{f_\theta : x \mapsto \langle w, \phi(Ux) \rangle \mid ||w||_2 \leq B_w, ||\boldsymbol{u}_i||_2 \leq B_u \forall i \in \{1, 2, \ldots, m\} \tag{94}$$

and suppose $\mathbb{E}\left[||x||_2^2\right] \leq C^2$*. Then*

$$R_n(\mathcal{H}) \leq 2 B_w B_u C \sqrt{\frac{m}{n}} \tag{95}$$

Strong Bound. *Let* $C(\theta) = \sum_{j=1}^m |w_j| \, ||u_j||_2$*. Restrict* $\mathcal{H}$ *from above to satisfy* $C(\theta) \leq B$ *for some constant* $B > 0$*.* ***Assume*** $\left|\left|x^{(i)}\right|\right|_2 \leq C$ *for all* $i \in \{1, \ldots, n\}$*. Then*

$$R_S(\mathcal{H}) \leq 2 \frac{B \, C}{\sqrt{n}} \tag{96}$$

Summaries **October 30, 2021**

Nonconvex Optimization

Written by Brandon McKinzie

Local Minimum of a Function

*We say that x is a **local minimum** of a function f if there exists an open neighborhood N around x such that in N, the function values are at least $f(x)$.*

Strict-Saddle Condition

*For **positive** α, β, γ, we say that $f : \mathbb{R}^d \to \mathbb{R}$ is (α, β, γ)-strict-saddle if $\forall x \in \mathbb{R}^d$, at least one of the following is true:*

- $||\nabla f(x)||_2 \geq \alpha$
- $\lambda_{min}\left(\nabla^2 f(x)\right) \leq -\beta$
- $\min_{x^*} ||x - x^*||_2 \leq \gamma$, *where x^* is a local minimum.*

Strict-Saddle Convergence

Suppose f is a function that satisfies the following condition:

> *$(\exists \epsilon_0, \tau_0, c > 0)$ such that if $x \in \mathbb{R}^d$ satisfies $||\nabla f(x)||_2 \leq \epsilon < \epsilon_0$ and $\nabla^2 f(x) \succeq -\tau_0 I$, then x is ϵ^c-close to a **global minimum** of f.*

*Then many optimizers can converge to a **global minimum** of f up to δ-error in Euclidean distance in time poly$\left(\frac{1}{\delta}, \frac{1}{\tau}_0, d\right)$.*

TODO *figure out if $A \succeq B$ means elemwise $\geq$ or if it means $(A - B)$ is psd.*

PCA / Matrix Factorization / Linearized NN

Let $M \in \mathbb{R}^{d\times d}$ be symmetric and psd. We want to find the best rank-1 approximation of the matrix M. The **non-convex** objective is

$$\min_{x\in\mathbb{R}^d} g(x) \triangleq \frac{1}{2}||M - xx^T||_F^2 \tag{97}$$

Theorem 8.7

All local minima of g are global minima (even though g is non-convex).

Proof:

1. Show that all stationary points must be eigenvectors of M.
   (a) $\nabla g(x) = -2(M - xx^T)x$ (from HW0[a])
   (b)
   $$\nabla g(x) = 0 \implies Mx = ||x||_2^2 x \tag{98}$$
   (c) Therefore, for all stationary points x, we have that x is an eigenvector of M with eigenvalue $||x||_2^2$.
2. Show that all local minima must be eigenvectors [of M] of the largest eigenvalue.
   (a) We now also require $\nabla^2 g(x) \succeq 0$.
   (b) To obtain the expression for $\langle v\nabla^2 g(x)v\rangle$, expand out $g(x+v)$ and collect terms that are quadratic in v, to obtain:
   $$\langle v, \nabla^2 g(x)v\rangle = 2\langle x, v\rangle^2 + ||x||_2^2\,||v||_2^2 - v^T M v \tag{99}$$
   (c) Recall that we know from step 1 that x is an eigenvector of M. Furthermore, since 99 must hold for all $v \in \mathbb{R}^d$, then it must hold for the eigenvector v_1 with largest eigenvalue λ_1. In other words, for any local minimum x, the following must be true:
   $$2\langle x, v_1\rangle^2 + ||x||_2^2\,||v_1||_2^2 - v_1^T M v_1 \geq 0 \tag{100}$$
   We then have two cases to consider:
      i. x has eigenvalue λ_1. Then, by the **Eckart-Young-Mirsky Theorem**, x is a *global* minimum.
      ii. x has eigenvalue $\lambda < \lambda_1$. Remember that $\lambda = ||x||_2^2$.(eq 98). Then $\langle x, v_1\rangle = 0$, and 100 says
      $$-\lambda_1 + \lambda \geq 0 \implies \lambda \geq \lambda_1 \tag{101}$$
      which is a contradiction, since we know that $\lambda_1 > \lambda$.
   (d) Therefore, if a stationary point x also satisfies $\nabla^2 g(x) \succeq 0$, then x is an eigenvector of M with the largest eigenvalue.
3. Since all local minima are eigenvectors of M with the largest eigenvalue, by the Eckart-Young-Mirsky Theorem, **all local minima are global minima**.

[a] Approach is to expand $g(x+\delta)$ and exploit properties of Frobenius norm, like $||A||_F^2 = tr(A^T A)$.

Matrix Completion

We consider rank-1 matrix completion. Let $M = zz^T$ be a rank-1 symmetric and psd matrix for some $z \in \mathbb{R}^d$ with $||z||_2 = 1$. Also assume that the ground truth vector z satisfies the **incoherence condition**, $||z||_\infty \leq \frac{\mu}{\sqrt{d}}$ for some constant μ. Given random entries of M, taken by zeroing out each entry with probability p, our goal is to recover the rest of the entries[a]

Let $\Omega \in [d] \times [d]$ denote the set of indices of M that we observe, and let $P_\Omega(M)$ denote the matrix M with all entries outside of Ω set to zero. Our objective function is[b]

$$\min_{x \in \mathbb{R}^d} f(x) \triangleq \frac{1}{2}||P_\Omega\left(M - xx^T\right)||_F^2 \tag{102}$$

Theorem 8.14

Assume $p = \frac{poly(\mu, \log d)}{\epsilon^2 d}$ *for sufficiently small constant* ϵ *and* z *is incoherent. Then w.h.p. all local minima of* f *(eq 102) are* $O(\sqrt{\epsilon})$*-close to* $\pm z$ *(the global minima of* f*).*

Lemma 8.17

Assume $p = \frac{poly(\mu, \log d)}{\epsilon^2 d}$ *for sufficiently small constant* $\epsilon > 0$*. For any two matrices* $A = uu^T$ *and* $B = vv^T$*, for some* u *and* v *both satisfying* $|| \cdot ||_2 \leq 1$ *and* $|| \cdot ||_\infty \leq \mu/\sqrt{d}$*, we have*

$$\left|\frac{1}{p}\langle P_\Omega(A), B\rangle - \langle A, B\rangle\right| \ \leq \ \epsilon \qquad (w.h.p) \tag{103}$$

In all that follows, f denotes 102, and g denotes 97.

Lemma 8.18

$$\nabla g(x) = 0 \implies \langle x, z\rangle^2 = ||x||_2^4 \tag{104}$$

Lemma 8.19

Assume $||x||_\infty \leq 2\mu/\sqrt{d}$.

$$\nabla f(x) = 0 \implies \langle x, z\rangle^2 \geq ||x||_2^4 - \epsilon \qquad (w.h.p.) \tag{105}$$

Lemma 8.20

$$\nabla^2 g(x) \succeq 0 \implies ||x||_2^2 \geq \frac{1}{3} \tag{106}$$

Lemma 8.21

Assume $||x||_\infty \leq \mu/\sqrt{d}$.

$$\nabla^2 f(x) \succeq 0 \implies ||x||_2^2 \geq \frac{1}{3} - \frac{\epsilon}{3} \qquad (w.h.p.) \tag{107}$$

Lemma 8.22

All local minima of g *are global minima.*

[a] Note that d parameters are needed to fully identify/specify an arbitrary rank-1 matrix. The number of expected observed entries is pd^2, so we **assume** here that $p \gg \frac{1}{d}$.

[b] It's important to always keep in mind our running assumption that the structure of M is such that $M = zz^T$ for some $z \in \mathbb{R}^d$.

Summaries **November 09, 2021**

Neural Tangent Kernel

Table of Contents Local *Written by Brandon McKinzie*

Motivation. For a given initialization $\theta^0 \in \mathbb{R}^p$, denote by $B(\theta^0)$ the neighborhood around θ^0 for which the loss function is convex and its *global* minimum is attained. Notably, we are **assuming** that $\forall (x, y)$, there exists $\theta^* \in B(\theta^0)$ such that $y = f_{\theta^*}(x)$ exactly, where $f_\theta : \mathbb{R}^d \mapsto \mathbb{R}$ is our parametric model. We can approximate f with a Taylor expansion around θ^0:

$$f_\theta(x) = \underbrace{f_{\theta^0}(x) + \langle \nabla_\theta f_{\theta^0}(x), \theta - \theta^0 \rangle}_{\triangleq g_\theta(x)} + \text{higher order terms} \tag{108}$$

$$\hat{L}(g_\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - \phi(x^{(i)})^\top \Delta\theta \right)^2 = \frac{1}{n} ||\boldsymbol{y} - \boldsymbol{\Phi} \cdot \Delta\theta||_2^2 \tag{109}$$

$\phi(x) \triangleq \nabla_\theta f_{\theta^0}(x)$

WLOG, **assume** we choose our initial θ^0 such that $f_{\theta^0}(x) = 0 \ \forall x$. In this case, and ignoring higher-order terms,

$$f_\theta(x) \approx \langle \nabla_\theta f_{\theta^0}(x), \theta^* - \theta^0 \rangle \tag{110}$$

$$\underbrace{\ell(f_\theta(x), y)}_{\text{not necessarily convex}} \approx \underbrace{\ell(g_\theta(x), y)}_{\text{convex}} \tag{111}$$

assuming our loss function ℓ is convex[6]. We see that y can be approximated as a linear function of the difference between the global optimum θ^* and initialization θ^0, with coefficients $\phi(x) \triangleq \nabla_\theta f_{\theta^0}(x)$. The **neural tangent kernel** K is given by

$$K(x, x') = \langle \phi(x), \phi(x') \rangle = \langle \nabla_\theta f_{\theta^0}(x), \nabla_\theta f_{\theta^0}(x') \rangle \tag{112}$$

[6] Recall that any convex function composed with a linear function is still convex.

Henceforth, we will also **assume** that $y^{(i)} = O(1)$ and $||\boldsymbol{y}||_2 = O(\sqrt{n})$.

Lemma 8.25: It suffices to optimize in B_{θ^0}

Assume *the following:*

- $p \geq n$
- $rank(\boldsymbol{\Phi}) = n$
- $\sigma_{min}(\boldsymbol{\Phi}) = \sigma > 0$

Then, letting $\Delta\hat{\theta}$ denote the minimum norm solution of $\boldsymbol{\Phi v} = \boldsymbol{y}$[7]*, we have*

$$\left|\left|\Delta\hat{\theta}\right|\right|_2 \leq O\left(\frac{\sqrt{n}}{\sigma}\right) \tag{113}$$

Remark 8.26*. Note that this essentially characterizes how large the ball B_{θ^0} must be in order to contain a global minimum, where*

$$B_{\theta^0} = \{\theta = \theta^0 + \Delta\theta : ||\Delta\theta||_2 \leq O\left(\frac{\sqrt{n}}{\sigma}\right)\} \tag{114}$$

Proof

Note that $\Delta\hat{\theta} = \boldsymbol{\Phi}^+\boldsymbol{y}$, and $\left|\left|\boldsymbol{\Phi}^+\right|\right|_{op} = \frac{1}{\sigma_{min}(\boldsymbol{\Phi})} = \frac{1}{\sigma}$. Then

$$\begin{aligned} \left|\left|\Delta\hat{\theta}\right|\right|_2 &= \left|\left|\boldsymbol{\Phi}^+\boldsymbol{y}\right|\right|_2 & (115)\\ &\leq \left|\left|\boldsymbol{\Phi}^+\right|\right|_{op} ||\boldsymbol{y}||_2 \quad \textit{By def of op norm} & (116)\\ &\leq O\left(\frac{\sqrt{n}}{\sigma}\right) \quad ||\boldsymbol{y}||_2 \leq O(\sqrt{n}) & (117)\end{aligned}$$

Lemma 8.28: $\forall\theta \in B(\theta^0)$, $f_\theta \approx g_\theta$ and $\hat{L}(f_\theta) \approx \hat{L}(g_\theta)$

Assume *$\nabla_\theta f_\theta(x)$ is β-Lipschitz in θ, i.e. $\forall x, \theta, \theta'$:*

$$||\nabla_\theta f_\theta(x) - \nabla_\theta f_{\theta'}(x)||_2 \leq \beta \cdot ||\theta - \theta'||_2 \tag{118}$$

NB: the above is equivalent to the statement $\left|\left|\nabla^2_\theta f_\theta\right|\right|_{op} \leq \beta$ if f_θ is twice-differentiable.

Then $\forall\theta$

$$(\forall\theta) \quad |f_\theta(x) - g_\theta(x)| \leq O\left(\beta\,||\Delta\theta||_2^2\right) \tag{119}$$

$$(\forall\theta \in B_{\theta^0})|f_\theta(x) - g_\theta(x)| \leq O\left(\frac{\beta n}{\sigma^2}\right) \tag{120}$$

where $\Delta\theta = \theta - \theta^0$.

[7] In other words, $\Delta\hat{\theta}$ is the smallest vector $\boldsymbol{v}$ satisfying $\boldsymbol{\Phi v} = \boldsymbol{y}$.

The NTK Regime. The **NTK regime** refers to the situation where $\beta / \sigma^2 \to 0$. The following are two such examples:

1. **Reparameterize with a scalar**. Let $f_\theta = \alpha \bar{f}_\theta$ for arbitrary neural net $\bar{f}$ with fixed width and depth. Then

$$\frac{\beta}{\sigma^2} = \frac{\bar{\beta}}{\bar{\sigma}^2}\frac{1}{\alpha} \;\to\; 0 \qquad \text{as } \alpha \to \infty \tag{121}$$

2. **Overparametrization** (with specific initialization). Consider a two-layer network with m neurons:

$$\hat{y} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \sigma(w_i^\top x) \tag{122}$$

$|\sigma(x) - \sigma(x')| \leq \left|\left|x - x'\right|\right|_2$

$a_i \sim \{\pm 1\}$

$W \in \mathbb{R}^{m \times d}$

$w_i^0 \sim \mathcal{N}(0, I_d)$

$||x||_2 = \Theta(1)$

$\theta \equiv \text{vec}(W) \in \mathbb{R}^{dm}$

where the $a_i \sim \{\pm 1\}$ are constants (not optimized). The following is a condensed derivation for the behavior of β/σ^2 in the infinite width limit ($m \to \infty$):

(a) Key initial observations:

$$\sigma({w_i^0}^\top x) = O(1) \tag{123}$$

$$\left|\sum_{i=1}^{m} a_i \sigma({w_i^0}^\top x)\right| = \Theta(\sqrt{m}) \tag{124}$$

$$f_{\theta^0}(x) = \Theta(1) \tag{125}$$

(b) Show that overall *scale* of the gradients doesn't depend on m in the limit $m \to \infty$.

$$||\nabla_\theta f_\theta(x)||_2^2 = \frac{1}{m}\sum_{i=1}^{m} \left|\left|\sigma'(\boldsymbol{w}_i^T x)\right|\right|_2^2 \tag{126}$$

$$\lim_{m\to\infty} ||\nabla_\theta f_\theta(x)||_2^2 = \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[\sigma'\left(w^\top x\right)^2\right] \cdot ||x||_2^2 \tag{127}$$

$$= O(1) \tag{128}$$

(c) Note that[8] $\sigma_{\min}(\Phi) = \sqrt{\sigma_{\min}(\Phi\Phi^\top)}$. The NTK in the infinite width limit:

$$\left(\Phi\Phi^\top\right)_{i,j} = \left\langle \nabla_\theta f_{\theta^0}\left(x^{(i)}\right), \nabla_\theta f_{\theta^0}\left(x^{(j)}\right)\right\rangle \tag{129}$$

$$K^\infty \triangleq \lim_{m\to\infty}\left(\Phi\Phi^\top\right)_{i,j} = \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[\sigma'\left(w^\top x^{(i)}\right)\sigma'\left(w^\top x^{(j)}\right)\right]\left\langle x^{(i)}, x^{(j)}\right\rangle \tag{130}$$

(d) It can be shown that K^∞ is full rank, and that

$$\sigma_{\min}\left(\Phi\Phi^\top\right) > \frac{1}{2}\sigma_{\min}(K^\infty) > 0 \tag{131}$$

From this and the definition of K^∞, then, σ is constant in the infinite width limit.

[8] Yes, we are overloading σ here. It should be clear from context whether we're referring to the activation function σ or the smallest singular value $\sigma_{\min}(\Phi)$.

(e) Next we need to find β, the Lipschitzness of the gradients.

$$||\nabla_\theta f_\theta(x) - \nabla_\theta f_{\theta'}(x)||_2^2 = O\left(\frac{1}{m}||\theta - \theta'||_2^2\right) \tag{132}$$

$$\implies \beta = O(\frac{1}{\sqrt{m}}) \tag{133}$$

(f) We now arrive at our final result

$$\lim_{m\to\infty} \frac{\beta}{\sigma^2} \approx \lim_{m\to\infty} \frac{1}{\sqrt{m}} \frac{1}{\sigma_{\text{min}}(K^\infty)^2} = 0 \tag{134}$$

Which says that the gradient becomes more smooth (smaller β) as we increase the number of neurons. It also says that our linear approximation $g_\theta(x) \to f_\theta(x)$ in the infinite width limit.

Optimizing $\hat{L}(g_\theta)$ vs. $\hat{L}(f_\theta)$. Here we show that optimizing $\hat{L}(f_\theta)$ amounts to optimizing $\hat{L}(g_\theta)$ inside the ball $B(\theta^0)$. First, denote by $g_\theta^t(x)$ the Taylor expansion of f_θ around θ^t (the value of θ at timestep t)[9][10]

$$g_\theta^t(x) \triangleq f_{\theta^t}(x) + \left\langle \nabla_\theta f_{\theta^t}(x), \theta - \theta^t \right\rangle \tag{137}$$

$$\nabla \hat{L}(g_\theta^t)\Big|_{\theta=\theta^t} = \nabla \hat{L}(f_\theta)\Big|_{\theta=\theta^t} \tag{138}$$

TODO

[9] In this notation, so far we've been working with $g_\theta^0(x)$.

[10] Note that we are using really sloppy notation:

$$\nabla_\theta f_{\theta^t}(x) \equiv \nabla_\theta f_\theta(x)\Big|_{\theta=\theta^t} \tag{135}$$

$$\nabla \hat{L}(f_{\theta^t}) \equiv \nabla_\theta \hat{L}(f_\theta)\Big|_{\theta=\theta^t} \tag{136}$$

Summaries **Nov 12, 2021**

Implicit/Algorithmic Regularization Effect

Table of Contents Local *Written by Brandon McKinzie*

Overparametrized Linear Regression (d > n). **Assume** $X \in \mathbb{R}^{n \times d}$ is full rank.

$$\hat{L}(\beta) = \frac{1}{2} ||y - X\beta||_2^2 \tag{139}$$

Lemma 9.1

β is a global minimizer of $\hat{L}$ iff $\beta = X^+ y + \zeta$ for some ζ such that $\zeta \perp x_1, \ldots, x_n$.

Corollary 9.2. *The minimum norm solution is $\beta^* = X^+ y$.*

Theorem 9.3

Suppose GD on $\hat{L}(\beta)$ with $\beta^0 = 0$ converges to $\hat{\beta}$ s.t. $\hat{L}(\hat{\beta}) = 0$. Then $\hat{\beta} = \beta^*$.

And thus we see an "implicit regularization" effect of gradient descent on overparametrized linear regression, in that it converges to the minimum-norm solution β^*.

Algorithmic Regularization in Non-Linear Models. Now, instead of $X\beta$, let $f_\beta := \langle \beta \odot \beta, x \rangle$, where $\odot$ is the Hadamard (aka elementwise) product, and

$x^{(i)} \overset{iid}{\sim} \mathcal{N}(0, I_d)$

$$\hat{L}(\beta) := \frac{1}{4n} \sum_{i=1}^{n} \left(y^{(i)} - f_\beta(x^{(i)}) \right)^2 \tag{140}$$

where the ground truth $y^{(i)} = f_{\beta^*}(x^{(i)})$ is r-sparse ($||\beta^*||_0 = r$). For simplicity, **assume** $\beta_i^* = \mathbb{1} i \in S$ for some $S \subset [d]$ such that $|S| = r$ [11] , and that $n \geq \widetilde{\Omega}(r^2)$.

[11] aka exactly r elements equal to 1, all others equal to 0.

Summaries **October 30, 2021**

Important Inequalities and Theorems

Table of Contents Local *Written by Brandon McKinzie*

Here I list the key theorems/facts/etc that we repeatedly use throughout the course.

Miscellaneous Norm Inequalities

$$
\begin{align}
||x||_2 \leq ||x||_1 \leq \sqrt{n}\,||x||_2 \tag{141}\\
||x||_\infty \leq ||x||_2 \leq \sqrt{n}||x||_\infty \tag{142}\\
||x||_\infty \leq ||x||_1 \leq n||x||_\infty \tag{143}\\
\tag{144}
\end{align}
$$

Convex Function

A function $f : \mathbb{R}^n \to \mathbb{R}$ that has convex dom$(f) \subseteq \mathbb{R}^n$ is called a ***convex function*** *iff*

$$
f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) \tag{145}
$$

If f is differentiable on $\mathbb{R}^n$, then the following statements are equivalent:

- *f is a convex function*
- $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$ $(\forall x, y \in \mathbb{R}^n)$
- $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq 0$ $(\forall x, y \in \mathbb{R}^n)$

Cauchy-Schwarz Inequality

$$
|\langle \boldsymbol{u}, \boldsymbol{v}\rangle|^2 \leq \langle \boldsymbol{u}, \boldsymbol{u}\rangle \cdot \langle \boldsymbol{v}, \boldsymbol{v}\rangle \tag{146}
$$

Jensen's Inequality

For any convex function ϕ,

$$
\phi\left(\mathbb{E}\left[X\right]\right) \leq \mathbb{E}\left[\phi(X)\right] \tag{147}
$$

Holder's Inequality

For integers p, q in the open interval $(1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$:

$$
\begin{align}
|\langle x, y\rangle| \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \left(\sum_{j=1}^{n} |y_j|^q\right)^{\frac{1}{q}} \tag{148}\\
\mathbb{E}\left[|XY|\right] \leq \left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}} \left(\mathbb{E}\left[|Y|\right]^q\right)^{\frac{1}{q}} \tag{149}
\end{align}
$$

Spectral Decomposition

If A is a symmetric matrix that is orthogonally diagonalized by $\boldsymbol{P} = \begin{bmatrix} \boldsymbol{u}_1 & \dots & \boldsymbol{u}_n \end{bmatrix}$, where the $\boldsymbol{u}_i$ are unit eigenvectors, and let $\boldsymbol{D} = diag(\boldsymbol{\lambda})$. Then

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T \tag{150}$$

Polar Decomposition

Suppose $T \in \mathcal{L}(V)$. Then there exists an isometry $S \in \mathcal{L}(V)$ such that $T = S\sqrt{T^*T}$.

Singular Value Decomposition

Suppose $T \in \mathcal{L}(V)$ has singular values $s_1, \dots, s_n$. Then there exists orthonormal bases $e_1, \dots e_n$ and $f_1, \dots, f_n$ of V such that

$$Tv = s_1\langle v, e_1\rangle f_1 + \cdots + s_n\langle v, e_n\rangle f_n \tag{151}$$

Theorems.

Spectral Theorem

Suppose $T \in \mathcal{L}(V)$. Then the following are equivalent:

- *T is self-adjoint*
- *V has an orthonormal basis consisting of eigenvectors of T*
- *T has a diagonal matrix wrt some orthonormal basis of V*

Basic Definitions and Terminology.

Adjoint

*Suppose $T \in \mathcal{L}(V, W)$. The **adjoint** of T is the function $T^* : W \to V$ such that*

$$\langle Tv, w\rangle = \langle v, T^*w\rangle \tag{152}$$

Isometry

*An operator $S \in \mathcal{L}(V)$ is called an **isometry** if $||Sv||_2 = ||v||_2 \ \forall v \in V$.*

Singular Values

*Suppose $T \in \mathcal{L}(V)$. The **singular values** of T are the eigenvalues of $\sqrt{T^*T}$, with each eigenvalue λ repeated dim $E(\lambda, \sqrt{T^*T})$ times.*

Recall that $E(\lambda, T) = null(T - \lambda I)$ denotes the eigenspace of T corresponding to λ.

Characterization of Positive Semi-Definite Operators

Let $T \in \mathcal{L}(V)$. Then the following are equivalent:

- *T is psd*
- *T is self-adjoint and all $\lambda \geq 0$*
- *T has a [unique] psd square root (some R st $R^2 = T$)*
- *T has a self-adjoint square root*
- *$\exists R \in \mathcal{L}(V)$ st $T = R^* R$.*

Operator Norm

Let $T \in \mathcal{L}(V)$ be a linear operator on a normed vector space V with some norm $||\cdot||$. The ***operator norm*** *of T, denoted $||T||_{op}$, is*

$$||T||_{op} \triangleq \sup_{||v||=1} ||Av|| \tag{153}$$

Note that, if $\hat{\sigma}, \sigma$, denote the smallest and largest singular values of T, respectively, then

$$\hat{\sigma}||v|| \leq ||T||_{op} \leq \sigma||v|| \qquad \forall v \in V \tag{154}$$

Papers

Contents

Papers **October 30, 2021**

Deep Learning without Poor Local Minima

Table of Contents Local *Written by Brandon McKinzie*

Kenji Kawaguchi, "Deep Learning without Poor Local Minima" *MIT*, (Dec 2016).

Linear Regression Review

$$
\begin{aligned}
\hat{\boldsymbol{y}} &= X\boldsymbol{w} && (155)\\
L(\boldsymbol{w}) &= \frac{1}{2}||X\boldsymbol{w}-\boldsymbol{y}||_2^2 && (156)\\
\nabla L(\boldsymbol{w}) &= \sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}^{(i)}-y_i)\boldsymbol{x}^{(i)} = X^T(X\boldsymbol{w}-\boldsymbol{y}) && (157)\\
\nabla L(\boldsymbol{w}) &= 0 && (158)\\
\implies \boldsymbol{w}^* &= (X^TX)^{-1}X^T\boldsymbol{y} && (159)
\end{aligned}
$$

Note that we rely on X^TX being invertible. One way to ensure this is to ensure that the$p$ columns of X are linearly independent (i.e. X has full column rank). If so, then X^TX is a diagonal matrix with strictly positive elements (check this) and thus is positive definite. Since any p.d. matrix is invertible[a], X^TX must be invertible when X is full column rank.

[a] Nearly trivial to prove: assume the matrix A is *not* invertible. Then, by definition $\exists x \neq 0$ s.t. $Ax = 0$. This means that x is an eigenvector with eigenvalue 0. This is a contradiction, since A is positive definite. Therefore, A must be invertible.

Model and Notation (2.1).

- H: number of hidden layers.
- $(X, Y) \in (\mathbb{R}^{d_x \times n}, \mathbb{R}^{d_y \times n})$: training dataset of n examples.
- $\Sigma = YX^T(XX^T)^{-1}XY^T$.
- $p = \min(d_H, \ldots, d_1)$ is the smallest width of a hidden layer.
- $\hat{Y} = \prod_{i=1}^{H+1} W_i X$ is the output prediction of a deep MLP, with $\hat{Y} \in \mathbb{R}^{d_y \times n}$.
- $\mathcal{L}(W) = \frac{1}{2}||\hat{Y} - Y||_F^2$.

Papers **November 23, 2021**

Neural Tangent Kernel

Table of Contents Local *Written by Brandon McKinzie*

Jacot et al., "Neural Tangent Kernel: Convergence and Generalization in Neural Networks" *Polytechnique*, (Feb 2020).

Neural Networks (2). **Assume** all parameters are initialized as iid Gaussians $\mathcal{N}(0, 1)$.

$W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell+1}}$

$b^{(\ell)} \in \mathbb{R}^{n_{\ell+1}}$

- **Function space** $\mathcal{F} = \{f : \mathbb{R}^{n_0} \to R^{n_L}\}$
- **ANN realization function** $F^{(L)} : \mathbb{R}^p \to \mathcal{F}$, where $P = \sum_{\ell=0}^{L-1}(n_\ell + 1)n_{\ell+1}$.
- **Input space seminorm**[12][13]

$$\langle f, g \rangle_{p^{in}} \triangleq \mathbb{E}_{x \sim p^{in}} \left[f(x)^T g(x) \right] \tag{160}$$

- **Empirical distribution**

$$p^{in} := \frac{1}{N} \sum_{i=1}^{N} \delta(x^{(i)}) \tag{161}$$

- **Network function**

$$f_\theta(x) := \tilde{\alpha}^{(L)}(x; \theta) \tag{162}$$

$$\tilde{\alpha}^{(\ell)} = \frac{1}{\sqrt{n_{\ell-1}}} W^{(\ell-1)} \alpha^{(\ell-1)}(x; \theta) + \beta b^{(\ell-1)} \tag{163}$$

$$\alpha^{(\ell)}(x; \theta) = \sigma\left(\tilde{\alpha}^{(\ell)}(x; \theta)\right) \tag{164}$$

- **Functional cost** $C : \mathcal{F} \to \mathbb{R}$
- **Multi-dimensional kernel** function $K : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_L \times n_L}$[14] defines bilinear form

$$\langle f, g \rangle_K \triangleq \mathbb{E}_{x, x' \sim p^{in}} \left[f(x)^T K(x, x') g(x') \right] \tag{165}$$

We say K is *positive definite* wrt $||\cdot||_{p^{in}}$ if

$$||f||_{p^{in}} > 0 \implies ||f||_k > 0$$

- **Dual of** $\mathcal{F}$ wrt p^{in} is denoted $\mathcal{F}^*$:

$$\mathcal{F}^* = \{\mu : \mathcal{F} \to \mathbb{R}\} \tag{166}$$

$$\text{where} \quad \mu(f) = \langle d, f \rangle_{p^{in}} \text{ for some } d \in \mathcal{F} \tag{167}$$

[12] A *seminorm* is literally just a norm that doesn't require the usual implication of $||x|| = 0 \implies x = 0$.

[13] Wait wtf how could $||f||_{p^{in}} = 0$ for any function other than $f(x) \triangleq \mathbf{0}\ \forall \boldsymbol{x}$?? O wait technically it could true for the identity function over degenerate $p^{in}(x) = \delta(0)$, but that's dumb.

[14] NB: this is NOT the familiar gram matrix (or anything analogous to it), it is more like a multi-dimensional analogue of what *each element* of the Gram matrix $G_{i,j} \triangleq k(x^{(i)}, x^{(j)})$ represented.

Condensed Setup/Notation Summary.

$$
\begin{aligned}
\text{[function space]} \quad & \mathcal{F} = \{f : \mathbb{R}^{n_0} \to R^{n_L}\} && (168) \\
\text{[empirical dist.]} \quad & p^{in}(x) := \frac{1}{N} \sum_{i=1}^{N} \delta(x - x^{(i)}) && (169) \\
\text{[network function]} \quad & f_\theta(x) := \tilde{\alpha}^{(L)}(x; \theta) && (170) \\
& \tilde{\alpha}^{(\ell)} = \frac{1}{\sqrt{n_{\ell-1}}} W^{(\ell-1)} \alpha^{(\ell-1)}(x; \theta) + \beta b^{(\ell-1)} && (171) \\
& \alpha^{(\ell)}(x; \theta) = \sigma\left(\tilde{\alpha}^{(\ell)}(x; \theta)\right) && (172) \\
\text{[input seminorm]} \quad & \langle f, g \rangle_{p^{in}} \triangleq \mathbb{E}_{x \sim p^{in}} \left[f(x)^T g(x) \right] && (173) \\
\text{[kernel bilinear form]} \quad & \langle f, g \rangle_K \triangleq \mathbb{E}_{x, x' \sim p^{in}} \left[f(x)^T K(x, x') g(x') \right] && (174) \\
\text{[dual space]} \quad & \mathcal{F}^* = \{\mu : f \to \langle d, f \rangle_{p^{in}} \mid \forall f \in \mathcal{F}\} && (175) \\
& && (176)
\end{aligned}
$$

Kernel Gradient (3). Define the map $\Phi_K : \mathcal{F}^* \to \mathcal{F}$ as mapping some function $\mu(\cdot) \in \mathcal{F}^*$ to another function denoted $f_\mu(\cdot) \in \mathcal{F}$[15]

$$
\begin{aligned}
& \Phi_k(\mu) = f_\mu && (177) \\
\text{s.t.} \quad & f_{\mu,i}(x) = \mu\left(K_{i,\cdot}(x, \cdot)\right) && (178) \\
& = \langle d, K_{i,\cdot}(x, \cdot) \rangle_{p^{in}} && (179)
\end{aligned}
$$

where we've used the fact that the ith row of $K(x, \cdot)$ is a vector in $\mathbb{R}^{n_L}$ (i.e. $K_{i,\cdot}(x, \cdot) \in \mathcal{F}$). Recall that our function cost $C : \mathcal{F} \to \mathbb{R}$, combined with the empirical definition for p^{in}, only depends on the values of any $f \in \mathcal{F}$ at the n training data points. I'm not sure how to phrase this, but basically the derivative of C is somehow itself a function in $\mathcal{F}^*$[16]. Accordingly, the derivative of C evaluated at some specific $f = f_0$ will in practice be a vector of n elements, with element i representing the change in C corresponding to a change in $f_0(x^{(i)})$. We can interpret that vector (for some reason), as itself a function in $\mathcal{F}^*$ such that

$$
\left.\frac{\partial C(f)}{\partial f}\right|_{f=f_0} \triangleq \left\langle d\big|_{f_0}, \cdot \right\rangle_{p^{in}} \tag{180}
$$

$\left.\frac{\partial C(f)}{\partial f}\right|_{f=f_0} \in \mathcal{F}^*$

where $d\big|_{f_0} \in \mathcal{F}$ is the corresponding dual element. The authors denote the above derivative as $\partial_f^{in} C\big|_{f_0}$ but I think that's disgusting notation.

[15] It will be useful to recall the **Riesz Representation Theorem**: *Suppose V is finite-dimensional and φ is a linear functional on V. Then there exists a unique vector $u \in V$ such that $\varphi(v) = \langle v, u \rangle$ for every $v \in V$.*

[16] My confusion is that, although I have no issue with interpretating the derivative as a function evaluated at each of the n data points, I don't understand how to interpret that returned function, when evaluated, as a scalar value in $\mathbb{R}$.

The **kernel gradient**, denoted $\nabla_K C\big|_{f_0} \in \mathcal{F}$ is defined as

$\nabla_K C\big|_{f_0} \in \mathcal{F}$

$$\nabla_K C\big|_{f_0} \triangleq \Phi_K \left(\frac{\partial C(f)}{\partial f} \bigg|_{f=f_0} \right) \tag{181}$$

$$\nabla_K C\big|_{f_0}(x) = \frac{1}{N} \sum_{i=1}^{N} K(x, x^{(i)}) d\big|_{f_0}(x^{(i)}) \tag{182}$$

With a confuse abuse of notation, let $f(t)$ denote a time-dependent function that, for any given time t, return a function in $\mathcal{F}$. We say such a function follows the **kernel gradient descent with respect to K** if it satisfies the differential equation:

$$\frac{\partial f(t)}{\partial t} = -\nabla_K C\big|_{f(t)} \tag{183}$$

$$\frac{\partial C(f)}{\partial t}\bigg|_{f=f(t)} = -\left\langle d\big|_{f(t)}, \nabla_K C\big|_{f(t)} \right\rangle_{p^{in}} \tag{184}$$

$$= -\left|\left| d\big|_{f(t)} \right|\right|_K^2 \tag{185}$$

where 184 is clearly a consequence of some chain-rule-ish thing that I'm having trouble formalizing, and 185 is by plugging in 182[17] If we assume the kernel K is pd wrt $||\cdot||_{p^{in}}$, then 185 is always negative, aka the cost is strictly decreasing (until it hits a critical point and thus converges).

TODO: ok but wtf is the interpretation of $d\big|_{f(t)}$ here?

Random Functions Approximation (3.1). We can approximate a kernel K by sampling P random functions iid from any distribution on $\mathcal{F}$, denoted $P_{\mathcal{F}}$, that satisfies

$$\mathbb{E}_{f^{(p)} \sim P_{\mathcal{F}}} \left[f_k^{(p)}(x) f_{k'}^{(p)}(x') \right] = K_{k,k'}(x, x') \tag{186}$$

i.e. any distribution whose covariance equals the kernel. We can then define a parameterized linear function $F^{lin} : \mathbb{R}^P \to \mathcal{F}$ as

$$F^{lin}(\theta) \triangleq f_\theta^{lin} \tag{187}$$

$$\text{where} \quad f_\theta^{lin}(x) \triangleq \frac{1}{\sqrt{P}} \sum_{p=1}^{P} \theta_p f^{(p)}(x) \tag{188}$$

$$\frac{\partial F^{lin}(\theta)}{\partial \theta_p} = \frac{1}{\sqrt{P}} \sum_{p'=1}^{P} \frac{\partial}{\partial \theta_p} \left[\theta_{p'} f^{(p')} \right] \tag{189}$$

$$= \frac{1}{\sqrt{P}} f^{(p)} \tag{190}$$

[17] Note that 185 makes me want to vomit because it is so misleading: $||\cdot||_K^2$ is not guaranteed to be positive (!!) (just plug in the definitions to see).

OH, ok so the entire point of this sub-section, btw, is to show a simple case where the realization function $F^L := F^{lin}$ is linear wrt the parameters θ. It's analogous to, instead of having one architecture with trainable weights, you have p fixed architectures with random weights, and all you do is train their weighted combination $\sum_p \theta_p f^{(p)}$. In real life, however, the realization function is not linear in the parameters.

Another analogy is freezing all weights except the last layer, where the last layer is a linear layer with no activation.

Now consider the time-evolution of element $\theta_p = \theta_p(t)$ during gradient descent:

$$\begin{aligned}
\frac{\partial \theta_p(t)}{\partial t} &= -\frac{\partial}{\partial \theta_p} C(F^{lin}(\theta(t))) && (191)\\
&= -\frac{\partial C(f)}{\partial f}\bigg|_{f_\theta^{lin}(t)} \frac{\partial F^{lin}(\theta)}{\partial \theta_p} && (192)\\
&= -\frac{1}{\sqrt{P}} \frac{\partial C(f)}{\partial f}\bigg|_{f_\theta^{lin}(t)} f^{(p)} && (193)\\
&= -\frac{1}{\sqrt{P}} \left\langle d|_{f^{lin}_{\theta(t)}}, f^{(p)} \right\rangle_{p^{in}} && (194)\\
\frac{\partial f}{\partial t} &= \sum_{p=1}^{P} \frac{\partial f}{\partial \theta_p} \frac{\partial \theta_p}{\partial t} && (195)\\
&= -\frac{1}{P} \sum_{p=1}^{P} \left\langle d|_{f^{lin}_{\theta(t)}}, f^{(p)} \right\rangle_{p^{in}} f^{(p)} && (196)\\
&= -\nabla_{\tilde{K}} C && (197)\\
\text{where} \quad \tilde{K} &\triangleq \frac{1}{P} \sum_{p=1}^{P} f^{(p)} \otimes f^{(p)} && (198)
\end{aligned}$$

where $\tilde{K}$ is the **tangent kernel**. Since the sampled function $f^{(p)}$ are random variables, $\tilde{K}$ is a random n_L-dimensional kernel with values

$$\tilde{K}_{i,i'}(x, x') = \frac{1}{P} \sum_{p=1}^{P} f_i^{(p)}(x) f_{i'}^{(p)}(x') \tag{199}$$

In the limit $P \to \infty$, we have $\widetilde{K} \to K$.

Papers **November 24, 2021**

NTK Eigenvalues Accurately Predict Generalization

Table of Contents Local *Written by Brandon McKinzie*

Simon et al., “Neural Tangent Kernel Eigenvalues Accurately Predict Generalization” *Redwood Center for Theoretical Neuroscience*, (Oct 2021).

Review of the NTK. Consider $\hat{f}_\theta : \mathcal{X} \to \mathbb{R}$. One step of gradient descent on training point (x, y) with small learning rate η[18]

$$\ell_\theta(x, y) \triangleq (\hat{f}_\theta(x) - y)^2 \tag{200}$$

$$\theta \to \theta + \delta\theta \tag{201}$$

$$\delta\theta = -\eta \nabla_\theta \ell_\theta(x, y) \tag{202}$$

$$= -2\eta(\hat{f}_\theta(x) - y)\nabla_\theta \hat{f}_\theta(x) \tag{203}$$

We are interested in how much that single update changed the predictions of our network on some test input x':

$$\hat{f}_{\theta+\delta\theta}(x') = \underbrace{\hat{f}_\theta(x')}_{\text{original prediction}} + \underbrace{\left\langle \nabla_\theta \hat{f}_\theta(x'), \delta\theta \right\rangle}_{\text{linearized change in pred}} + \mathcal{O}(\delta\theta^2) \tag{204}$$

$$= \hat{f}_\theta(x') - \eta(\hat{f}_\theta(x) - y)\left\langle \nabla_\theta \hat{f}_\theta(x'), \nabla_\theta \hat{f}_\theta(x) \right\rangle + \mathcal{O}(\delta\theta^2) \tag{205}$$

$$= \hat{f}_\theta(x) - \eta(\hat{f}_\theta(x) - y)K(x, x') + \mathcal{O}(\delta\theta^2) \tag{206}$$

My interpretation:

1. The parameter update moved θ in the direction of steepest descent in the loss landscape *for the given training input* x.
2. The prediction of the network on some unseen test x' can be expanded as its prediction before the update plus a term depending on the inn (ok how tf to word this TODO)

Some weird implications:

- If $\left\langle \nabla_\theta \hat{f}_\theta(x'), \nabla_\theta \hat{f}_\theta(x) \right\rangle = 0$, then the *test prediction is unchanged/unaffected* by the weight update. Consider that $\nabla_\theta \hat{f}_\theta(x)$ is a vector in the same space as θ. If, for example, $\theta \in \mathbb{R}^p$, then there exists a set of $p - 1$ orthonormal vectors v_i that are orthogonal to this direction. In fact, these $p - 1$ vectors span a $p - 1$ dimensional subspace of $\mathbb{R}^p$ for which, if we had simply set $\delta\theta$ to any vector in that subspace, the prediction on the training point x wouldn’t have changed at all. It kind of blows my mind that I hadn’t fully realized this until now, that there are a *huge* number of changes to the parameters θ we can make that won’t alter the predictions of the network on a given input x.
- If $\left\langle \nabla_\theta \hat{f}_\theta(x'), \nabla_\theta \hat{f}_\theta(x) \right\rangle = \left|\left|\nabla_\theta \hat{f}_\theta(x')\right|\right|_2 \left|\left|\nabla_\theta \hat{f}_\theta(x)\right|\right|_2$ (i.e. perfectly parallel), that means the test prediction will be changed by the same exact amount as the training prediction was changed as a result of the weight update. A worst-case scenario of catastrophic forgetting would basically be if the ground truth for x' is $-y$, since that means we just updated our weights in the worst possible direction for improving the test prediction on x'. Note the even bigger implication is that a “perfect” task/distribution for this network is when all inputs $\{x^{(i)}\}$ that share the same target y have identical gradients $\nabla_\theta \hat{f}_\theta(x^{(i)})$, *and* for which any other set of inputs $\{x^{(j)}\}$ that have *different* target values y have orthogonal gradients $\nabla_\theta \hat{f}_\theta(x^{(j)})$ to the others.

[18] Never really appreciated how elegant/simple the gradient for MSE loss is:

> *If the prediction $\hat{f}_\theta(x)$ is larger (smaller) than y, that means we need to change θ s.t. $\hat{f}$ will decrease (increase). Therefore, move in the opposite (same) direction as $\nabla_\theta \hat{f}_\theta(x)$.*

Figures of Merit of $\hat{f}$

First, the inner product we'll be using is defined as

$$\langle f, g \rangle \triangleq \frac{1}{M} \sum_{x \in \mathcal{X}} g(x) h(x) \tag{207}$$

$$\text{[MSE]} \quad \mathcal{E}^{\mathcal{D}}(f) \triangleq \left\langle f - \hat{f}, f - \hat{f} \right\rangle \quad \text{and} \quad \mathcal{E}(f) \triangleq \mathbb{E}_{\mathcal{D}}\left[\mathcal{E}^{\mathcal{D}}(f)\right] \tag{208}$$

$$\text{[Orthogonality]} \quad \textbf{TODO: wut is this??} \tag{209}$$

$$\text{[Learnability]} \quad \mathcal{L}^{\mathcal{D}}(f) \triangleq \frac{\left\langle f, \hat{f} \right\rangle}{\langle f, f \rangle} \quad \text{and} \quad \mathcal{L}(f) \triangleq \mathbb{E}_{\mathcal{D}}\left[\mathcal{L}^{(D)}(f)\right] \tag{210}$$

The Kernel Eigensystem. NB: authors assume hereafter that $m = 1$ (scalar-output functions). The authors are basically treating the entire input space $\mathcal{X}$ like we usually do for just the training data. Let $M = |\mathcal{X}|$ denote the number of possible inputs x to the network.

Function-Space Perspective.

By definition, any kernel function K is *symmetric* and *positive-semidefinite*: Recall that the definition of a kernel function $K(x, x')$ is that it must be expressible as $\langle \phi(x), \phi(x') \rangle$ for some feature function $\phi : \mathcal{X} \to V$ where $V \subseteq \mathcal{X}$.

1. Clearly, this is symmetric wrt the arguments x, x'.
2. To show it is psd, notice that from the definition we see we can write $K = \Phi\Phi^\top$ for the matrix Φ defined as $\Phi_i \equiv \phi(x^{(i)})$. Therefore, for any function $f : \mathcal{X} \to \mathbb{R}$:

$$\langle f | K | f \rangle = \langle f | \Phi\Phi^\top | f \rangle \tag{211}$$

$$= \left|\left| \Phi^\top | f \rangle \right|\right|_2^2 \geq 0 \tag{212}$$

Recall that any linear Hermitian operator H has a set of orthonormal eigenfunctions that form a basis for the Hilbert space that the operator acts upon[19].

$$\langle K(x, \cdot), \phi_i \rangle = \frac{1}{M} \sum_{x' \in \mathcal{X}} K(x, x') \phi_i(x') = \lambda_i \phi_i(x) \tag{216}$$

[19] Furthermore, H admits a **spectral decomposition** in this eigenbasis

$$H = \sum_i \lambda_i |\phi_i\rangle \langle \phi_i| \tag{213}$$

$$\Longrightarrow \forall \psi \quad \langle \psi | H | \psi \rangle = \sum_i \lambda_i \langle \psi \mid \phi_i \rangle \langle \phi_i \mid \psi \rangle = \sum_i \lambda_i \langle \phi_i \mid \psi \rangle^2 \tag{214}$$

which implies that, if $||\psi||_2 = 1$, we have $\langle \psi | H | \psi \rangle \geq \min_i \lambda_i$.

$$\langle \phi_j | K | \phi_i \rangle = \langle \phi_j | \left(\sum_k \lambda_k |\phi_k\rangle \langle \phi_k| \right) |\phi_i\rangle = \lambda_i |\phi_i\rangle \qquad [\langle \phi_k \mid \phi_i \rangle = \delta_{k,i}] \tag{215}$$

which is an equivalent way of saying “K is an operator on functions of $\boldsymbol{x} \in \mathcal{X}$ with eigenfunctions $\{\phi_i\}$ s.t. $K\left|\phi_i\right\rangle = \lambda_i \left|\phi_i\right\rangle$. Next, note that we can express both f and $\hat{f}$ in the eigenbasis via

$$\left|f\right\rangle = \sum_{i=1}^{M} v_i \left|\phi_i\right\rangle \tag{217}$$

$$\left|\hat{f}\right\rangle = \sum_{i=1}^{M} \hat{v}_i \left|\phi_i\right\rangle \tag{218}$$

It is straightforward to verify/check that $\left\langle f, \hat{f} \right\rangle = \boldsymbol{v}^\top \hat{\boldsymbol{v}}$. Letting $\boldsymbol{\Phi}(\mathcal{D}) := \phi_i(\boldsymbol{x}^{(j)})$ denote the $M \times n$ matrix of eigenfunctions evaluated at the n training points. Then we can write/define $K(\mathcal{D}, \mathcal{D}) = \boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})$. Plugging this in directly:

$$\hat{f}(x) = K(x, \mathcal{D})K(\mathcal{D}, \mathcal{D})^{-1} f(\mathcal{D}) \tag{219}$$

$$= \begin{bmatrix} \phi_1(x) & \cdots & \phi_M(x) \end{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\Phi} \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi(\mathcal{D})}^\top \boldsymbol{v} \tag{220}$$

$$\langle \phi_i, f \rangle \triangleq \frac{1}{M} \sum_{x \in \mathcal{X}} \phi_i(x) f(x) \tag{221}$$

$$= \frac{1}{M} \sum_{x \in \mathcal{X}} \sum_{i'=1}^{M} \phi_i(x)\phi_{i'}(x) \left(\boldsymbol{\Lambda}\boldsymbol{\Phi} \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi(\mathcal{D})}^\top \boldsymbol{v}\right)_{i'} \tag{222}$$

$$= \frac{1}{M} \sum_{x \in \mathcal{X}} \sum_{i'=1}^{M} \phi_i(x)\phi_{i'}(x)\lambda_{i'} \left(\boldsymbol{\Phi} \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi(\mathcal{D})}^\top \boldsymbol{v}\right)_{i'} \tag{223}$$

$$= \sum_{i'} \lambda_{i'} \left(\boldsymbol{\Phi} \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi(\mathcal{D})}^\top \boldsymbol{v}\right)_{i'} \underbrace{\frac{1}{M} \sum_{x \in \mathcal{X}} \phi_i(x)\phi_{i'}(x)}_{\delta_{i,i'}} \tag{224}$$

$$= \lambda_i \left(\boldsymbol{\Phi} \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi(\mathcal{D})}^\top \boldsymbol{v}\right)_{i} \tag{225}$$

$$= \lambda_i \underbrace{\phi_i(\mathcal{D})^\top}_{1 \times n} \underbrace{\left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi(\mathcal{D})}^\top \boldsymbol{v}}_{n \times 1} \tag{226}$$

and therefore

$$\hat{\boldsymbol{v}} = \underbrace{\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D}) \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi}^\top(\mathcal{D})}_{\triangleq \boldsymbol{T}^{(\mathcal{D})}} \boldsymbol{v} \tag{227}$$

where $\boldsymbol{T}^{(\mathcal{D})}$ is the **learning transfer matrix**.

Exact Results (2.4).

Lemma 1

(a) $\mathcal{L}^{\mathcal{D})(\phi_i)} = \boldsymbol{T}^{\mathcal{D}}_{ii}$ *and* $\mathcal{L}(\phi_i) = \boldsymbol{T}_{ii}$.

▶ *uh*

(e) Let $\mathcal{D}_+ = \mathcal{D} \cup x$, *where* $x \in \mathcal{X}$, $x \notin \mathcal{D}$ *is a new data point. Then* $\mathcal{L}^{\mathcal{D}_+}(f) \geq \mathcal{L}^{\mathcal{D}}(f)$

Proof Sketch: Property (e) of Lemma 1

1. Rewrite $\boldsymbol{T}^{\mathcal{D}}$ with $\boldsymbol{\Lambda} \to \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}$ and observe it is a bunch of products of the $m \times n$ matrix $\boldsymbol{A} \triangleq \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Phi}$:

$$\boldsymbol{T}^{\mathcal{D}} = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top\boldsymbol{\Lambda}^{\frac{1}{2}} = \boldsymbol{\Lambda}^{\frac{1}{2}}(\boldsymbol{A}\boldsymbol{A}^\top)(\boldsymbol{A}\boldsymbol{A}^\top)^{+}\boldsymbol{\Lambda}^{\frac{1}{2}}$$

   Note that the above implies that the $n \times n$ matrix $\boldsymbol{A}^\top\boldsymbol{A}$ is invertible (a consequence of the fact that $n \leq M$ and $\boldsymbol{A}$ has full column rank), but not necessarily $\boldsymbol{A}\boldsymbol{A}^\top$, hence our use of the pseudoinverse.
2. Notice that the effect of appending one more data point is appending one more column to $\boldsymbol{\Phi}$[a], which we will denote as the M-dimensional vector ξ.
3. The Sherman-Morrison formula tells us how we can evaluate an inverse of the form $(\boldsymbol{B} + uu^\top)^{-1}$ if $\boldsymbol{B}$ is invertible. Here, we'll set $\boldsymbol{B} := \boldsymbol{A}\boldsymbol{A}^\top + \delta\boldsymbol{I}_M$ and $u := \boldsymbol{\Lambda}^{\frac{1}{2}}\xi$. Combining this with the limit definition of the pseudoinverse gives us

$$\boldsymbol{T}^{\mathcal{D}_+} = \boldsymbol{T}^{\mathcal{D}} + \lim_{\delta \to 0^+} \delta \frac{\boldsymbol{B}^{-1}\xi\xi^\top\boldsymbol{B}^{-1}}{1 + \xi^\top\boldsymbol{B}^{-1}\xi}$$

   ...although I'm not entirely sure how the exact derivation works...
4. Since

$$\begin{aligned}\mathcal{L}^{(\mathcal{D}_+)}(f) &\propto \boldsymbol{v}^\top\boldsymbol{T}^{(D_+)}\boldsymbol{v} && (228)\\ &= \mathcal{L}^{(\mathcal{D})} + \lim_{\delta \to 0^+} \delta\boldsymbol{v}^\top(\cdots)\boldsymbol{v} && (229)\end{aligned}$$

   and the matrices inside the $(\cdots)$ are psd, we have the desired result.

[a] Somehow we know that this additional column will be orthonormal to the others, but I don't see how that is guaranteed. Obvi there exists a column whose values would satisfy it, but I don't see how the column of ϕ_i evaluated on the new point x is automatically orthonormal wrt the other columns.

Deriving a Closed-Form Expression for $\boldsymbol{T}$ (2.5). To motivate the following derivations, recall the resolution of the identity using Dirac notation:

$$\boldsymbol{I}_M = \sum_{i=1}^{M} |\phi_i\rangle\langle\phi_i| = \boldsymbol{\Phi}^\top\boldsymbol{\Phi} \tag{230}$$

where it's important to emphasize that, for now, we should interpret the above solely from the perspective of some abstract Hilbert space with some orthonormal basis of eigenfunctions ϕ_i, and to view $\boldsymbol{\Phi}$ as an associated linear operator. The point is that $\boldsymbol{\Phi}$ is [clearly] a *unitary operator*. Now, if we restrict to the n training points in some dataset $\mathcal{D}$, we should observe that things like the above become an *approximation*, e.g.

$$\langle \phi_i \mid \phi_j \rangle \triangleq \frac{1}{M} \sum_{x \in \mathcal{X}} \phi_i(x) \phi_j(x) = \delta_{i,j} \tag{231}$$

$$\approx \frac{1}{n} \sum_{x \in \mathcal{D}} \phi_i(x) \phi_j(x) \tag{232}$$

$$= \phi_i^\top(\mathcal{D}) \phi_j(\mathcal{D}) \tag{233}$$

This motivates re-writing $\boldsymbol{T} \triangleq \mathbb{E}_{\mathcal{D}} \left[\boldsymbol{T}^{(\mathcal{D})} \right]$ as

$$\boldsymbol{T} = \lim_{n \to \infty} \mathbb{E}_{\substack{\boldsymbol{\Phi} \sim \mathbb{R}^{M \times n} \\ \boldsymbol{\Phi}^\top \boldsymbol{\Phi} = \boldsymbol{I}_n}} \left[\boldsymbol{\Lambda} \boldsymbol{\Phi} \left(\boldsymbol{\Phi}^\top \boldsymbol{\Lambda} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^\top \right] \tag{234}$$

and thus we can *approximate* $\boldsymbol{T}$ by computing the above expectation for some reasonably large value of n.

OH JESUS I'M DUMB. M is the dimension of the feature space. We usually call this P. But this is a generalization of that from "P features" to "P basis feature functions".

Vector-Space Perspective: *the following is how I interpreted it by converting function stuff into vectors. Have since learned it is probably best to think purely in functional terms, but I think the following perspective is still helpful.* We can think of K as a massive $M \times M$ matrix with values $K(x, x') = \left\langle \nabla_\theta \hat{f}_\theta(x), \nabla_\theta \hat{f}_\theta(x') \right\rangle$. Since any kernel K is symmetric and psd by definition, this means there exists a set of orthonormal eigenvectors ϕ_i such that[20]

$$K\phi_i = \lambda_i \phi_i \tag{235}$$

where, again, we are interpreting functions ϕ_i as M-dimensional vectors with elements $\phi_i(x)$ (assuming some agreed upon ordering over $x \in \mathcal{X}$). In other words, for some specific $x \in \mathcal{X}$, this expands/corresponds to:

$$\sum_{x' \in \mathcal{X}} K(x, x')\phi_i(x') = \lambda_i \phi_i(x) \tag{236}$$

TODO: how do the above equations change when our network outputs an m-dimensional vector (instead of a scalar, as assumed above).

It follows that K has an eigendecomposition

$$K = \Phi\Lambda\Phi^T = \sum_{i=1}^{M} \lambda_i \phi_i \phi_i^\top \tag{237}$$

$$\implies K(x_1, x_2) = \sum_{i=1}^{M} \lambda_i \phi_i(x_1)\phi_i(x_2) \tag{238}$$

where Φ is an $M \times M$ matrix whose ith column is ϕ_i. If instead we restrict to the training data, denoted $K(\mathcal{D}, \mathcal{D})_{i,j} \triangleq K(x_i, x_j)$ for $x_i, x_j \in \mathcal{D}$, and $|\mathcal{D}| = n$, then $\Phi^\top(\mathcal{D})_{i,j} = \phi_i(x_j)$.

Recall from regular OLS linear regression, in the case where we can assume $X \in \mathbb{R}^{n \times p}$ is full row rank, then $(XX^\top)^{-1}$ exists and

$$\hat{f} \triangleq Xw \tag{239}$$

$$\text{[preds on train set]} \quad \hat{f}^* = XX^\top(XX^\top)^{-1}f(\mathcal{D}) \tag{240}$$

$$\text{[pred on test point]} \quad \hat{f}^*(x) = x^\top X^\top(XX^\top)^{-1}f(\mathcal{D}) \tag{241}$$

If we generalize the simple kernel here $K = XX^\top$ to any symmetric positive-definite K, we get **kernel regression**:

$$\hat{f}(x) = K(x, \mathcal{D})K(\mathcal{D}, \mathcal{D})^{-1}f(\mathcal{D}) \tag{242}$$

[20] The authors introduce a normalization constant of $\frac{1}{M}$ that I'm omitting for now.

Since we can also view f and $\hat{f}$ as themselves being M-dimensional vectors[21], we can express f and $\hat{f}$ in the eigenbasis as

$$f(x) = \sum_{i=1}^{M} v_i \phi_i(x) \tag{243}$$

$$\hat{f}(x) = \sum_{i=1}^{M} \hat{v}_i \phi_i(x) \tag{244}$$

3.3.1 Relevant Maths

Review: Vectors in Function Spaces

An operator is a mapping between functions in its **domain** and functions in its **range**. Assume we are only interested in **linear operators** on **Hilbert spaces**[a]. Let f and g denote functions, and let k denote a constant. If A and B are linear operators, then

$$(A+B)f = Af + Bf, \quad A(f+g) = Af + Ag, \quad Ak = kA \tag{245}$$

Furthermore, we assume that any linear operator applied to a member φ_μ of some orthonormal basis $\{\varphi_i\}$ can itself be expanded in that basis,

$$A\varphi_\mu = \sum_\nu a_{\nu\mu}\varphi_\nu \tag{246}$$

which is true because we are assuming that the domain and range of our linear operators are in the same Hilbert space. Using dirac notation gives us an extremely sexy way of looking at application of a linear operator A to a function $\psi = \sum_\mu c_\mu \varphi_\mu$:

$$A\left|\psi\right\rangle = \sum_\mu c_\mu A\left|\varphi_\mu\right\rangle = \sum_\mu c_\mu \sum_\nu a_{\nu\mu}\left|\varphi_\nu\right\rangle = \sum_\nu \left(\sum_\mu a_{\nu\mu}c_\mu\right)\left|\varphi_\nu\right\rangle \tag{247}$$

$$= \sum_{\nu,\mu}\left|\varphi_\nu\right\rangle \underbrace{\left\langle\varphi_\nu\right| A \left|\varphi_\mu\right\rangle}_{a_{\nu\mu}} \underbrace{\left\langle\varphi_\mu|\psi\right\rangle}_{c_\mu} \tag{248}$$

$$\implies A = \sum_{\nu,\mu}\left|\varphi_\nu\right\rangle\left\langle\varphi_\nu\right| A\left|\varphi_\mu\right\rangle\left\langle\varphi_\mu\right| \tag{249}$$

[a] A**Hilbert space** is a vector space (closed under addition and scalar multiplication) that has a scalar product $\langle f \mid g\rangle$ that exists for all pairs of its members f and g.

The *rank* of a general $n \times m$ matrix $\boldsymbol{X}$ is the dimension spanned by its columns (aka the

[21] Since apparently we've still forgotten that f and $\hat{f}$ should be outputting m-dimensional vectors (not scalar)

column space). Hence, the rank of $\boldsymbol{X}$ is the smallest r for which we can express

$$\begin{aligned}\underbrace{\boldsymbol{X}}_{n\times m} &= \underbrace{\boldsymbol{R}}_{n\times r}\underbrace{\boldsymbol{S}}_{r\times m} && (250)\\ &= \underbrace{\begin{bmatrix}\boldsymbol{r}_1 & \boldsymbol{r}_2 & \cdots & \boldsymbol{r}_r\end{bmatrix}}_{\text{linearly independent}}\boldsymbol{S} && (251)\\ &= \begin{bmatrix}\sum_{i=1}^{r} S_{i,1}\left|\boldsymbol{r}_i\right\rangle & \cdots & \sum_{i=1}^{r} S_{i,m}\left|\boldsymbol{r}_i\right\rangle\end{bmatrix} && (252)\end{aligned}$$

and thus we can interpret the m'th column of $\boldsymbol{X}$ as $\sum_i^r \left|\boldsymbol{r}_i\right\rangle\left\langle\boldsymbol{r}_i\right|\boldsymbol{S}\left|\boldsymbol{e}_m\right\rangle$. In other words, the columns of $\boldsymbol{S}$ express the columns of $\boldsymbol{X}$ in the basis defined by the linearly independent columns of $\boldsymbol{R}$. We say that $\boldsymbol{X}$ is **full rank** if its rank is equal to $\min(n, m)$.

If any $n\times n$ matrix $\boldsymbol{A}$ is non-singular (i.e. invertible), then its n columns are linearly independent. If there exists $\boldsymbol{x}\neq 0$ and $\lambda\neq 0$ (since $\boldsymbol{A}$ is non-singular) such that $\boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x}$, then the **Rayleigh quotient**

$$\frac{\left\langle\boldsymbol{x}\right|\boldsymbol{A}\left|\boldsymbol{x}\right\rangle}{\left\langle\boldsymbol{x}\mid\boldsymbol{x}\right\rangle}=\lambda \qquad (253)$$

evaluates that eigenvalue. We can extend this to obtain the following:

Courant-Fisher Theorem

If $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ is symmetric, then for $k\in[1..n]$,,

$$\lambda_k(\boldsymbol{A}) = \min_{\substack{T\\ dim(T)=n-k+1}}\ \max_{\substack{\boldsymbol{v}\in T\\ \boldsymbol{v}\neq 0}}\frac{\left\langle\boldsymbol{v}\right|\boldsymbol{A}\left|\boldsymbol{v}\right\rangle}{\left\langle\boldsymbol{v}\mid\boldsymbol{v}\right\rangle} = \max_{\substack{T\\ dim(T)=k}}\ \min_{\substack{\boldsymbol{v}\in T\\ \boldsymbol{v}\neq 0}}\frac{\left\langle\boldsymbol{v}\right|\boldsymbol{A}\left|\boldsymbol{v}\right\rangle}{\left\langle\boldsymbol{v}\mid\boldsymbol{v}\right\rangle} \qquad (254)$$

with the extrema achieved by the corresponding eigenvector.

Symmetric matrices. If $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ is symmetric, and $\boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x}$ and $\boldsymbol{A}\boldsymbol{z}=\mu\boldsymbol{z}$ with $\mu\neq\lambda$, then

$$\lambda\left\langle\boldsymbol{x},\boldsymbol{z}\right\rangle = \left\langle\boldsymbol{x}\boldsymbol{A}\mid\boldsymbol{z}\right\rangle = \left\langle\boldsymbol{x}\right|\boldsymbol{A}^{\top}\left|\boldsymbol{z}\right\rangle = \left\langle\boldsymbol{x}\right|\boldsymbol{A}\left|\boldsymbol{z}\right\rangle = \mu\left\langle\boldsymbol{x},\boldsymbol{z}\right\rangle \implies \left\langle\boldsymbol{x},\boldsymbol{z}\right\rangle = 0 \qquad (255)$$

since $\mu\neq\lambda$. Therefore, eigenvectors of symmetric matrices with different eigenvalues are orthogonal. Since there can be at most n orthogonal eigenvectors for an $n\times n$ matrix, we have that any symmetric matrix A has at most n distinct eigenvalues. Let $\boldsymbol{V}$ denote the $n\times n$ matrix whose columns are the orthonormal eigenvectors of A. Note that $\boldsymbol{V}^{\top}\boldsymbol{V}=\boldsymbol{V}\boldsymbol{V}^{\top}=\boldsymbol{I}_n$, and that

$$\boldsymbol{A}\boldsymbol{V} = \begin{bmatrix}\boldsymbol{A}\boldsymbol{v}_1 & \cdots\boldsymbol{A}\boldsymbol{v}_n\end{bmatrix} = \begin{bmatrix}\lambda_1\boldsymbol{v}_1 & \cdots\lambda_n\boldsymbol{v}_n\end{bmatrix} = \boldsymbol{V}\boldsymbol{\Lambda} \implies \boldsymbol{A}=\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\top} \qquad (256)$$

Notice that the final implication is only valid since we know $\boldsymbol{V}^{-1}=\boldsymbol{V}^{\top}$ exists. If the symmetric matrix $\boldsymbol{A}$ has k [distinct] nonzero eigenvalues, then we can use the simpler

$$\boldsymbol{A} = \underbrace{\boldsymbol{V}_k}_{n\times k}\underbrace{\boldsymbol{\Lambda}_k}_{k\times k}\underbrace{\boldsymbol{V}_k^{\top}}_{k\times n} \qquad (257)$$

to show this, note that obviously $\boldsymbol{A} = \sum_{i=1}^{k} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$. The result immediately follows. Since the k eigenvectors are orthonormal (as a result of $\boldsymbol{A}$ being symmetric and the k eigenvalues being distinct), we also have that $\text{rank}(\boldsymbol{A}) = k$.

Characterization of Kernels

A function

$$\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

which is either continuous or has a finite domain, can be decomposed

$$\kappa(\boldsymbol{x}, \boldsymbol{z}) = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{z}) \rangle \tag{258}$$

for feature map ϕ into a Hilbert space F if and only if it satisfies the ***finitely positive semi-definite property****: it is a symmetric function for which the matrices formed by restriction to any finite subset of the space $\mathcal{X}$ are positive semi-definite.*

Compact proof of reverse implication

Assume κ satisfies the finitely psd property. We need to show that this implies κ can be decomposed via 258. Let $\mathcal{F}$ denote a vector space over functions of the form

$$f(\boldsymbol{x}) := \sum_{i=1}^{\ell} \alpha_i \kappa(\boldsymbol{x}^{(i)}, \boldsymbol{x}) \tag{259}$$

for some $\ell \in \mathbb{N}$, $\boldsymbol{x}^{(i)} \in \mathcal{X}$, $\alpha_i \in \mathbb{R}$[a]. Let $f, g \in \mathcal{F}$, with f denoted as above, and $g(\boldsymbol{z}) := \sum_j \beta_j \kappa(\boldsymbol{z}^{(j)}, \boldsymbol{z})$. Define the scalar product of this Hilbert space as

$$\langle f \mid g \rangle \triangleq \sum_{i,j} \alpha_i \beta_j \kappa(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(j)}) \tag{260}$$

$$= \sum_i \alpha_i \left(\sum_j \beta_j \kappa(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(j)}) \right) = \sum_i \alpha_i g(\boldsymbol{x}^{(i)}) \tag{261}$$

$$= \sum_j \beta_j \left(\sum_i \alpha_i \kappa(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(j)}) \right) = \sum_j \beta_j f(\boldsymbol{z}^{(j)}) \tag{262}$$

Note that if $g(\boldsymbol{x}) = \kappa(\boldsymbol{z}, \boldsymbol{x})$, then $\langle f, g \rangle = \sum_i \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{z}) = f(\boldsymbol{z})$ which is known as the **reproducing property** of the kernel. Next, define $\boldsymbol{\phi} : \mathcal{X} \to F_\kappa$, where $F_\kappa \subseteq \mathcal{F}$ is just the restriction that $\ell = 1$, $\alpha_1 = 1$, as

$$\boldsymbol{\phi}(\boldsymbol{x}) \triangleq \kappa(\boldsymbol{x}, \cdot) \tag{263}$$

where it's worth emphasizing again that $\boldsymbol{\phi}$ returns a *function* itself. Then we have our desired result:

$$\kappa(\boldsymbol{x}, \boldsymbol{z}) = \langle \kappa(\boldsymbol{x}, \cdot), \kappa(\boldsymbol{z}, \cdot) \rangle = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{z}) \rangle \tag{264}$$

[a] Note that this means each function in this space is associated with some ℓ number of vectors in $\mathcal{X}$.

3.3.2 Extended Notes for canatar2021

A **reproducing kernel Hilbert space** (RKHS) $\mathcal{H}$ living on $\mathcal{X} \subset \mathbb{R}^D$ is a subset of *square integrable functions*[22] $L_2(\mathcal{X}, p)$ for measure p equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a kernel K satisfying the **reproducing property**:

$$f(x) = \langle f(\cdot), K(\cdot, x) \rangle_{\mathcal{H}} \qquad (\forall x \in \mathcal{X})(\forall f \in \mathcal{H}) \tag{265}$$

Define the **integral operator** $T_K : L_2(\mathcal{X}, p) \rightarrow L_2(\mathcal{X}, p)$, which is a linear map from functions to functions:

$$T_K[f](x') \triangleq \int p(x) K(x, x') f(x) \mathrm{d}x \tag{266}$$

$$\textbf{[Mercer Decomposition]} \qquad K(x, x') = \sum_{\ell=0}^{\infty} \lambda_\ell \phi_\ell(x) \phi_\ell(x') \tag{267}$$

$$\begin{aligned}
\textbf{[Eigenfunction Property]} \qquad T_K[\phi_\ell](x') &= \int p(x) K(x, x') \phi_\ell(x) \mathrm{d}x && (268)\\
&= \int p(x) \left(\sum_{\ell'=0}^{\infty} \lambda_{\ell'} \phi_{\ell'}(x) \phi_{\ell'}(x') \right) \phi_\ell(x) \mathrm{d}x && (269)\\
&= \sum_{\ell'=0}^{\infty} \lambda_{\ell'} \phi_{\ell'}(x') \int p(x) \phi_{\ell'}(x) \phi_\ell(x) \mathrm{d}x && (270)\\
&= \sum_{\ell'=0}^{\infty} \lambda_{\ell'} \phi_{\ell'}(x') \delta \ell, \ell' && (271)\\
&= \lambda_\ell \phi_\ell(x') && (272)
\end{aligned}$$

A function f is said to be a member of the RKHS $\mathcal{H}$ if and only if $||f||^2_{\mathcal{H}} < \infty$, where

$$||f||^2_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}} = \sum_{\ell, \ell'} a_\ell a_{\ell'} \langle \phi_\ell, \phi_{\ell'} \rangle_{\mathcal{H}} = \sum_{\ell=0}^{\infty} \frac{1}{\lambda_\ell} a_\ell^2 \tag{273}$$

where the dimension of the RKHS equals the number of nonzero eigenvalues λ_ℓ.

[22] $\int |f(x)|^2 \mathrm{d}x < \infty$